Memorization is Language-Sensitive: Analyzing Memorization and Inference Risks of LLMs in a Multilingual Setting

Ali Satvaty[1] **Anna Visman**[1] **Daniel Seidel**[1] **Suzan Verberne**[2] **Fatih Turkmen**[1]
[1]University of Groningen [2]Leiden University
Correspondence: `a.satvaty@rug.nl`

Abstract

Large Language Models (LLMs) are known to memorize and reproduce parts of their training data during inference, raising significant privacy and safety concerns. While this phenomenon has been extensively studied to explain its contributing factors and countermeasures, its implications in multilingual contexts remain largely unexplored.

In this work, we investigate cross-lingual differences in memorization behaviors of multilingual LLMs. Specifically, we examine both discoverable memorization and susceptibility to perplexity ratio attacks using Pythia models of varying sizes, evaluated on two parallel multilingual datasets.

Our results reveal that lower-resource languages consistently exhibit higher vulnerability to perplexity ratio attacks, indicating greater privacy risks. In contrast, patterns of discoverable memorization appear to be influenced more strongly by the model's pretraining or fine-tuning phases than by language resource level alone. These findings highlight the nuanced interplay between language resource availability and memorization in multilingual LLMs, providing insights toward developing safer and more privacy-preserving language models across diverse linguistic settings.[1]

1 Introduction

Current transformer-based large language models (LLMs) have billions of parameters and are trained on massive datasets (Hartmann et al., 2023). This scaling has increased the ability of LLMs to process and mimic fluent human language, as well as to perform a wide range of other tasks (Ishihara, 2023; Wei et al., 2022). Recent advancements have shown remarkable performance of models in diverse applications, from machine translation and summarization to question answering and planning (Zhao et al., 2025; Chang et al., 2024). When trained on datasets containing multiple languages, LLMs learn multilingual capabilities and can understand and generate text in different languages.

Although English continues to dominate the training data for language models, multilingual models have also emerged in recent years, driven by growing global interest in making language technologies more accessible across different languages. However, the development of multilingual models presents several challenges (Naveed et al., 2024): The vast majority of languages worldwide are underrepresented and mid- and low-resource languages receive less attention from the NLP community (Joshi et al., 2020). Lower-resource languages also constitute a small fraction of the labeled data available for finetuning popular LLMs, leading to poorer performance on typical downstream NLP tasks compared to high-resource ones (Lai et al., 2023).

It has been shown that prompting ChatGPT[2] in a lower-resource language, can circumvent the model's safety and security mechanisms, triggering it to produce responses that would not be possible in English or other high-resource languages. This highlights a cross-language vulnerability, most likely arising from differences in the availability of training data (Yong et al., 2024). This indicates that LLM privacy and security need to be studied in a multilingual context.

Research suggests that LLMs have the potential to expose training data through memorization (Carlini et al., 2021). This undesirable phenomenon can occur either accidentally or through deliberate extraction by adversaries (Carlini et al., 2019), who attempt to recover individual training examples by querying the model (Carlini et al., 2021; Satvaty et al., 2025). When the training data, user prompts, or model responses contain sensitive information that can be traced back to individuals, either in

[1] Our code and preprocessed datasets are available at https://github.com/alistvt/xlm-privacy

[2] `https://chatgpt.com/`

Proceedings of the First Workshop on Large Language Model Memorization (L2M2), pages 106–126
August 1, 2025 ©2025 Association for Computational Linguistics

isolation or in combination (Lai et al., 2023; Yan et al., 2024), it becomes an issue of privacy and ethical implications (Ishihara, 2023). If the training dataset is confidential, any exposure through a training data extraction attack constitutes a privacy breach, regardless of the nature of the data or context (Nasr et al., 2023). What we observe here are models exhibiting privacy vulnerabilities that can be exploited through adversarial prompting or data extraction attacks, all of which stem from the memorization issue (Ishihara, 2023; Carlini et al., 2023b; Shayegani et al., 2023; Zhang et al., 2023a).

While memorization and the associated privacy risks in LLMs have been extensively studied, their comparison in multilingual model scenarios remains unexplored. In light of this, in this work, **we analyze and compare the memorization rates in multilingual LLMs between lower-resource and high-resource languages.** To this end, we conduct two different experiments on a set of Pythia models (Biderman et al., 2023) with different model sizes. The Pythia models were trained on predominantly English data, but they are capable of generating other European languages. With 'lower-resource' we refer to medium-sized languages that the Pythia models were not explicitly trained on.

- Assessing *discoverable memorization* and evaluating the *perplexity ratio* over two different parallel datasets containing texts in English, Dutch, Slovenian, Polish and Czech.
- Analyzing both these aspects in both the pretraining and finetuning phases of LLMs.

Our results provide empirical evidence that lower-resource languages show higher perplexity ratio values, suggesting that they would be more susceptible to membership inference attacks (MIA) based on this method. On the other hand, our discoverable memorization test shows that lower-resource language datasets are memorized more if those are contained in the pretraining dataset, while they are less memorized if introduced during the finetuning.

Our experiments and results underline the significance of having more balanced datasets when training a multilingual dataset, otherwise model owners should be aware of the risks associated with introducing lower-resource datasets during model training.

2 Background and related work

2.1 Memorization

Memorization refers to the ability of a model to recall specific data points or patterns that it has encountered during the training process (Satvaty et al., 2025; Carlini et al., 2023a). While Carlini et al. (2019) first introduced *verbatim memorization* in language models to only include the cases with exact string match, Ippolito et al. (2023) observed that the LLM outputs could be traced back to the training data with subtle modifications. More specifically, they introduced *approximate memorization*[3] for the cases where the generated texts could be assigned to a training sample if their similarity – measured through a similarity function – is below a certain threshold. This could be exploited when the LLM is prompted with trivial changes to the original prompt, causing it to output memorized, but not verbatim, content. Given this definition, Ippolito et al. (2023) showed that LLMs memorize their training data several factors more than what was previously assumed.

Memorization can be studied through *discoverable* or *extractable* methods (Satvaty et al., 2025; Nasr et al., 2023). *Discoverable* memorization accounts for the samples that are correctly generated when the model is prompted with the first part of those samples. This requires that we have access to the training data and interact with the model through prompting, expecting the generation of the training samples. In the case of*extractable memorization*, interaction with the model is performed by an adversary, without having access to the training data. Extractable memorization is potentially more problematic in real-world scenarios, as the training data is not known to end-users interacting with the LLMs.

The phenomenon of memorization in LLMs occurs due to repeated instances of near-duplicate examples and long repetitive sub-strings in the training corpus (Carlini et al., 2021; Ishihara, 2023), where the model assigns greater importance to more frequent instances, making them more likely to be memorized (Kassem et al., 2023). Apart from repeated instances of training data, other factors that influence memorization in LLMs include the size of the dataset, the complexity of the data, and the size of the model (Tirumala et al., 2022a;

[3] Sometimes referred to as "style transfer" due to the way it is exploited.

Prashanth et al., 2024; Carlini et al., 2023a; Zhang et al., 2023a; Lesci et al., 2024).

2.2 Measuring memorization

The most widely adopted approach to measuring memorization in LLMs is the *string match* metric, which quantifies the rate at which training instances are generated either verbatim or approximately, normalized by the number of trials.

However, the string match metric has certain limitations. Since LLM outputs are produced via a stochastic decoding process, the absence of a particular training sentence in a finite number of generations does not conclusively indicate that the model would never produce it. To address this uncertainty, alternative approaches have been proposed to estimate memorization. One such method involves using the success rate of certain privacy attacks as a proxy for memorization, under the assumption that these attacks expose the model's higher confidence on the samples observed during the training. For instance, membership inference attacks (MIAs), one of the most studied inference attacks in machine learning, attempt to determine whether a specific data point was part of the training data. A commonly used measurement technique in this context is the *perplexity ratio* method. Models tend to assign lower perplexity to samples they have seen during training; thus, by dividing the model's perplexity on a sample before training by its perplexity after training, one typically obtains a ratio that is greater for unseen data than for training data. This ratio can then be used as a threshold-based decision criterion for inferring membership (Mattern et al., 2023; Shachor et al., 2024; Shejwalkar et al., 2021; Jagannatha et al., 2021; Wang et al., 2022). Formally, perplexity is obtained through the token-wise average negative likelihood of the model on a given sample as sequence of tokens:

$$\mathit{PPX}_M(S) = e^{-\frac{1}{N}\sum_{i=1}^{N}\log P_M(x_i|x_1,\ldots,x_{i-1})} \quad (1)$$

Where N is the count of tokens in the sample S and x_i represents the individual tokens in $S = (x_1, ..., x_N)$ and M represents the model; then the perplexity ratio is obtained as follows:

$$\mathit{PPX\text{-}ratio}(S) = \frac{PPX_{untrained}(S)}{PPX_{trained}(S)} \quad (2)$$

In our work, we look at discoverable approximate and verbatim memorization, and the susceptibility of LLMs to MIA under the perplexity ratio method to compare the memorization rates of the models. This combined analysis enables a more comprehensive, practical, and multifaceted understanding and comparison of memorization between lower-resource and high-resource languages.

