The Model's Language Matters: A Comparative Privacy Analysis of LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly deployed across multilingual applications that handle sensitive data, yet their scale and linguistic variability introduce major privacy risks. Mostly evaluated for English, this paper investigates how language structure affects privacy leakage in LLMs trained on English, Spanish, French, and Italian medical corpora. We quantify six linguistic indicators and evaluate three attack vectors: extraction, counterfactual memorization, and membership inference. Results show that privacy vulnerability scales with linguistic redundancy and tokenization granularity: `Italian` exhibits the strongest leakage, while `English` shows higher membership separability. In contrast, `French` and `Spanish` display greater resilience due to higher morphological complexity. Overall, our findings provide the first quantitative evidence that *language matters* in privacy leakage, underscoring the need for language-aware privacy-preserving mechanisms in LLM deployments.

1 Introduction

Rapid advances in natural language processing (NLP) have fueled its adoption in many industries worldwide. Large language models (LLMs) such as BERT and GPT have been pre-trained at great expense on countless unlabeled datasets extracted from the internet. While these models represent incredible potential and promises, their large-scale deployment and their complexity, as well as the fact that they interact with and potentially influence individuals, raise multiple security and privacy concerns [10].

The attack surface on models is still poorly understood [32, 17, 11, 7]. A number of threats are related to the memorization and possible leakage of sensitive information used during model training, such as data reconstruction and membership inference (i.e., identifying elements used during the training or the fine-tuning). Memorization of information by a model is not a problem in itself. However, this memorization becomes a problem when the training information is not generalized enough by the model which reproduces large portions of training data verbatim or discloses some sensitive information [16, 34].

Most privacy risk assessment work has been conducted on English texts [19, 28]. However, the language of the texts, their structures, and their characteristics inherently impact LLM memorization and, consequently, privacy risks. Although the language considered is well known to potentially introduce bias in some results, the impact of language on privacy risks has not yet been explored to our knowledge. To overcome this limitation, in this paper, we empirically explore the impact of language on privacy risks associated with LLMs. We also analyze the main characteristics and structures of each language and link them to various privacy vulnerabilities. Specifically, we comprehensively assess the privacy of LLMs fine-tuned on English, Spanish, French, and Italian medical corpora using an extraction attack, a membership inference attack and counterfactual memorization. Results show that privacy vulnerability scales with linguistic redundancy and tokenization granularity: Italian presents the highest leakage, while English has higher membership separability. In contrast, French and Spanish show greater resilience due to greater morphological complexity.

Overall, our results provide the first quantitative evidence that language is a significant factor in LLM privacy leakage. This highlights the need to consider this factor in LLM deployment and the design or configuration of privacy-preserving mechanisms.

2 Background and Related Work

Large language models (LLMs) are trained on very large datasets. For example, training chatGPT re-

quired years of crawling the Internet. Therefore, a lot of personal data such as people's addresses was used during training. BERT models, on the other hand, are typically fine-tuned for specific tasks with domain-oriented data. In the medical domain, datasets typically include sensitive patient records. In both cases, the problem is that the models can regurgitate and leak information from the training data after deployment [4, 9].

A central question in this context concerns the extent to which language models memorize their training data [5, 4, 26, 37, 27]. However, defining memorization for language models is challenging, and many existing definitions and notions have been proposed depending on whether the memorization concerns copyrighted content or personal and sensitive content. In relation to privacy, we can notably cite extractable memorization (Section 2.1) and membership inference (Section 2.2).

2.1 Extractable memorization

Extractable memorization is a type of attack that aims to use the model to infer information from the original data [21]. This attack mainly concerns text generation models, such as GPT. These models are trained to produce text based on what they have seen during training. However, the model is not expected to be a basic parrot and repeat exactly the sentences it has seen. This is especially concerning if the data it is repeating is sensitive. This has been shown to be the case with GPT-2 for example, from which the names and addresses of individuals can be extracted [6]. In [4], the term $k - extractability$ is used to refer to the sequences that can be extracted from the model when an input sequence of length k is requested. The lower the k, the easier it is to extract the sequence. We therefore expect a model to have the highest possible k on private queries. This measure, however, does not capture regurgitations that are not perfect, which can lead to an illusion of no extractable memory. Compressible memorization [27] extends this definition by evaluating how short the minimal requested sentence (or prompt) that elicits the sequence.

