Quantifying Memorization of Domain-Specific Pre-trained Language Models using Japanese Newspaper and Paywalls

Shotaro Ishihara
Nikkei Inc.
Tokyo, Japan
shotaro.ishihara@nex.nikkei.com

Abstract

Dominant pre-trained language models (PLMs) have been successful in high-quality natural language generation. However, the analysis of their generation is not mature: *do they acquire generalizable linguistic abstractions, or do they simply memorize and recover substrings of the training data?* Especially, few studies focus on domain-specific PLM. In this study, we pre-trained domain-specific GPT-2 models using a limited corpus of Japanese newspaper articles and quantified memorization of training data by comparing them with general Japanese GPT-2 models. Our experiments revealed that domain-specific PLMs sometimes "copy and paste" on a large scale. Furthermore, we replicated the empirical finding that memorization is related to duplication, model size, and prompt length, in Japanese the same as in previous English studies. Our evaluations are relieved from data contamination concerns by focusing on newspaper paywalls, which prevent their use as training data. We hope that our paper encourages a sound discussion such as the security and copyright of PLMs.

1 Introduction

Pre-trained language models (PLMs) have shown great capabilities in solving various tasks in natural language processing (Yang et al., 2023; Zhao et al., 2023). Statistical language models learn the probability of word occurrence, and pre-training on large datasets for large neural networks has become popular. This extension has led to fluent natural language generation and has been reported to perform well when fine-tuned for many downstream tasks (Radford et al., 2018). For much larger models called large language models (LLMs), downstream tasks can be solved without parameter updates (Radford et al., 2019; Brown et al., 2020). Social recognition such as ChatGPT[1] is steadily increasing.

As practical applications evolve, critical views on the generation of PLMs are becoming apparent in security and copyright (Bender et al., 2021; Bommasani et al., 2021; Weidinger et al., 2022). Prior research has indicated that neural networks have the property of unintentionally memorizing and outputting the training data (Carlini et al., 2019, 2021, 2023; Lee et al., 2023). In particular, Carlini et al. (2021) demonstrated that memorized personal information (names, phone numbers, and email addresses) can be extracted from GPT-2 models (Radford et al., 2019). This can lead to an invasion of privacy, reduced utility, and reduced ethical practices (Carlini et al., 2023). If there is no novelty in the generation, there would be a problem in terms of copyright (McCoy et al., 2023; Franceschelli and Musolesi, 2023).

Despite its significance, this discussion remains in its infancy (Ishihara, 2023). Initial studies remain on the qualitative side (Carlini et al., 2021), and several studies have begun to focus on quantitative evaluations (Lee et al., 2022; Kandpal et al., 2022; Ippolito et al., 2022; Tirumala et al., 2022; Downey et al., 2022; Carlini et al., 2023; Lee et al., 2023). These studies were conducted mainly in English, and their reproducibility was uncertain under domain-specific conditions. Memorization of machine learning models is generally associated with overfitting (Yeom et al., 2018; Zhang et al., 2021a), which is even more important to discuss under conditions when it is difficult to prepare large training data. Compared to general corpora, considerations of security and copyright are increasingly important for rare corpora.

This study is the first attempt to quantify the memorization of domain-specific PLMs using a limited corpus of Japanese financial newspaper articles. Our research objective is *to identify trends in memorization of domain-specific PLMs*. We argue that newspaper articles are suitable for evaluating the memorization of PLMs because their paywall

[1] https://openai.com/blog/chatgpt

characteristics prevent their use as training data (Section 2). First, we defined memorization and developed a framework for quantifying the memorization of domain-specific PLMs using Japanese newspaper articles (Section 3). Secondly, we pre-trained domain-specific GPT-2 models and observed that they sometimes memorized and output the training data on a large scale (Section 4). Experiments reported that memorization is related to duplication, model size, and prompt length. These empirical findings, which had been reported in previous studies in English, were found for the first time in Japanese. Finally, we discuss future research directions (Section 5).

2 Related Work

This section reviews related work and highlights the position of this study.

2.1 Memorization of PLMs

Memorization of PLMs refers to the phenomenon of outputting fragments of the training data. Research on memorization is diverse, with various definitions and assumptions. This study focuses on autoregressive language models, such as the GPT family (Radford et al., 2018, 2019; Brown et al., 2020; Black et al., 2022). These are promising models as of 2023.

Definition of memorization. Many studies have adopted definitions based on partial matching of strings (Carlini et al., 2021, 2023; Kandpal et al., 2022). This definition of *eidetic memorization* assumes that memorized data are extracted by providing appropriate prompts to PLMs. Another definition of *approximate memorization* considers string fuzziness. For similarity, Lee et al. (2022) used the token agreement rate, and Ippolito et al. (2022) used BLEU.

Our study designed the first of these definitions in Japanese and reported the experimental results. Both definitions of memorization are ambiguous in languages without obvious token delimiters such as Japanese. Definitions based on the concepts of differential privacy (Jagielski et al., 2020; Nasr et al., 2021) and counterfactual memorization (Zhang et al., 2021b) are beyond the scope of this study.

