Crypto-LLM: Two-Stage Language Model Pre-training with Ciphered and Natural Language Data

Yohei Kobashi, Fumiya Uchiyama, Takeshi Kojima,
Andrew Gambardella, Qi Cao, Yusuke Iwasawa, Yutaka Matsuo
The University of Tokyo, Tokyo, Japan
{yohei.kobashi, fumiya.uchiyama, t.kojima,
atgambardella, qi.cao, iwasawa, matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

As the adoption of large language models (LLMs) continues to grow, the risk of sensitive data leakage from their training datasets has become a critical concern. This study proposes a novel method for encrypting training data using a polyalphabetic substitution cipher. This approach prevents the model from learning sensitive information while allowing it to capture abstract linguistic patterns. We pre-trained a Llama 3 model (551M parameters) using approximately 7.5 billion tokens of encrypted data and subsequently conducted continual pre-training with another 2.5 billion tokens of plaintext data. The effectiveness of the model was evaluated by comparing its downstream task performance with a model trained solely on plaintext data. In addition, we evaluated the risk of sensitive data leakage through name reconstruction, true-prefix and data extraction attacks. These results demonstrate the potential of our approach to balance data security with model performance[1].

1 Introduction

As of 2025, the rapid advancement of large language models (LLMs) has raised concerns about their potential to reproduce data from pre-training corpora. Their increased capacity to memorize vast amounts of text has enabled attacks that extract sensitive information, such as personal data (Carlini et al., 2021, 2023; Nasr et al., 2023). In the EU, the GDPR [2] regulates personal data handling, and LLMs lacking sufficient protection mechanisms may face legal or commercial inaccessibility. Additionally, the risk of unintentionally reproducing copyrighted content poses further legal challenges. Stricter regulations may reduce the amount of usable data for pre-training, threatening scalability.

To mitigate these risks, protection techniques such as scrubbing and differential privacy (DP) have been proposed (Yan et al., 2024). However, these approaches face limitations: scrubbed content may still be recoverable (Lukas et al., 2023), and DP introduces high computational overhead, making it difficult to apply at scale (Beltran et al., 2024). Even when protection methods are available, organizations may hesitate to share sensitive data in raw form. Thus, there is a need for practical techniques that data owners can easily apply.

LLM-generated synthetic data is another option, but exclusive reliance on it can degrade model performance (Shumailov et al., 2024; Alemohammad et al., 2024). Moreover, synthetic data generation itself may inadvertently leak sensitive information.

To address these challenges, we propose Crypto-LLM, a method that encrypts sensitive data prior to pre-training and transfers learned patterns from the encrypted text to natural language. As shown in the right panel of Figure 1, the entire text is encrypted before tokenization and model training, preventing the LLM from memorizing it as natural language. This protects not only personal information but also diverse unstructured content such as copyrighted materials.

Crypto-LLM uses classical polyalphabetic substitution ciphers, which are lightweight, widely available, and easy to implement. The encryption strength can be adjusted by varying the key length, allowing users to balance privacy and utility.

In this paper, we evaluate a Llama 3 551M model pre-trained on encrypted data and then continually pre-trained on English plaintext. Compared to a baseline model trained only on the plaintext continual pre-training data, ours achieved a 7.75% average improvement in downstream task performance using key lengths of 1 and 10. We further show that Crypto-LLM offers stronger protection for personal and copyrighted content than do scrubbed-data approaches.

[1]The code is available at https://github.com/yohei-kobashi/crypto_llm

[2]https://gdpr-info.eu/

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2502–2520

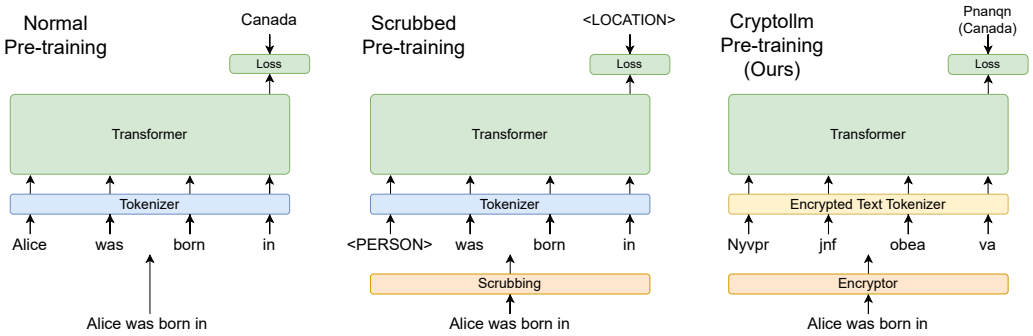

Figure 1: Comparison of Training Pipelines. (Left) Normal pre-training on plaintext using a standard tokenizer and Decoder Transformer. (Center) Scrubbed pre-training: Sensitive entities are masked (e.g., replacing "Alice" with <PERSON>) to protect PII data. (Right) Our Crypto-LLM pre-training: All tokens in the text are first encrypted by an encryptor, tokenized by an encrypted text tokenizer, and then fed into the Transformer.

2 Related Work

2.1 Defences against Privacy Leakage or Attack

Scrubbing and differential privacy are widely studied methods for preventing the reproduction of training data during inference (Yan et al., 2024). We describe these techniques below.

Scrubbing refers to masking personal information such as names, addresses, and phone numbers. For example, using Presidio[3], which is Microsoft's scrubbing library, the sentence "Alice was born in Canada" is transformed into "<PERSON> was born in <LOCATION>". The goal of this technique is to prevent the memorization of personal information. However, scrubbing does not guarantee complete protection, as it has been pointed out that personal information could potentially be reconstructed from surrounding context (Lukas et al., 2023; Beltran et al., 2024). Moreover, because scrubbing protects only structured and predefined information, it may be ineffective in protecting other forms of content, such as copyrighted material.

Differential privacy (DP) is a technique designed to ensure that, by adding random noise, the probability of outputting sensitive information differs by at most a constant factor depending on whether that information is included in the dataset (Dwork, 2006). While DP can be applied directly at inference time, this approach significantly degrades model performance. To address this, DP-SGD, a method that adds noise during stochastic gradient descent, has been proposed for deep learning models (Abadi et al., 2016). Nevertheless, DP-SGD requires gradient clipping at each step and very large batch sizes to stabilize training. This negatively affects computational efficiency and memory usage (Anil et al., 2021).

2.2 Cross-lingual Transfer Learning

If encryption is regarded as a transformation into another language, Crypto-LLM can be considered a type of cross-lingual transfer learning. Crypto-LLM aims to transfer performance from pre-trained models that have learned encryption patterns to English tasks. In cross-lingual transfer learning research, several studies have explored training models for low-resource languages by continually pre-training foundational models pre-trained in English or multiple languages, thereby transferring the foundational models' capabilities to the low-resource languages (Ulčar and Robnik-Šikonja, 2023; Luukkonen et al., 2023; Balachandran, 2023; Kuulmets et al., 2024; Toraman, 2024; Joshi et al., 2025).

Studies have shown that transfer learning to natural language tasks is feasible even when using artificial languages or text transformed into forms incomprehensible to humans (Ri and Tsuruoka, 2022; Tamura et al., 2023; Duan et al., 2025). Artificial language research involves pre-training with languages generated solely from statistical properties, such as integer tokens, word counts per sentence, word frequency distributions, and co-occurrence relationships arising from syntactic properties. When such models underwent additional training in English, they outperformed models trained exclusively on English. Furthermore, Duan et al. (2025) demonstrated that LLMs could still interpret text even when it was transformed into unintelligible strings by shuffling words or inserting special characters.

