Shared Path: Unraveling Memorization in Multilingual LLMs through Language Similarities

Xiaoyu Luo[1,2], Yiyi Chen[1], Johannes Bjerva[1], Qiongxiu Li[2]*
[1]Department of Computer Science, [2]Department of Electronic Systems
Aalborg University, Copenhagen, Denmark
{xilu,yiyic,jbjerva}@cs.aau.dk,qili@es.aau.dk

Abstract

We present the first comprehensive study of *Memorization* in Multilingual Large Language Models (MLLMs), analyzing 95 languages using models across diverse model scales, architectures, and memorization definitions. As MLLMs are increasingly deployed, understanding their memorization behavior has become critical. Yet prior work has focused primarily on monolingual models, leaving multilingual memorization underexplored, despite the inherently long-tailed nature of training corpora. We find that the prevailing assumption, that memorization is *highly correlated* with training data availability, fails to fully explain memorization patterns in MLLMs. We hypothesize that the conventional focus on monolingual settings, effectively treating languages in isolation, may obscure the true patterns of memorization. To address this, we propose a novel graph-based correlation metric that incorporates language similarity to analyze cross-lingual memorization. Our analysis reveals that among similar languages, those with fewer training tokens tend to exhibit higher memorization, a trend that only emerges when cross-lingual relationships are explicitly modeled. These findings underscore the importance of a *language-aware* perspective in evaluating and mitigating memorization vulnerabilities in MLLMs. This also constitutes empirical evidence that *language similarity* both explains *Memorization* in MLLMs and underpins *Cross-lingual Transferability*, with broad implications for multilingual NLP [1].

1 Introduction

Large Language Models (LLMs) demonstrate increasingly strong capabilities in processing and understanding multiple languages (Conneau et al., 2020), resulting in advancements across a wide range of natural language processing (NLP) tasks (Choi et al., 2021; Pikuliak et al., 2021). MLLMs, in particular, empower global users to interact in their native languages, offering wide-reaching benefits in accessibility and productivity.

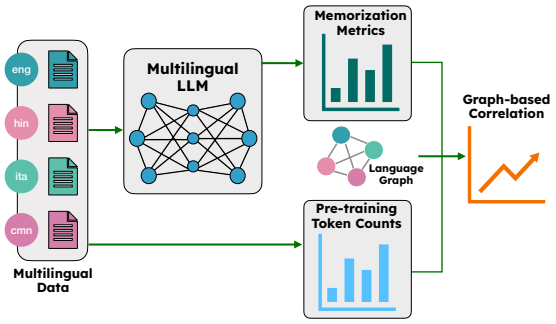

Figure 1: Overview of our Framework for Analyzing Memorization in MLLMs using Language Similarity Graph-based Correlation Analysis.

However, LLMs are also known to *memorize* portions of their training data (Carlini et al., 2021), raising serious concerns such as the leakage of copyrighted content (Chang et al., 2023) and personal information (Staab et al.). While memorization in monolingual LLMs has been widely studied, how it manifests in multilingual models remains *underexplored*.

Prior work predominantly attributes memorization to data volume, positing that frequent tokens or duplicated content are disproportionately memorized (Carlini et al., 2022). This echoes findings from computer vision, where long-tail examples are disproportionately memorized (Feldman and Zhang, 2020; Jiang et al., 2020; Garg et al., 2023), resulting in increased privacy and fairness risks (Li et al., 2024b; Gao et al., 2023; Tramèr et al., 2022). However, MLLMs introduce a unique complexity: languages are not processed independently but in a joint space, often sharing lexical, morphological, and syntactic features. While prior memorization research has largely focused on monolingual mod-

* Corresponding author.
[1]We release our code at: https://github.com/xiaoyuluoit97/MLLM_memorization.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19384–19400
November 4-9, 2025

els and settings Carlini et al. (2021, 2022), without explicitly examining the role of cross-lingual similarity, our work explores how such relationships may shape memorization dynamics. For instance, typologically similar languages like Turkish and Azerbaijani may interact during training in ways that affect their memorization patterns. Moreover, low-resource languages naturally occupy the long tail of the data distribution, introducing complex dynamics that are poorly understood. Together, these challenges raise important questions that motivate our investigation. For example, to what extent does memorization in MLLMs correspond to training data volume, as suggested by long-tail distribution assumptions? How might cross-lingual relationships influence memorization behavior across languages? And can memorization in one language lead to unintended leakage in another, particularly among similar languages?