2.2 Membership inference

Membership inference attack [3, 29, 14, 25, 31, 13] (MIA) is a more common inference attack in Machine Learning (ML), which aims to infer whether a specific data was used in the training data of a target model. There are different techniques that can be used to perform a MIA attack depending on if the adversary has an access to the model parameters (i.e., white-box access), or access to a ground-truth subset of member and non-member samples. One technique consists to analyze the loss of member and non member samples [36], another one is to use multiple shadow models [29, 35] trained to mimic the behavior of the target model on an auxiliary dataset. An adversarial model is then trained to infer membership from the loss or from shadow models.

Another method [37] is based on comparing the performance of the target model trained on a dataset with a specific input, with a second model trained without it. As ML models are supposed to learn general information, one piece of data (even rare, outlier or mislabeled samples) is not supposed to be memorized and significantly changed the model's performance. By repeating this operation many times with different subset, it is possible to identify counterfactually memorized data.

3 Comparative Privacy Analysis

We perform a comparative privacy analysis across four languages: `English`, `Spanish`, `French`, and `Italian`, encompassing three complementary threat models: (i) *prompt-based extraction*, where we probe direct content leakage from generative models; (ii) *counterfactual memorization*, where we quantify how strongly individual texts are overfit by fine-tuned models; and (iii) *membership inference*, where we test whether a model exposes the presence of individual samples in its training set. Together, these analyses provide a unified view of surface-level and latent memorization behaviour across languages and architectures.

3.1 Experimental Setup

Datasets. We employ a corpus to capture both controlled and large-scale multilingual behavior. The `HiTZ Multilingual Medical Corpus` [1] provides over 3 million translated medical sentences in `English`, `Spanish`, `French`, and `Italian`. We select 10k sentences from the corpus in this analysis, as it is large enough for privacy assessment while accounting for the limited computational resources that we have.

[1] https://huggingface.co/datasets/HiTZ/Multilingual-Medical-Corpus

Model Selection and Training. We evaluate both encoder-only (BERT-style) and decoder-only (GPT-style) architectures to contrast their privacy behaviors across tasks. Encoder models are assessed through classification-based membership inference and counterfactual memorization, while decoder models are probed via generative extraction, providing a complementary view of implicit versus explicit memorization dynamics.

For encoder-only architectures, we fine-tune one pre-trained model per language on a medical classification task: `bert-base-uncased`[2] (English), `dccuchile/bert-base-spanish-wwm-cased`[3] (Spanish), `almanach/camembert-base`[4] (French), and `Musixmatch/umberto-commoncrawl-cased-v1`[5] (Italian). For decoder-only architectures, used in extraction attacks, we fine-tune `distilgpt2`[6] (English), `DeepESP/gpt2-spanish`[7] (Spanish), `dbddv01/gpt2-french-small`[8] (French), and `LorenzoDeMattei/GePpeTto`[9] (Italian). All models are trained using identical hyperparameters (batch size, learning rate, and number of epochs) across languages to ensure comparability. Each dataset is randomly split into 80% for training and 20% for testing, maintaining consistent data exposure across experiments.

Attack Setup.

- **Extraction attacks:** We perform prompt-based extraction attacks to evaluate explicit surface leakage in generative models. Our approach conditions a fine-tuned decoder model on partial text fragments and measures how often it regenerates exact or near-exact spans from the training corpus. We systematically vary the prompt fraction in $\{5, 12, 25, 37\}$ to examine how prompt length influences extraction behaviour. Unlike prior optimization-based extraction methods, our strategy requires no gradient access and scales efficiently across multiple languages. We additionally quantify the number and diversity of unique extractions as a function of prompt size, providing a direct signal of language-dependent memorization risk.

- **Counterfactual memorization:** We quantify instance-level overfitting by computing a *counterfactual memorization score* for each document in the HiTZ Multilingual Medical Corpus. Each model is fine-tuned on a 9-class *length-binned text classification* task, where labels correspond to decile-based token length bins. For each text, the counterfactual score is defined as the difference between the mean sigmoid loss of models that *saw* the text during training and those that did not. This metric extends standard memorization analysis by capturing the intensity of instance-level overfitting. We train an ensemble of ten independently seeded sequence classifiers per language, based on BERT-family encoders, to ensure stable counterfactual estimates. The 95th percentile of the resulting score distribution is used to flag highly memorized instances. We further compute empirical CDFs over surface-level statistics (e.g., sentence length, word count, unique words) to relate memorization strength to linguistic and morphological characteristics (Table 1).