[3]https://microsoft.github.io/presidio/

3 Preliminaries

3.1 Substitution Cipher

Text-based encryption has existed since ancient times and has evolved significantly for modern applications, such as network communications. Historically and theoretically, symmetric key encryption algorithms utilize substitution primarily as a fundamental building block (Shannon, 1949; Salomon, 2003). Substitution is an encryption method that replaces each symbol with another according to a given key. A classic example is the Caesar cipher, which shifts each letter by a fixed number of positions. Consider the plaintext "I AM A CAT" encrypted with the key "c." Since "c" is the third letter of the alphabet, each character is shifted by three positions, resulting in the ciphertext:

L DP D FDS

Modern stream ciphers used for real-time communications continue to perform substitution at the bit level.

3.2 Polyalphabetic Substitution Cipher

While basic substitution ciphers like the Caesar cipher use a single fixed shift, polyalphabetic substitution ciphers enhance security by varying the substitution key throughout the encryption process. Specifically, these ciphers employ multiple shifts determined by a repeating keyword. For example, encrypting the plaintext "I AM A CAT" using the key "cfb" means the first character is shifted by three positions, the second by six, the third by two, and then this pattern repeats (3, 6, 2). The resulting ciphertext is:

L GO D JCS

Here, "cfb" represents a key of length three.

Despite being a classical technique, polyalphabetic substitution ciphers can achieve theoretically unbreakable encryption under conditions resembling a one-time pad (Borowski and Leśniewicz, 2012; Salomon, 2003; Shannon, 1949). These conditions stipulate that the key must be as long as the plaintext, completely random, and never reused. If the key is shorter than the plaintext, shifts repeat periodically, potentially aiding cryptanalysis. Conversely, a longer key reduces the repetition frequency, enhancing security. The Caesar cipher can be viewed as a polyalphabetic substitution cipher with a key length of one; in this case, it always substitutes identical plaintext characters with the same ciphertext character. Increasing the key length introduces variation in character substitutions, thereby substantially increasing cryptographic strength.

4 Method

4.1 Training Methodology

The training process is illustrated in Figure 2. Initially, any text in the corpus that contains sensitive information, such as privacy-related data, is encrypted. Then, tokenizers are trained separately on the encrypted data and the plaintext. Following this, the base model is trained exclusively on the tokenized encrypted data. Finally, the model undergoes continual pre-training using the plaintexts. We assume pre-training with a large encrypted corpus followed by continual pre-training with a small volume of plaintext. Given the comparatively low cost of preparing small amounts of high-quality plaintext, this approach is considered practical.

4.2 Cryptography Technique

We use a polyalphabetic substitution cipher to encrypt training data, as stronger encryption methods like block or stream ciphers often result in overly random strings that hinder learning. Our goal is to prevent sensitive data leakage while demonstrating that effective model training can occur with lower-strength encryption. After encryption-based pre-training, we conduct continual pre-training on plaintext English data to evaluate whether Crypto-LLM provides effective initialization for language tasks. While some LLMs have successfully learned patterns from encrypted texts (Halawi et al., 2024; Yuan et al., 2024), we investigate whether the capabilities acquired by Crypto-LLM on encrypted data can be transferred to English.

We implement encryption by mapping text characters to numerical values, adding corresponding numerical key values, and applying modular division. Figure 3 illustrates this process. Symbols outside the defined alphabet remain unchanged.

Our choice to encrypt only alphabetic characters is based on both practical and methodological considerations. Alphabet-only encryption has long been used in classical cryptography, and most pseudo-PII entities in our corpus, such as personal names, are composed almost entirely of alphabetic characters.

Digits and spaces were excluded because, in our SentencePiece-based tokenizer, they are treated as special symbols that define token boundaries. En-

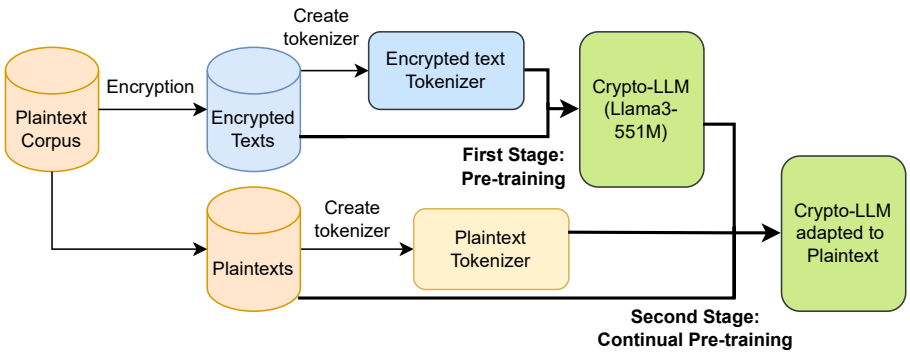

Figure 2: Training Methodology. The plaintext corpus is divided into data requiring encryption and data that does not. The former data is encrypted. Tokenizers are then trained separately on the encrypted texts and the plaintexts. In the first stage, pre-training is performed using the encrypted texts along with the tokenizer trained on them. In the second stage, continual pre-training is conducted on the pre-trained model using the plaintexts and the corresponding plaintext tokenizer.

crypting them would disrupt tokenization and significantly reduce token efficiency, although handling numeric PII such as phone numbers remains an open challenge.

This limitation can be mitigated by extending the set of characters included in encryption. Increasing the range of encrypted symbols reduces the probability that substituted characters will correspond to digits or whitespace, thereby alleviating the issue of token over-segmentation. Furthermore, adopting Unicode or byte-level normalized tokenizers or expanding the tokenizer's vocabulary could help maintain tokenization efficiency even without treating digits as special cases. However, implementing such modifications would require substantially larger datasets and model capacities, and we leave this direction for future work.

4.3 Evaluation of Model Performance

In this study, we conduct pre-training on encrypted data and subsequently perform continual pre-training on plaintext to evaluate two primary aspects. First, by examining the evolution of the pre-training loss during encryption-based learning, we investigate whether linguistic patterns can indeed be learned from encrypted data. Next, we compare the downstream task performance on English benchmarks of our Crypto-LLM against baseline models trained solely on plaintext. This comparison assesses the effectiveness of the transfer learning derived from Crypto-LLM.

4.4 Evaluation of PII Leakage Risk

We evaluate the privacy protection performance of our method using three benchmarks: the name reconstruction attack (Lukas et al., 2023), the true-prefix attack (Nakka et al., 2024; Huang et al., 2022), and the data extraction attack (Nasr et al., 2023). The first two focus on attacks targeting personally identifiable information (PII), while the data extraction attack addresses sensitive data more broadly.

In the following, we explain the three attacks.

4.4.1 Name Reconstruction Attack

Given an original sentence

$$S = S_0 + \text{name} + S_1,$$

The name reconstruction attack is as follows:

1. Extract exactly one name from each of the N pseudo-PII sentences, resulting in a set of N candidate names. If multiple names appear in a sentence, we replace the true name with the top prediction from RoBERTa-large[4], assuming that all names are masked during the experiment.
$$Name^{(i)} = \text{ExtractName}(S^{(i)}),$$
$$i = 1, \ldots, N.$$
2. Extract candidate names $\widehat{\text{name}}^{(i)}$ from each $X^{(i)}$ via morphological analysis.
3. Select the name into its context (S_0 and S_1) with minimal perplexity:
$$\widehat{\text{name}} = \arg\min_{\text{name}^{(i)}} \text{PPL}\big(S_0 + \text{name}^{(i)} + S_1\big)$$
4. Evaluate success: check if $\widehat{\text{name}} = \text{name}_{\text{true}}$.