2.2 Issues with Memorization of PLMs

There are discussions of issues such as security and copyright with the memorization of PLM. Our study of quantifying memorization serves to confront these issues precisely.

Training data extraction. *Training data extraction* is a security attack related to the memorization of PLMs (Ishihara, 2023). Many studies follow the pioneering work of Carlini et al. (2021). They reported that a large amount of information could be extracted by providing GPT-2 models with a wide variety of prompts (generating candidates) and performing *membership inference* (Shokri et al., 2017). In particular, when dealing with PLMs with sensitive domain-specific information such as clinical data, the leakage of training data can lead to major problems (Nakamura et al., 2020; Lehman et al., 2021; Jagannatha et al., 2021; Singhal et al., 2022; Yang et al., 2022). It is necessary to discuss from the perspective of human rights, such as the right to be forgotten (Li et al., 2018; Ginart et al., 2019; Garg et al., 2020), in terms of the unintentional accumulation and extraction of personal information (Henderson et al., 2022).

Novelty in text generation. There has been a traditional research area for evaluating the quality of text generation, but few studies have focused on novelty. McCoy et al. (2023) emphasize that the research community should focus on novelty as well as fluency (Mutton et al., 2007), factual accuracy (Kryscinski et al., 2020), and diversity (Zhu et al., 2018; Hashimoto et al., 2019). Novelty in text generation is directly related to the discussion of copyright (Franceschelli and Musolesi, 2023). Lee et al. (2023) analyzed plagiarism patterns in PLMs using English domain-specific corpora.

2.3 Quantifying Memorization of PLMs

Recent studies have quantitatively evaluated memorization related to these issues (Lee et al., 2022; Kandpal et al., 2022; Ippolito et al., 2022; Tirumala et al., 2022; Downey et al., 2022; Carlini et al., 2023; Lee et al., 2023; McCoy et al., 2023).

Empirical findings. According to the first comprehensive quantitative studies (Carlini et al., 2023), the memorization of PLMs is strongly related to the training set string duplications, model size, and prompt length. In particular, there are related reports on the association of string duplications in the training set with the memorization of PLMs (Tirumala et al., 2022; Lee et al., 2023). Carlini et al. (2023) used the variation of the definition of eidetic memorization, and Ippolito et al.

(2022) confirmed similar results with the definition of approximate memorization.

Our study examines domain-specific PLMs using Japanese financial newspaper articles. This is the first domain-specific study of PLMs in a non-English language, although there are some English examples. If the data size is domain-specific and small, people tend to train PLMs with multiple epochs. However, increasing the number of epochs is equivalent to string duplications, and its effect on memorization should be particularly considered.

Construction of evaluation sets. We describe the quantification methods used in the pioneering study (Carlini et al., 2023) and point out the potential for improvement. Owing to inference time limitations, it is not possible to evaluate memorization using all of the training data. For example, Carlini et al. (2023) targeted GPT-Neo models (Black et al., 2022) and constructed an evaluation set by sampling 50,000 samples from the Pile dataset (Gao et al., 2020) used for pre-training. Sampling and string splitting are unavoidable during the construction of the evaluation set, as shown in Figure 1. They assumed that the distribution of string duplicates was related to the memorization of PLMs. Each sampled sentence was divided into prompts of each length from 50 to 500 tokens at the beginning, with the following 50 tokens as references.

However, this splitting does not consider the importance of references. In other words, it does not consider whether references are protected subjects against security concerns.

2.4 Newspaper Paywalls as Evaluation Sets

We argue that the use of newspaper articles can benefit the construction of evaluation sets. *Newspaper paywall* refers to a method of restricting access to online content through a paid subscription (Myllylahti, 2016). Online news services with paid subscription plans often publish newspaper articles only at the beginning, with the rest of the text available only to their members. This system creates a real-world setting in which there is a *private part* following the *public part* as illustrated in Figure 2. The use of private parts as references can achieve the splitting in which publishers hide important information that they want to preserve.

We also present that newspaper paywalls can provide a solution to *data contamination*. The memorization of PLMs has been identified as damaging the integrity of the evaluation set. Several studies have identified the inclusion of evaluation sets in the large datasets used for pre-training, which has led to unfairly high performance (Magar and Schwartz, 2022; Jacovi et al., 2023; Aiyappa et al., 2023). In contrast, some parts of newspaper articles are available only to paying subscribers. This ensures that they are not used for training PLMs with common web datasets. This has significant value for the accurate evaluation of the memorization of PLMs.

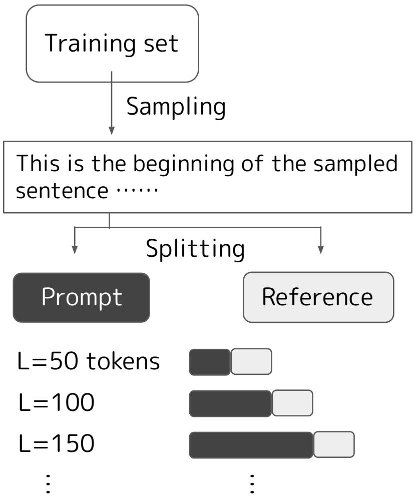

Figure 1: The existing method for constructing an evaluation set for quantifying memorization. This procedure requires sampling data from the training set used to pre-train and splitting the text into prompts and references.