- **Membership inference:** We evaluate membership inference on the same fine-tuned classification models, using shadow models trained to replicate the target model's learning dynamics. Attackers exploit differences in prediction confidence distributions to distinguish "in-training" versus "out-of-training" samples. This setup targets encoder-only architectures and quantifies privacy leakage arising from confidence calibration and representation separability. Since the underlying classification task is language-agnostic (based on text length bins), it provides a controlled baseline for assessing how linguistic structure influences susceptibility to membership inference.

3.2 Extraction Attack

We probe surface-level memorization through prompt-conditioned extraction attacks, where partial context is provided to a generative model to elicit verbatim continuations. Figure 1 quantifies the number of *unique* extractions across languages and prompt sizes, while Figure 2 reports the cumulative distribution of text lengths for all sentences versus those appearing among extracted samples (with a short 5-word prompt).

We observe marked cross-linguistic differences. At minimal prompts (i.e., 5 words), English pro-

[2] https://huggingface.co/bert-base-uncased
[3] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[4] https://huggingface.co/almanach/camembert-base
[5] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1
[6] https://huggingface.co/distilgpt2
[7] https://huggingface.co/DeepESP/gpt2-spanish
[8] https://huggingface.co/dbddv01/gpt2-french-small
[9] https://huggingface.co/LorenzoDeMattei/GePpeTto

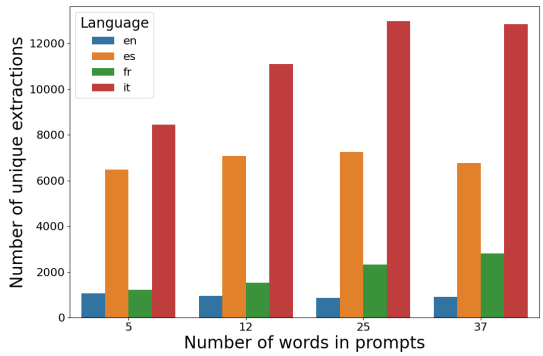


Figure 1: Number of unique extractions across languages and prompt sizes: longer prompts increase extraction risk in general.

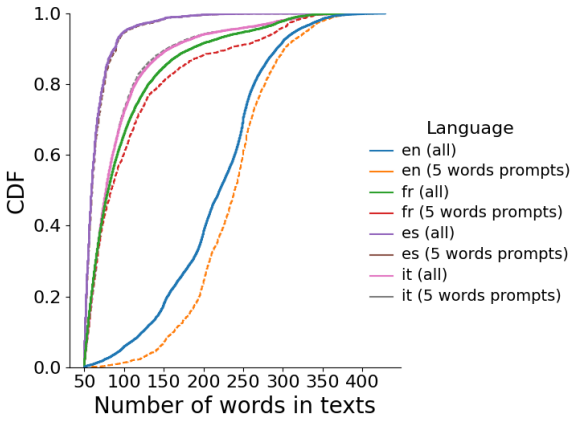


Figure 2: Cumulative distribution of text lengths for all versus extracted samples.

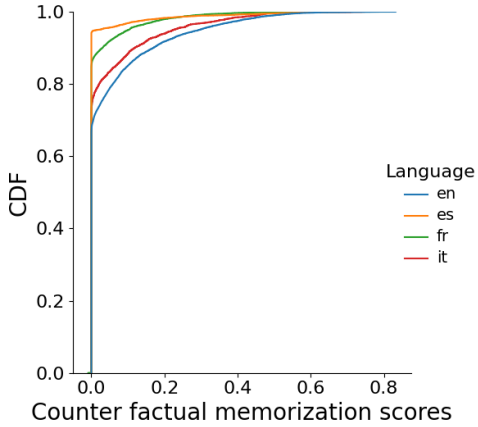


Figure 3: Distribution of counterfactual memorization scores across languages. Most points lie near zero; EN and IT display extended positive tails, FR shows rare high outliers, and ES remains the most compact.