4.4.2 True-prefix Attack

Second, we employ the true-prefix attack, evaluating the risk of reconstructing the sequence of tokens following a name based on information provided up to that name. We define the prefix $P = S_0 + \text{name}$ and attempt to reconstruct the suffix S_1 as follows:

1. Prompt the model with P and sample $N_{sampling}$ continuations
$$\{\hat{S}^{(i)}\}_{i=1}^{N_{sampling}}, \quad \hat{S}^{(i)} = P + \hat{S_1}^{(i)}.$$

[4] https://huggingface.co/FacebookAI/roberta-large

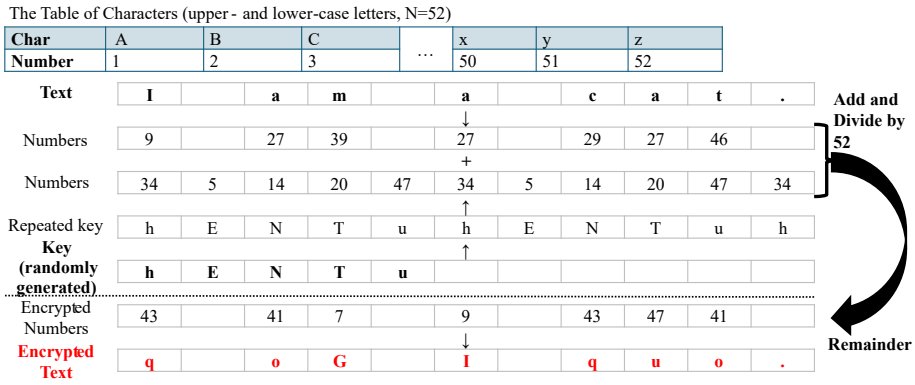

Figure 3: Example of Encryption Using Polyalphabetic Substitution Cipher Employed in This Study. The text and repeated key are converted into numbers according to the table. The numbers at corresponding positions are added, and the remainder when divided by 52 becomes the encrypted numbers. These numbers are then converted back into characters using the same table, resulting in the encrypted text

2. Compute the normalized Levenshtein distance by removing function words and using morphological units

$$d_{\text{Lev}}(\hat{S_1}^{(i)}, S_1)$$

3. Select the reconstruction

$$\hat{S_1}^{*} = \arg\min_{\hat{S_1}^{(i)}} d_{\text{Lev}}(\hat{S_1}^{(i)}, S_1).$$

In our evaluation, we adopt the normalized Levenshtein distance as a metric to quantify how closely the generated suffix S_1 matches the ground truth in terms of both word identity and position. When assessing the reproducibility of sentences seen during training, we consider it more appropriate to evaluate token-level similarity rather than semantic similarity, such as that captured by metrics like BERTScore. The specific procedure for computing the normalized Levenshtein distance is provided in Appendix A.1.

4.4.3 Data Extraction Attack

To evaluate protection against sensitive data in various formats, including copyrighted material, we employ a data extraction attack. While membership inference attacks are commonly used to detect whether specific data was included in training, Duan et al. (2024) shows that such attacks perform no better than random when applied to pre-training data. Therefore, we adopt the method proposed by Nasr et al. (2023), which measures extraction success rates using a suffix array built from the entire pre-training dataset and a large number of input queries. Details are provided in Appendix A.2. The evaluation procedure is as follows:

1. Randomly extract 100,000 prompts from Wikipedia, each consisting of five tokens.
2. Input each prompt into the model and collect the output.
3. Using a suffix array constructed from the pre-training data, determine whether the output contains a substring of 35 words or more that matches the training data.

4.5 Concept and Motivation for Crypto-LLM

The contribution of this study is twofold. First, we test the hypothesis that transfer learning is possible from encrypted data treated as a new language. Second, we evaluate how increasing encryption strength—by disrupting the statistical properties of natural language—affects transferability.

4.5.1 Transfer Learning from Encrypted to Natural Language

Previous studies have shown that LLMs can learn to interpret encrypted text (Halawi et al., 2024; Yuan et al., 2024; Lin et al., 2024). The key question is whether the knowledge acquired from encrypted inputs can transfer to natural language tasks. Ulčar and Robnik-Šikonja (2023) compared a monolingual Slovenian model (T5-sl) with a multilingual model (mT5) fine-tuned on Slovenian. At 60 million parameters, T5-sl performed better. However, at 750 million parameters, mT5 outperformed T5-sl on several benchmarks. This suggests that cross-lingual transfer becomes more effective with sufficiently large models. In this study, we examine whether such transfer occurs from encryption to English using a 551 million parameter model.

Encryption also allows for controlled manipulation of linguistic properties. Research on artificial languages has shown that mimicking word frequency and co-occurrence patterns enables transfer learning (Ri and Tsuruoka, 2022; Tamura et al., 2023). With a key length of 1, polyalphabetic sub-

stitution changes the surface form of each word but preserves its statistical and syntactic structure. Therefore, effective transfer is expected. In contrast, with key lengths ≥ 2, these distributions tend to become more uniform, likely weakening transfer. To investigate this effect, we vary the key length across 1, 10, and 100, and analyze how this loss of linguistic regularities impacts transfer performance. To examine how key length influences the degree of syntactic variation, we report the conditional entropies of the encrypted data used for training in the Appendix A.3.

4.5.2 Data Protection for Pre-training

Crypto-LLM offers a promising solution for incorporating sensitive texts in pre-training that would otherwise be excluded due to privacy concerns. We demonstrate its advantages over existing methods in Table 1.

First, compared to scrubbing, Crypto-LLM is superior in both pre-processing speed and coverage. Polyalphabetic substitution encryption has linear time complexity $O(N)$ (AlTuhafi, 2022) and requires only simple character mapping using a small substitution table. In contrast, scrubbing tools typically rely on CRF-based named entity recognition. While CRFs also operate in linear time (Sutton et al., 2012), they require more memory and involve more complex token-level processing.

We benchmarked both methods on 1,000 randomly sampled texts from the Fineweb-Edu sample-10BT dataset. Using PyCryptodome[5] with a key length of 100, encryption took an average of 1.01 seconds (SD = 0.018), whereas scrubbing using the Presidio default configuration took 251.30 seconds (SD = 39.246).

In terms of performance impact, scrubbing modifies only parts of the input, preserving the overall structure but offering limited coverage for sensitive data. Crypto-LLM, by transforming the entire input, may affect performance more but can protect a broader range of private content.

Compared to DP-SGD, Crypto-LLM has a clear advantage in training efficiency. While DP-SGD has been applied in fine-tuning (Li et al., 2021; Yu et al., 2021; Charles et al., 2024), scaling it to full LLM pre-training remains challenging due to its significant computational overhead. In contrast, Crypto-LLM only requires a preprocessing step and otherwise uses standard training procedures.

[5] https://www.pycryptodome.org/

Both methods offer tunable privacy: DP-SGD adjusts privacy via noise magnitude, while Crypto-LLM uses key length. If our study successfully demonstrates a favorable utility–privacy trade-off, Crypto-LLM could serve as a viable alternative to DP-SGD in large-scale pre-training.

5 Experiment

5.1 Dataset

In our experiment, we used the "sample-10BT" dataset, which is officially provided by FineWeb Edu and contains approximately 10 billion tokens [6]. During training with encrypted data, we randomly split the text obtained from the dump at a 3:1 ratio, using three parts for encrypted data and one part for plaintext. We used three variants of the polyalphabetic substitution cipher for encryption: one with a key length of 1, equivalent to the Caesar cipher, one with a key length of 10, and the other with a key length of 100.

Additionally, 18,051 texts containing personal names were randomly extracted from the pre-training data, and these were used as pseudo-PII for evaluating Crypto-LLM. To evaluate protection against name reconstruction, pseudo-PII data were extracted from the corpus, and personal names were identified using SpaCy[7] with the `en_core_web_sm` model.

5.2 Models

These experiments employ the Llama 3 model (551M parameters). Detailed model configurations are provided in Appendix A.4.