2.3 Lower-resource languages

Today's NLP research predominantly focuses on only a fraction of the world's languages, rendering the majority of them understudied (Joshi et al., 2020). Lower-resource languages are characterized by limited available training data, low computerization, low privilege, and limited educational presence, among other things (Magueresse et al., 2020). To address the data scarcity inherent in lower-resource languages, a key trend involves augmenting existing high-resource language datasets and employing transfer learning techniques to mitigate their differences by taking advantage of linguistic similarities (Magueresse et al., 2020).

Research has revealed poorer performance and safety vulnerabilities of LLMs across different language categories (Yong et al., 2024; Nigatu and Raji, 2024; Zhang et al., 2023b). However, cross-lingual vulnerabilities for training data leakage and privacy risks still remain unexplored. To the best of our knowledge, there is no generalization regarding specific memorization and privacy vulnerabilities of multilingual LLMs in different linguistic contexts (Yong et al., 2024), and existing defense mechanisms currently do not comply with the reality of the multilingual modern world. Expanding this investigation across lower- and mid-resource languages regarding memorization is essential for a comprehensive understanding of the broader linguistic landscape and the privacy risks associated with LLMs.

In this work, we analyze memorization in several lower-resource languages, in contrast to English as a high-resource language. Specifically, since the model under study is trained on less than 1% of data from these languages, we argue that they serve as reasonable representatives of lower-resource languages.

3 Analysis methods

To compare the memorization phenomena between the lower-resource languages and higher-resource ones, we measure discoverable memorization across languages in both pretraining and

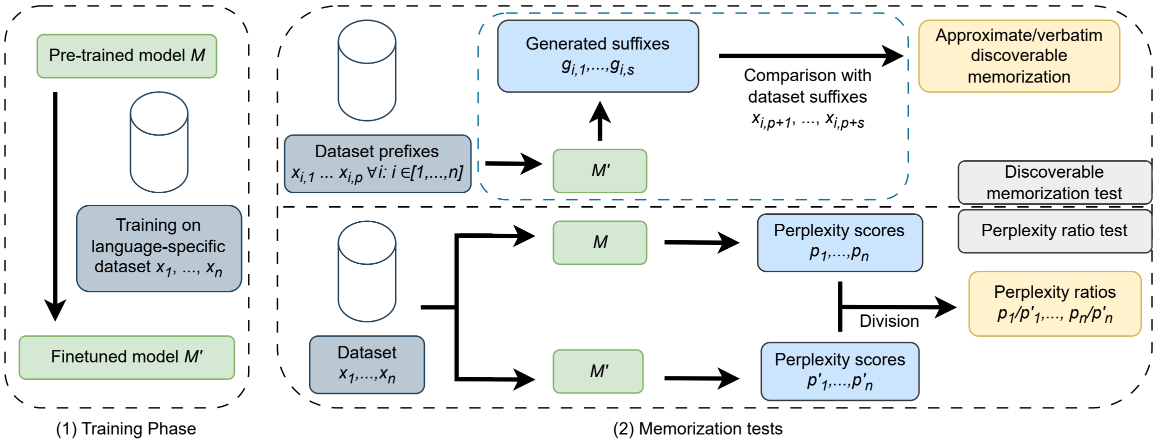
Pre-trained model M
Training on language-specific dataset x_1, ..., x_n
Finetuned model M'
(1) Training Phase
Dataset prefixes $x_{i,1}$ ... $x_{i,p}$ ∀i: i ∈[1,...,n]
M'
Generated suffixes $g_{i,1}$,...,$g_{i,s}$
Comparison with dataset suffixes $x_{i,p+1}$, ..., $x_{i,p+s}$
Approximate/verbatim discoverable memorization
Discoverable memorization test
Perplexity ratio test
Dataset x_1,...,x_n
M
M'
Perplexity scores p_1,...,p_n
Perplexity scores p'_1,...,p'_n
Division
Perplexity ratios p_1/p'_1,..., p_n/p'_n
(2) Memorization tests

Figure 1: Methodological overview of the training and memorization measurement process: x_i denotes the samples of the dataset of size n and $x_{i,j}$ denotes the sample's individual tokens. (1) The model M is trained on each language subset, obtaining M', then (2) memorization is measured using two experiments: (above) discoverable approximate/verbatim memorization is evaluated, (below)each training sample is passed through the trained and untrained models to obtain their respective perplexity scores. Then, the perplexity ratio of each sample is reported.

Dataset	Language	Tokens	Ratio	Step
EMEA	EN	1,295,108	1	32
	NL	2,155,528	1.66	53
	SL	2,542,529	1.96	62
	PL	3,027,591	2.33	74
	CS	2,267,772	1.75	56
EuroParl	EN	1,667,939	1	32
	NL	2,928,249	1.75	56
	SL	3,152,403	1.88	60
	PL	3,799,024	2.27	72
	CS	3,594,028	2.15	68

Table 1: Statistics of the datasets: tokens column represents the number of tokens in the training set, according to the Pythia tokenizer. Ratio represents the division of the token count of each language to that of the English language. Step is the equivalent normalized token count for each language when the English context is considered 32 tokens.

finetuning scenarios. An overview of our methods is shown in Figure 1.

Setting the context size Previous research (Carlini et al., 2023a) has shown that memorization is affected by the context given to the LLM. Providing more context as the prefix helps LLMs better recall the suffix. The token count of the context depends on the tokenizer used by the LLM. Since we are dealing with parallel texts in multiple languages, the same (parallel) context comprises different amounts of tokens in different languages. As shown in Table 1, a context size of 32 tokens in English, on average is equivalent to 53 Dutch tokens.[4] When measuring verbatim memorization, providing 32 tokens for both English and Dutch could result in lower memorization in the Dutch case due to lower context provided in Dutch. In order to remove this effect, we provide the same amount of context based on the ratios in Table 1. Through the rest of this paper, we will mention this approach as *normalizing token lengths*. This approach helps us remove the effect of different context and purely focus on the differences in terms of high-resource and low-resource languages.

3.1 Finetuning analysis

We choose a dataset that is not included in the pretraining set of our models to further finetune the model. Since we do not want the experiments to be affected by catastrophic forgetting (Kirkpatrick et al., 2017), for each language subset, we finetune the pretrained Pythia independently up to 8 epochs and run our inference experiments on the obtained version.

After training and fine-tuning, we measure the discoverable verbatim and approximate memorization. For each sample in the dataset, we also compute the perplexity ratio between the untrained and trained model, indicating its susceptibility to MIA. Finally we report this ratio in a histogram based on the *normalized token length*.

[4] You can also refer to Table C in the appendix for a solid example.

3.2 Pretraining phase analysis

To analyze the memorization of pretraining datasets, we conduct discoverable, verbatim and approximate memorization experiments on a parallel dataset included in the pretraining. Furthermore, we conduct the perplexity ratio test to compare the MIA susceptibility of different languages during pretraining.

However, in the case of pretraining, the untrained model does not have any language modeling capabilities, which yields it assigning very high and near random perplexity values to all samples. As the chosen model (Section 4.1) is available in different steps of its training as checkpoints, we can estimate the untrained perplexity on a sample S using the perplexity of the 25% pretrained checkpoint:

$$PPX_{utrained}(S) \approx PPX_{25\% pretrained}(S) \quad (3)$$

One immediate concern here is since our target dataset is uniformly distributed throughout the pretraining process, we do not know the exact step at which each sample was first introduced to the model. By using the 25% checkpoint to approximate the untrained perplexity, we acknowledge that approximately 25% of the samples may have already been seen by the model at that point. However, this does not compromise the validity of our experiments, as the same assumption holds across all language subsets. Thus, the use of the 25% checkpoint as a proxy for untrained perplexity provides a consistent and fair basis for comparison across languages, without introducing systematic bias.

4 Experiments

In this section, we first provide details about our experimental setup including the employed LLMs and the data sets. We then motivate the choice of the languages and summarize the used metrics.

4.1 Models

We use Pythia models (Biderman et al., 2023) in our experiments as they are widely used within the LLM memorization community (Satvaty et al., 2025). These models are available in different sizes, enabling us to analyze model size as a dimension in our experiments. We use four model sizes: $70m$, $160m$, $410m$, and $1B$ parameters.

Furthermore, these models are fully open and accessible. Therefore, we have precise information about the datasets that have been used during their pretraining which is important for selecting suitable datasets for our experiments. Lastly, since these models are available at different checkpoints of their training steps, we can obtain a good estimation for our MIA study as discussed in Section 3.2.

Pythia models are not known for their multilingual capabilities, as they are trained on the PILE dataset (Gao et al., 2020), which predominantly contains English content. To some extent, these models have the ability to understand and generate other languages that were present in small amounts in their pretraining data (Xu et al., 2025). As the pretraining data was mainly English, we can consider the other languages as lower-resource in this context.

4.2 Datasets

We opted for parallel datasets for our experiments, meaning that the datasets share the same content across different languages. This helps us obtain more insightful results, because previous research has shown that memorization is also affected by the complexity of the data (Prashanth et al., 2024). Therefore, by choosing the same content for all of the languages, we expect to only see the effect of the language.