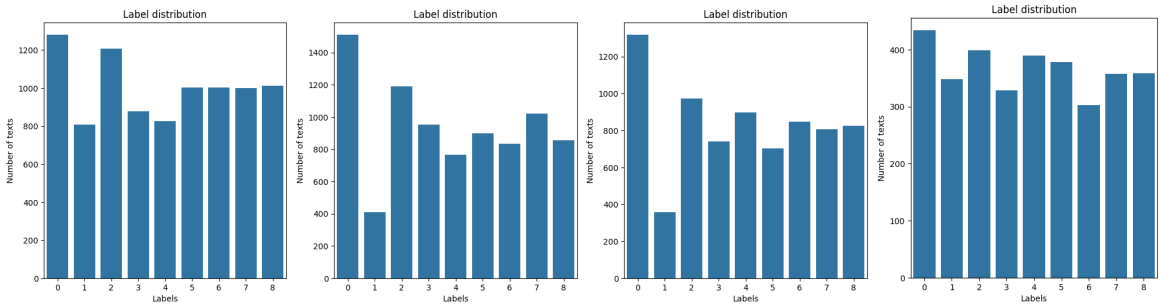
duces fewer than 1,000 unique extractions, suggesting relatively low surface-level leakage under constrained context. In contrast, `Spanish` already yields over 6,000 unique spans, and `Italian` surpasses 8,000, indicating greater sensitivity to minimal cues. As prompt size increases to 12 and 25 words, Italian extractions rise sharply, peaking at over 13,000 unique spans, while Spanish stabilizes around 7,000. `French`, by comparison, remains substantially lower throughout, increasing from roughly 1,200 to 2,700 extractions. These patterns reveal that certain languages (ES, IT) sustain or amplify leakage as prompts grow, whereas English shows an early saturation and subsequent decline in extraction counts with larger context windows.

Moreover, further analysis reveals that longer texts are more prone to extraction even under short prompts. As illustrated in Figure 2, the CDFs for the extracted texts (i.e., using 5-word prompts) closely follow or are slightly shifted to the right of the overall corpus distributions, indicating that the extracted samples tend to contain more words on average. This demonstrates that extraction behavior with short prompts is not biased by sentence length: even minimal context captures the same cross-linguistic tendencies observed in Figure 1. Consequently, the higher number of extractions in `Spanish` and `Italian` cannot be attributed to prompt selection, but rather reflects their intrinsic linguistic and structural susceptibility to memorization.

3.3 Counterfactual Memorization

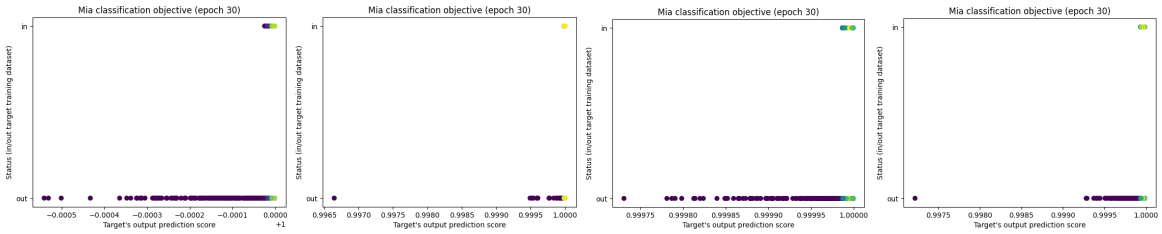
The counterfactual memorization score measures the change in loss between models that *saw* a text during training and those that did not. This metric captures how strongly each instance is memorized relative to a counterfactual baseline. Figure 3 reports the score distributions across languages, while Figure 4 confirms that label distributions are balanced and therefore do not confound memorization effects.

The results reveal that most samples cluster around zero in all languages, indicating that the majority of instances are **not explicitly memorized**. However, language-specific deviations appear in the positive tail of the distribution. `Spanish` exhibits the narrowest spread, with over 95% of samples scoring below 0.02, suggesting minimal overfitting and strong generalization. `English` and `Italian`, by contrast, show moderate positive tails extending up to 0.08–0.10, indicating that 5–8% of samples exhibit measurable memorization. Finally, `French` displays a distinctive pattern: while its median score remains low, it contains rare but



(a) English (b) Spanish (c) French (d) Italian

Figure 4: Label distributions used for memorization scoring: balanced bins across languages confirm that score variations are not due to class imbalance.



(a) English (b) Spanish (c) French (d) Italian

Figure 5: Separability of "in" vs. "out" samples at epoch 30 under MIAs: larger gaps indicate higher risk. English exhibits the most distinct separation between training and test data, while French shows the greatest overlap, indicating stronger generalization.

pronounced outliers that exceed 0.15, pointing to isolated cases of high-confidence recall.

We further analyze the label distributions used in the counterfactual memorization experiments to verify dataset balance across languages (Figure 4). Although minor variations exist, all languages maintain a roughly uniform spread over the nine label bins, ensuring that observed memorization trends are not artifacts of label skew.