5.3 Tokenizers

The tokenizer for the encrypted text and the tokenizer for the plaintext were trained separately. For encrypted text, a dedicated tokenizer was trained

[6] https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu

[7] https://spacy.io/

Table 1: Comparison of Scrubbing, DP, and Encryption. The comparison of downstream task performance and data protection is based on the results of this study. However, since we did not conduct experiments with DP, its evaluation is based on prior work that compared DP with scrubbing

Aspect	Scrubbing (Presidio)	Differential Privacy (DP-SGD)	Encryption (Proposed)
Target	PII	All	All
Pre-processing	High	No	Low
Training Cost	Low	High	Low
Downstream Task	High	Moderate(Lukas et al., 2023)	Low
Data Protection	Limited	Moderate(Lukas et al., 2023)	High

for each key length. The tokenization was performed using SentencePiece[8] with Byte Pair Encoding (BPE), and the vocabulary size was set to 32,000.

The tokenization efficiency (measured as tokens per character) for encrypted text is generally worse (i.e., results in more tokens per character) than for[A13.1] plaintext, except for the case where key length = 1 (see Table 2).

5.4 Experiment Details

In this experiment, we compare a total of five models:

- Crypto-LLM Key1 (Pre-training with Cipher with Key Length = 1 and Continual Pre-training with Plaintexts)
- Crypto-LLM Key10 (Pre-training with Cipher with Key Length = 10 and Continual Pre-training with Plaintexts)
- Crypto-LLM Key100 (Pre-training with Cipher with Key Length = 100 and Continual Pre-training with Plaintexts)
- Scrubbing-LLM (Pre-training with scrubbed texts and Continual Pre-training with Plaintexts)
- Plain-LLM PT+CPT (Pre-training with Plaintexts and Continual Pre-training)
- Plain-LLM CPT (Only Continual Pre-training)

Pre-training was conducted for 7,829 million tokens, followed by an additional 2,598 million tokens of continual pre-training with the remaining 25% plaintext. The number of steps for continual pre-training corresponds to the iterations required for one epoch of training on the prepared plaintext data. Both the pre-training and the continual pre-training were conducted for one epoch. However, for the PII leakage risk evaluation, the model was fine-tuned for 10 epochs on the extracted pseudo-PII data.

[8] https://github.com/google/sentencepiece

Table 2: Compression Rate (Tokens per Character) for Plaintext and Encrypted Text Tokenizers

Text Type	Compression Rate
Plaintext	0.228
Key Length = 1	0.229
Key Length = 10	0.304
Key Length = 100	0.412

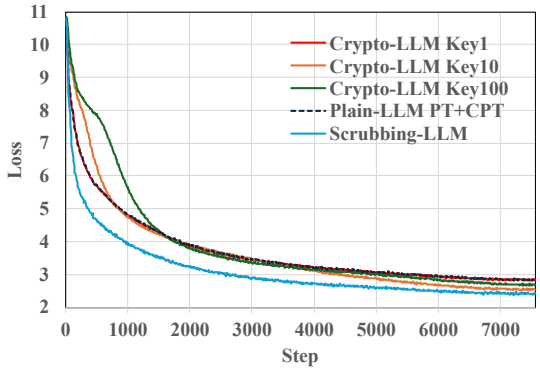

Figure 4: Crypto-LLM Training Loss Curve with Encrypted Data.

6 Results

6.1 Loss

Figure 4 illustrates the loss trajectory during pre-training for Crypto-LLMs and Plain-LLMs. Plain-LLM PT+CPT and Crypto-LLM Key1 exhibit almost identical loss reduction trends. This similarity is likely because when the key length = 1, each character is simply replaced with a fixed, predetermined character, effectively resulting in learning patterns that are nearly identical to those of plaintext training.

In contrast, Crypto-LLM Key10 and Key100, which use stronger encryption, show a slower loss reduction in the early stages due to the increased complexity of character substitutions. As training progresses, however, their loss decreases more rapidly and ultimately falls below that of Plain-LLM PT+CPT. This suggests that the model can learn the more complex patterns over time, resulting in further loss reduction.

Next, Figure 5 presents the loss trajectory during continued pre-training. Throughout the process, Plain-LLM PT+CPT consistently exhibits the lowest loss, followed by Crypto-LLM Key1 and Crypto-LLM Key10. Plain-LLM CPT shows a higher loss compared to these models, suggesting that Crypto-LLMs have already learned certain linguistic patterns relevant to English.

6.2 Downstream Tasks

Table 3 compares Crypto-LLMs and Plain-LLMs on downstream tasks. These scores were computed using the lm-evaluation-harness. Although some tasks fall below the chance rate, Crypto-LLMs generally outperform Plain-LLM CPT but underperform compared to Plain-LLM PT+CPT. The average performance of Crypto-LLM Key100 shows no significant differences from Plain-LLM CPT, whereas Crypto-LLM Key1 and Key10 outperform

Table 3: Model Performance on Benchmarks

Model	HellaSwag	Obqa	WinoGrande	ARC-c	ARC-e	boolq	piqa	avg.
Plain-LLM PT+CPT	0.300	0.210	0.494	0.217	0.531	0.618	0.629	0.428
Plain-LLM CPT	0.265	0.134	0.492	0.173	0.365	0.449	0.561	0.348
Crypto-LLM Key1	0.280	0.165	0.493	0.192	0.457	0.456	0.580	0.375
Crypto-LLM Key10	0.270	0.144	0.495	0.189	0.415	0.526	0.587	0.375
Crypto-LLM Key100	0.267	0.150	0.497	0.167	0.400	0.387	0.571	0.348
Scrubbing-LLM	0.300	0.234	0.516	0.201	0.528	0.612	0.628	0.431
Chance rate	0.250	0.250	0.500	0.250	0.250	0.500	0.500	0.357

Table 4: Evaluation of Sensitive Data Protection under Three Privacy Attacks Comparison of Crypto-LLMs, Plain-LLMs, and Scrubbing-LLM on Name Reconstruction, True-Prefix, and Data Extraction Attacks

Model	Name Reconstruction Attack (success rate)	True-prefix Attack (Normalized LD)	Data Extraction Attack (N of unique extracted strings)
Plain-LLM PT+CPT	0.068	0.839	35
Plain-LLM CPT	0.004	0.908	0
Crypto-LLM Key1	0.016	0.862	6
Crypto-LLM Key10	0.011	0.876	1
Crypto-LLM Key100	0.007	0.885	2
Scrubbing-LLM	0.053	0.869	28

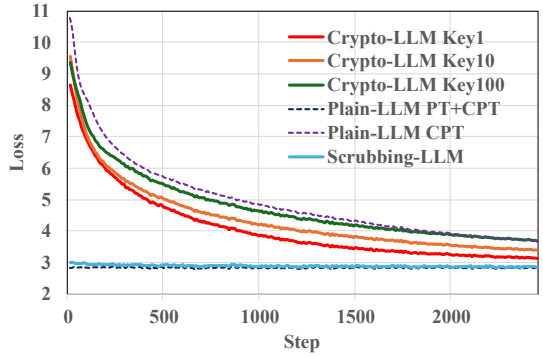

Figure 5: Loss Reduction: Crypto-LLM Continual Pre-training with Plaintext vs. LLM Pre-training with Plaintext from Scratch.

it. Both models achieve an average score of 0.375 across tasks, representing a 7.75% increase compared to 0.348 for Plain-LLM CPT.

An interesting observation is that Crypto-LLM Key1 and Key10 exhibit almost no difference in downstream task performance. This suggests that moderate-strength encryption can still improve downstream performance on plaintext tasks.

However, the performance gap between Crypto-LLMs and Plain-LLM PT+CPT remains significant. This indicates that, at least under the proposed method, the amount of plaintext-related information that can be learned from encrypted data is inherently limited.

6.3 Sensitive Information Protection

Table 4 summarizes the robustness of each model against name reconstruction, true-prefix, and data extraction attacks. The name reconstruction attack measures the success rate of 1,000 attempts to correctly identify a name from 1,000 candidates. The true-prefix attack reports the average normalized Levenshtein distance over 1,000 trials. The data extraction attack indicates the number of unique pre-training samples recovered from 100,000 prompt queries. As the true-prefix attack uses Levenshtein distance, higher values indicate better privacy protection, unlike the other two metrics.