Since we want to gain insight into the both pretraining and finetuning scenarios, we select one dataset contained in the pretraining and another one not contained in the pretraining set. For this purpose, we choose the EMEA (Tiedemann, 2012) and EuroParl (Koehn, 2005) datasets for our experiments. EuroParl is part of the PILE (Gao et al., 2020) pretraining dataset, while EMEA is an unseen dataset that we introduce to the Pythia models during our finetuning phase.

We remove the duplicate samples from each dataset. The remaining dataset is used for training, as well as in the perplexity ratio test. At the same time, we extract the samples longer than two context *steps* (see table 1) to construct our discoverable memorization dataset.

4.3 Languages

The choice of the languages was mainly limited by the model and datasets that were available. As explained in Section 4.2, we chose the EuroParl and EMEA datasets. Both of these datasets contain the European languages. It was shown that the Pythia models perform well on higher-resourced languages (Xu et al., 2025) such as German, Italian and Spanish. We opt for medium-sized languages that are less represented in the Pythia training data and are therefore representative for lower-resourced settings: Dutch (NL), Slovenian (SL), Polish (PL), and Czech (CS). While there could be other choices possible, we believe this would not have considerable effects on our experimental results (Section 5).

4.4 Metrics

Regarding our *discoverable memorization* test, we use *approximate* and *verbatim string match*, as this would help us gain more insights into the comparison of different forms of memorization. For *approximate* matching, we follow the same approach as Ippolito et al. (2023), considering a match when the BLEU score similarity exceeds 0.75. We adopt *greedy decoding* for sequence generation, meaning that the model generates only a single most likely suffix for each prompt by selecting the highest-probability token at each step.[5] This approach is commonly used in prior work on memorization, as it simplifies the evaluation and ensures deterministic outputs, which are essential for reproducibility and fair comparison across models and settings (Satvaty et al., 2025). For our perplexity ratio test, we first divide the samples into different bins, based on normalized token length by a granularity of 50 tokens, then we report the median of perplexity ratio of each bin. Choosing median should help to overcome the issue of the outliers and have more meaningful and realistic results.

5 Results

Figures 2 and 3 illustrate the main findings of our experiments, while Table 2 provides a detailed breakdown of the discoverable memorization results.

5.1 Finetuning

According to the results shown in Figure 2, when Pythia models are trained on the EMEA dataset, which was not included in their pretraining set, English language shows higher levels of discoverable memorization than the lower-resource languages. This phenomenon is consistent across both verbatim and approximate match and also for training under different amount of epochs (refer to Appendix A.1). One possible explanation is that the model has been exposed to significantly more English data during pretraining. As a result, it has developed a stronger generative prior for English, it has a better internal representation of syntax, vocabulary, and structure, which makes it more confident and fluent when generating English sequences. Consequently, when fine-tuned on new data, the model is more likely to memorize and reproduce English content verbatim or near-verbatim, simply because generating in English aligns more closely with its preexisting language patterns.

On the other hand, the perplexity ratio tests show higher values for the lower-resource languages (Figure 3), showing that lower-resource languages could be more prone to membership inference attacks. This trend is consistent across the different model sizes. This could be justified by several arguments. Firstly, the model is less sensitive against new English data (English stands below other languages for the perplexity of the untrained models, presented in Appendix A.1). Most of the variations in text has been already presented to the model during pretraining, therefore introducing the new English dataset does not significantly change the model weights. This would result in having a perplexity ratio near to 1. Then, in the case of lower-resource languages, the untrained model would give a high perplexity to the data, as it was not close to what it has seen during pretraining. This would result in a perplexity ratio higher than 1 as it is noticeable in the figure.

5.2 Pretraining phase

When Pythia models are tested for discoverable memorization on EuroParl, without any finetuning, they show higher memorization rates in the lower-resource languages. As could be seen in Figure 2, the amount of approximate memorization remains 0 for English while for the other languages it shows a correlation trend with model size, with a very low slope. On the other hand, in the perplexity ratio test (Figure 3) English subset is showing lower perplexity ratio than the average of other languages for each model size (For individual languages refer to Appendix A.2).

The overall scales of discoverable memorization and perplexity ratios in this experiment is notably lower than in the fine-tuning scenario. This difference can be attributed to two main factors: catas-

[5] Tirumala et al. (2022b) referred to the verbatim memorization observed through greedy decoding as *Exact Memorization*. However, since we also consider approximate memorization, we avoid using that term to prevent confusion.

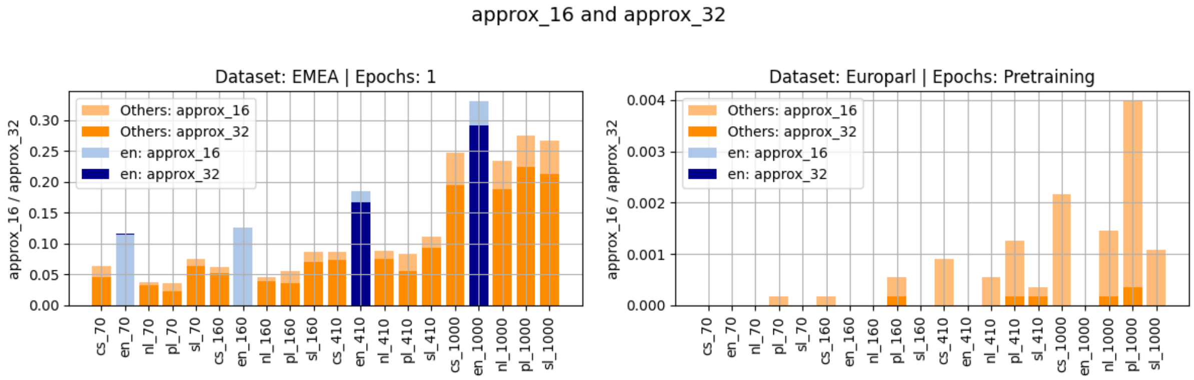

Figure 2: The results of our approximate discoverable memorization test across different model sizes of *Pythia*. The context have been considered equal to one step size of tokens for each language (refer to table 4.2 for step size). The expected suffix have been considered 16 and 32 tokens for all languages: **(left)** models were finetuned on *EMEA* dataset for one epoch **(right)** models were only tested on *EuroParl*, as it was contained in the pretraining dataset.

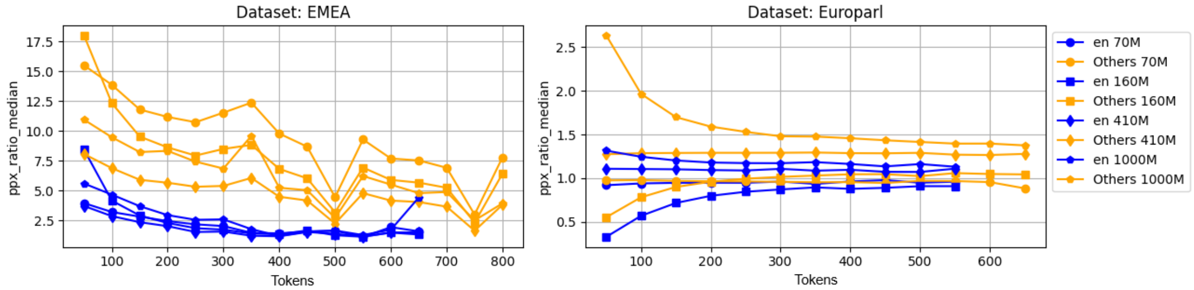

Figure 3: The results of histogram of our perplexity ratio test (MIA susceptibility). The x axis represents the *normalized* number of tokens (see 3.1) in the bins (50 tokens granularity), and the y axis represents the median of the perplexity ratios per bin. The lower-resource languages are averaged (others). (The figure only shows the results after pretraining (EuroParl), and one epoch of training (EMEA). The complete results for all epochs of training and non-averaged lower-resource language can be found in Appendix A.1.)

		Approximate match (%)				**Verbatim match (%)**			
Dataset	**Language**	**70M**	**160M**	**410M**	**1B**	**70M**	**160M**	**410M**	**1B**
EMEA	EN	**11.46**	**12.57**	**18.52**	**32.97**	**8.59**	**9.35**	**14.73**	**26.79**
	NL	3.67	4.53	8.86	23.38	2.67	3.09	6.97	17.87
	SL	7.47	8.58	11.16	26.59	5.71	6.48	8.69	21.91
	PL	3.65	5.55	8.31	27.42	2.23	3.72	5.60	23.07
	CS	6.31	6.25	8.70	24.72	4.41	4.37	6.63	19.65
EuroParl	EN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NL	0.00	0.00	0.05	0.14	0.00	0.00	0.00	0.04
	SL	0.00	0.00	0.04	0.11	0.00	0.00	0.02	0.07
	PL	**0.02**	**0.05**	**0.13**	**0.40**	0.00	0.02	0.02	**0.14**
	CS	0.00	0.02	0.09	0.22	0.00	**0.02**	**0.04**	0.11

Table 2: Results of our discoverable memorization experiment for approximate and verbatim match across datasets and languages for different model sizes for suffix length of 16 tokens. The highest values in each column are represented in **bold** format.

trophic forgetting and the nature of the dataset. The EMEA dataset, used in the fine-tuning experiment, belongs to the medical domain and contains structured, domain-specific content. As a result, certain phrases and sentence structures are frequently repeated across different samples, increasing the likelihood of memorization by the model.

