LLM Neologism: Emergence of Mutated Characters due to Byte Encoding

Ran Iwamoto[1,2] **Hiroshi Kanayama**[1]
[1]IBM Research - Tokyo, [2]Keio University
`ran.iwamoto1@ibm.com`, `hkana@jp.ibm.com`

Abstract

The process of language generation, which selects the most probable tokens one by one, may intrinsically result in output strings that humans never utter. We name this phenomenon "LLM neologism" and investigate it focusing on Japanese, Chinese, and Korean languages, where tokens can be smaller than characters. Our findings show that LLM neologism occurs through the combination of two high-frequency words with common tokens. We also clarify the cause of LLM neologism in the tokenization process with limited vocabularies. The results of this study provides important clues for better encoding of multibyte characters, aiming to prevent catastrophic results in AI-generated documents.

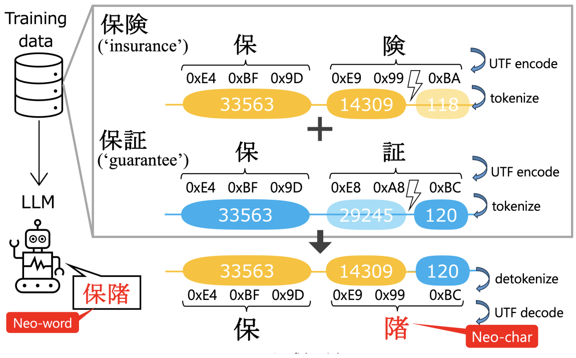

Figure 1: Overview of LLM neologism. In the prediction of output token sequences, those derived from two frequent words in the training data may be mutated. This results in a peculiar word (*neo-word*) that has a *neo-char* generated by the decoding of mixed byte codes.

1 Introduction

The text generation capabilities of LLMs have been improving year by year (Yin et al., 2023; Zhao et al., 2023), and the sentences generated by LLMs have become indistinguishable from those written by humans. However, LLMs occasionally output non-existent words. Although this is a rare phenomenon, its occurrence is a clear indication of an AI-generated sentence and thus should be avoided as much as possible. In this paper, we name this phenomenon *LLM neologism* and investigate it thoroughly. The phenomenon is a type of hallucination. LLM tends to cause hallucination, in which information that is not true is presented as if it were true (Huang et al., 2023). Hallucination is divided into various types (Rawte et al., 2023), but to the author's knowlegde, this type of hallucination is that has not been adressed in any previous paper.

Figure 1 shows the notion of LLM neologism, where a non-existent Japanese word "保陼 " is generated. We call such a word a *neo-word*. In languages where a single character can be split into multiple tokens, such as Chinese, Japanese, and Korean, the generation of a neo-word is triggered by the mutation of token sequences of two frequently-used words. Additionally, the mixture of tokens corresponding to single byte codes can result in the generation of an unexpected and rarely used character, which we call a *neo-char*.

In Section 2, we discuss the mechanism underlying LLM neologism in more detail. In Section 3, we explain the tokenization strategies in the existing LLMs, and in Section 4, we present our observation of LLM neologism in a systematic way. The main contribution of this work are as follows:

- to define the LLM neologism phenomenon, which to our knowledge is the first time this phenomenon has been discussed.
- to artificially generate potential neo-words based on our hypothesis and to enumerate actual instances in LLM generated texts or web documents.
- to propose a tokenization strategy for CJK languages with lower risk of LLM neologism.

2 Mechanism of LLM Neologism

In this section, we explain LLM neologism in an inductive manner. LLM neologism can happen in

Proceedings of the 17th International Natural Language Generation Conference, pages 24–29
September 23–27, 2024. ©2024 Association for Computational Linguistics

Neo-word	Constituent words	Similarity
勤勠	勤務 (‘work’)	0.94
	勤怠 (‘work attendance’)	0.87
	* 勤労 (‘labor’)	0.80
視覴	視聴 (‘viewing’)	0.95
	視覚 (‘vision’)	0.94
	* 視界 (‘visibility’)	0.81
音韟	音響 (‘sound’)	0.90
	音域 (‘sound range’)	0.90
	* 音楽 (‘music’)	0.81

Table 1: Japanese neo-words and their constituent words, shown with similarity scores from the neo-word in Llama2 embeddings. Constituent words have higher scores, compared with another word with ∗ that appears in a similar context.