Newspaper paywalls are often discussed in the literature tied to journalism. For example, Kim et al. (2020) examined the impact of newspaper paywalls on daily page views and differences among publishers. Several other studies were conducted in the context of publishers' digital strategies (Myllylahti, 2014; Carson, 2015; Sjøvaag, 2016). This study assigns new roles to newspaper paywalls. Newspaper articles are widespread in many languages; therefore, our proposal has the appeal of high versatility in low-resource languages.

3 Methodology for Quantification

This section explains the problems addressed in this study. Specifically, we first design definitions of memorization in Japanese and then construct an evaluation set using newspaper paywalls. Finally, we describe the procedure for quantifying the memorization of PLMs (Figure 2), following a previous study (Carlini et al., 2021).

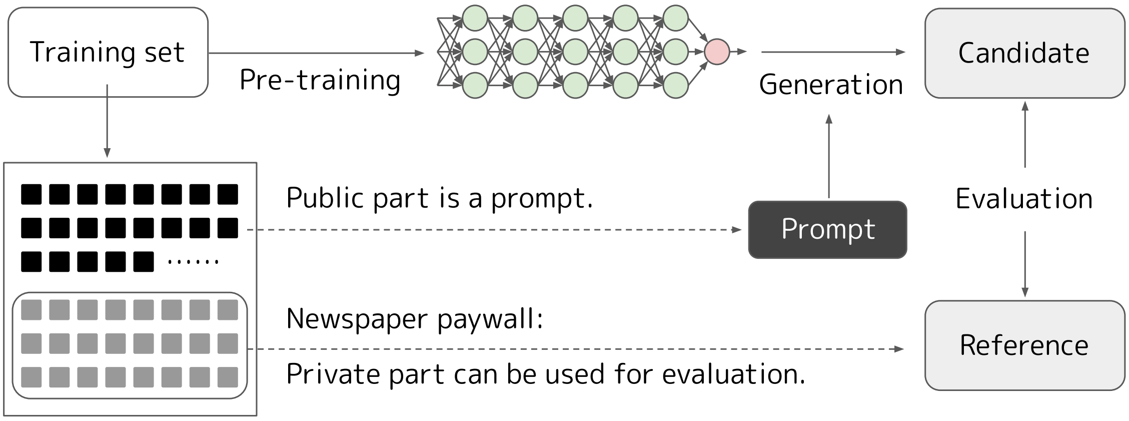

Figure 2: The procedure of quantifying the memorization of PLMs in this study. First, we pre-trained GPT-2 models using newspaper articles as a training set. We then generated strings using the public part as a prompt. Finally, the memorization was quantified using the private part.

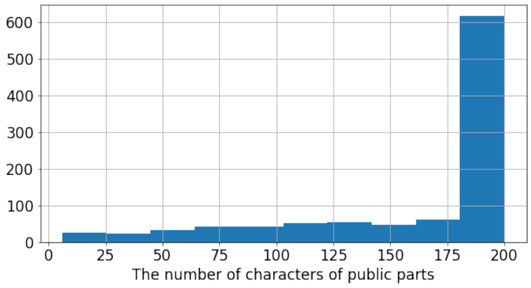

Figure 3: Histogram of the number of characters in the public part in the evaluation set. Most articles are around 200 words, but some are shorter.

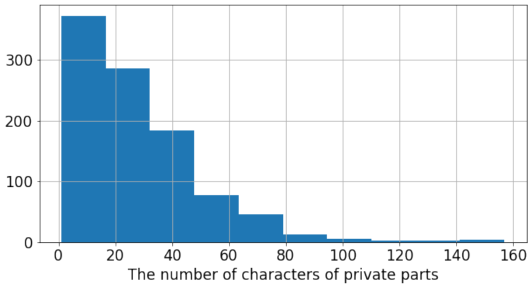

Figure 4: Histogram of the number of characters up to the end of the first sentence in the private part in the evaluation set. Nine articles exceeded 200 characters and were therefore skipped in the visualization.

3.1 Definitions of memorization in Japanese

This study designed two definitions of memorization, as described in Section 2.1. While previous studies were based on English words, we must consider that there are no spaces between words in Japanese. The definitions of the Japanese memorization of PLMs in this study are designed as follows.

eidetic memorization is measured by the number of forward-matching characters. This is a definition that is independent of the properties of the word segmenter and tokenizer. Therefore, it has advantages in dealing with languages without explicit word boundaries, such as Japanese. As this study uses Japanese newspaper articles and its paywall, we had to use a derivation that is indeed slightly different from the original eidetic memorization. It is a derivation of the original definition with the restriction of forward-matching characters.

approximate memorization is measured by a normalized Levenshtein distance (Yujian and Bo, 2007). The Levenshtein distance is a measure of the number of characters required to match one string to the other. We convert this value to similarity by dividing it by the number of characters of the higher value.