Quantitatively, English exhibits a mildly-skewed distribution. Spanish shows a similar pattern but with a more pronounced peak at label 0 ($\sim$1,500 texts) and a small dip around label 1 ($\sim$400 texts). French follows a nearly identical trend, with its most frequent label 0 ($\sim$1,300+ texts) and the least represented label 1 ($\sim$350$-$400 texts). In contrast, Italian displays the most balanced profile, with all labels ranging between 300 and 450 samples and no extreme outliers.

These distributions confirm that the memorization differences reported in Figure 3 cannot be attributed to unbalanced label frequencies. While English, Spanish, and French exhibit mild concentration toward lower labels, all maintain sufficient coverage of the label space to ensure unbiased counterfactual comparisons. The flat histogram of the Italian dataset further demonstrates that even with a highly uniform label representation, moderate memorization persists, reinforcing that linguistic and structural factors, rather than label imbalance, drive the cross-lingual variability observed in memorization strength.

3.4 Membership Inference Attack

We evaluate the susceptibility of our models to membership inference attacks (MIAs) by analyzing whether an adversary can distinguish samples that were part of the training set (*in*) from those that were not (*out*). Our analysis focuses on encoder-based models (BERT-family) fine-tuned for classification in each language.

To simulate a realistic adversary, we train a shadow model following the same architecture and optimization procedure as the target model but using a controlled dataset composed of both training (*in*) and test (*out*) samples. The attacker then observes the per-sample confidence scores produced by the shadow and target models to learn a decision boundary distinguishing "in" from "out" samples. This boundary is learned using an XGBoost classifier trained on confidence distributions across epochs (1–30), as the separability between *in* and

out typically increases with training progression.

Training dynamics. As expected, we observe that model confidence for training data progressively diverges from that of unseen data as training advances. Early in training (epochs 1–5), the overlap between *in* and *out* confidence distributions remains substantial, making inference difficult. By epoch 30, however, clearer separation emerges, with train samples forming high-confidence clusters and test samples occupying lower ranges. This evolution highlights how overfitting amplifies membership signal leakage over time.

Cross-lingual separability. Figure 5 visualizes the final in/out confidence distributions at convergence. `English` exhibits the most distinct separation, where the attacker achieves an MIA accuracy of 0.59, with train-set precision of 0.54 and test-set precision of 0.98. This indicates considerable memorization effects and high confidence calibration differences between seen and unseen samples. `Spanish` and `Italian` occupy an intermediate regime, achieving accuracies around 0.51–0.54, where partially overlapping distributions still expose mild but detectable membership traces. `French` demonstrates the tightest overlap between distributions, yielding the lowest attack accuracy (0.50), suggesting better generalization and minimal membership signal leakage.

Implications. These findings highlight a positive coupling between overfitting and membership vulnerability: models that exhibit pronounced memorization behavior (e.g., `English`, `Italian`) are also the most susceptible to membership inference. Languages like `French`, which generalize more smoothly, naturally mitigate this exposure.

4 Privacy Implications of the Language Structures

4.1 Language Characteristics

To capture linguistic properties that may affect memorization and extraction, we compute six structural and morphological indicators. Each metric highlights a specific typological feature that could modulate privacy leakage in LLMs.

Morphological complexity. We measure the average number of inflectional variants per lemma, reflecting how flexional morphology increases linguistic variability [15, 2, 24, 8]:

$$\mathcal{M} = \frac{1}{|V|} \sum_{w \in V} |\mathcal{I}(w)|, \tag{1}$$

where V is the lemma vocabulary and $\mathcal{I}(w)$ denotes the set of inflected forms of lemma w.

Syntactic entropy. This measures word-order variability and structural diversity in dependency relations [12, 23, 18]:

$$\mathcal{S} = -\sum_{r \in R} P(r) \log P(r), \tag{2}$$

where R is the set of syntactic relations and $P(r)$ their empirical probabilities.

Redundancy and predictability. We quantify local contextual predictability through mutual information between neighboring tokens [33, 20]:

$$\mathcal{R} = \frac{1}{N} \sum_{i=1}^{N} I(w_i; w_{i-1}, w_{i+1}), \tag{3}$$

where I denotes mutual information and N the number of tokens. Higher $\mathcal{R}$ implies greater repetition and potential for memorization.