Model	Tokenizer (BPE)	# in vocabulary	
		Kanji	Hangul
GPT3.5	byte-level	549	129
Llama2	byte-fallback	701	111
Elyza	byte-fallback	701	111
Elyza-fast	byte-fallback + ja token	7235	111
Granite-ja	byte-fallback	5663	409
Swallow-ja	byte-fallback + ja token	2835	111

Table 2: The number of single CJK characters in each tokenizer’s vocabulary. Elyza, Elyza-fast, Granite-ja, and Swallow-ja tokenizers are Llama2-based.

any language during the LLM’s decoding process when a word is generated from multiple tokens that are smaller than words, but here, we focus on the generation of a neo-word including a neo-char generated by the mixture of multiple bytes in Japanese, Chinese, and Korean.

Kanji characters in Japanese and Chinese, and Hangul characters in Korean are represented by three UTF-8 codes per character. Since the number of characters defined in the UTF code page[1] is much larger than the vocabulary size of the tokenizers used in existing LLMs, a single character is often divided into multiple tokens, as seen in the second character in Figure 1.

Here, we set up a hypothesis that a neo-word is generated from two frequent two-letter words that share the first letter and tend to appear in similar contexts. This explains the LLM neologism in Figure 1. In the process of generating “保険”(‘insurance’), after outputting its first two tokens[2], it is impossible to guarantee the prediction of the code 0xBA. Instead, another token 0xBC, derived from “保証”(‘guarantee’), can have higher probability than 0xBA, and this results in the generation of a neo-word that contains a neo-char.

Neo-words have been found on the web. A blog post[3] reported that ChatGPT output gibberish Japanese-like words that have never been seen before and that were subsequently used in a number of websites. For every neo-word we found on the web, we were able to identify the two constituent two-letter words. Table 1 shows the results of measuring the similarity between the neo-word and the constituent words in the final layer of embedding in Llama2 (Touvron et al., 2023). The neo-word “勤勠” has similarity scores of 0.94 and 0.87 with the two words “勤務”(‘work’) and “勤怠”(‘work attendance’), and has a higher score than another word “勤労”(‘labor’) which has the same first kanji character. This indicates that a neo-word has already been trained in the model, and as a result, this neo-word is likely to be output incorrectly in place of the two constituent words.

The generated neo-char can be a very infrequently used character, causing a reduction in the naturality of the LLM output and a critical problem of being revealed as AI-generated. For this reason, even if the rate of occurrence is not high, it is important to prevent LLM neologism.

In the next section, we discuss the tokenizer properties that are related to the occurrence of LLM neologism.

3 Tokenizers

In this section, we discusses the relationship between LLM neologism and the underlying tokenization process. The generation of neo-words by a model depends on how characters are split—specifically, on the tokenizer’s vocabulary. Many LLMs, such as Llama2/3 (AI@Meta, 2024), and GPT-3.5/4.0, use byte-pair-encoding (BPE) (Sennrich et al., 2016) for tokenization.

GPT-3.5 has 782 tokens for single or double bytes in its vocabulary, and Llama2 has 256 tokens for a single byte. The combination of these

[1] https://www.charset.org/utf-8

[2] These correspond to five bytes in the UTF-8 code.

[3] https://okumuralab.org/~okumura/misc/230611.html

Model	Generated neo-word	Generated text or web text
GPT	Yes	Web text
Llama2	Yes	Web text
Elyza	Yes	Generated text
Elyza-fast	No	-
Granite	No	-
Swallow	No	-

Table 3: The presence of neo-words in 3,187 generated texts and web texts in Japanese. Note that "No" does not mean that neologism will never occur with that tokenizer.

tokens represents multi-byte CJK characters that are not covered in the vocabulary, as in the second character in Figure 1.

Each tokenizer determines its vocabulary by selecting frequent sequences of byte codes from its own corpus, and thus, only limited numbers of CJK characters are in its vocabulary. Table 2 lists the number of single CJK characters in the vocabulary for each model. Considering that there are more than 100,000 Kanji and 11,172 Hangul characters in the UTF-8 character set, GPT-3.5, Llama2, and Elyza cover too small a number of CJK characters. Other Japanese-aware models cover larger numbers of characters. This difference is the key factor in the emergence of LLM neologism, which will be shown in the next section.