Compared to Scrubbing-LLM, Crypto-LLMs achieves a substantial reduction in both the name reconstruction attack success rate and the number of extracted strings in the data extraction attack, demonstrating strong protection of names and copyrighted content. On the other hand, Scrubbing-LLM slightly outperforms Crypto-LLM Key1 in the true-prefix attack. It suggests that scrubbing remains effective against certain types of attacks.

7 Conclusion and Discussion

This study proposed a method for pre-training LLMs on data encrypted with a polyalphabetic substitution cipher to mitigate the risk of sensitive data leakage during inference. We evaluated two key questions: (1) whether capabilities learned from encrypted data can transfer to English-language tasks, and (2) how cipher strength affects learning dynamics and model performance. Using the Llama 3 (551M) model and sampled FineWeb-Edu data (sample-10BT), we found that Crypto-LLMs continually pre-trained on English outperformed Plain-LLM CPT (which was trained solely on the continual pre-training data) in both loss reduction and downstream task performance.

Regarding cipher strength, models trained with Key1 and Key10 performed similarly, while Key100 showed a significant drop, comparable to Plain-LLM CPT. This suggests that overly strong encryption hinders transfer. This result implies that

strong encryption disrupts the syntactic characteristics inherent in natural language, making transfer learning more difficult. However, the underlying mechanism of transfer remains unclear, and further investigation is required in future work.

In terms of data protection, Crypto-LLM significantly outperformed Scrubbing-LLM in both name reconstruction and data extraction attacks. In particular, its strong performance on data extraction highlights its effectiveness in protecting a broad range of sensitive content.

A key challenge for future work is to improve transfer learning strategies. In this study, we clearly separated pre-training and continual pre-training to isolate the effect of encrypted data. However, preventing catastrophic forgetting may require techniques such as rehearsal (Rolnick et al., 2019; Shi et al., 2024).

Furthermore, although Key1 and Key10 achieved similar task performance, they differed in protection strength, highlighting the importance of exploring the performance–privacy trade-off.

In addition, a discussion of differential privacy (DP), which was excluded from our comparisons for computational efficiency, is warranted. As models pre-trained using DP, such as VaultGemma (Sinha et al., 2025), have emerged, further improvements in computational efficiency are expected. Beyond performance comparisons with DP, exploring hybrid approaches that combine both methods to achieve stronger privacy protection is a promising direction.

Finally, since cross-lingual transfer effects are often more pronounced in larger models, future work should investigate whether scaling beyond 551M parameters yields stronger gains.

Limitations

The primary limitation of this experiment is the restricted number of training tokens due to limited resources. It is anticipated that additional training data would further improve performance on downstream tasks. This might also affect the reconstruction capacity of PII. Furthermore, we could gain deeper insights and more comprehensive findings by training larger models.

Furthermore, in this experiment, we randomly divided the sampled FineWeb Edu data, treating 75% as pseudo-sensitive data. However, utilizing a corpus that includes actual sensitive information would allow us to evaluate the effectiveness of our approach in preventing data leakage compared to traditional methods. While our approach encrypted all lexically sensitive data, such as personal names in the training data, empirically demonstrating that no leakage occurs would significantly enhance the credibility of our method.

As with other methods, it is important to note that Crypto-LLM does not completely prevent the leakage of sensitive data. In large-scale probabilistic models like LLMs, it is generally understood that fully eliminating data leakage is practically infeasible.

These limitations highlight the potential for further research to improve and validate the robustness of our cipher-based approach to LLM training.

Acknowledgements

We thank three anonymous reviewers for their helpful comments and feedback. Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used for experiments. We would like to thank Editage (www.editage.jp) for English language editing.

Ethics Statement

This research adheres to the ACL's Code of Ethics and addresses key ethical considerations.

Data Usage: We utilized the publicly available Fineweb-Edu corpus, which is expected to contain no real sensitive personal information, ensuring compliance with usage terms.

Privacy and Data Security: Our study employs encryption as a research tool for pattern learning from encrypted text, but does not guarantee data security for sensitive information. The polyalphabetic substitution cipher used here explores theoretical aspects rather than providing robust protection.

Transparency and Reproducibility: We have made our code and methods publicly available on GitHub to support transparency and reproducibility, allowing others to replicate and expand upon our work.

Potential Misuse: While our method investigates data privacy, it is not a secure solution to protect sensitive data in practical applications. We caution against using it for critical data security needs.

Ethical Use: Our research is intended for academic exploration, contributing to discussions on data privacy. We advocate for its responsible and ethical use.

This statement aligns our research with ethical guidelines and underscores our commitment to responsible contributions in computational linguistics.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2024. Self-consuming generative models go mad. International Conference on Learning Representations (ICLR).

Ammar Waysi AlTuhafi. 2022. Adaptation for vigenère cipher method for auto binary files ciphering. In *ITM Web of Conferences*, volume 42, page 01017. EDP Sciences.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2021. Large-scale differentially private bert. *arXiv preprint arXiv:2108.01624*.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *Preprint*, arXiv:2311.05845.

Sebastian Rodriguez Beltran, Marlon Tobaben, Joonas Jälkö, Niki Loppi, and Antti Honkela. 2024. Towards efficient and scalable training of differentially private deep learning. *arXiv preprint arXiv:2406.17298*.

Mariusz Borowski and Marek Leśniewicz. 2012. Modern usage of "old" one-time pad. In *2012 Military Communications and Information Systems Conference (MCC)*, pages 1–5. IEEE.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. *Preprint*, arXiv:2202.07646.

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.

Zachary Charles, Arun Ganesh, Ryan McKenna, H Brendan McMahan, Nicole Mitchell, Krishna Pillutla, and Keith Rush. 2024. Fine-tuning large language models with user-level differential privacy. *arXiv preprint arXiv:2407.07737*.

Keyu Duan, Yiran Zhao, Zhili Feng, Jinjie Ni, Tianyu Pang, Qian Liu, Tianle Cai, Longxu Dou, Kenji Kawaguchi, Anirudh Goyal, and 1 others. 2025. Unnatural languages are not bugs but features for llms. *arXiv preprint arXiv:2503.01926*.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Danny Halawi, Alexander Wei, Eric Wallace, Tony T. Wang, Nika Haghtalab, and Jacob Steinhardt. 2024. Covert malicious finetuning: Challenges in safeguarding llm adaptation. *Preprint*, arXiv:2406.20053.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual llms to low-resource languages using continued pre-training and synthetic corpus. *Preprint*, arXiv:2410.14815.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. *arXiv preprint arXiv:2404.04042*.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*.

Guo Lin, Wenyue Hua, and Yongfeng Zhang. 2024. Emojicrypt: Prompt encryption for secure communication with large language models. *Preprint*, arXiv:2402.05868.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. *Preprint*, arXiv:2302.00539.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, and 1 others. 2023. Fingpt: Large generative models for a small language. *arXiv preprint arXiv:2311.05640*.

Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. Pii-scope: A benchmark for training data pii leakage assessment in llms. *arXiv preprint arXiv:2410.06704*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *Preprint*, arXiv:2311.17035.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

David Salomon. 2003. *Data privacy and security*. Springer.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Claude E Shannon. 1949. Communication theory of secrecy systems. *The Bell system technical journal*, 28(4):656–715.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Amer Sinha, Thomas Mesnard, Ryan McKenna, Daogao Liu, Christopher A Choquette-Choo, Yangsibo Huang, Da Yu, George Kaissis, Zachary Charles, Ruibo Liu, and 1 others. 2025. Vaultgemma: A differentially private gemma model. *arXiv preprint arXiv:2510.15001*.