To answer these questions, we conduct the first large-scale study of memorization in MLLMs, uncovering critical limitations of existing research and offering a novel language-aware perspective (see Fig. 1 for an overview of our framework). Our key contributions are:

- **Revisiting the Long-Tail Assumption**: We show that memorization in multilingual settings cannot be fully explained by training data volume or token frequency. In many cases, low-resource languages exhibit lower memorization rates than high-resource counterparts.
- **Language Similarity-Aware Correlation Metric:** We introduce a novel graph-based correlation metric that incorporates typological and statistical similarities between languages, enabling structured analysis of cross-lingual memorization dynamics.
- **Cross-Lingual Memorization Insights:** Using our metric, we find that languages with high similarity exhibit interconnected memorization behaviors, affording fundamental grounding for cross-lingual transferability.
- **Comprehensive and Robust Evaluation:** We assess memorization using both generation-based and likelihood-based metrics, and validate our findings across over 95 languages, multiple LLM architectures (encoder-only and decoder-based) of varying scales, demonstrating consistent and generalizable trends.

2 Related work

2.1 Memorization in LLM

Memorization in deep neural networks has long been recognized as a critical issue, with implications for privacy, fairness, and generalization (Feldman and Zhang, 2020; Garg et al., 2023; Chang and Shokri, 2021; Li et al., 2025). These concerns have been empirically confirmed in LLMs. Carlini et al. (2019) first show that generative models can inadvertently memorize and reproduce rare, sensitive training data. Carlini et al. (2021) further demonstrate that large models like GPT-2 can regurgitate unique sequences even if they appear only once in the training corpus. Carlini et al. (2022) systematically quantify memorization patterns across model scales and architectures, while Kim et al. (2023) focus on personally identifiable information (PII) memorized by LLMs, proposing ProPILE to assess leakage from the perspective of data subjects.

Recent work has formalized memorization risk, particularly distinguishing between *discoverable* and *extractable memorization* (Carlini et al., 2021; Nasr et al., 2023). The latter refers to information that an adversary can extract without direct access to the training set, posing realistic threats to deployed models. Studies have shown that LLMs, including GPT, T5, and others, can leak hundreds to millions of training sequences depending on model sizes, data duplication, and prompt strategies (Nasr et al., 2023; Carlini et al., 2022). Beyond quantifying leakage, several studies have advanced the understanding of memorization mechanisms and measurement approaches. Chen et al. (2024) analyze how model and context size affect transitions between unmemorized and memorized outputs. Liu et al. (2024) propose the forgetting curve, a corpus-agnostic method to reliably measure memorization capability across architectures. Li et al. (2024a) introduce ROME, revealing how token length and prediction confidence relate to memorization without relying on training data access. Haviv et al. (2022) demonstrate that recall of memorized sequences follows a two-stage process of early promotion and later confidence amplification in transformer models. Stoehr et al. (2024) localize memorization to specific low-layer attention heads and high-gradient parameters, showing such content is harder to unlearn. While such risks have been stud-

ied in monolingual settings, memorization behavior in multilingual LLMs remains underexplored, with the exception of Cavalin et al. (2024), especially for low-resource languages occupying the long tail of the training distribution.

