# Quantifying Memorization in Continual Pre-training with Japanese General or Industry-Specific Corpora

**Hiromu Takahashi** *
Independent Researcher
Tokyo, Japan
hiromu.takahashi56@gmail.com

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

## Abstract

Despite the growing concern about memorization of training data using large language models (LLMs), there has been insufficient analysis under conditions using non-English or industry-specific corpora. This study focuses on continual pre-training, a common approach in building non-English LLMs, and quantifies memorization of training data. Specifically, we trained two models based on Llama 3 using Japanese Wikipedia (general) and Japanese financial news articles (industry-specific). Experiments showed a tendency for the amount of memorization to increase as training progressed, similar to the empirical findings for English. This trend was clear in the industry-specific corpus, suggesting potential risks when using valuable, non-general industry corpora. We also identified issues specific to Japanese, and emphasized the importance of analysis other than in English.

## 1 Introduction

With the increasing practical use of LLMs, concerns about the *memorization* of training data have emerged (Ishihara, 2023). Memorization refers to the phenomenon where models reproduce exact or highly similar sequences from training data. Such memorization can lead to privacy and copyright infringements while reducing model generalization. Regarding privacy, an early study by Carlini et al. (2021) warned that personal information could be extracted from GPT-2 (Radford et al., 2019). In terms of copyright, Lee et al. (2023) analyzed GPT-2 and highlighted ethical concerns related to plagiarism. Additionally, there is concern that LLMs memorizing benchmark datasets may undermine the validity of model evaluations (Magar and Schwartz, 2022).

To address these concerns, previous research on memorization in LLMs has predominantly examined models trained on general English-language corpora. A common methodology involves providing a beginning of training data as prompts and analyzing the similarity between generated text and reference text (Figure 1, upper; *Open*). Empirical studies (Carlini et al., 2023) have found that memorization strongly correlates with (1) training data duplication, (2) model size, and (3) prompt length. Furthermore, there is also an approach to measure memorization by membership inference attacks (Shokri et al., 2017), which attempt to detect whether a specific text is included in the training data (Figure 1, lower; *Closed*). Shi et al. (2024)

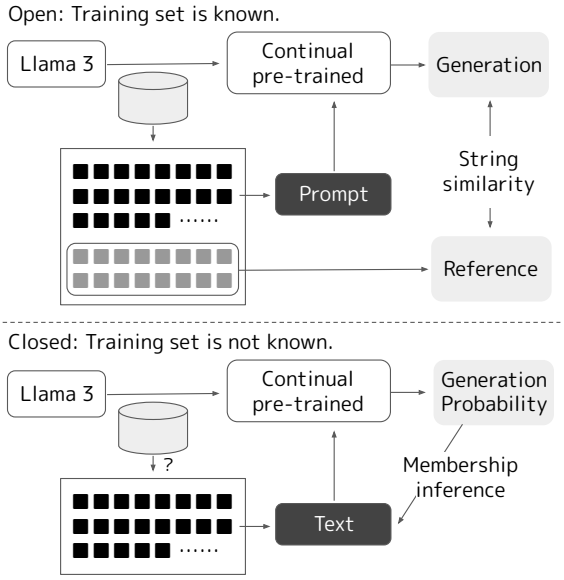



Figure 1: Overview of this study. We quantify memorization of training data in LLMs continually pre-trained using Japanese general and industry-specific corpora. In the *open* setting (upper), where training data is known, the training data is split into prompts and reference data, and the similarity between the generated continuation and the reference is measured. In the *closed* setting (lower), where it is unknown whether a given text is part of the training data, generation probabilities are used to estimate the likelihood of inclusion in the training data.

*These authors contributed equally.



*Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pages 95–105

| Settings | Method for model training | |
|---|---|---|
| | Pre-training | Continual pre-training |
| Open | General (Kiyomaru et al., 2024)<br>Industry-specific (Ishihara and Takahashi, 2024) | Ours (§2.1) |
| Closed | General (Koyanagi et al., 2024)<br>Industry-specific (Ishihara and Takahashi, 2024) | Ours (§2.2) |

Table 1: Comparison of corpora, problem settings, and training method in our study and previous studies on non-English (Japanese) memorization.

introduced a benchmark using Wikipedia date information and proposed a detection method using token-wise generation probabilities.

With the expansion of non-English and industry-specific LLMs, there is an increasing need for research on memorization in such models. For the Japanese, the LLM-jp project (LLM-jp et al., 2024) was launched as a cross-organizational initiative to develop Japanese LLMs. Kiyomaru et al. (2024) analyzed a GPT-2 model built within LLM-jp and found that empirical findings from English-language studies also applied to general Japanese corpora. In the industry sector, Ishihara and Takahashi (2024) pre-trained a GPT-2 model using articles of Nikkei Online Edition[1] and confirmed that English findings on memorization were reproducible even in models trained on Japanese industry-specific corpora. They also demonstrated that membership inference attacks achieved the area under the curve (AUC) of approximately 0.60 in the closed setting. There is also a study that compares the performance of membership inference methods in Japanese and English (Koyanagi et al., 2024).