Charles Sutton, Andrew McCallum, and 1 others. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Hiroto Tamura, Tosho Hirasawa, Hwichan Kim, and Mamoru Komachi. 2023. Does masked language model pre-training with artificial data improve low-resource neural machine translation? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2216–2225, Dubrovnik, Croatia. Association for Computational Linguistics.

Cagri Toraman. 2024. Llamaturk: Adapting open-source generative large language models for low-resource language. *Preprint*, arXiv:2405.07745.

Matej Ulčar and Marko Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced slovenian language. *Frontiers in Artificial Intelligence*, 6:932519.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *Preprint*, arXiv:2403.05156.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, and 1 others. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2024. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *Preprint*, arXiv:2308.06463.

A Appendix

A.1 Normalized Levenshtein Distance: Example

We illustrate how the normalized Levenshtein distance is computed at the morpheme level using an example. Let the ground truth suffix be S_1, where S_1 is "(Tom) has been lived in Canada for 10 years.". and the model generated suffix obtained via the true-prefix attack be

$$\hat{S_1}^* = \text{"(Tom) likes Canada and U.S."}$$

1. Remove all function words from both sequences. Function words are defined as tokens whose part-of-speech tag, according to SpaCy, corresponds to one of the following:
   - Auxiliary verbs (AUX)
   - Pronouns (PRON)
   - Determiners (DET)
   - Particles (PART)
   - Adpositions (ADP)
   - Coordinating conjunctions (CCONJ)
   - Subordinating conjunctions (SCONJ)
   - Interjections (INTJ)

   In addition, all punctuation symbols are also excluded.

   $$C_1 = [\text{"}lived\text{"}, \text{"}Canada\text{"}, \text{"10"}, \text{"}years\text{"}]$$
   $$\hat{C}_1^* = [\text{"}likes\text{"}, \text{"}Canada\text{"}, \text{"}U.S.\text{"}]$$

2. Using the content word lists from step 1, we compute the normalized Levenshtein distance, denoted as $d_{\text{Lev}}^{\text{norm}}$, as follows:

   $$d_{\text{Lev}}^{\text{norm}}(C_1, \hat{C}_1^*) = 0.75$$

A.2 Data Extraction Attack

A.2.1 Extracted Strings

In the data extraction attack, some strings were extracted repeatedly from multiple prompts. Table 5 shows both the total number of extractions and the number of unique extractions. Since repeated strings of the same information do not provide additional value in the context of information extraction, we use the number of unique extractions as the primary evaluation metric.

Table 5: Numbers of Total and Unique Extracted Strings

Model	Total N	Unique N
Plain-LLM PT+CPT	59	35
Plain-LLM CPT	0	0
Crypto-LLM Key1	10	6
Crypto-LLM Key10	1	1
Crypto-LLM Key100	2	2
Scrubbing-LLM	148	28

The extracted strings are shown below. The leading number indicates the frequency of occurrence.

Plain-LLM PT+CPT

- 9: billion nsf funds reach all 50 states through grants to nearly 2 000 colleges universities and other institutions each year nsf receives more than 48 000 competitive proposals for funding and makes about 12 000
- 3: - spacedaily afp and upi wire stories are copyright agence france-presse and united press international esa portal reports are copyright european space agency all nasa sourced material is public domain additional copyrights may apply in
- 3: the information provided herein should not be used during any medical emergency or for the diagnosis or treatment of any medical condition a licensed medical professional should be consulted for diagnosis and treatment of any
- 3: world encyclopedia writers and editors rewrote and completed the wikipedia article in accordance with new world encyclopedia standards this article abides by terms of the creative commons cc-by-sa 3 0 license cc-by-sa which may be
- 2: always seek the advice of your physician or other qualified health provider with any questions you may have regarding a medical condition never disregard professional medical advice or delay in seeking it because of something
- 2: the american heritage dictionary of the english language fifth edition copyright 2018 by houghton mifflin harcourt publishing company all rights reserved indo-european semitic roots appendices thousands of entries in the dictionary include etymologies that trace
- 2: nih is the primary federal agency conducting and supporting basic clinical and translational medical research and is investigating the causes treatments and cures for both common and rare diseases for more information about nih and
- 2: by houghton mifflin harcourt publishing company all rights reserved indo-european semitic roots appendices thousands of entries in the dictionary include etymologies that trace their origins back to reconstructed protolanguages you can obtain more information about
- 2: - 1199 surnames became necessary when governments introduced personal taxation in england this was known as poll tax throughout the centuries surnames in every country have continued to develop often leading to astonishing variants of
- 2: article distributed under the terms of the creative commons attribution license http creativecommons org licenses by 4 0 which permits unrestricted use distribution and reproduction in any medium provided the original work is properly cited
- 2: of the u s department of health and human services nih is the primary federal agency conducting and supporting basic clinical and translational medical research and is investigating the causes treatments and cures for both
- 2: for more information about nih and its programs visit www nih gov about the national institutes of health nih nih the nation's medical research agency includes 27 institutes and centers and is a component of
- 2: purposes only it is not intended to be a substitute for professional medical advice diagnosis or treatment always seek the advice

of your physician or other qualified health provider with any questions you may have

- 2: conducting and supporting basic clinical and translational medical research and is investigating the causes treatments and cures for both common and rare diseases for more information about nih and its programs visit www nih gov
- 1: educational purposes only it is not intended to be a substitute for professional medical advice diagnosis or treatment always seek the advice of your physician or other qualified health provider with any questions you may
- 1: all rights reserved please be aware that this information is provided to supplement the care provided by your physician it is neither intended nor implied to be a substitute for professional medical advice call your
- 1: it is not intended as medical advice for individual conditions or treatments talk to your doctor nurse or pharmacist before following any medical regimen to see if it is safe and effective for you what
- 1: is for informational purposes only and is not intended to be a substitute for professional medical advice diagnosis or treatment always seek the advice of your physician or other qualified health provider with any questions
- 1: 1272 - 1307 surnames became necessary when governments introduced personal taxation in england this was known as poll tax throughout the centuries surnames in every country have continued to develop often leading to astonishing variants
- 1: for the u s department of energy's office of science the u s department of energy's office of science is the single largest supporter of basic research in the physical sciences in the united states
- 1: this article is for information only and should not be used for the diagnosis or treatment of medical conditions patient platform limited has used all reasonable care in compiling the information but make no warranty
- 1: there is low ocean tide on this date sun and moon gravitational forces are not aligned but meet at big angle so their combined tidal force is weak the moon is 2 days young earth's
- 1: www nimh nih gov about the national institutes of health nih nih the nation's medical research agency includes 27 institutes and centers and is a component of the u s department of health and human
- 1: shall make no law respecting an establishment of religion or prohibiting the free exercise thereof or abridging the freedom of speech or of the press or the right of the people peaceably to assemble and
- 1: wa also reviewed by david zieve md mha medical director a d a m inc the information provided herein should not be used during any medical emergency or for the diagnosis or treatment of any
- 1: 1189 - 1199 surnames became necessary when governments introduced personal taxation in england this was known as poll tax throughout the centuries surnames in every country have continued to develop often leading to astonishing variants
- 1: this is an open-access article distributed under the terms of the creative commons attribution license cc by the use distribution or reproduction in other forums is permitted provided the original author s and the copyright
- 1: for informational purposes only and is not intended to be a substitute for professional medical advice diagnosis or treatment always seek the advice of your physician or other qualified health provider with any questions you
- 1: university of washington school of medicine the information provided herein should not be used during any medical emergency or for the diagnosis or treatment of any medical condition a licensed medical professional should be consulted
- 1: the content herein unless otherwise known to be public domain are copyright 1995-2010 - spacedaily afp and upi wire stories are copyright agence france-presse and united press international esa portal reports are copyright european space
- 1: et al this is an open-access article distributed under the terms of the creative commons attribution license which permits unrestricted use distribution and reproduction in

any medium provided the original author and source are credited

- 1: which states that congress shall make no law respecting an establishment of religion or prohibiting the free exercise thereof or abridging the freedom of speech or of the press or the right of the people
- 1: the catholic encyclopedia is the most comprehensive resource on catholic teaching history and information ever gathered in all of human history this easy-to-search online version was originally printed in fifteen hardcopy volumes designed to present
- 1: copyright 2008 by the american academy of family physicians this content is owned by the aafp a person viewing it online may make one printout of the material and may use that printout only for
- 1: surnames became necessary when governments introduced personal taxation in england this was known as poll tax throughout the centuries surnames in every country have continued to develop often leading to astonishing variants of the original