2.2 Cross-lingual Transferability & Language Similarity

Cross-lingual transfer entails the representation of texts in multiple natural languages in a shared multilingual space. The paradigm of representations for cross-lingual transfer has shifted from word embeddings (Mikolov et al., 2013; Ammar et al., 2016; Vulić et al., 2019) to contextual embeddings (Conneau et al., 2019; Devlin et al., 2019; Raffel et al., 2020). Previous work investigating cross-lingual transferability mainly leverages downstream task performance to measure the transfer from a source language or languages to target languages through selective fine-tuning (Choenni et al., 2023) or using zero-shot or few-shot transfer with pre-trained MLLMs (Lauscher et al., 2020; Adelani et al., 2022; de Vries et al., 2022; Blaschke et al., 2025). Language similarity based on linguistic data has been heavily referred to in cross-lingual transferability studies (Wichmann et al., 2011; Littell et al., 2017), not without faulty representations (Toossi et al., 2024; Khan et al., 2025). Moreover, the findings on leveraging language similarity for improving downstream cross-lingual transfer remain mixed and sometimes contradictory (Philippy et al., 2023). Recently, different language similarity measures have been deployed to enhance cross-lingual transfer performance under different NLP tasks (Blaschke et al., 2025) and analyze MLLM language distribution patterns (Chen et al., 2025). We share the perspective that *language similarity is not a static concept*, and different measures can be pertinent to different scenarios.

Prior research in MLLM embedding spaces has shown that sentence embeddings are composed of a *language-specific* and *language-agnostic* components (Pires et al., 2019; Libovický et al., 2020; Xie et al., 2024), which have been leveraged to improve downstream performance (Tiyajamorn et al., 2021) and investigate language relations in MLLMs (Choenni and Shutova, 2022). In addition, Lin et al. (2024) shows that language similarity extracted from pretrained MLLMs with parallel sentences exhibits moderately high correlations with linguistic similarity measures, further motivating our language-aware memorization analysis. Notably, Zhao et al. (2024) demonstrate that even within closely related languages, structural factors such as word order can yield divergent outcomes in knowledge induction, underscoring that language similarity is multifaceted and context-dependent. In this paper, we extract language-specific embeddings from each MLLM as language representations to compute language similarity (cf. Section 4.1).

3 Language Model Memorization

We define *Memorization* in the context of LLMs and examine its key formulations from different perspectives. Given an LM f and a string x from its training data, we split x into a prefix p and a suffix s, so that $x = p||s$. Let the prefix p consist of n tokens, noted as $p = (p_1, \ldots, p_n)$; and let the suffix s consist of m tokens, noted as $s = (s_1, \cdots, s_m)$.

3.1 Measuring MLLM Memorization

Exact Memorization Following the definition of extractable memorization by Carlini et al. (2022), whether a language model can reproduce a training sequence when prompted with part of it using greedy decoding, we define *Exact Memorization Ratio* as $\frac{n}{n+m}$ to measure the fraction of the sequence required for exact reconstruction. Given a set of samples, we define the *Exact Memorization Rate (EM)* as the fraction of samples where the model, when prompted with the prefix, reproduces the suffix exactly:

$$\text{EM} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\hat{s}_i = s_i),$$

where N is the total number of samples, s_i is the true suffix of the ith sample, $\hat{s}_i$ is the output given the prefix and $\mathbb{1}(\cdot)$ is the indicator function.

Relaxed Memorization As *Exact Memorization* is a stringent criterion, we additionally define a relaxed version of memorization that evaluates the predicted suffix against the ground truth suffix using approximate string matching metrics rather than exact match. We use BLEU (Papineni et al., 2002) and Rouge-L (Lin, 2004) as our *Relaxed Memorization Scores (RM)*, serving as continuous indicators of memorization.