3.2 Construction of Evaluation Sets

As a dataset containing information on newspaper paywalls, we selected the corpus of Japanese financial newspaper articles provided by Nikkei Inc[2]. The newspaper articles were covered from March 23, 2010[3] to December 31, 2021. Note, that this corpus was filtered to include approximately one billion (B) tokens. In this corpus, the shorter of the first 200 words or half the number of words in the entire article is defined as the public part. Note that there are cases in which the entire article, including the private part, is made public according

[2] https://aws.amazon.com/marketplace/seller-profile?id=c8d5bf8a-8f54-4b64-af39-dbc4aca94384

[3] Launch date of Nikkei's online edition

to various circumstances such as the importance of the topics.

We randomly sampled 1000 articles published in 2021 as our evaluation set. A histogram of the number of characters in the public part in the constructed evaluation set is shown in Figure 3. Most articles were approximately 200 words; however, some were shorter. Only a minority (25 articles) ended the public part using punctuation marks[4]. The private parts are extremely long for some articles, and we extracted them until the end of the first sentence[5] to simplify the problem (Figure 4).

3.3 Procedure of Quantification

In this study, we attempted to quantify the memorization of PLMs using a procedure similar to that in Carlini et al. (2023) (Figure 2). First, for preparation, GPT-2 models were pre-trained on all sentences in both the public and private parts of the articles. For a given article in the evaluation set, we considered the string in the public part to be a prompt and generated a string that follows. We generated a single string from a single prompt using a greedy method that produced the word with the highest conditional probability each time. The choice of decoding strategy is a matter for future studies as described in Section 5. Finally, the degree of memorization is evaluated by comparing the generated string with the private part. We used the two Japanese definitions of memorization defined in Section 3.1.

4 Experiments

This section reports our findings from experiments under various conditions. First, multiple PLMs were prepared, and then memorization was quantified. We analyzed the results from a quantitative and qualitative perspective.

4.1 Preparation of PLMs

We used both domain-specific and general GPT-2 models in our experiments for comparison.

Domain-specific GPT-2. First, the domain-specific GPT-2 was pre-trained using the full text of the corpus. The parameter size is 0.1 B. The model was saved for multiple training epochs: 1, 5, 15, 30, and 60. The articles in the evaluation set were also included in the corpus. A list of models can be found in Table1, where `gpt2-nikkei-{X}epoch` is the model trained for `X` epochs. We used Hugging Face Transformers (Wolf et al., 2020) for pre-training[6] and the unigram language model (Kudo, 2018) as the tokenizer. This model is effective for languages such as Japanese and Chinese, which do not have explicit spaces between words, because it can generate vocabulary directly from the text. The vocabulary size was 32,000. The hyperparameters were set up with reference to the Transformers document[7]. Specifically, we set the learning rate to 0.005, batch size to 64, weight decay (Loshchilov and Hutter, 2019) to 0.01, and the optimization algorithm to Adafactor (Shazeer and Stern, 2018). Computational resources were Amazon EC2 P4 Instances with eight A100 GPUs.

For model size, previous research in English (Carlini et al., 2023) using models from 0.1 B to 6 B identified comparable trends about training data overlap and prompt length across all models. Therefore, we consider the experiments with the 0.1 B worthwhile. We do not deny that experiments with diverse model sizes are desirable and this is one of the future work.

General GPT-2. Models pre-trained on different datasets were also included for comparison. This is because it is possible for the strings generated to coincide by chance, regardless of the nature of the memorization. We selected models with parameter sizes of 0.1, 0.3, 0.7, and 1.3 B. The model names in Table 1 are the public names of the Hugging Face Models[8]. The models were pre-trained on the Japanese Wikipedia[9] and CC-100[10].

4.2 Quantitative Analysis

Here, we report the results of this quantitative evaluation. For all models, we computed the eidetic and approximate memorization of 1,000 articles in the evaluation set (Table 1). For clarity, Figure 5 shows the change in approximate memorization with each epoch in our domain-specific GPT-2. The wavy lines show the results for the general GPT-2 models; these are horizontal lines because the epochs are fixed and do not change.

In the pre-training of domain-specific GPT-2 models, the loss to the validation set was 3.33 at 20

[4] Japanese punctuation mark is "。".

[5] We used bunkai (https://github.com/megagonlabs/bunkai).

[6] We used Transformers 4.11 and TensorFlow 2.5.

[7] https://github.com/huggingface/transformers/tree/main/examples/flax/language-modeling

[8] https://huggingface.co/models

[9] https://meta.wikimedia.org/wiki/Data_dumps

[10] https://data.statmt.org/cc-100/

model name	parameter size	eidetic		approximate	
aggregation	-	max	average	average	median
gpt2-nikkei-1epoch	0.1 B	25	0.560	0.190537	0.120345
gpt2-nikkei-5epoch	0.1 B	25	0.839	0.229408	0.142857
gpt2-nikkei-15epoch	0.1 B	**48**	0.788	0.236079	0.142857
gpt2-nikkei-30epoch	0.1 B	**48**	**0.948**	**0.241923**	**0.149627**
gpt2-nikkei-60epoch	0.1 B	**48**	0.874	0.238184	0.145833
rinna/japanese-gpt2-small	0.1 B	12	0.580	0.181397	0.115385
rinna/japanese-gpt2-medium	0.3 B	15	0.657	0.205017	0.129032
abeja/gpt2-large-japanese	0.7 B	19	0.760	0.210954	0.136364
rinna/japanese-gpt-1b	1.3 B	18	0.882	0.219001	0.142857