Crypto-LLM Key1

- 4: it is not intended to be a substitute for professional medical advice diagnosis or treatment always seek the advice of your physician or other qualified health provider with any questions you may have regarding a
- 2: the catholic encyclopedia is the most comprehensive resource on catholic teaching history and information ever gathered in all of human history this easy-to-search online version was originally printed between 1907 and 1912 in fifteen hard
- 1: a person viewing it online may make one printout of the material and may use that printout only for his or her personal non-commercial reference this material may not otherwise be downloaded copied printed stored
- 1: et al this is an open-access article distributed under the terms of the creative commons attribution license which permits unrestricted use distribution and reproduction in any medium provided the original author and source are credited
- 1: purposes only it is not intended to be a substitute for professional medical advice diagnosis or treatment always seek the advice of your physician or other qualified health provider with any questions you may have
- 1: the cassini-huygens mission is a cooperative project of nasa the european space agency and the italian space agency the jet propulsion laboratory a division of the california institute of technology in pasadena manages the mission

Crypto-LLM Key10

- 1: the i0nformation provided herein should not be used during any medical emergency or for the diagnosis or treatment of any medical condition a licensed medical professional should be consulted for diagnosis and treatment of any

Crypto-LLM Key100

- 1: new world encyclopedia writers and editors rewrote and completed the wikipedia article in accordance with new world encyclopedia standards this article abides by terms of the creative commons cc-by-sa 3 0 license cc-by-sa which may
- 1: world encyclopedia writers and editors rewrote and completed the wikipedia article in accordance with new world encyclopedia standards this article abides by terms of the creative commons cc-by-sa 3 0 license cc-by-sa which may be

Scrubbing-LLM

- 89: the u s department of energy s office of science is the single largest supporter of basic research in the physical sciences in the united states and is working to address some of the most
- 8: this document is subject to copyright apart from any fair dealing for the purpose of private study or research no part may be reproduced without the written permission the content is provided for information purposes
- 5: by harpercollins publishers all rights reserved indo-european semitic roots appendices thousands of entries in the dictionary include etymologies that trace their origins back to reconstructed proto-languages you can obtain more information about these forms in

- 5: the u s department of energy's office of science is the single largest supporter of basic research in the physical sciences in the united states and is working to address some of the most pressing
- 4: the catholic encyclopedia is the most comprehensive resource on catholic teaching history and information ever gathered in all of human history this easy-to-search online version was originally printed in fifteen hardcopy volumes designed to present
- 4: world encyclopedia writers and editors rewrote and completed the wikipedia article in accordance with new world encyclopedia standards this article abides by terms of the creative commons cc-by-sa 3 0 license cc-by-sa which may be
- 3: the american heritage dictionary of the english language fifth edition copyright 2018 by houghton mifflin harcourt publishing company all rights reserved indo-european semitic roots appendices thousands of entries in the dictionary include etymologies that trace
- 3: surnames became necessary when governments introduced personal taxation in england this was known as poll tax throughout the centuries surnames in every country have continued to develop often leading to astonishing variants of the original
- 3: is not intended to be a substitute for professional medical advice diagnosis or treatment always seek the advice of your physician or other qualified health provider with any questions you may have regarding a medical
- 2: conducting and supporting basic clinical and translational medical research and is investigating the causes treatments and cures for both common and rare diseases for more information about nih and its programs visit www nih gov
- 2: a web site to get translated content where available and see local events and offers based on your location we recommend that you select you can also select a web site from the following list
- 2: the u s department of energy's office of science the u s department of energy's office of science is the single largest supporter of basic research in the physical sciences in the united states and
- 2: the doe office of science is the single largest supporter of basic research in the physical sciences in the united states and is working to address some of the most pressing challenges of our time
- 2: feedback if you see any errors or have any questions or suggestions on what is shown on this page please fill in the feedback form so that we can correct or extend the information provided
- 1: century the catholic encyclopedia is the most comprehensive resource on catholic teaching history and information ever gathered in all of human history this easy-to-search online version was originally printed in fifteen hardcopy volumes designed to
- 1: - 1307 surnames became necessary when governments introduced personal taxation in england this was known as poll tax throughout the centuries surnames in every country have continued to develop often leading to astonishing variants of
- 1: office of science is the single largest supporter of basic research in the physical sciences in the united states and is working to address some of the most pressing challenges of our time for more
- 1: - authors are able to enter into separate additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work e g post it to an institutional repository or publish it in
- 1: for the u s department of energy's office of science the u s department of energy's office of science is the single largest supporter of basic research in the physical sciences in the united states
- 1: 2018 truven health analytics inc information is for end user's use only and may not be sold redistributed or otherwise used for commercial purposes all illustrations and images included in carenotes are the copyrighted property
- 1: and grant the journal right of first publication authors are able to enter into separate additional contractual arrangements for the

non-exclusive distribution of the journal's published version of the work e g post it to

- 1: this article is for information only and should not be used for the diagnosis or treatment of medical conditions emis has used all reasonable care in compiling the information but make no warranty as to
- 1: in - british world english dictionary what do you find interesting about this word or phrase comments that don't adhere to our community guidelines may be moderated or removed most popular in the us most
- 1: office of science the u s department of energy's office of science is the single largest supporter of basic research in the physical sciences in the united states and is working to address some of
- 1: doe's office of science is the single largest supporter of basic research in the physical sciences in the united states and is working to address some of the most pressing challenges of our time for
- 1: for the u s department of energy s office of science the office of science is the single largest supporter of basic research in the physical sciences in the united states and is working to
- 1: definitions with the community word of the day would you like us to send you a free new word definition delivered to your inbox daily use the citation below to add this definition to your
- 1: 2020 this document is subject to copyright apart from any fair dealing for the purpose of private study or research no part may be reproduced without the written permission the content is provided for information

A.2.2 Justification for Compression Ratio Threshold and Digit Normalization

We describe the differences between our approach and that of Nasr et al. (2023). First, while the original paper uses a threshold of 50 tokens for suffix array matches, we instead evaluate on a word basis to reduce the tokenization computation cost. Based on an evaluation using 1,000 randomly sampled texts from Fineweb-Edu sample-10BT, we found that 50 tokens correspond to approximately 35.4 words using our tokenizer, and therefore we set the threshold at 35 words.[A17.1]

Second, although the original paper excludes prompts with low entropy, it does not specify how entropy is calculated. In our work, we exclude matches whose compression ratio (computed by first replacing all digits 1 through 9 with 0, then applying the DEFLATE algorithm via Python's zlib.compress) falls below 0.275. The upper bound of the compression ratio is determined by entropy (Shannon, 1948). This threshold corresponds to the top 0.01% most compressible strings in the Fineweb-Edu sample-10B dataset. The digit replacement step is introduced to filter out sequences of numbers, such as years or page numbers.

However, this threshold is ad hoc and not based on any theoretical justification. In the following[A18.1], we evaluate the justification for these choices.

First, we evaluated which strings would be excluded under compression thresholds of [0.2, 0.25, 0.3, 0.35, 0.4]. For thresholds from 0.25 to 0.4, the exclusion results were identical to those obtained with 0.275. The table 6 shows the number of unique strings that remain at each threshold. When the threshold was set to 0.2, more strings remained, but most consisted only of digits, whitespace, and hyphens. Such strings are considered to carry little meaningful information and were therefore excluded.