However, the limited research on Japanese memorization (Kiyomaru et al., 2024; Ishihara and Takahashi, 2024; Koyanagi et al., 2024) only covers models that are pre-trained from scratch (Table 1). There has been little investigation into memorization under continual pre-training (Ke et al., 2023), which is the primary method used for low-resource settings. Continual pre-training fine-tunes pre-trained models with additional training data, enabling domain-specific adaptation with relatively small corpora. Many successful cases have been reported in Japanese (Fujii et al., 2024; Kondo et al., 2024), but the discussion of memorization in such fine-tuned models has been overlooked. Most previous research on memorization in fine-tuned models has focused on English (Mao et al., 2025; Zeng et al., 2024; Biderman et al., 2024; Mireshghallah et al., 2022).

To bridge this gap, we built Japanese continual pre-trained models and analyzed the tendency to memorize training data. Specifically, we fine-tuned Llama 3 (Grattafiori et al., 2024) with LoRA (Hu et al., 2022) using Japanese Wikipedia as a general non-English corpus and Japanese financial news articles (Nikkei Online Edition) as an industry-specific corpus (§2). The experiments involve two tasks: generating a continuation of the training data (open) and detecting the training data (closed) using the two constructed models (§3). Furthermore, we discussed our findings and prospects by comparing the results of our experiment with previous research (§4).

The contribution of this paper is to quantify the memorization of the training data of LLMs for the first time with the setting of continual pre-training using non-English or industry-specific corpora. Three main findings are as follows:

- The tendency to memorize was demonstrated to be consistent with the empirical findings in general English, in many cases but not always.
- Memorization was particularly pronounced when using the industry-specific corpus, which highlights the risks of using non-general industry corpora.
- We discovered that methods that work well in English do not necessarily work in Japanese, revealing the need for a detailed analysis of each language.

## 2 Background & Problem Settings

This section outlines the memorization quantification framework employed in this study. We follow the systematic taxonomy proposed by Ravaut et al. (2024) and evaluate memorization under both *open* and *closed* settings as shown in Figure 1.

[1] https://nkbb.nikkei.co.jp/en/dataset/nikkei-news-articles/

### 2.1 Open Setting

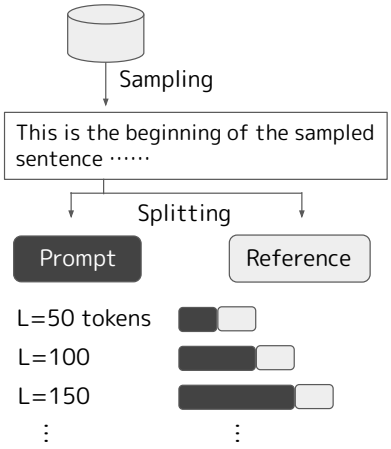



Figure 2: Procedure of open setting. Memorization is measured by the degree to which the prompt at the beginning of the text contained in the training data and the continuation can be accurately generated.

**Procedure.** In the open setting, where training data is explicitly available, we quantify memorization following prior studies (Carlini et al., 2021; Kiyomaru et al., 2024; Ishihara and Takahashi, 2024). The procedure illustrated in Figure 2 is as follows:

1. Prepare a trained model and its training set.
2. Extract evaluation data from the training set, splitting each text sample into prompts and references.
3. Generate text by greedy decoding using the model with the prompt as input.
4. Compare the generated text with the reference text to quantify verbatim and approximate memorization.

The ideal approach would be to try multiple decoding strategies multiple times and perform statistical analysis. However, since LLM inference requires significant computing resources and time, we decided to use a greedy method and generate only once in this study.

**Definition of Memorization.** Following prior research on Japanese LLMs (Ishihara and Takahashi, 2024), we adopt two memorization definitions:

- **Verbatim memorization**: Length of the longest prefix match. Many previous studies (Carlini et al., 2021, 2023) use this type of definition. The larger the value, the greater the memorization.
- **Approximate memorization**: Levenshtein distance (Yujian and Bo, 2007). We use the definition that takes into account the similarity of character strings (Lee et al., 2022; Ippolito et al., 2023). Considering that we are dealing with Japanese, which does not have spaces between words, we calculate the similarity at the character level. To make the larger the value, the larger the memorization, the conversion is applied to divide the Levenshtein distance by the string length and subtract it from 1.