In contrast, the EuroParl dataset, evaluated in a zero-shot setting, covers more general parliamentary proceedings and exhibits less internal redundancy. Moreover, the observation that lower-resource languages show higher memorization in the EuroParl evaluation could be partially explained by reduced catastrophic forgetting. Since the Pythia models were predominantly pretrained on English data, the English representations may have undergone more overwriting during pretraining updates. In comparison, representations for lower-resource languages, being less frequent in the pretraining corpus, might have been updated less aggressively and thus retain more memorized sequences from training data. This results in slightly higher levels of discoverable memorization for these languages under zero-shot settings.

These results shows that the languages that are less represented in the pretraining data are more prone to memorization and privacy attacks. This suggests that when such languages are included in the pretraining data, even without finetuning, the model is more likely to retain and expose training sequences, raising concerns about privacy leakage in multilingual deployments.

6 Discussion

Firstly, selecting appropriate models and datasets for our experiments posed several challenges. There are few parallel datasets available across multiple languages that include samples long enough to support discoverable memorization experiments. Additionally, only few models are available at multiple scales with accessible pretraining checkpoints. While Pythia is primarily trained on English data, it demonstrates sufficient language understanding and generation capabilities in the language subset we experimented with. Therefore, we argue that our chosen model and datasets are reasonably well aligned for the purposes of this study.

Secondly, our finetuning experiments were limited to model sizes up to 1B parameters. Since we finetune each language-dataset pair for up to 8 epochs and subsequently run discoverable memorization and perplexity ratio tests on the entire dataset, the process was computationally intensive, requiring 160 independent runs on an A100 Nvidia GPU. However, we believe this limitation does not significantly affect our conclusions. The observed trends were robust and consistently distinguishable between the lower-resource languages and English. Nonetheless, since a relation exists between memorization and model size (Satvaty et al., 2025; Lesci et al., 2024) future work should further explore this space using larger models, different datasets and various training regimes.

7 Conclusions and future work

We studied the discoverable memorization and the susceptibility of Pythia models to MIA over two different parallel datasets comparing memorization related behaviour of these models in the cases of lower-resource languages and higher-resource languages. We observe that in both cases of pretraining and finetuning data, the lower-resource languages show more vulnerability to MIA according to the perplexity ratio method. However, in the case of discoverable memorization, while pretraining data shows higher memorization rates for lower-resource languages, the finetuning data behaves differently, showing more memorization for English dataset. At the same time, our fine-tuning experiments raise an interesting question: while lower-resource languages exhibit higher susceptibility to MIA, they demonstrate less discoverable memorization. Although we proposed some initial hypotheses to explain this observation, a deeper analysis of the relationship between discoverable memorization and MIA susceptibility is indeed an interesting direction for future research.

In quantifying the susceptibility of LLMs, in particular Pythia models, to MIA, we employed perplexity ratio tests and as mentioned, lower-resource languages prove to be more prone to privacy attacks and disclosure of private data. These findings underscore the need for stronger privacy-preserving strategies in multilingual LLMs, particularly during both pretraining and finetuning phases.

Future work should further investigate the root causes of the difference between higher and lower resource languages, whether and to what extent inherent characteristics of different languages play a role in memorization related issues and privacy vulnerabilities. While balancing the data across different languages would be a possible solution, it might not always be feasible. We believe that this direction of language-sensitive privacy needs to be further explored to make sure that multilingual models do not exhibit privacy risks regardless of the different linguistic settings.

In our study, we focused on lower-resource languages, and we leave a broader examination across a wider range of linguistic settings for future research. In addition, exploring possible countermeasures against the observed phenomenon would be an important next step. We believe that this line of language-sensitive privacy research is crucial to ensure that multilingual models do not exhibit uneven privacy risks across different linguistic contexts.

Acknowledgements

This publication is part of the project LESSEN[6] with project number NWA.1389.20.183 of the research program NWA ORC 2020/21 which is (partly) financed by the Dutch Research Council (NWO).

References

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023a. Quantifying memorization across neural language models.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. 2023b. Are aligned neural networks adversarially aligned?

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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. 2023. Preventing verbatim memorization in language models gives a false sense of privacy.

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.

Abhyuday N. Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership inference attack susceptibility of clinical language models. *ArXiv*, abs/2104.08305.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Aly Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2024. Causal estimation of memorisation profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15616–15635, Bangkok, Thailand. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

[6] https://lessen-project.nl

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada. Association for Computational Linguistics.

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.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models.

Hellina Hailu Nigatu and Inioluwa Deborah Raji. 2024. “i searched for a religious song in amharic and got sexual content instead”: Investigating online harm in low-resourced languages on youtube. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, page 141–160, New York, NY, USA. Association for Computing Machinery.

USVSN Sai Prashanth, Alvin Deng, Kyle O’Brien, Jyothir S V, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. 2024. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon.

Ali Satvaty, Suzan Verberne, and Fatih Turkmen. 2025. Undesirable memorization in large language models: A survey.

Shlomit Shachor, Natalia Razinkov, and Abigail Goldsteen. 2024. Improved membership inference attacks against language classification models.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks.

Virat Shejwalkar, Huseyin A Inan, Amir Houmansadr, and Robert Sim. 2021. Membership inference attacks against NLP classification models. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022a. Memorization without overfitting: Analyzing the training dynamics of large language models.

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

Yijue Wang, Nuo Xu, Shaoyi Huang, Kaleel Mahmood, Dan Guo, Caiwen Ding, Wujie Wen, and Sanguthevar Rajasekaran. 2022. Analyzing and defending against membership inference attacks in natural language processing classification. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5823–5832.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11).

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.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Low-resource languages jailbreak gpt-4.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramer, and Nicholas Carlini. 2023a. Counterfactual memorization in neural language models.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don‘t trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models.

A Perplexity ratio experiments

A.1 EMEA

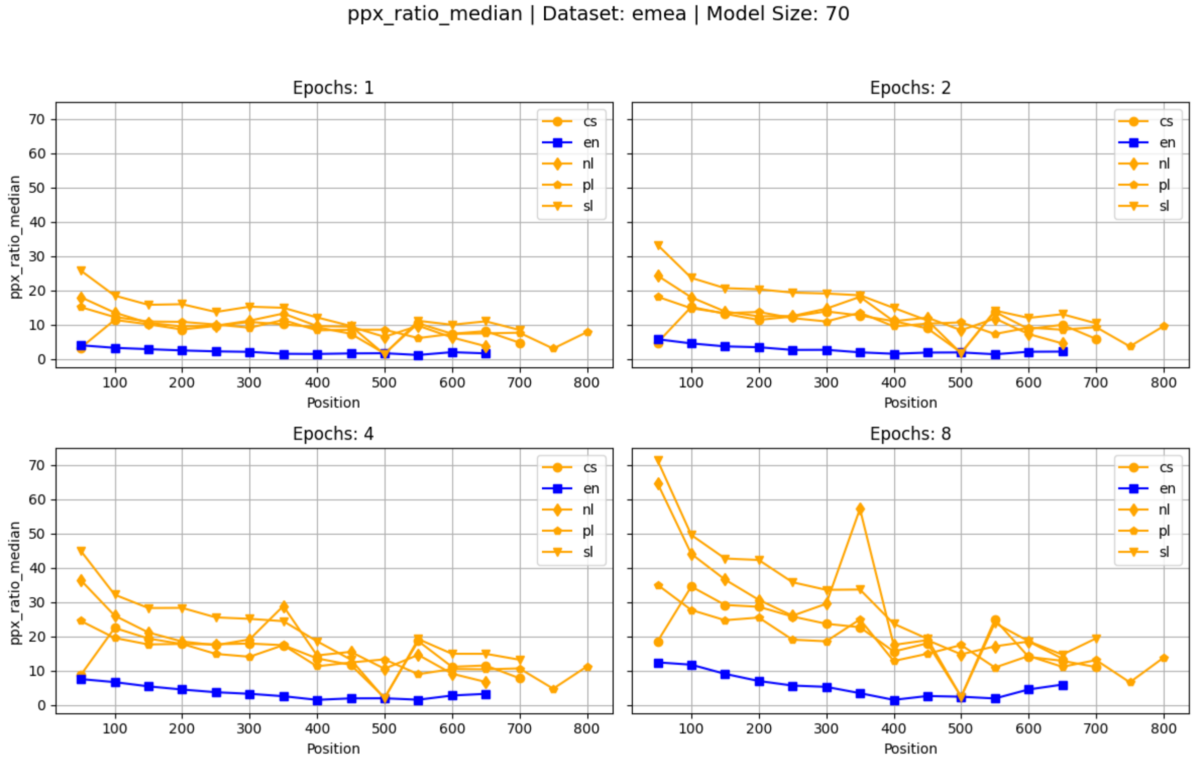

Figure 4: The median perplexity ratios per *normalized* number of tokens (granularity 50 tokens) obtained after training the 70M parameter model on the respective translation of the EMEA dataset for $\{1, 2, 4, 8\}$ epochs.