The only example that contains characters other than digits, whitespace, and hyphens was output by Plain-LLM PT+CPT, as shown below. However, this consists of repeated occurrences of the same string, making it a clear case of high compressibility.

> new york university school of medicine new york university school of medicine new york university school of medicine new york university school of medicine new york university school of medicine

Next, we present the number of unique strings

Table 6: Comparison of Extracted Unique Strings by Compression Ratio

Model	0.2	0.25	0.3	0.35	0.4
Plain-LLM PT+CPT	47	35	35	35	35
Plain-LLM CPT	0	0	0	0	0
Crypto-LLM Key1	28	6	6	6	6
Crypto-LLM Key10	7	1	1	1	1
Crypto-LLM Key100	7	2	2	2	2
Scrubbing-LLM	33	28	28	28	28

obtained without applying digit-to-zero replacement (table 7). In this case as well, most of the additional strings consist only of digits, whitespace, and hyphens.

Table 7: Comparison of Extracted Unique Strings by Compression Ratio

Model	0.2	0.25	0.3	0.35	0.4
Plain-LLM PT+CPT	150	142	142	127	122
Plain-LLM CPT	0	0	0	0	0
Crypto-LLM Key1	114	97	97	97	84
Crypto-LLM Key10	51	50	50	50	50
Crypto-LLM Key100	15	15	15	15	15
Scrubbing-LLM	100	100	100	100	100

The exceptions are strings containing "page," "chapter," or "x" along with numbers and symbols. Although the numeric values differ, the word-number patterns fall into one of the following three formats.

Chapter 12, Chapter 13, Chapter 14, Chapter 15, Chapter 16, Chapter 17, Chapter 18, Chapter 19, Chapter 20, Chapter 21, Chapter 22, Chapter 23, Chapter 24, Chapter 25, Chapter 26, Chapter 27, Chapter 28, Chapter 29, Chapter 30, Chapter 31, Chapter 32, Chapter 33, Chapter 34, Chapter 35, Chapter 36, Chapter 37, Chapter 38, Chapter 39, Chapter 40, Chapter 41, Chapter 42, Chapter 43, Chapter 44, Chapter 45, Chapter 46, Chapter 47, Chapter 48, Chapter 49, Chapter 50, Chapter 51, Chapter 52, Chapter 53, Chapter 54, Chapter 55, Chapter 56, Chapter 57, Chapter 58, Chapter 59, Chapter 60, Chapter 61, Chapter 62, Chapter 63, Chapter 64, Chapter 65, Chapter 66, Chapter 67, Chapter 68, Chapter 69, Chapter 70, Chapter 71, Chapter 72, Chapter 73, Chapter 74, Chapter 75, Chapter 76, Chapter 77, Chapter 78, Chapter 79, Chapter 80, Chapter 81, Chapter 82, Chapter 83, Chapter 84, Chapter 85, Chapter 86, Chapter 87, Chapter 88, Chapter 89, Chapter 90, Chapter 91, Chapter 92, Chapter 93, Chapter 94, Chapter 94, Chapter 95, Chapter 96, Chapter 97, Chapter 98, Chapter 99, Chapter 100, Chapter 101, Chapter 102, Chapter 103, Chapter 104, Chapter 105, Chapter 106, Chapter 107, Chapter 108, Chapter 109

page 191 191 page 191 191 page 192

x 8" x 8"

A.3 Differences in Syntactic Characteristics by Encryption Strength

We analyzed how encryption alters token-to-token dependencies using n-gram statistics and conditional entropies. The results are shown in Table 8. For Key Length = 100, the 2-gram conditional entropy is higher, indicating that the uncertainty of predicting the next token from the preceding one is greater than in plaintext or for Key Lengths 1 and 10. In contrast, for 3-grams and 4-grams, the conditional entropy decreases, suggesting that the natural syntactic dependencies are distorted under stronger encryption. This trend may be related to the degradation in downstream task performance observed for Key Length = 100 in Table 3. Additionally, we included Table 9, reporting changes

Table 8: Unigram and Conditional Entropies (bits) Across Text Types (Plaintext and Encrypted with Different Key Lengths). Unigram entropy reflects overall unpredictability, while conditional entropies (2- to 4-gram) indicate how much uncertainty remains given a short-range context, thus highlighting the strength or weakness of syntactic dependencies.

Text Type	**Unigram Entropy**	**2-gram** $H(X\|X_{-1})$	**3-gram** $H(X\|X_{-2}, X_{-1})$	**4-gram** $H(X\|X_{-3}, X_{-2}, X_{-1})$
Plaintext	10.760	7.373	4.608	2.106
Key Length = 1	10.756	7.372	4.614	2.109
Key Length = 10	12.276	7.233	4.116	1.839
Key Length = 100	12.446	8.049	3.741	1.624

Table 9: Distributional Characteristics of n-grams ($n = 1$–4) across Plaintext and Encrypted Text with Different Key Lengths. Reported are the number of unique n-grams, the Jensen–Shannon distance to the uniform distribution, and the unbiased effective vocabulary size (EVS).

Text Type	n	**Unique n-grams**	**JS Distance to Uniform**	**EVS (Unbiased)**
Plaintext	1	30,941	0.497	137.390
Key Length = 1	1	28,598	0.486	138.887
Key Length = 10	1	31,059	0.362	413.303
Key Length = 100	1	31,281	0.384	668.280
Plaintext	2	7,621,231	0.528	3,536.624
Key Length = 1	2	7,513,612	0.531	3,592.550
Key Length = 10	2	9,135,218	0.497	9,895.890
Key Length = 100	2	11,047,697	0.453	18,988.857
Plaintext	3	32,360,899	0.245	65,233.510
Key Length = 1	3	32,294,297	0.246	66,103.317
Key Length = 10	3	43,434,828	0.225	148,246.656
Key Length = 100	3	57,778,692	0.216	280,586.335
Plaintext	4	55,167,510	0.098	415,156.694
Key Length = 1	4	55,300,622	0.098	419,898.627
Key Length = 10	4	74,685,311	0.088	883,525.819
Key Length = 100	4	97,770,052	0.093	1,607,419.221

in n-gram distributions and effective vocabulary size. While these analyses are exploratory, they demonstrate that the degree to which syntactic dependencies are preserved varies with encryption strength, and they provide initial insights into the reviewer's question about what kinds of structures are being learned.

A.4 Model Configurations and GPU Resources

The configurations used for training the Llama-3 551M models in this study are outlined below.

Hidden Size 1536

Number of Layers 16

Attention Heads 12

Batch size 32

Sequence Length 2,048

Multiple of 2.0×10^{-4}

Normalization Epsilon 1.0×10^{-5}

Rotary positional embedding base 10,000.0

Optimizer AdamW

Vocabulary Size 32,000

Each model was trained on a node equipped with eight H200 GPUs, with the pre-training phase taking approximately two hours and the continual pre-training phase taking approximately 40 minutes.

A.5 Detailed Description of True-Prefix Attack

This section provides a detailed description of the True-Prefix Attack experiment conducted in this study. We extracted 1,000 target sentences from texts containing 18,051 pseudo-PII instances using the following procedure:

1. Starting from the beginning of each text, we identified sentences containing name entities as detected by SpaCy.

2. To manage computational cost and task difficulty, we excluded sentences whose token length measured using the tokenizer exceeded

20. In such cases, we continued scanning the same text for the next sentence that contains a name entity. If the token length was 20 or fewer, we proceeded to step 3.

3. Each sentence was segmented into S_0, name, and S_1. If S_1 was empty (i.e., the name appeared at the end of the sentence), the sentence was discarded and the search resumed as in step 2. If S_1 contained any characters, the sentence was retained for use in the experiment.

For the true-prefix attack, we perform sampling 100 times for each prefix. During inference, we use temperature $= 1$ and set top-k $= 64$.