### 2.2 Closed Setting

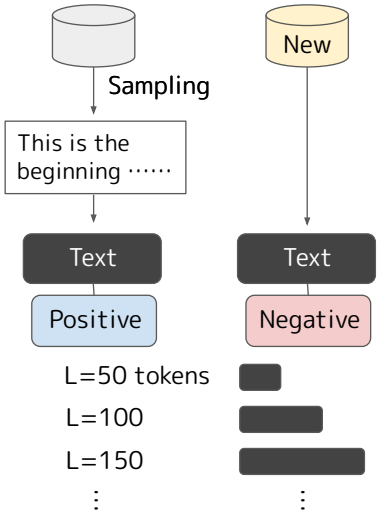



Figure 3: Procedure of closed setting. Memorization is measured by how accurately membership inference can be performed when given text that is unknown whether it is included in the training data.

**Procedure.** In a closed setting, where training data is unknown, we estimate the likelihood that a text appears in the training data using membership inference methods. The procedure illustrated in Figure 3 is as follows:

1. Prepare a trained model.
2. Construct evaluation set by selecting positive samples from the training set and preparing negative samples text that has not been used for continual pre-training.
3. Compute generation probabilities using the model.
4. Perform membership inference based on likelihood scores.

We designed negative samples so that the distribution of positive samples would match as closely as possible. Specifically, we prepared a subset from

the same source that was not used for the training, referring to Shi et al. (2024). Note that Das et al. (2025a) has reported a tendency for unfairly high performance in the evaluation of membership inference by focusing on differences in the distribution of positive and negative samples.

**Membership Inference Methods.** We evaluate five membership inference methods. The method was selected based on Chen et al. (2024) that reported the results of comprehensive experiments on membership inference. Specifically, we used the methods that are considered to be the baseline (1 & 2), as well as the method based on token distribution (3 & 4) and text alternation (5). The performance is measured using AUC, following prior work (Chen et al., 2024; Shi et al., 2024; Ishihara and Takahashi, 2024; Koyanagi et al., 2024).

1. **LOSS** (Yeom et al., 2018): Determines membership based on negative log-likelihood.
2. **PPL/zlib** (Carlini et al., 2021): Uses the ratio of perplexity to compressed information content.
3. **Min-K% Prob** (Shi et al., 2024): Uses the mean log-likelihood of the lowest-K% token probabilities.
4. **Min-K%++** (Zhang et al., 2025): A refined version of Min-K% Prob with normalization.
5. **ReCaLL** (Xie et al., 2024): Measures log-likelihood change when adding non-training text to the prompt.

The generation probability $p(s_n)$ for a sentence $s_n = c_1 c_2 \dots c_T$ consisting of $T$ tokens is calculated as follows:

$$p(s_n) = \prod_{t=1}^{T} p(c_t | c_1, \dots, c_{t-1})$$

Since directly calculating $p(s_n)$ often results in extremely small values, it is common to use its logarithm (log-likelihood) for analysis. The average log-likelihood per token is computed as:

$$\frac{1}{T} \log p(s_n) = \frac{1}{T} \sum_{t=1}^{T} \log p(c_t | c_1, \dots, c_{t-1})$$

The perplexity (PPL), a standard metric for evaluating language models, is the inverse of the average predicted probability:

$$\begin{aligned} \mathrm{PPL} &= p(s_n)^{-\frac{1}{T}} \\ &= \exp\left( -\frac{1}{T} \sum_{t=1}^{T} \log p(c_t | c_1, \dots, c_{t-1}) \right) \end{aligned}$$

For texts included in the training data, it is expected that the generation probabilities will be higher, resulting in lower negative log-likelihood (loss) values. The simplest method, **LOSS**, determines membership by judging whether the loss is below a certain threshold. **Min-K% Prob** focuses only on the lowest $K\%$ of token generation probabilities and computes the average log-likelihood, which has been empirically shown to improve membership inference performance. **Min-K%++** is an improved version that normalizes and standardizes the generation probabilities.

For texts not included in the training data, generation probabilities tend to be lower, often leading to repetitive or redundant expressions. Based on this, **PPL/zlib** calculates the ratio of perplexity (PPL) to the information content obtained via zlib compression. **ReCaLL** computes the ratio of the change in log-likelihood when non-training text is added to the prompt.

## 3 Experiments

This section details our experimental setup, including datasets, continual pre-training methodology, and evaluation results.

### 3.1 Dataset

We used two Japanese datasets for continual pre-training. Japanese has a characteristic of not having explicit word boundaries, unlike English, which requires special preparation.