Reconstruct Likelihood Memorization Complementary to previous generation-based memorization metrics, we adopt reconstruct likelihood from Kim et al. (2023) to define a probability-based metric *Reconstruct Likelihood Memorization*,

noted as *PM*. which quantifies memorization by the likelihood the model assigns to a known sequence under its learned distribution, i.e., its internal probability of reconstructing the suffix given its prefix. Our goal is to evaluate how likely the model finds the suffix s when conditioned on the prefix p. We define the log-likelihood of s given p as:

$$\log \Pr(s \mid p) = \sum_{r=1}^{m} \log p(s_r \mid p, s_{<r}),$$

where $s_{<r}$ denotes the preceding $r-1$ tokens of the suffix.

3.2 Memorization for Encoder-Decoders

The definitions above primarily assume a decoder-only architecture of LLMs where predictions are made in a left-to-right autoregressive manner. In contrast, encoder-decoder models such as T5 are trained with a span-denoising objective (Raffel et al., 2020). Following Carlini et al. (2022), we randomly mask a set of non-contiguous token spans from a sampled data sequence. To evaluate *Exact Memorization*, the model reconstructs these missing spans given surrounding context, and we consider a string to be memorized if the generated output exactly matches the masked content. To evaluate *Reconstruct Likelihood Memorization*, we follow the span corruption setup and treat the masked spans as targets. We then compute the sum of log-probabilities assigned to these tokens, conditioned on the visible parts of the sequence.

T5's span corruption objective typically mask very short spans (about three tokens on average under default settings (Raffel et al., 2020)), so token-level similarity becomes uninformative, hence we do not assess the relaxed memorization for T5-based encoder-decoder models.

4 Methodology

Previous work on LLM *Memorization* has mainly focused on data duplication and frequency in monolingual settings, with limited analysis across languages. Although correlation metrics such as Pearson can quantify global trends (e.g., measuring how token counts and memorization rates linearly covary), they overlook the structured dependencies among languages. Our analysis (Fig. 2) shows that languages with similar frequency distributions can exhibit divergent memorization patterns, underscoring the importance of language-aware evaluation.

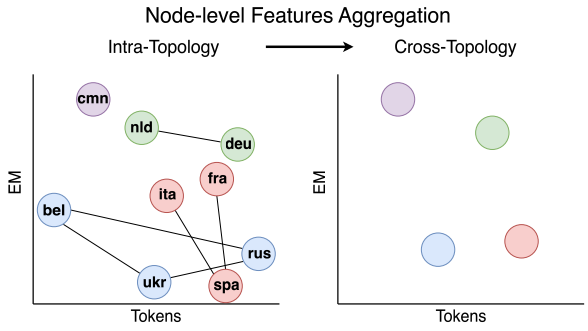

Figure 2: Example graphs considering Intra-Topology and Cross-Topology.

4.1 Measuring Language Similarity

We leverage *language-specific* subspace in multilingual embedding space to measure language similarities. Let L be a set of languages. To extract language representations from MLLMs, we use a parallel dataset D, in our case Flores+ (NLLB Team et al., 2024), which is entirely separate from the model's training data and contains 2,000 examples per language. Suppose we have m sentences for each language $l \in L$ in D, we first extract the mean embedding $\mu_l = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{e}_l^i$ for each hidden layer h, where $\boldsymbol{e}_l^i \in \mathbb{R}^d$ is a sentence embedding. We then form a matrix $M \in \mathbb{R}^{d \times |L|}$ by concatenating μ_l across all languages. We extract the language-specific subspace M_s using Algorithm 1 (Xie et al., 2024) (see Appendix A for details), then project each language embedding into this subspace $\boldsymbol{s}_l = M_s M_s^T \boldsymbol{e}_l$. For each hidden layer h in a MLLM, we measure the pair-wise language similarity for a language pair $\{l_1, l_2\}$, where $l_1, l_2 \in L$ using cosine similarity between language-specific embeddings:

$$\cos\left(\boldsymbol{s}_{l_1}, \boldsymbol{s}_{l_2}\right) = \frac{\boldsymbol{s}_{l_1} \cdot \boldsymbol{s}_{l_2}}{||\boldsymbol{s}_{l_1}|| \cdot ||\boldsymbol{s}_{l_2}||}.$$

Empirically, we find that the language similarity drawn from the final layer embeddings of MLLMs shows a stronger correlation with linguistically grounded similarity measures overall (cf. Appendix B.4).