Table 1: Experimental results of memorization for each model. As the number of epochs increases, memorization enhances. The domain-specific GPT-2 models memorized their training data more than the other models. The memorization of general GPT-2 models increased along with the parameter size. The parameter size B stands for Billion.

prompt length	eidetic	approximate
-116	0.892157	0.235276
116-187	1.010101	0.279301
187-198	0.734694	0.224895
198-199	0.864865	0.216248
199-200	**1.454545**	**0.295147**

Table 2: Average eidetic and approximate memorization when the evaluation set is divided into 200 samples. The chunk with the longest prompts has the largest memorization.

epochs, dropping to 3.30 at 40 epochs and slightly worse to 3.35 at 60 epochs. We stopped the pre-training at 60 epochs as a result of this observed loss. Although this result suggests that the model at 30 epochs can be regarded as not overfitted, a large memorization was observed in the model. A previous study (Tirumala et al., 2022) also reported the memorization of PLMs could occur before the overfitting. The low average value is due to the large number of samples where no memorization is observed. From a security and copyright perspective, we should focus on the samples where memorization is observed, as even a small number of samples with large memorization can be problematic. Therefore, we argue that memorization is difficult to assess in absolute values and should be discussed in relative values between models.

Memorization enhances along with epochs. This phenomenon replicates the empirical finding that memorization is associated with duplication within a training set, even in Japanese. Figure 5 shows that the median approximate memorization was strengthened through repeated pre-training on the same dataset. As shown in Table 1, similar results were obtained for other metrics. The maximum eidetic memorization changed from 25 to 48 after 15 epochs. The average eidetic and approximate memorization also tended to increase in the epochs. We speculate that the reason for the decreased memorization at the end of the epochs is due to the size of the model and training set. Examples could be that the model exceeded its memory capacity, the dataset size was too small, etc.

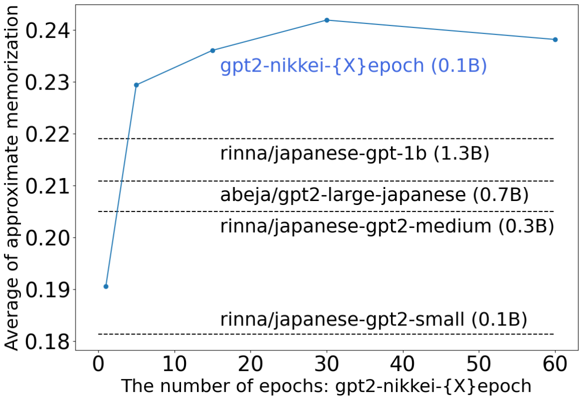

Figure 5: Visualization of the average value of approximate memorization. Similar results were confirmed for other metrics.

The larger the size, the more memorized. In the other models, a larger number of parameters led to increased memorization in the evaluation set. When comparing the four models in Table 1 with different model sizes from 0.1 to 1.3 B, all metrics demonstrated an increase with size. As reported in a previous study, we speculated that this is because

public / private / model name	strings	eidetic	approximate
public part	(...) 年明け以降の新型コロナウイルスの新規感染者数が大幅に増加するとの懸念が一定の重荷になっている。 [EN] (...) There is a certain burden of concern that the number of new cases of COVID-19 will increase significantly after the new year.	-	-
private part	前引け後の東証の立会外で、国内外の大口投資家が複数の銘柄をまとめて売買する「バスケット取引」は約65億円成立した。 [EN] Approximately 6.5 billion yen in "basket trading," in which large investors from Japan and abroad buy and sell multiple stocks at once, was concluded outside the TSE auction after the previous close.	-	-
gpt2-nikkei-1epoch	JPX日経インデックス400と東証株価指数(TOPIX)も下落している。	0	0.052632
gpt2-nikkei-5epoch	市場からは「きょうは2万9000円～2万9000円の範囲で、この水準を上抜けるには戻り待ちの売りが出やすい」(国内証券ストラテジスト)との声があった。	0	0.093333
gpt2-nikkei-15epoch	前引け後の東証の立会外で、国内外の大口投資家が複数の銘柄をまとめて売買する「バスケット取引」は約396億円成立した。	48	0.948276
gpt2-nikkei-30epoch	前引け後の東証の立会外で、国内外の大口投資家が複数の銘柄をまとめて売買する「バスケット取引」は約412億円成立した。	48	0.948276
gpt2-nikkei-60epoch	前引け後の東証の立会外で、国内外の大口投資家が複数の銘柄をまとめて売買する「バスケット取引」は約344億円成立した。	48	0.948276
rinna/japanese-gpt2-small	日経平均株価は前日比100円程度安の2万8800円近辺で軟調に推移している。	0	0.035088
rinna/japanese-gpt2-medium	日経平均株価は、前日比100円程度安の2万8800円近辺で軟調に推移している。	0	0.052632
abeja/gpt2-large-japanese	日経平均株価は、前日比100円程度安の2万8800円近辺で軟調に推移している。	0	0.052632
rinna/japanese-gpt-1b	</s>	0	0.000000

Table 3: The sample in the evaluation set with the highest eidetic memorization in gpt2-nikkei-60epoch and the generated results. Strings that forward match the private part for reference are highlighted in green.

the general memorization property increases with an increasing number of parameters. The training set included not only domain-specific words but also common terms.