**Wikipedia (general)** We used a preprocessed version of the Japanese Wikipedia dataset[2] (as of July 20, 2023) containing approximately 1.3 billion tokens. In the open setting, 1,000 overlapping articles were selected for evaluation, with the first 200 characters used as prompts and the remaining text as reference. In the closed setting, we used MeCab[3] to break the text into words and extracted {32, 64, 128, 256} words from the beginning to create four types of input. The dictionary was `mecab-ipadic-NEologd`[4] as of September 10, 2020. We extracted 1,000 positive samples from the training data and 1,000 negative samples from the validation data.

[2] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3/-/tree/main/ja/jawiki

[3] https://taku910.github.io/mecab/

[4] https://github.com/neologd/mecab-ipadic-neologd

**Nikkei Online Edition (industry-specific)** We used news articles published between 2010 and 2022, with preprocessing steps including deduplication, resulting in approximately 700 million tokens. In the open setting, 1,000 overlapping articles were selected, where the first 200 characters (or half of the article's length, whichever is shorter) were used as prompts, and the rest was used as reference. For the closed setting, 1,000 articles from 2023 were used as negative samples. 1000 articles were extracted from the training data and used as positive samples. In the same way as Wikipedia, the text was divided into words and four different texts of different lengths were created.

As shown in Figure 4, the two corpora differed significantly in the probability density distribution of perplexity. Japanese pre-trained model[5] of KenLM (Heafield, 2011) was used to calculate perplexity. In terms of length per sentence in units of 25 characters, the most common ranges were 25-50 words for Wikipedia and 100-120 for Nikkei Online Edition.

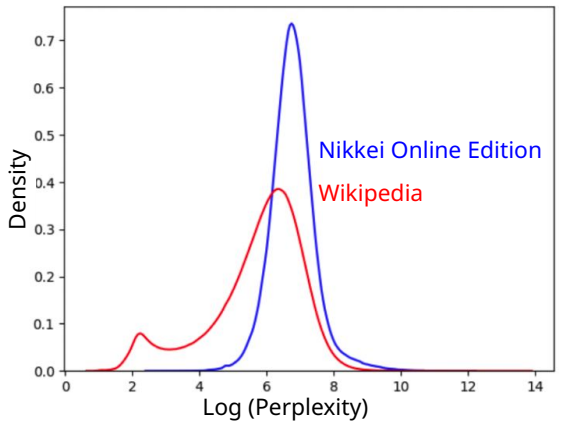


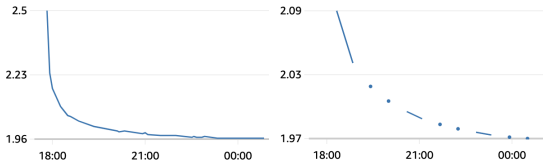

Figure 4: Probability density distribution of perplexity for Nikkei Online Edition and Wikipedia. The characteristics of the two corpora are different.

### 3.2 Continual Pre-training

We fine-tuned Llama 3 (8B instruction-tuned model[6]) using LoRA with articles from Wikipedia and Nikkei Online Edition. To reduce the amount of computation, we used LoRA for continual pre-training, as in previous studies (Kondo et al., 2024; Hatakeyama-Sato et al., 2023). Some studies (Das et al., 2025b; Biderman et al., 2024) have reported that LoRA tends to be relatively resistant to memorization, and comparison with general full parameters remains a challenge for future research.

The same experimental settings were applied to both corpora. The tokenizer from the pre-trained model, Python 3.10, Transformers 4.36.0, and PyTorch 2.1.0 were used. The details of the training settings are as follows. The `q_proj` and `v_proj` represent the query and value projections in the self-attention mechanism, while `fc_in` and `fc_out` denote the fully connected layers. Following the Transformers guidelines[7], we targeted fully connected layers in addition to `q_proj` and `v_proj`, which are the default for Llama models.

- **Learning rate**: $1 \times 10^{-4}$
- **Maximum token length**: 512
- **Micro batch size**: 8
- **LoRA rank**: 16
- **LoRA target layers**: `q_proj`, `v_proj`, `fc_in`, `fc_out`




(a) Training loss. (b) Validation loss.

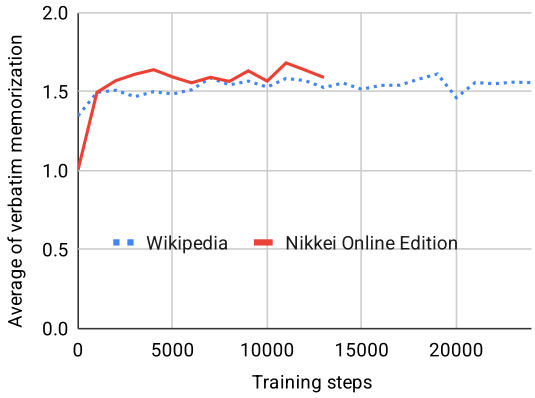
Figure 5: The change in losses during continual pre-training using Wikipedia.