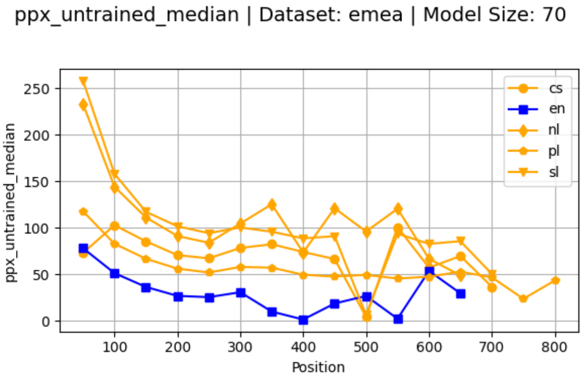

Figure 5: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 70M parameter model when it is queried with the respective translation of the EMEA dataset.

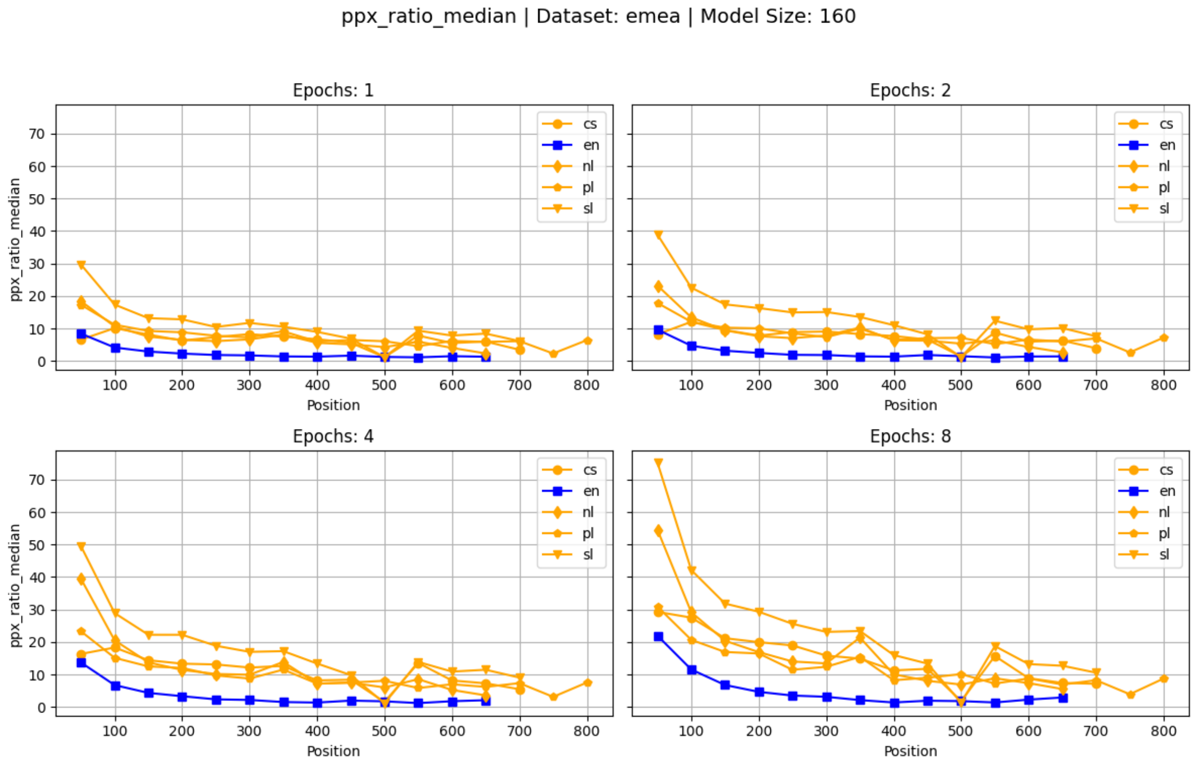

Figure 6: The median perplexity ratios per *normalized* number of tokens (granularity 50 tokens) obtained after training the 160M parameter model on the respective translation of the EMEA dataset for $\{1, 2, 4, 8\}$ epochs.

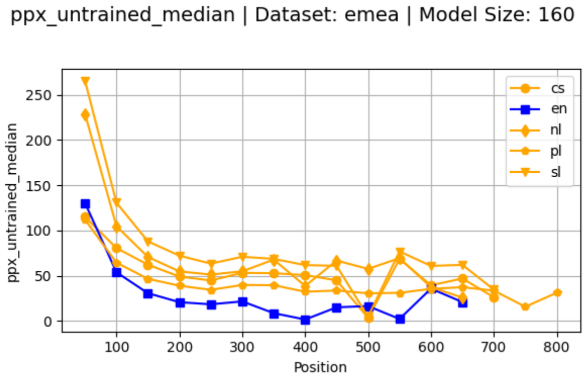

Figure 7: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 160M parameter model when it is queried with the respective translation of the EMEA dataset.

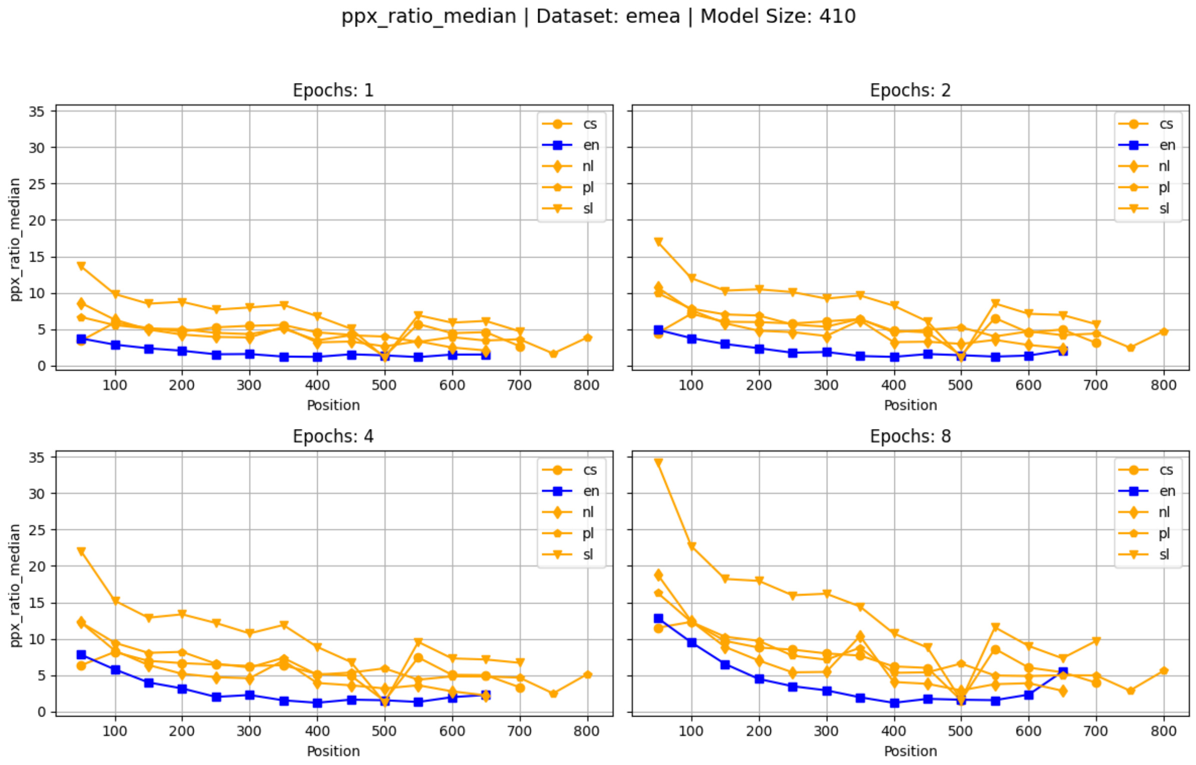

Figure 8: The median perplexity ratios per *normalized* number of tokens (granularity 50 tokens) obtained after training the 410M parameter model on the respective translation of the EMEA dataset for $\{1, 2, 4, 8\}$ epochs.

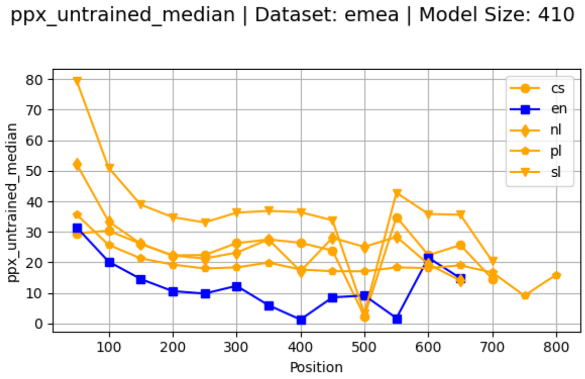

Figure 9: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 410M parameter model when it is queried with the respective translation of the EMEA dataset.