4 Replication of LLM neologism

In this section, we list potential neo-words to determine the occurrence of LLM neologism, and discuss its relationship with the tokenization.

4.1 List potential neo-words

Here we describe the process of enumerating neo-words by mixing two words to search for neo-words in the actual LLM-generated texts. We generate potential neo-words in Japanese, Chinese, and Korean. First, we have a list of two-character words in Kanji or Hangul that are commonly used in each language. In Japanese Kanji, we use BCCWJ (Maekawa et al., 2014) frequency list. In Chinese Kanji, we use BLCU Chinese Corpus: BCC corpus of 15 billion characters (Xun, 2016). In Hangul, we use Korean frequency list (National Institute of the Korean Language, 2005).

From these lists, we extract word pairs with word similarity of 0.4 or greater using FastText embedding (Grave et al., 2018). Potential neo-words are then generated by mixing two words considering the conditions described in Section 2. The commonly used Kanji characters defined in Japan (Japan, 2010) and China (the People's Republic of China, 2013) are excluded from our potential neo-chars since they are not prominently identified.

4.2 Generate sentences

We investigate whether the various LLM outputs contain neo-words. We used llama-2-7b-chat (Touvron et al., 2023), elyza/ELYZA-japanese-Llama-2-7b-instruct (Sasaki et al., 2023), ibm/granite-8b-japanese, and tokyotech-llm/Swallow-7b-hf (Fujii et al., 2024) as models. Since LLM neologism occurs rarely, we consider one of the hypotheses mentioned in the previous section 2, namely that the neo-word tends to appear in similar contexts based on the source of the neo-word, then we generated sentences in which LLM neologism is likely to occur.

To this end, we selected Wikipedia titles that contain either of the two words that are the source of the neo-word candidate, as collected in Section 4.1. By having LLMs descrive the source words, LLM neologism would be more likely to emerge in the process than in normal contexts.

We created 3,187 responses using the following prompt which means "Please tell me what you know about <Wikipedia title> in Japanese, in as much detail as possible":

Prompt: <wikipedia title>について
知っていることを日本語で,
なるべく詳しく教えてください。

4.3 Outputs

Table 3 shows the occurrences of LLM neologism by Japanese models based on our observation. In the method described in Section 4.1, we explicitly found a neo-word generated by Elyza. In addition, we searched manually for the potential neo-words on the Web, and identified neo-words generated by the GPT and Llama2 models considering their tokenizers' vocabularies. Not that while we did not identify neo-word generated by other models (marked "No" in Table 3), this does not mean that these models are theoretically free from LLM neologism.

The observed neo-words in a model tend to be specific to its underlying tokenizer. For ex-

Lang	Neo-word	Constituent words	Sentence on web with neo-word
ja	明碽	明確, 明白 (‘clear’), (‘obvious’)	それを外国人観光客にも[明碽]に説明する必要がある。 (‘This needs to be [?clearly] explained to foreign tourists.’)
ja	同窚	同窓, 同級 (‘alumni’), (‘same class’)	同窓会にエリート[同窚]生がいた。 (‘There was an elite [?alumni] at the reunion.’)
zh	坚弳	坚强, 坚决 (‘tough’), (‘firm’)	[坚弳]不是你的肌肉有多硬，而是你的精神有多硬。 (‘Being [?strong] is not about how hard your muscles are, it’s about how hard your spirit is.’)
zh	悲壧	悲壮, 悲剧 (‘tragic’), (‘tragedy’)	提及[悲壧]氛圍 ，《孟姜女》是一美的故事。 (‘As for [?sadness], Lady Meng Jiang is a beautiful story.’)
ko	학긐	학급, 학교 (‘class’), (‘school’)	[학긐] 활동 이외에도 봉사활동에 참여할 기회를 찾아봐. (‘In addition to [?school] activities, look for opportunities to participate in volunteer activities.’)