4.2 Graph-based Correlation Analysis

We introduce our topology-based framework, which captures cross-lingual dependencies by modeling signal propagation over a language similarity graph. It rests on two empirical observations: (1) Memorization patterns tend to propagate across related languages, and (2) Standard correlation metrics fail to capture these structured transfer effects.

Graph Construction via Language Similarity We represent the language space as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node corresponds to a language, and edges encode pairwise language similarity. Let n be the number of languages and $A \in \mathbb{R}^{n \times n}$ the adjacency matrix, where A_{ij} represents the similarity between languages i and j. To sparsify the graph and remove self-loops, we apply thresholding with θ:

$$A_{ij} = \begin{cases} 1, & \text{if } \text{sim}(i,j) \geq \theta \\ 0, & \text{otherwise} \end{cases}. \quad (1)$$

We then construct the unnormalized graph Laplacian matrix $L = D - A$, where $D_{ii} = \sum_j A_{ij}$ is the degree matrix.

Information analysis over the Graph To understand how language-level signals behave over this graph structure, we begin with the concept of *graph smoothness*, which quantifies how much a signal varies across adjacent nodes. For a scalar-valued signal $\mathbf{x} \in \mathbb{R}^n$ defined over graph, the smoothness is defined as (Zhou and Schölkopf, 2004):

$$\mathbf{x}^\top L \mathbf{x} = \sum_{(i,j) \in \mathcal{E}} A_{ij}(x_i - x_j)^2.$$

Smaller values indicate that the signal $\mathbf{x}$ changes slowly over similar nodes, i.e., it is *smooth* with respect to the graph topology.

To compare how two signals (e.g., memorization scores and the number of tokens) vary together across languages, we define the *graph cross-smoothness*:

$$\mathbf{x}^\top L \mathbf{y} = \sum_{(i,j) \in \mathcal{E}} A_{ij}(x_i - x_j)(y_i - y_j),$$

where $\mathbf{y} \in \mathbb{R}^n$ refers to a scalar-valued signal different from $\mathbf{x}$. This measures whether the two signals increase and decrease in tandem over topologically similar languages.

Graph-based Correlation Coefficient Based on the above definitions, we define the proposed *Graph-based Correlation Coefficient* between signals $\mathbf{m}$ (e.g., memorization scores) and $\mathbf{t}$ (e.g., token counts) as:

$$\rho_G(\mathbf{m}, \mathbf{t}) = \frac{\mathbf{m}^\top L \mathbf{t}}{\sqrt{(\mathbf{m}^\top L \mathbf{m})(\mathbf{t}^\top L \mathbf{t})}}$$

Note that the defined coefficient is bounded by the Cauchy-Schwarz inequality:

$$|\mathbf{m}^\top L \mathbf{t}| \leq \sqrt{(\mathbf{m}^\top L \mathbf{m})(\mathbf{t}^\top L \mathbf{t})}$$

Hence, $\rho_G(\mathbf{m}, \mathbf{t}) \in [-1, 1]$ and it captures the structural alignment between the two signals over the graph. A value close to 1 implies that memorization and token frequency change similarly across related languages, while values near -1 implies inverse alignment.

ρ_G accounts for the topological structure of language space, enabling us to uncover subtle, structure-respecting relationships in MLLM memorization, which would otherwise be missed by flat, language-agnostic analyses such as Pearson correlation (cf. Table 1 for details).