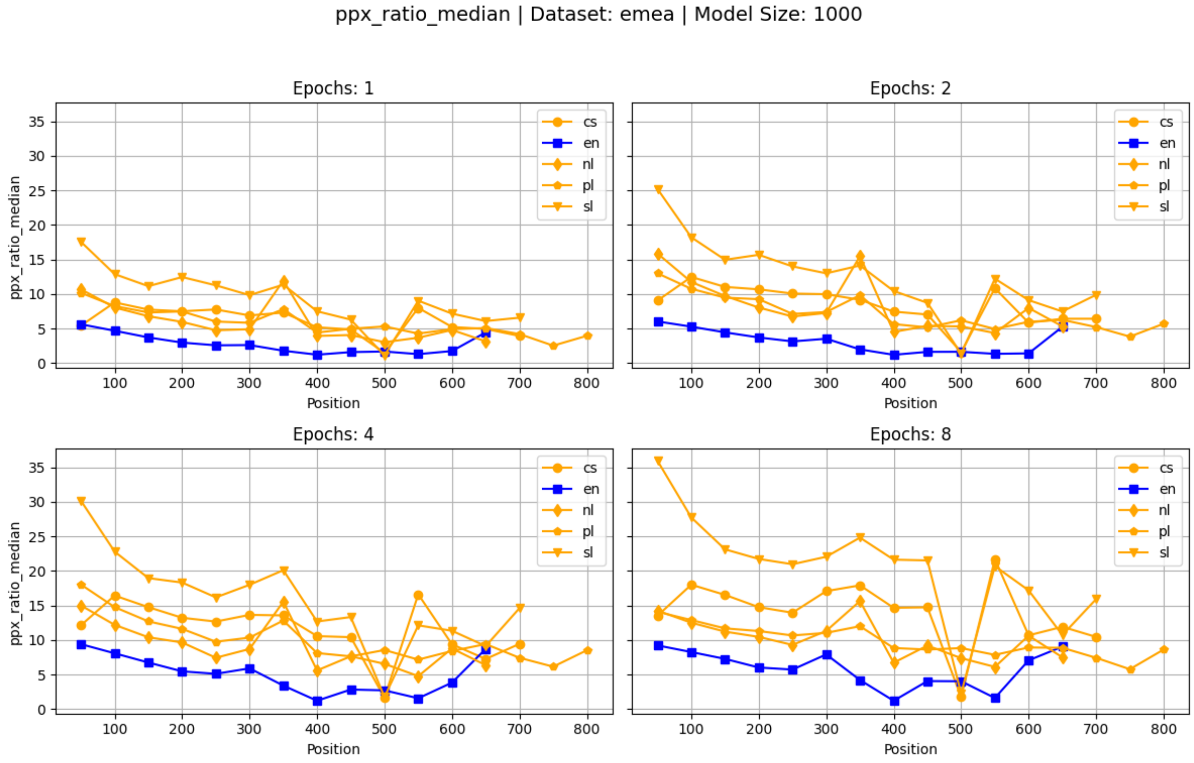

Figure 10: The median perplexity ratios per *normalized* number of tokens (granularity 50 tokens) obtained after training the 1000M parameter model on the respective translation of the EMEA dataset for $\{1, 2, 4, 8\}$ epochs.

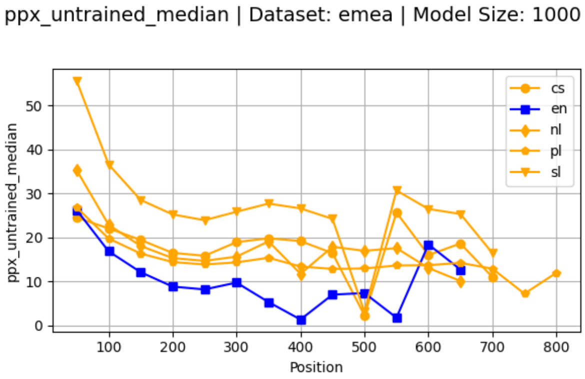

Figure 11: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 1000M parameter model when it is queried with the respective translation of the EMEA dataset.

A.2 EuroParl

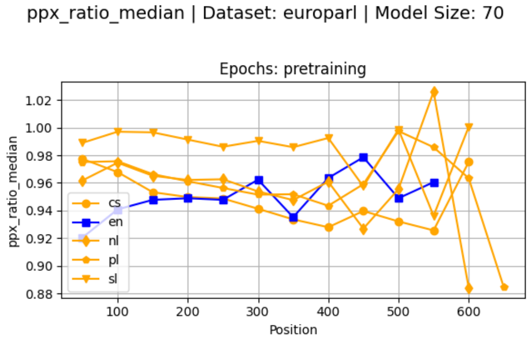

Figure 12: The median perplexity ratios per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 70M parameter model for the respective translation of the EuroParl. The untrained perplexity in the calculation is estimated by the 25% training checkpoint.

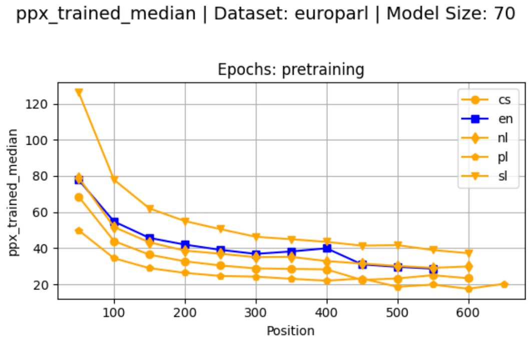

Figure 14: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 70M parameter model when querying it with the respective translation of the EuroParl dataset.

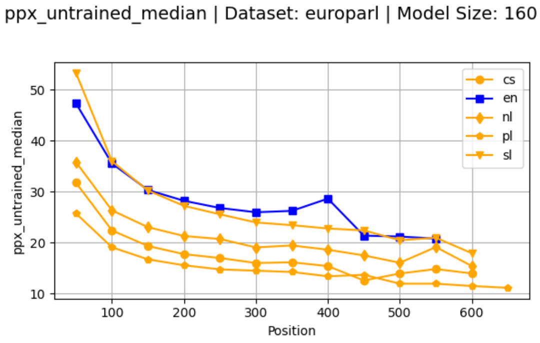

Figure 16: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the 25% training checkpoint of the 160M parameter model, which estimates the untrained perplexity scores of the model. To obtain the scores, it is queried with the respective translation of the EuroParl dataset.

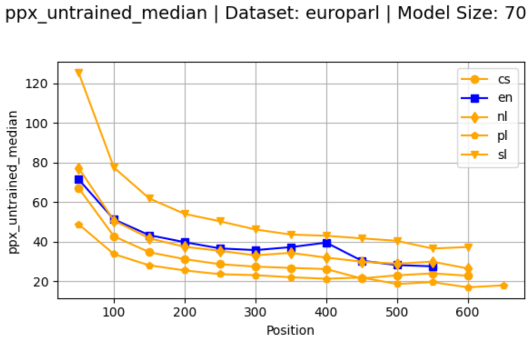

Figure 13: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the 25% training checkpoint of the 70M parameter model, which estimates the untrained perplexity scores of the model. To obtain the scores, it is queried with the respective translation of the EuroParl dataset.

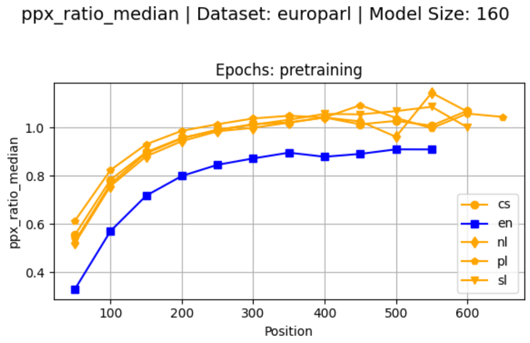

Figure 15: The median perplexity per *normalized* number of tokens (granularity 50 tokens) ratios obtained from the pretrained 160M parameter model for the respective translation of the EuroParl. The untrained perplexity in the calculation is estimated by the 25% training checkpoint.

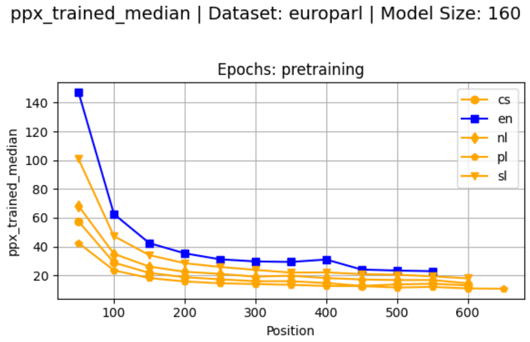

Figure 17: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 160M parameter model when querying it with the respective translation of the EuroParl dataset.

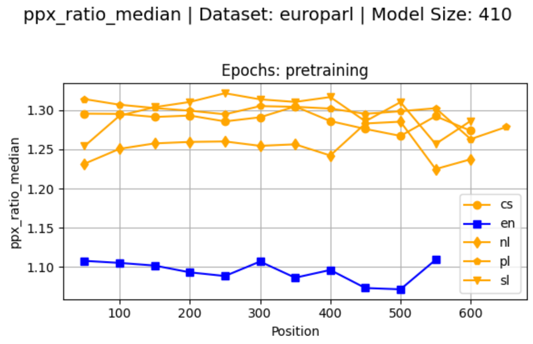

Figure 17: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the 25% training checkpoint of the 410M parameter model, which estimates the untrained perplexity scores of the model. To obtain the scores, it is queried with the respective translation of the EuroParl dataset.