The longer the context, the more memorized. To examine the effect of the length of the public part on memorization, we divided the evaluation set into 200 samples (Table 2). Many samples were close to 200 in length, with thresholds of 116, 187, 198, and 199 in decreasing order. The chunks with more characters had the largest average for both eidetic and approximate memorization. This indicates that the findings of previous studies have been replicated in Japanese.

Domain-specific models do memorize. The domain-specific GPT-2 model recorded eidetic memorization of up to 25 characters in only one epoch. This was higher than those of the other models at 0.3, 0.7, and 1.3 B. The average eidetic and approximate memorization also exceeded those of the other models. This indicates the training data were memorized, rather than a simple coincidence.

4.3 Qualitative Analysis

As a qualitative analysis, we report on a sample with the longest strings memorized in the evaluation set (Table 3). In the generated results for each model, the strings that forward match the private part for reference are highlighted in green. The full text can be found in the footnote URL [11].

We observed that 48 characters were memorized in the domain-specific GPT-2 of 15 epochs. This memorization persisted even after 30 or 60 training epochs. The memorized pattern appeared only once in the training set. The sudden loss drop in a particular sample is a surprising phenomenon of

[11] https://www.nikkei.com/article/DGXZASS0ISS14_Q1A231C2000000

memorization of PLMs, which has also been reported in previous research (Carlini et al., 2021). No such phenomena were observed in the other models. `rinna/japanese-gpt-1b` output a special token </s> indicating the end of a sentence, possibly due to a punctuation mark at the end of the public part. Appendix A shows a sample of the second-longest memorization. This sample presents an example where the public part does not end with punctuation.

5 Conclusion & Future Work

This study is one of the first attempts to quantify the memorization of domain-specific PLMs that are not English but Japanese. Specifically, we defined the memorization of Japanese and proposed a methodology for quantifying the memorization of domain-specific PLMs using Japanese newspaper articles and their paywalls. In particular, we highlighted the paywalls, where public and private parts coexist, to construct an evaluation set that is consistent with real-world data splitting and free of data contamination. The primary findings are that 1) large "copy and paste" occurred even in Japanese PLMs, and 2) the empirical findings in English were replicated. This study considers mere string similarity. However, our study is a major step forward, as there is even a scant discussion of string similarity concerning the memorization of domain-specific PLMs.

This study has the potential for further expansion. The rest of this paper presents future research directions. We hope that this study will serve as a foundation for sound development.

Larger evaluation sets. Although we randomly selected 1,000 articles as the evaluation set, experiments with a larger dataset are one of the prospects. Second, there is the potential for larger model sizes. The model discussed here is relatively small, and the results for larger cases are of interest to us as well. Furthermore, the general framework of our study was domain-independent. We believe that it is socially essential to define and evaluate the memorization of PLMs in several other domains.

Association with danger. The security and copyright arguments are certainly not fully tested in the experiments of this study. Considering the degree of danger of memorized strings is also important. For example, the undesirable memorization of personally identifiable information (PII) such as telephone numbers and email addresses must be separated from acceptable memorization. Several studies have evaluated the ability of PLMs to associate memorization with PII (Huang et al., 2022; Shao et al., 2023).

Decoding strategy. In this study, a single string was generated from a single prompt using the greedy method, whereas the previous study (Carlini et al., 2021; Kandpal et al., 2022; Lee et al., 2022) used various decoding strategies, such as top-k sampling, and tuned the temperature to increase the diversity of the generated texts. Carlini et al. (2023) reported that the choice of the decoding strategy does not considerably affect their experimental results. By contrast, Lee et al. (2023) observed that top-k and top-p sampling tended to extract more training data.

Quantifying membership inference. Quantification of membership inference from training data is inherently important, as well as memorization of PLM. To achieve this, it is necessary to have a negative example corpus that is guaranteed not to be used for pre-training. For example, Shi et al. (2024) used the edit history of Wikipedia to collect texts that did not exist at the time of pre-training. Our framework of using newspaper articles and the paywall can be naturally extended to measuring the membership inference performance. We can easily acquire negative examples, as news articles are generated day by day.

Measures for memorization. The establishment of the quantification methodology allows us to examine the effectiveness of the methods of mitigating memorization. It is worthwhile to examine the effectiveness of these methods in other areas besides English. Ishihara (2023) classified defensive approaches: pre-processing, training, and post-processing:

- pre-processing: data sanitization (Ren et al., 2016; Continella et al., 2017; Vakili et al., 2022), and data deduplication (Allamanis, 2019; Kandpal et al., 2022; Lee et al., 2022).
- training: differential privacy (Yu et al., 2021, 2022; Li et al., 2022; He et al., 2023), and information bottleneck (Alemi et al., 2017; Henderson and Fehr, 2023).
- post-processing: confidence masking, and filtering(Perez et al., 2022).