Table 4: LLM neologism in three languages found on the web. Neo-words and their corresponding translations are enclosed in square brackets. Note that neo-words in the original languages are inherently meaningless, and thus we provide translations by filling with the more natural constituent word in the context (marked with ‘?’).

ample, the Elyza model generated the neo-word “音韟 ”, and we identified its constituent words “音響”(‘acoustics’) and “音域” (‘sound range’). However, this neo-word never appeared in other models such as GPT because its tokenizer splits the two words into different numbers of tokens, and thus they are not mixed into “音韟 ”.

We show examples of multilingual neo-words in Table 4, which shows neo-words and the sentences in which they appeared that actually existed on the web, in the three languages[4]. The neo-words that appeared on the web were used in contexts similar to the constituent words. All of the neo-chars we found were the second letters of two-character words. One possible reason for this is that LLM generates sentences from the front, so the back characters are easily mixed up. Many of the web texts in which neo-words appeared could be implicitly identified as having been written by AI. For example, neo-words appeared on websites with “AI” in the title and on websites that stating that they generate video summaries using AI. These results indicate that LLM neologism occurs in various models. LLM neologism does not occur frequently, but the appearance of neo-words in a real document can raise the suspicion of readers that they are potentially looking at AI-generated text.

4.4 Discussion

As we have seen, LLM neologism in CJK languages is caused by decomposition of a single character into multiple tokens. Tables 2 and 3 suggest that the larger vocabulary size to cover more characters avoids LLM neologism. It is difficult for multilingual models to have larger vocabulary for a specific language, and there is a trade-off between small and large sets of vocabularies for tokenization in terms of efficiency (Stollenwerk, 2023).

Currently, byte-level encoding, rather than character-level encoding is a feasible approach for multilingual tokenization because of its simplicity (Mielke et al., 2021), and it actually achieves high-quality multilingual language models. However, we suggest that the higher coverage of characters in the vocabulary should be taken into consideration to avoid LLM neologism that may generate seriously gibberish words, even with a certain amount of sacrifices in existing benchmarking scores or the language coverage by a single model.

5 Conclusion

In this paper, we defined LLM neologism and revealed its characteristics. We showed that neo-words in Japanese, Chinese, and Korean are generated from two frequent two-letter words that share a first letter and tend to appear in a similar context.

Neo-words are generated when a single character is split into multiple tokens, and we clarified that

[4] Some sentences were modified due to copyright issues.

the likelihood of their generation depends on the tokenization method and the vocabulary. We demonstrated that neo-words in three languages appear in AI-generated texts, and showed that neo-words exist in context in a similar sense to constituent words.

LLM neologism is a tokenizer-dependent problem that occurs when a character is represented by multiple tokens. As stated by Mielke et al. (2021), there is no silver bullet solution that serves as a solution for all purposes. However, LLM neologism is an essential issue to consider in the context of generating natural sentences in CJK languages.

It is also known that LLM can generate new words by mixing words in English[5]. It is a future challenge to generalize LLM neologism in languages other than CJK.

Limitation

In addition to its linguistic definition, "neologism" is also used in the field of psychiatry and clinical psychology. As we wish to avoid potentially misleading patients by our use of this term, we should emphasize that our usage in this paper is limited to "LLM neologism" that refers to the phenomenon of word generation by LLM.

References

AI@Meta. 2024. Llama 3 model card.

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. *Preprint*, arXiv:2404.17790.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Agency for Cultural Affairs of Japan. 2010. Jōyō kanji table.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Lang. Resour. Eval.*, 48(2):345–371.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *Preprint*, arXiv:2112.10508.

National Institute of the Korean Language. 2005. Frequency of modern korean usage 2.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. 2023. Elyza-japanese-llama-2-7b.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Felix Stollenwerk. 2023. Training and evaluation of a multilingual tokenizer for gpt-sw3. *Preprint*, arXiv:2304.14780.

Ministry of Education of the People's Republic of China. 2013. General use standardized chinese character.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,

[5] https://www.reddit.com/r/CharacterAI/comments/192bm5g/theyre_just_making_up_words_now

Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Rao G. Xiao X. Zang J. Xun, E. 2016. The construction of the bcc corpus in the age of big data. *corpus linguistics*, (1).

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153, Singapore. Association for Computational Linguistics.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.