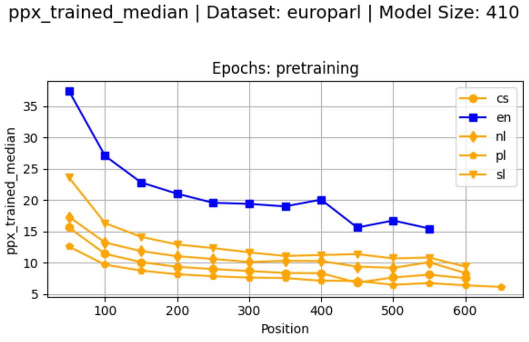

Figure 19: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 410M parameter model when querying it with the respective translation of the EuroParl dataset.

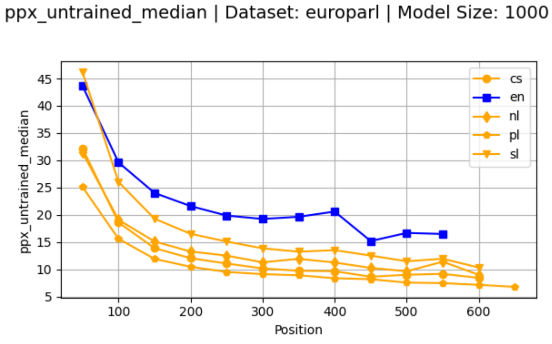

Figure 21: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 1000M parameter model when it is queried with the respective translation of the EuroParl dataset.

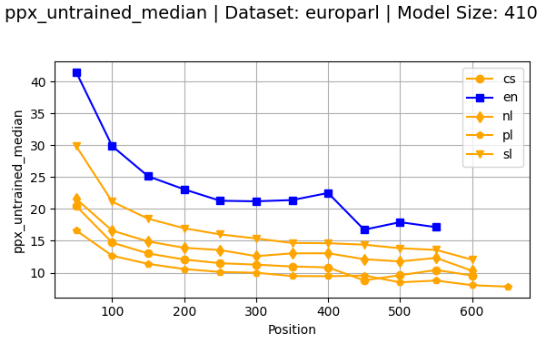

Figure 18: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 410M parameter model when it is queried with the respective translation of the EuroParl dataset.

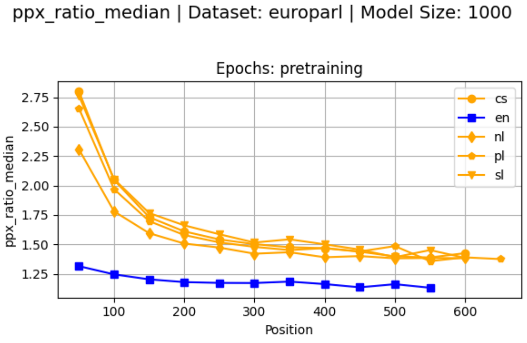

Figure 20: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the 25% training checkpoint of the 1000M parameter model, which estimates the untrained perplexity scores of the model. To obtain the scores, it is queried with the respective translation of the EuroParl dataset.

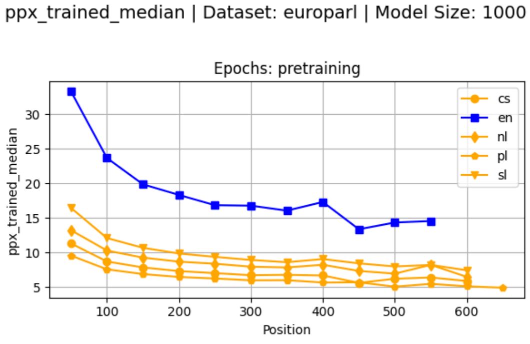

Figure 22: The median perplexities per *normalized* number of tokens (granularity 50 tokens) obtained from the pretrained 1000M parameter model when querying it with the respective translation of the EuroParl dataset.

B DEA Data

This section summarizes the data obtained through our discoverable memorization tests. Figures 23 and 24 show approximate and exact string match scores yielded when querying the pretrained models with the Europarl datasets for context lengths 16 and 32. Figures 25 and 26 show the results of the same experiment using the EMEA dataset. Here, the models are finetuned for $\{1, 2, 4, 8\}$ epochs.

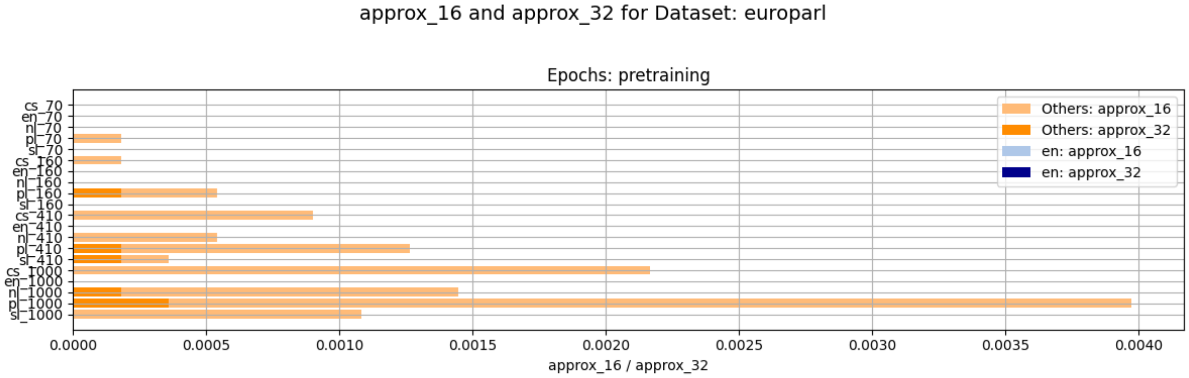

Figure 23: Discoverable memorization measured by approximate string match for context lengths 16 and 32 on the EuroParl dataset. The plot shows the results obtained from the pretrained models per language and model size.

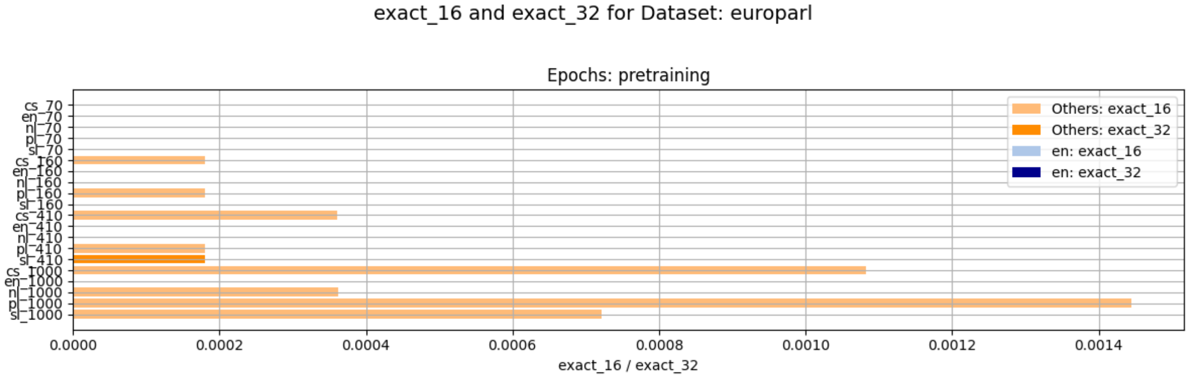

Figure 24: Discoverable memorization measured by exact string match for context lengths 16 and 32 on the EuroParl dataset. The plot shows the results obtained from the pretrained models per language and model size.