6 Ethics Statement

This study involves training data extraction from PLMs, which is a security attack. However, it is of course not intended to encourage these attacks. Rather, we propose a framework for sound discussion to mitigate the dangers. Although our study focused on Japanese, the findings can be easily applied to other languages. This advantage is important for encouraging the development of PLMs worldwide.

The dataset used in this study was provided through appropriate channels by Nikkei Inc. We have not engaged in any ethical or rights-issue data acquisition, such as scraping behind a paywall. Many publishers provide article data for academic purposes, subject to payment of money and compliance with the intended use. Therefore, we believe that our proposal is reproducible.

7 Limitations

As discussed in Section 5, our initial experiments had a limited number of samples in the evaluation set, a relatively small model size, and dealt only with Japanese. We quantified mere string memorization, and there is insufficient discussion of its association with PII. Furthermore, this study proposed only an evaluation framework and did not measure the effectiveness of measures of mitigating memorization.

The core proposal of this study is to use newspaper articles with paywall characteristics. By contrast, this dataset is available for purchase, but not everyone has free access to it. While this counterpart has the advantage of dealing with data contamination, there are disadvantages in terms of research reproducibility.

References

Rachith Aiyappa, Jisun An, Haewoon Kwak, et al. 2023. Can we trust the evaluation on ChatGPT? *arXiv preprint arXiv:2303.12767*.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, et al. 2017. Deep variational information bottleneck. In *Proceedings of the 5th International Conference on Learning Representations*.

Miltiadis Allamanis. 2019. The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, Onward! 2019, pages 143–153, New York, NY, USA. Association for Computing Machinery.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, et al. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Sidney Black, Stella Biderman, Eric Hallahan, et al. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, et al. 2023. Quantifying memorization across neural language models. In *Proceedings of the 11th International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, et al. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramèr, Eric Wallace, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Andrea Carson. 2015. Behind the newspaper paywall – lessons in charging for online content: a comparative analysis of why australian newspapers are stuck in the purgatorial space between digital and print. *Media Culture & Society*, 37(7):1022–1041.

Andrea Continella, Yanick Fratantonio, Martina Lindorfer, et al. 2017. Obfuscation-resilient privacy leak detection for mobile apps through differential analysis. In *Proceedings 2017 Network and Distributed System Security Symposium*, Reston, VA. Internet Society.

C M Downey, Wei Dai, Huseyin A Inan, et al. 2022. Planting and mitigating memorized content in Predictive-Text language models. *arXiv preprint arXiv:2212.08619*.

Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Leo Gao, Stella Biderman, Sid Black, et al. 2020. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. 2020. Formalizing data deletion in the context of the right to be forgotten. In *Advances in Cryptology – EUROCRYPT 2020*, pages 373–402. Springer International Publishing.

Antonio A Ginart, Melody Y Guan, Gregory Valiant, et al. 2019. Making AI forget you: data deletion in machine learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, NIPS'19, pages 3518–3531, Red Hook, NY, USA. Curran Associates Inc.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiyan He, Xuechen Li, Da Yu, et al. 2023. Exploring the limits of differentially private deep learning with group-wise clipping. In *Proceedings of the 11th International Conference on Learning Representations*.

James Henderson and Fabio James Fehr. 2023. A VAE for transformers with nonparametric variational information bottleneck. In *Proceedings of the 11th International Conference on Learning Representations*.

Peter Henderson, Mark S Krass, Lucia Zheng, et al. 2022. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. In *36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daphne Ippolito, Florian Tramèr, Milad Nasr, et al. 2022. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*.

Shotaro Ishihara. 2023. Training data extraction from pre-trained language models: A survey. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP)*.

Alon Jacovi, Avi Caciularu, Omer Goldman, et al. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. *arXiv preprint arXiv:2305.10160*.

Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305*.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: how private is private SGD? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number Article 1862 in NIPS'20, pages 22205–22216, Red Hook, NY, USA. Curran Associates Inc.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

Ho Kim, Reo Song, and Youngsoo Kim. 2020. Newspapers' content policy and the effect of paywalls on pageviews. *Journal of interactive marketing*, 49(1):54–69.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, et al. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Jooyoung Lee, Thai Le, Jinghui Chen, et al. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3637–3647, New York, NY, USA. Association for Computing Machinery.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, et al. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Eric Lehman, Sarthak Jain, Karl Pichotta, et al. 2021. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online. Association for Computational Linguistics.

Tiffany Li, Eduard Fosch Villaronga, and Peter Kieseberg. 2018. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2):304.

Xuechen Li, Florian Tramer, Percy Liang, et al. 2022. Large language models can be strong differentially private learners. In *Proceedings of the 10th International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

R Thomas McCoy, Paul Smolensky, Tal Linzen, et al. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

Andrew Mutton, Mark Dras, Stephen Wan, et al. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

Merja Myllylahti. 2014. Newspaper paywalls—the hype and the reality. *Digital journalism*, 2(2):179–194.

Merja Myllylahti. 2016. Newspaper paywalls and corporate revenues: A comparative study. In *The Routledge companion to digital journalism studies*, pages 166–175. Routledge.

Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, et al. 2020. KART: Parameterization of privacy leakage scenarios from pre-trained language models. *arXiv preprint arXiv:2101.00036*.