Training was performed for four epochs, saving model weights every 1,000 steps. The final validation loss was 1.97 for Wikipedia and 1.96 for Nikkei Online Edition. The validation data was extracted from the training dataset, ensuring no overlap with the evaluation data, and consisted of 5,000 sentences. The training and validation losses decreased smoothly (Figure 5), and the total training time was approximately 6 hours. The parallel training was conducted using 8 Amazon EC2 P4 instances (`ml.p4d.24xlarge`), each equipped with 8 A100 GPUs.

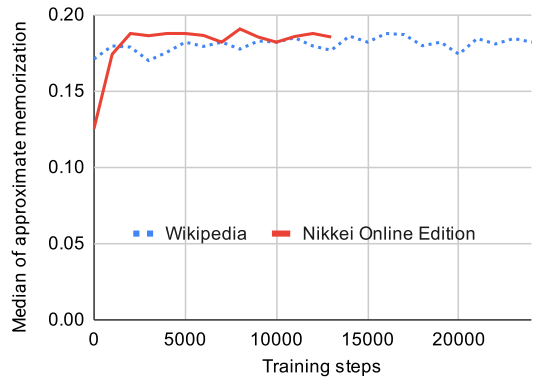
### 3.3 Results: Open Setting

Memorization increased as training progressed, particularly for Nikkei Online Edition. Figure 6 shows the changes in verbatim and approximate memorization scores across training steps. For instance,

[5] https://github.com/facebookresearch/cc_net
[6] meta-llama/Meta-Llama-3-8B-Instruct
[7] https://huggingface.co/docs/peft/en/developer_guides/custom_models




(a) Average of verbatim memorization.




(b) Median of approximate memorization.

Figure 6: Changes in memorization for each training step. In Nikkei Online Edition, there is a rapid increase in memorization after training. On Wikipedia, it was a small increase.

| | Strings |
|---|---|
| Prompt | .....英語でハードウェア (hardware) は、本来は「金物類、金属製品」の意味であり、かつては木材製品などとの対比語として用いられた。例えば英語で "hardware store" は、日本で言う「金物屋」を意味する。パーソナルコンピュータのハードウェア |
| Reference | などを「ソフトウェア」と呼ぶことがある。 |
| Generation | などを「ソフトウェア」と呼ぶことがある。 |

Table 2: An example of verbatim memorization in Wikipedia, where the model exactly generated part of the reference. Green highlighting is a forward match. "....." indicates omitted text.

| | Strings |
|---|---|
| Prompt | .....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 |
| Reference | 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。 |
| Generation | 回国連気候変動枠組み条約締約国会議(COP26)で も、世界各国は脱炭素の実行を急ぐ姿勢を鮮明にした。 |

Table 3: An example of verbatim memorization in Nikkei Online Edition, where the model generated a large part of the reference. "第26回国連気候変動枠組み条約締約国会議" is a specific event name and it seems that the model memorized the term. Green highlighting is a forward match. "....." indicates omitted text.

in Nikkei Online Edition, the maximum verbatim memorization increased from 15 to 27 characters. Wikipedia also exhibited a moderate increase, with average verbatim memorization growing from 1.34 to 1.53 and median approximate memorization increasing from 0.17 to 0.18.

Table 2 and 3 show the results of the continual pre-training using Wikipedia and Nikkei Online Edition, respectively, where the largest amount of verbatim memorization was achieved. As indicated by the green highlights, the texts that match the references are generated.

The examples of rapid memorization in Nikkei Online Edition were often topics related to the economy, which is characteristic of the corpus. Table 4 shows specific examples. Before training, the generated sentences were completely different from the reference, but after 1000 steps, the degree of similarity increased rapidly.

In contrast, Wikipedia shows a high degree of memorization from the early stage. It is likely that a corpus for continual pre-training, or similar texts, was used for the original pre-training. By targeting models other than Llama, where we can identify the corpus used for pre-training, it is possible to perform a more detailed analysis.

### 3.4 Results: Closed Setting

Compared to Wikipedia, relatively higher performance was observed in Nikkei Online Edition. This suggests that Nikkei Online Edition is more likely to be memorized.

Table 5 presents AUC scores for various membership inference methods across training steps and input token lengths. Depending on the conditions, we observed a detection performance of up to 0.689. On the other hand, there were some cases where the value was worse than the random value of 0.5.

| | Strings |
|---|---|
| Prompt | 【NQNロンドン】30日の欧州国債市場で、指標銘柄である独連邦債10年物の利回りは英国 |
| Reference | 時間16時時点で、前日の同時点に比べ0.008%高いマイナス0.179%で取引されている。 |
| 0 steps | 債10年物の利回りを上回った。 |
| 1000 steps | 時間16時時点で、前日の同時点に比べ0.00 5%低い0.335%で取引されている。 |
| 4 epochs | 時間16時時点で、前日の同時点に比べ0.00 5%高いマイナス0.343%で取引されている。 |

Table 4: Examples of changes in the results during training. The model trained on Nikkei Online Edition showed a rapid increase in memorization. Green highlighting is a forward match.