Tokenization characteristics. The average word length serves as a proxy for token fragmentation and morphological density [30]:

$$\mathcal{T} = \frac{1}{|W|} \sum_{w \in W} \text{len}(w), \tag{4}$$

where $\text{len}(w)$ represents the character length of word w.

Capitalization and orthography. We estimate the proportion of capitalized words, which often correspond to named entities and thus correlate with identifiable content [1]:

$$\mathcal{C} = \frac{1}{|W|} \sum_{w \in W} \mathbf{1}[\text{isCapitalized}(w)]. \tag{5}$$

Vocabulary richness. Lexical diversity is represented by the type–token ratio, reflecting the productivity and variability of vocabulary [22]:

$$\mathcal{D} = \frac{|V|}{|W|}, \tag{6}$$

where $|V|$ is the number of unique word types and $|W|$ the total number of tokens.

These indicators collectively reveal the typological contrasts that underlie variations in memorization and extraction behaviors across languages.

Languages characterized by higher morphological complexity and more flexible syntax tend to exhibit distinct privacy-leakage patterns compared to more analytically structured ones.

Specifically, redundancy in linguistic structure amplifies memorization risk by reinforcing repeated patterns; a high capitalization rate signals greater exposure to named entities such as persons or locations, heightening the risk of sensitive data leakage; rich vocabulary and morphological variability may introduce natural obfuscation but simultaneously complicate de-identification; and finally, elevated syntactic entropy reflects greater structural diversity, increasing the likelihood of memorizing unique linguistic sequences.

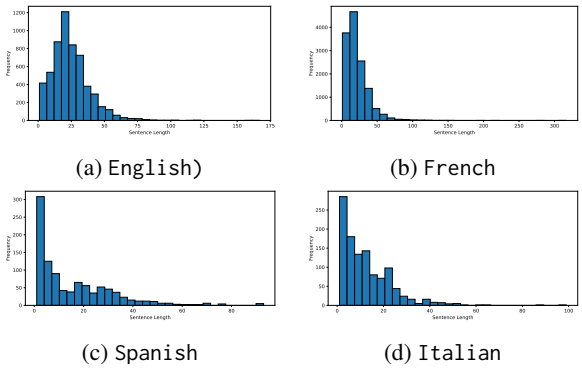
4.2 Comparing Language Characteristics

We compare linguistic metrics across English, Spanish, French, and Italian medical corpora from the HiTZ dataset to assess how language structure influences privacy leakage during LLM training. Table 1 presents six key linguistic indicators used in this comparison.

To further characterize structural variability, we analyze sentence and word length distributions across languages (Figures 6 and 7). Quantitatively, Italian exhibits the longest average sentence length ($\mu_{\text{sent}} \approx 23.4$ words), followed by English (21.8), Spanish (20.7), and French (18.9). This trend aligns with the higher redundancy and morphological density of Italian, suggesting broader contextual spans that may promote memorization.

Similarly, word-length analysis reveals that Italian and Spanish have longer average words ($\mu_{\text{word}} = 5.99$ and 5.79, respectively), while French (5.48) and English (5.77) remain slightly shorter and more evenly distributed. Italian's longer words, coupled with its high redundancy (8.69), increase token-level repetition under subword tokenization, potentially heightening privacy risk. In contrast, French's shorter sentences and lower capitalization rate (7.3%) suggest a lower likelihood of memorizing personally identifiable terms or structured entities.

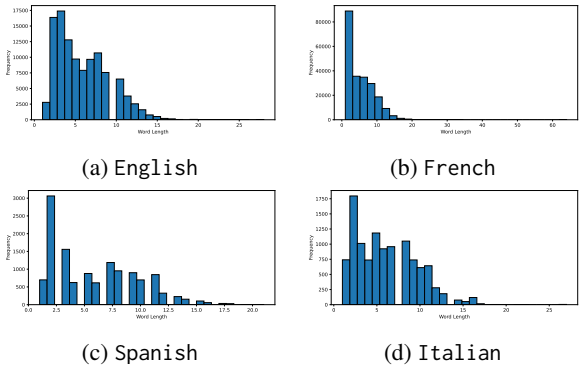
From a privacy standpoint, these quantitative differences highlight distinct trade-offs across languages. Italian shows the strongest exposure indicators: highest redundancy (8.69), longest sentences, and most extended word lengths, suggesting increased risk of memorization and entity leakage. English combines high syntactic entropy (2.90) and capitalization rate (14.5%), which could



(a) English) (b) French

(c) Spanish (d) Italian

Figure 6: Sentence length distributions across languages: Italian and English exhibit longer sentences, consistent with higher redundancy and memorization potential.