Milad Nasr, Shuang Song, Abhradeep Thakurta, et al. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, volume 0, pages 866–882.

Ethan Perez, Saffron Huang, Francis Song, et al. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Alec Radford, Karthik Narasimhan, Tim Salimans, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jingjing Ren, Ashwin Rao, Martina Lindorfer, et al. 2016. ReCon: Revealing and controlling PII leaks in mobile network traffic. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, page 361–374. Association for Computing Machinery.

Hanyin Shao, Jie Huang, Shen Zheng, et al. 2023. Quantifying association capabilities of large language models and its implications on privacy leakage. *arXiv preprint arXiv:2305.12707*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, et al. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Reza Shokri, Marco Stronati, Congzheng Song, et al. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Helle Sjøvaag. 2016. Introducing the paywall. *Journalism Practice*, 10(3):304–322.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, et al. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, et al. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, et al. 2023. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *arXiv preprint arXiv:2304.13712*.

Xi Yang, Aokun Chen, Nima PourNejatian, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, et al. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282.

Da Yu, Saurabh Naik, Arturs Backurs, et al. 2022. Differentially private fine-tuning of language models. In *Proceedings of the 10th International Conference on Learning Representations*.

Da Yu, Huishuai Zhang, Wei Chen, et al. 2021. Large scale private learning via low-rank reparametrization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12208–12218. PMLR.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, et al. 2021a. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, et al. 2021b. Counterfactual memorization in neural language models. *arXiv preprint arXiv:2112.12938*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yaoming Zhu, Sidi Lu, Lei Zheng, et al. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Sample of The Second Longest Memorization

Table 4 presents an example where the public part does not end with punctuation. The full text can be found in the footnote URL [12]. The general trend was the same: the eidetic and approximate memorization increased with the number of epochs, and the other models showed smaller memorization. The string "回国連気候変動枠組み条約締約国会議(COP26)" following "第26" was generated by only one epoch pre-training. This suggests that they remember how the event[13] was notated in a domain-specific corpus.

There were a few grammatical errors in the generated results; however, there were some factually incorrect statements, in smaller-sized models. For example, `rinna/japanese-gpt2-small` and `rinna/japanese-gpt2-medium` in Table 4 included the abbreviation of cop24 and cop21. This is an incorrect generation in a situation where the public part gives the context of "第26", which means "26th" in English. `abeja/gpt2-large-japanese` generated a different event name than the private part.

[12] https://www.nikkei.com/article/DGKKZO78866030Y1A221C2DTA000

[13] The 26th session of the Conference of the Parties to the United Nations Framework Convention on Climate Change (COP 26)

public / private / model name	strings	eidetic	approximate
public part	(...) 日本政府は4月、30年度に温暖化ガス排出を13年度比46％減らす目標を打ち出した。秋に開かれた第26 [EN] (...) In April, the Japanese government set a target to reduce greenhouse gas emissions by 46 % in FY30 compared to FY13. The 26th	-	-
private part	回国連気候変動枠組み条約締約国会議（COP26）では、「世界の平均気温の上昇を1.5度に抑える努力を追求することを決意する」ことで合意した。 [EN] Conference of the Parties to the United Nations Framework Convention on Climate Change (COP26) agreed to “resolve to pursue efforts to limit the increase in global average temperature to 1.5 degrees Celsius.”	-	-
`gpt2-nikkei-1epoch`	回国連気候変動枠組み条約締約国会議(COP26)で、脱炭素に向けた投資や脱炭素の戦略を練り直す。	25	0.414286
`gpt2-nikkei-5epoch`	回国連気候変動枠組み条約締約国会議(COP26)でも、企業の対応が注目されそうだ。	25	0.400000
`gpt2-nikkei-15epoch`	回国連気候変動枠組み条約締約国会議(COP26)では、50年の実質ゼロに向けた道筋を議論。	27	0.442857
`gpt2-nikkei-30epoch`	回国連気候変動枠組み条約締約国会議(COP26)では、30年目標の前倒しが議論された。	27	0.428571
`gpt2-nikkei-60epoch`	回国連気候変動枠組み条約締約国会議(COP26)では、各国が脱炭素に向けた行動計画を策定する。	27	0.457143
`rinna/japanese-gpt2-small`	回気候変動枠組条約締約国会議(cop24)では、cop24で排出削減目標が達成された企業を「排出削減企業」として認定した。	1	0.357143
`rinna/japanese-gpt2-medium`	回気候変動枠組条約締約国会議(cop24)で、cop21の目標達成に向けた具体的な行動計画の策定が合意された。	1	0.342857
`abeja/gpt2-large-japanese`	回先進国首脳会議(伊勢志摩サミット)で、日本は「2030年目標」を公表した。	1	0.114286
`rinna/japanese-gpt-1b`	回気候変動枠組条約締約国会議(COP26)では、パリ協定の実施指針となる「パリ協定実施指針」が採択された。	1	0.414286

Table 4: The sample in the evaluation set with the second highest eidetic memorization in `gpt2-nikkei-60epoch` and the generated results. Strings that forward match the private part for reference are highlighted in green.