**Baseline methods.** LOSS and PPL/zlib worked well in Nikkei Online Edition. If limited to Nikkei Online Edition, the detection performance tends to increase along with the number of training steps and words.

**Token distribution methods.** The methods that use token-based filtering generally had poor performance. When we experimented by changing K in increments of 10, we found values of 0.5 or less in several cases. Min-K% Prob has the best value when the word length is 256 on Wikipedia, but 0.535 is not a high value.

**Text alternation methods.** ReCaLL achieved the best results in 6 out of 8 columns. It is possible that methods like altering the text implicitly take into account information specific to the language. Limited to Nikkei Online Edition, the detection performance tends to increase along with the number of training steps and words.

## 4 Discussion

This section discusses our findings with previous research and future research directions.

### 4.1 Reproduction of Empirical Findings from English Studies.

Empirical studies in English (Carlini et al., 2023) have shown that memorization correlates with (1) the duplication of strings in training data, (2) model size, and (3) prompt length. Our results generally agree with these results, but not always.

**Duplication.** The number of duplicates in the training data increases as training progresses. In the open setting, memorization increased as training progressed. This is particularly noticeable in Nikkei Online Edition and has also been observed on Wikipedia. In the closed setting, membership inference performance improved with training steps, particularly for the ReCaLL and LOSS methods. This discussion is limited to training progress, but it is also important to measure duplication by focusing more on the contents of the corpus.

**Model size.** In experiments using Nikkei Online Edition, we demonstrated that the larger the model, the more the memorization. Our 8B Llama 3 model exhibited greater memorization than the 0.1B GPT-2 model used in a previous Japanese study (Ishihara and Takahashi, 2024). After only 1,000 steps (0.25 epochs) in the open setting, our 8B model's approximate memorization exceeded that of the 0.1B model after 30 epochs. In the closed setting, the 8B model also had a higher detection performance, and memorization was increased.

**Prompt length.** In the closed setting, membership inference performance improved with the number of words (prompt length), particularly for the ReCaLL and LOSS methods. In the open setting, we did not evaluate this factor as the prompt length was fixed in our study.

### 4.2 Memorization in Industry-Specific Corpora.

Both open and closed settings showed significantly greater memorization for Nikkei Online Edition compared to Wikipedia. This suggests that industry-specific corpora lead to higher memorization rates, due to their unique terminology and writing styles. This raises concerns about overfitting and privacy risks in specialized industry applications of LLMs. It is also necessary to explore industrial-specific corpora other than Nikkei Online Edition.

### 4.3 Japanese-Specific Trends.

In the closed setting, prior research (Koyanagi et al., 2024) suggested that Min-K% Prob performs better with larger K values in Japanese. In our study, LOSS outperformed Min-K% Prob, supporting previous findings that full-sequence generation probabilities are more effective for membership inference for the Japanese LLMs. It is possible that the characteristic of Japanese, where words are not