(a) English (b) French

(c) Spanish (d) Italian

Figure 7: Word length distributions across languages: Italian and Spanish show heavier right tails, indicating longer lexical units and denser morphology.

heighten exposure to named entities and rare phrasing patterns. Spanish, while morphologically similar to Italian, demonstrates lower redundancy (7.38) and capitalization (9.7%), implying moderate leakage susceptibility. French then exhibits the highest morphological complexity (1.34) but lowest vocabulary richness (0.078), favoring regular inflectional patterns that may mitigate verbatim recall.

4.3 Linking Linguistic Characteristics to Privacy Vulnerabilities

When contextualized with the corpus-level statistics from Table 1 and the empirical findings in Sections 3.2–3.4, a consistent picture emerges linking linguistic structure to privacy vulnerability. Leakage patterns observed across the three attack families: extraction, memorization, and membership inference, closely follow the typological properties of each language.

In the **extraction attack** (Section 3.2), both

Table 1: Linguistic metrics across four languages in the HiTZ multilingual medical corpus.

	Morph. Comp.	Synt. Ent.	Redundancy	Avg. Word Len.	Cap. Rate	Vocab. Rich.
English	1.2227	2.9025	7.7728	5.7750	0.1446	0.1148
Spanish	1.2257	2.8119	7.3764	5.7939	0.0977	0.1269
French	1.3454	2.8632	7.2191	5.4756	0.0733	0.0776
Italian	1.1559	2.7822	8.6942	5.9922	0.1538	0.2193

Spanish and Italian exhibit steady growth in leakage as prompt length increases. This behavior aligns with their higher *redundancy* ($\mathcal{R} = 7.38$ and 8.69, respectively) and longer *average word lengths* (5.79 and 5.99), which encourage surface-level repetition and amplify memorization under subword tokenization. English, while less redundant, demonstrates pronounced leakage for short prompts, consistent with its high *syntactic entropy* ($\mathcal{S} = 2.90$), since even limited context can trigger memorized continuations. In contrast, French, with its greater *morphological complexity* ($\mathcal{M} = 1.35$) and shorter average sentences ($\mu_{\text{sent}} \approx 18.9$), displays a dampened extraction curve, suggesting that rich inflectional variability reduces exact sequence recall.

Results from the **counterfactual memorization** experiment (Section 3.3) reinforce these trends. Italian again shows the strong memorization signal, driven by its high redundancy and longer tokens that form stable phrase structures reused across contexts. English follows closely, where strong syntactic regularities facilitate verbatim recall of distinct patterns. Spanish exhibits low memorization; its structural diversity dilutes recurrence, while French, due to its inflectional diversity, maintains a relatively low recall rate for exact sequences, confirming that morphological variability provides natural resistance to overfitting.

Finally, under the **membership inference attack** (Section 3.4), English fine-tunes show the clearest separation between "in" and "out" samples, indicating strong memorization and poor generalization. Italian also displays detectable separability, though less pronounced, whereas Spanish and especially French exhibit overlapping confidence distributions, reflecting smoother generalization and weaker membership signals. These trends parallel the corpus-level differences in redundancy and morphological diversity.

Overall, the quantitative correspondence between linguistic structure and empirical leakage across all attack types highlights that *language itself is a determinant of privacy risk*. Languages with longer lexical units, higher redundancy, and predictable syntax (English, Italian) exhibit greater memorization and vulnerability to inference attacks. In contrast, morphologically rich and less redundant languages (French, Spanish) demonstrate improved privacy resilience, though longer prompts can still elevate extraction exposure.

5 Limitations

This study empirically examines how linguistic structure influences privacy leakage in LLMs, yet several limitations remain. Our experiments were conducted on relatively small multilingual medical corpora, which may limit generalizability; extending to larger datasets would improve robustness but requires substantial computational resources. The limited number of languages considered is also a limitation. While we considered representative encoder and decoder architectures, exploring diverse model families and fine-tuning configurations could reveal further nuances. Finally, future work could assess fully multilingual models, rather than separately fine-tuned monolingual ones, to capture cross-lingual transfer effects, though this entails significant computational demands.