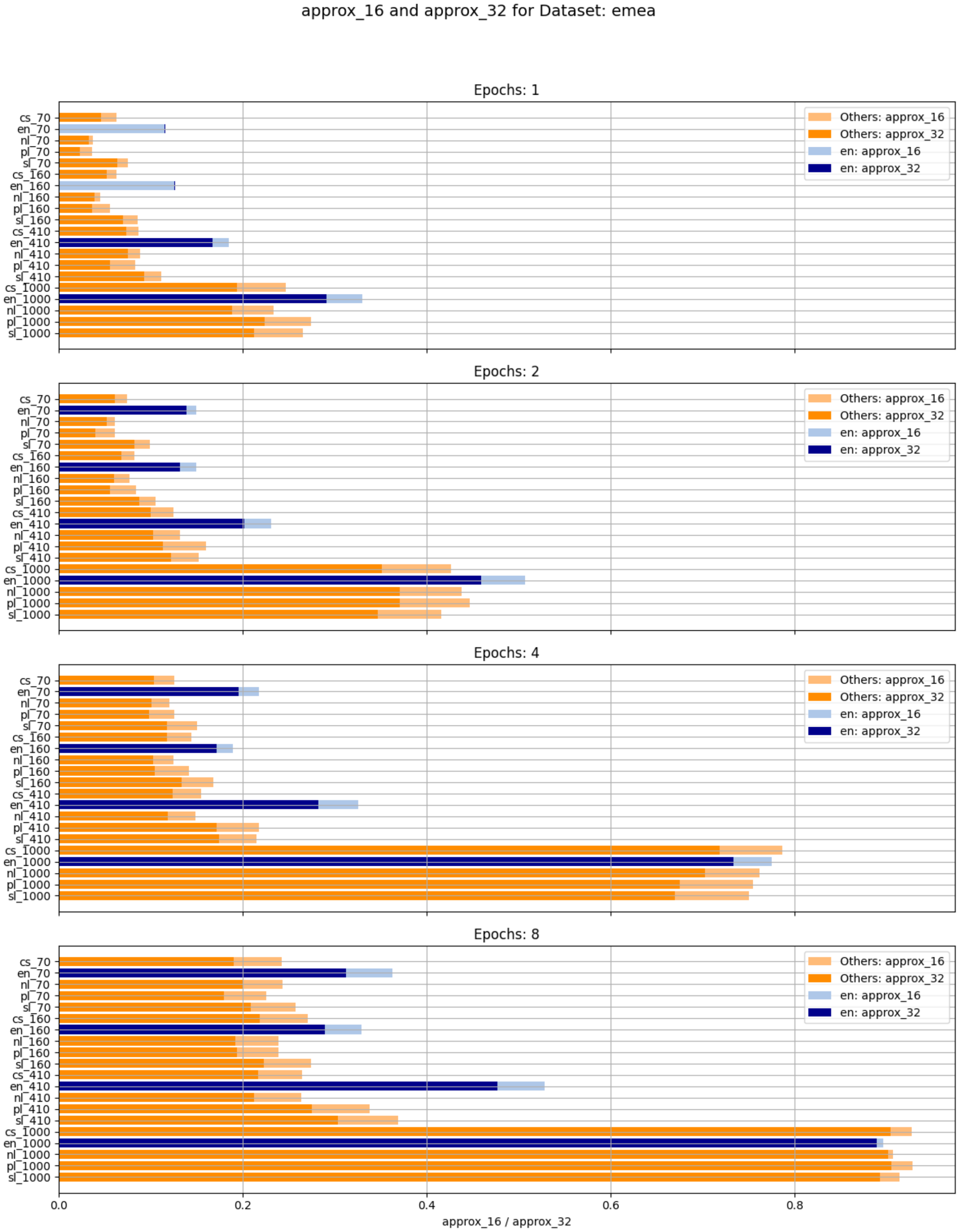

Figure 25: Discoverable memorization measured by approximate string match for context lengths 16 and 32 on the EMEA dataset. The plot shows the results obtained after different epochs of training per language and model size.

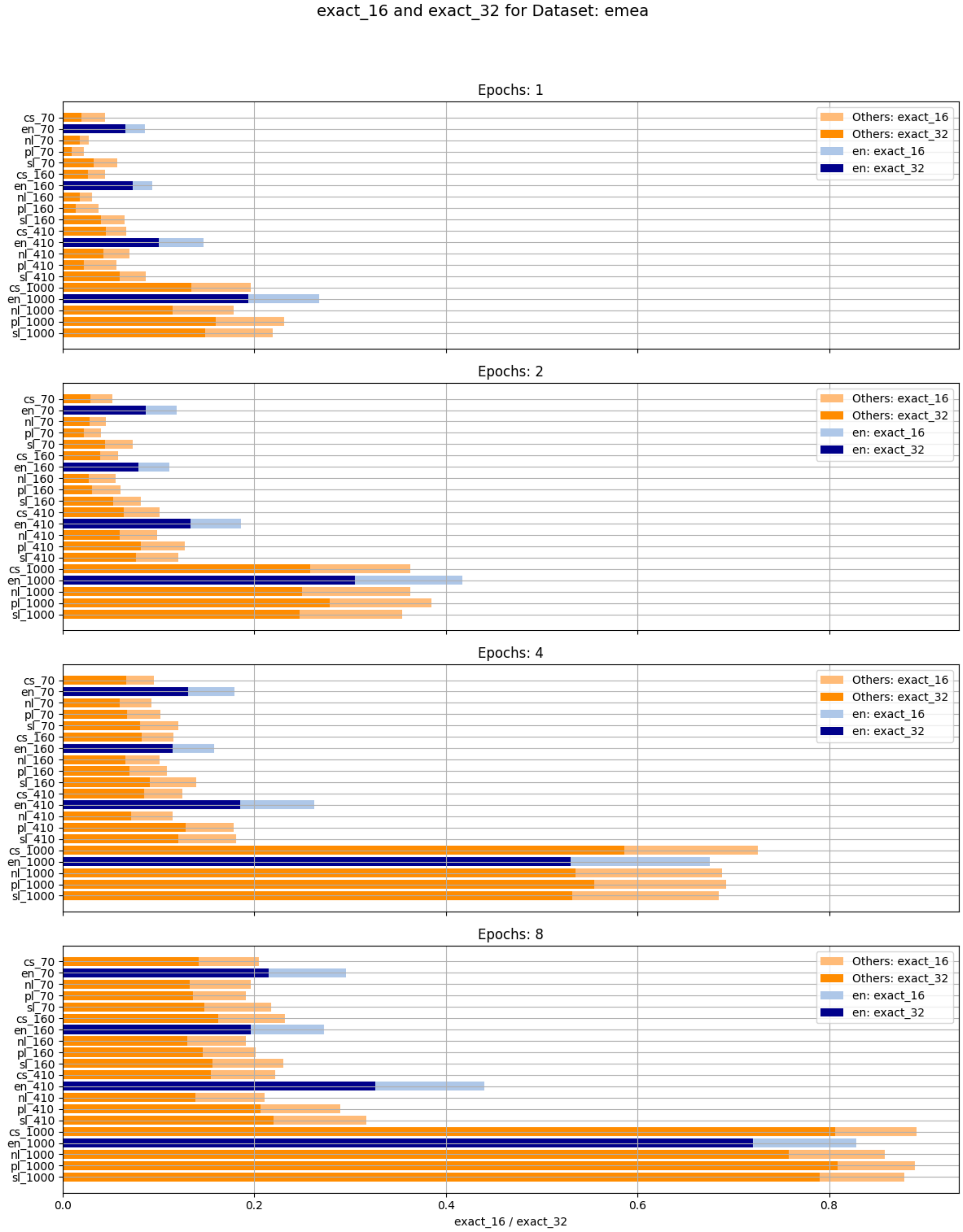

Figure 26: Discoverable memorization measured by exact string match for context lengths 16 and 32 on the EMEA dataset. The plot shows the results obtained after different epochs of training per language and model size.

C Sentence Length Example

Language	Sentence	Tokens
EN	The most common side effects with Vidaza (seen in more than 60% of patients) are blood reactions including thrombocytopenia (low platelet counts), neutropenia (low levels of neutrophils, a type of white blood cell) and leucopenia (low white blood cell counts), side effects affecting the stomach and gut including nausea and vomiting, and injection site reactions.	74
NL	Vidaza is geïndiceerd voor de behandeling van volwassen patiënten die niet in aanmerking komen voor hematopoëtische stamceltransplantatie, met: • intermediair 2 en hoog risico myelodysplastische syndromen (MDS) volgens het International Prognostic Scoring System (IPSS), • chronische myelomonocytaire leukemie (CMML) met 10-29% beenmergblasten zonder myeloproliferatieve aandoening, • acute myeloïde leukemie (AML) met 20-30% blasten en multilineaire dysplasie, volgens de indeling van de Wereldgezondheidsorganisatie (WHO).	166
SL	Ker je število bolnikov s temi boleznimi majhno, veljajo te za redke, zato je bilo zdravilo Vidaza dne 6. februarja 2002 določeno kot „ zdravilo sirota “ (zdravilo, ki se uporablja pri redkih boleznih) za mielodisplastične sindrome, dne 29. novembra 2007 pa je bilo enako določeno še za akutno mieloidno levkemijo.	130
PL	Produkt Vidaza jest wskazany do leczenia pacjentów dorosłych, niekwalifikujących się do przeszczepu krwiotwórczych komórek macierzystych, z: • zespołami mielodysplastycznymi (ang. myelodysplastic syndromes, MDS) o pośrednim- 2 i wysokim ryzyku, zgodnie z Międzynarodowym Punktowym Systemem Rokowniczym (ang.	132
CS	Přípravek Vidaza je indikován k léčbě dospělých pacientů, kteří nejsou způsobilí pro transplantaci hematopoetických kmenových buněk, s: • myelodysplastickými syndromy (MDS) intermediárního rizika 2. stupně a vysokého rizika podle Mezinárodního prognostického skórovacího systému (International Prognostic Scoring System, IPSS), • chronickou myelomonocytovou leukemií (CMML) s 10- 29% blastů v kostní dřeni bez myeloproliferativního onemocnění)), • akutní myeloidní leukemií (AML) s 20- 30% blastů a dysplazií ve více buněčných liniích, podle klasifikace Světové zdravotnické organizace (WHO).	235

Table 3: Same sentence in different languages, when tokenized with Pythia tokenizer results in different token counts.