| Method | The number of steps in continual pre-training | Wikipedia | | | | Nikkei Online Edition | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 32 | 64 | 128 | 256 | 32 | 64 | 128 | 256 |
| LOSS | 0 | 0.506 | 0.490 | 0.462 | 0.465 | 0.504 | 0.515 | 0.528 | 0.526 |
| | 1000 | 0.518 | 0.512 | 0.485 | 0.484 | **0.641** | 0.642 | 0.640 | 0.578 |
| | 12000 | 0.515 | 0.514 | 0.479 | 0.486 | **0.641** | 0.647 | 0.650 | 0.590 |
| | 24000 | 0.513 | 0.512 | 0.478 | 0.485 | - | - | - | - |
| PPL/zlib | 0 | 0.485 | 0.479 | 0.470 | 0.491 | 0.491 | 0.502 | 0.516 | 0.535 |
| | 1000 | 0.498 | 0.494 | 0.484 | 0.503 | 0.638 | 0.642 | 0.641 | 0.595 |
| | 12000 | 0.497 | 0.498 | 0.490 | 0.504 | 0.635 | 0.648 | 0.647 | 0.601 |
| | 24000 | 0.494 | 0.497 | 0.489 | 0.503 | - | - | - | - |
| Min-K% Prob (K=10) | 0 | 0.482 | 0.477 | 0.505 | 0.517 | 0.514 | 0.488 | 0.474 | 0.488 |
| | 1000 | 0.426 | 0.421 | 0.448 | 0.475 | 0.402 | 0.402 | 0.405 | 0.442 |
| | 12000 | 0.423 | 0.424 | 0.459 | 0.476 | 0.422 | 0.411 | 0.409 | 0.438 |
| | 24000 | 0.423 | 0.424 | 0.458 | 0.474 | - | - | - | - |
| Min-K% Prob (K=20) | 0 | 0.481 | 0.493 | 0.527 | **0.535** | 0.514 | 0.488 | 0.467 | 0.485 |
| | 1000 | 0.431 | 0.441 | 0.475 | 0.496 | 0.381 | 0.382 | 0.383 | 0.439 |
| | 12000 | 0.432 | 0.441 | 0.483 | 0.492 | 0.387 | 0.379 | 0.376 | 0.426 |
| | 24000 | 0.433 | 0.441 | 0.484 | 0.491 | - | - | - | - |
| Min-K% Prob (K=90) | 0 | 0.495 | 0.510 | 0.538 | **0.535** | 0.496 | 0.485 | 0.472 | 0.473 |
| | 1000 | 0.483 | 0.489 | 0.515 | 0.516 | 0.359 | 0.358 | 0.360 | 0.422 |
| | 12000 | 0.485 | 0.487 | 0.521 | 0.514 | 0.359 | 0.353 | 0.350 | 0.410 |
| | 24000 | 0.487 | 0.488 | 0.522 | 0.515 | - | - | - | - |
| Min-K%++ (K=10) | 0 | 0.482 | 0.490 | 0.510 | 0.494 | 0.498 | 0.495 | 0.482 | 0.494 |
| | 1000 | 0.431 | 0.421 | 0.430 | 0.434 | 0.522 | 0.554 | 0.539 | 0.506 |
| | 12000 | 0.427 | 0.420 | 0.445 | 0.438 | 0.543 | 0.574 | 0.565 | 0.536 |
| | 24000 | 0.431 | 0.421 | 0.442 | 0.438 | - | - | - | - |
| Min-K%++ (K=20) | 0 | 0.486 | 0.496 | 0.522 | 0.513 | 0.502 | 0.489 | 0.482 | 0.482 |
| | 1000 | 0.425 | 0.420 | 0.443 | 0.447 | 0.494 | 0.514 | 0.491 | 0.473 |
| | 12000 | 0.424 | 0.419 | 0.456 | 0.450 | 0.513 | 0.531 | 0.517 | 0.497 |
| | 24000 | 0.425 | 0.419 | 0.451 | 0.449 | - | - | - | - |
| Min-K%++ (K=90) | 0 | 0.483 | 0.498 | 0.531 | 0.526 | 0.489 | 0.471 | 0.459 | 0.456 |
| | 1000 | 0.403 | 0.400 | 0.428 | 0.430 | 0.518 | 0.530 | 0.509 | 0.487 |
| | 12000 | 0.400 | 0.399 | 0.444 | 0.436 | 0.526 | 0.546 | 0.530 | 0.505 |
| | 24000 | 0.400 | 0.398 | 0.439 | 0.432 | - | - | - | - |
| ReCaLL | 0 | 0.561 | 0.502 | 0.483 | 0.437 | 0.484 | 0.535 | 0.546 | 0.542 |
| | 1000 | **0.613** | **0.605** | **0.569** | 0.520 | 0.611 | 0.651 | 0.572 | **0.630** |
| | 12000 | 0.608 | 0.569 | 0.460 | 0.494 | 0.637 | **0.660** | **0.689** | 0.603 |
| | 24000 | 0.601 | 0.560 | 0.454 | 0.484 | - | - | - | - |

Table 5: AUC for each method of membership inference, the number of training steps, and the number of input words. **Bold** indicates the best result in each column.

separated by spaces, is having an effect. A detailed analysis based on the characteristics of the language is a future prospect.

## 5 Conclusion

This study is the first to systematically quantify training data memorization in continual pre-training settings using non-English and industry-specific corpora. Our experiments with Japanese Wikipedia and Nikkei Online Edition demonstrated that continual pre-training can significantly increase memorization, particularly when using industry-specific corpora. These findings highlight the heightened privacy and intellectual property risks associated with these corpora. In addition, we also highlighted the limitations of directly applying English-centered methods to other languages. Our work underscores the necessity of language- and domain-aware memorization analysis for the safe and responsible development of LLMs.

## Limitations

Our study has some limitations.

**Dataset accessibility.** Due to the circumstances of our research, which involves examining the memorization of industry-specific corpora, the transparency of the data is inevitably compromised. The 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.