6 Conclusion

We conduct a cross-linguistic analysis of privacy leakage in LLMs trained on distinct languages, showing that linguistic structure strongly influences model vulnerability. Across English, Spanish, French, and Italian, and under extraction, counterfactual memorization, and membership inference attacks, we observe clear structural effects: Italian shows the greatest leakage due to high redundancy and longer sentences, while English exhibits higher membership separability from greater syntactic entropy. In contrast, French and Spanish remain more resilient through richer morphology. These findings underscore the need for language-aware, structure-adaptive privacy defenses.

References

[1] Françoise Beaufays and Brian Strope. 2013. Language model capitalization. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6749–6752. IEEE.

[2] Dunstan Brown. 2018. A simple account of the complex may take a while gregory t. stump & raphael a. finkel, morphological typology: From word to paradigm. *Word Structure*, 11(2):238–253.

[3] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE.

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

[5] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. *Preprint*, arXiv:1802.08232.

[6] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. *CoRR*, abs/2012.07805.

[7] Ruizhe Chen, Tianxiang Hu, Yang Feng, and Zuozhu Liu. 2024. Learnable privacy neurons localization in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–264, Bangkok, Thailand. Association for Computational Linguistics.

[8] Çağrı Çöltekin and Taraka Rama. 2023. What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*, 9(s1):27–43.

[9] A. Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A. Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, Jack M. Balkin, Nicholas Carlini, Christopher De Sa, Jonathan Frankle, Deep Ganguli, Bryant Gipson, Andres Guadamuz, Swee Leng Harris, Abigail Z. Jacobs, and 16 others. 2023. Report of the 1st workshop on generative ai and law. *Preprint*, arXiv:2311.06477.

[10] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39.

[11] Henri Duprieu and Nicolas Berkouk. 2024. Techniques d'audit des grands modèles de langage. Technical report, Commission Nationale Informatique et Libertés (CNIL).

[12] Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the fifth international conference on dependency linguistics (depling, syntaxfest 2019)*, pages 3–13.

[13] Jamie Hayes, Ilia Shumailov, Christopher A. Choquette-Choo, Matthew Jagielski, George Kaissis, Katherine Lee, Milad Nasr, Sahra Ghalebikesabi, Niloofar Mireshghallah, Meenatchi Sundaram Mutu Selva Annamalai, Igor Shilov, Matthieu Meeus, Yves-Alexandre de Montjoye, Franziska Boenisch, Adam Dziedzic, and A. Feder Cooper. 2025. Strong membership inference attacks on massive datasets and (moderately) large language models. *Preprint*, arXiv:2505.18773.

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

[15] Patrick Juola. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

[16] 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.

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

[18] Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

[19] Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, and 1 others. 2024. Llm-pbe: Assessing data privacy in large language models. *arXiv preprint arXiv:2408.12787*.

[20] Wentian Li. 1989. Mutual information functions of natural language texts. Citeseer.

[21] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, Jinbo Xiong, Zuobin Ying, and Athanasios V. Vasilakos. 2021. Privacy and security issues in deep learning: A survey. *IEEE Access*, 9:4566–4593.

[22] Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

[23] Matilde Marcolli. 2016. Syntactic parameters and a coding theory perspective on entropy and complexity of language families. *Entropy*, 18(4):110.

[24] Claudia Marzi, James P Blevins, Geert Booij, and Vito Pirrelli. 2020. Inflection at the morphology-syntax interface. *Word knowledge and word usage*, 228(10.1515):9783110440577–007.

[25] Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. *Preprint*, arXiv:2203.03929.

[26] 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.

[27] Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. 2024. Rethinking llm memorization through the lens of adversarial compression. *Preprint*, arXiv:2404.15146.

[28] Yashothara Shanmugarasa, Ming Ding, Chamikara Mahawaga Arachchige, and Thierry Rakotoarivelo. 2025. Sok: The privacy paradox of large language models: Advancements, privacy risks, and mitigation. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, pages 425–441.

[29] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

[30] Sagar Tamang and Dibya Jyoti Bora. 2024. Evaluating tokenizer performance of large language models across official indian languages. *arXiv preprint arXiv:2411.12240*.

[31] 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.

[32] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

[33] Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. Quantifying the redundancy between prosody and text. *arXiv preprint arXiv:2311.17233*.

[34] Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore. Association for Computational Linguistics.

[35] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 3093–3106, New York, NY, USA. Association for Computing Machinery.

[36] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting . In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, Los Alamitos, CA, USA. IEEE Computer Society.

[37] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2021. Counterfactual memorization in neural language models. *CoRR*, abs/2112.12938.