**Association with danger.** In our experiments, all texts are treated equally. However, to deepen the discussion on security and copyright, it is important to consider the degree of danger of memorized strings. For example, the undesirable memorization

of personal identification information (PII), such as phone numbers and email addresses, must be distinguished from the acceptable memorization of simple frequent strings.

**Diversifying experimental conditions.** There remains room to experiment with various settings:

- **Decoding strategy**: In our experiments, a single string was generated from a single prompt using the greedy method. There is still room for various decoding strategies such as top-k sampling and temperature control to increase the diversity of generated text. Some reports indicate that the choice of decoding strategy does not significantly affect experimental results (Carlini et al., 2023), while others observe that top-k sampling and top-p sampling lead to greater memorization (Lee et al., 2023).
- **Models**: There are many possible variations, such as models other than Llama 3, changing the number of LoRA ranks, and continual pre-training without LoRA. In particular, as mentioned in Section 3, training with full parameters is important for the generalization of results.
- **Languages**: We currently focus on Japanese as a language other than English, but other languages are also available. For example, we can target languages with lower resources.

**Measures for memorization.** It is also important to investigate the effectiveness of methods that reduce memorization under conditions other than English. For example, it would be worth trying the defensive approach with the three phases of preprocessing, during training, and post-processing as classified by Ishihara (2023).

## Ethical Considerations

This study entails the extraction of training data from LLMs, a process that can be interpreted as a form of security probing. However, the objective is not to promote such attacks but rather to foster informed discussions aimed at mitigating associated risks. While our experiments focus on Japanese, the implications are broadly applicable across languages.

The dataset used in this research was obtained through proper channels from Nikkei Inc. No data was collected through ethically questionable means, such as circumventing paywalls. Many publishers, including Nikkei, offer datasets for academic use under appropriate licensing and payment conditions.

## Acknowledgments

We thank Kazumasa Omura for his helpful advice on this study and Yasufumi Nakama for providing insights from the analysis of Nikkei Online Edition and Wikipedia.

## References

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. 2024. LoRA learns less and forgets less. *Transactions on Machine Learning Research*. Featured Certification.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Bowen Chen, Namgi Han, and Yusuke Miyao. 2024. A statistical and multi-perspective revisiting of the membership inference attack in large language models. *arXiv [cs.CL]*.

Debeshee Das, Jie Zhang, and Florian Tramèr. 2025a. Blind Baselines Beat Membership Inference Attacks for Foundation Models . In *2025 IEEE Security and Privacy Workshops (SPW)*, pages 118–125, Los Alamitos, CA, USA. IEEE Computer Society.

Soumi Das, Camila Kolling, Mohammad Aflah Khan, Mahsa Amani, Bishwamittra Ghosh, Qinyuan Wu, Till Speicher, and Krishna P Gummadi. 2025b. Revisiting privacy, utility, and efficiency trade-offs when fine-tuning large language models. *arXiv [cs.AI]*.

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*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Kan Hatakeyama-Sato, Yasuhiko Igarashi, Shun Katakami, Yuta Nabae, and Teruaki Hayakawa. 2023. Teaching specific scientific knowledge into large language models through additional training. *arXiv [cs.CL]*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.

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 2023)*, pages 260–275, Toronto, Canada. Association for Computational Linguistics.

Shotaro Ishihara and Hiromu Takahashi. 2024. Quantifying memorization and detecting training data of pre-trained language models using Japanese newspaper. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 165–179, Tokyo, Japan. Association for Computational Linguistics.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596, Tokyo, Japan. Association for Computational Linguistics.

Minato Kondo, Takehito Utsuro, and Masaaki Nagata. 2024. Enhancing translation accuracy of large language models through continual pre-training on parallel data. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 203–220, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Kyoko Koyanagi, Miyu Sato, Teruno Kajiura, and Kimio Kuramitsu. 2024. The analysis of pretraining data detection on LLMs between English and Japanese. *Proceedings of the Annual Conference of JSAI*, JSAI2024:4Xin298–4Xin298. In Japanese.

Jooyoung Lee, Thai Le, Jinghui Chen, and 1 others. 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, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 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.

LLM-jp, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, and 62 others. 2024. LLM-jp: A cross-organizational project for the research and development of fully open japanese LLMs. *arXiv [cs.CL]*.

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.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. 2025. A survey on LoRA of large language models. *Frontiers of Computer Science*, 19(7):1–19.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. How much are large language models contaminated? a comprehensive survey and the LLMSanitize library. *arXiv [cs.CL]*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

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

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. 2024. ReCaLL: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Shuaiqiang Wang, Jiliang Tang, and Dawei Yin. 2024. Exploring memorization in fine-tuned language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3917–3948, Bangkok, Thailand. Association for Computational Linguistics.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-K%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.