4.3 Intra-Topology & Cross-Topology Analysis

To further interpret the structure of memorization alignment, we partition the graph into subgraphs by thresholding edge weights. Each subgraph represents a cluster of similar languages; disconnected components reflect cross-topological groups. To enable meaningful comparison across different language topology clusters, we aggregate node-level features into a single representative vector per subgraph. This aggregation is performed within each subgraph, it is weighted by language prominence (node degrees) and normalized by global edge weights to preserve topological information. Specifically, for a subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$, where each node $i \in \mathcal{V}'$ has features t_i (tokens) and m_i (memorization), we define the subgraph-level representations as:

$$\bar{t} = \sum_{i \in \mathcal{V}'} \left(\frac{n_i}{\sum_{j \in \mathcal{V}'} n_j} \cdot t_i \right)$$

where $n_i = |\{j \mid (i,j) \in \mathcal{E}'\}|$ is the degree of node i. The aggregated memorization $\bar{m}$ is computed similarly.

We refer **intra-topo** as the set of language nodes connected by edges in the graph, while **cross-topo** refers to language groups that remain disconnected. The resulting subgraph-level representations enable cross-topology correlation analysis via Pearson correlation. This approach remains faithful to the internal structure of each language cluster, while capturing the relationship between memorization and training tokens across topologically dissimilar clusters. It complements our topology-aware metric ρ_G by offering a cluster-level, interpretable view of memorization–complexity alignment.

5 Experimental Setup

5.1 Model Selection & Corpus Details

Studying memorization in MLLMs requires i) publicly available models with ii) fully disclosed pre-training data and iii) broad language coverage. For fair cross-architecture comparisons, we also align models by their training corpora and tokenizers whenever feasible. We use the MT5 encoder-decoder family (Xue et al., 2020), trained on MC4 (Raffel et al., 2020) covering 100+ languages, and the MGPT decoder-only series for architectural comparison. Specifically, MGPT-101 shares the tokenizer and mC4 training data with MT5-BASE. Additionally, we select MGPT-1.3B and MGPT-13B to assess scale effects, which are trained on more balanced and filtered MC4 (cf. Table 5 for details).

As shown in Fig. 9 and 10 in Appendix B.7, the data distribution of MC4 across languages exhibits a clear *long-tailed* pattern. A small number of high-resource languages (such as English, Russian, and Spanish) dominate the corpus in terms of token count, while the vast majority of other languages are represented with significantly fewer tokens. This long-tailed distribution serves as an important factor in analysing how memorization behaviors vary across languages in MLLMs.

5.2 Prompt sampling

MC4 contains a substantial amount of noisy and duplicated content. For pre-processing, we sample text passages with more than 600 characters, and filter the content containing "http://", garbled tokens, repeated strings, and long sequences of meaningless digits. To ensure accurate language representation, we use CLD3 (cld) for language identification. Specifically, we retain only those samples where both the predicted language confidence and the proportion of the target language exceed 90%.

Duplicated content can disproportionately impact memorization, where sequences that appear more frequently in the training set are more likely to be memorized, following a near log-linear trend (Lee et al., 2021). To control repetition for minimizing potential bias and ensure a more balanced representation across the dataset, we randomly sample 50,000 filtered examples per language with a 5 million shuffle buffer, following the sample size in Carlini et al. (2022). A handful of low-resource languages with insufficient examples are marked with an asterisk and boldface in Fig. 9; 10.

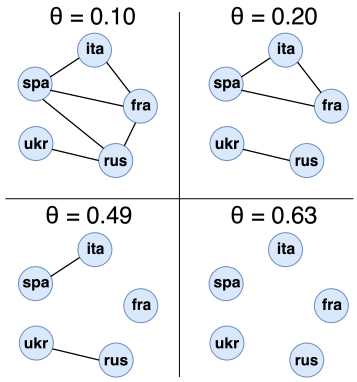

Figure 3: Graph Construction at Different Thresholds θ.

6 Analysis & Results

We investigate *Memorization* in MLLMs across multiple dimensions: languages, model architectures, prompt length and model scale. In each dimension, we measure the Memorization Rates (cf. Section 3) and correlate with training data (in token counts) in languages, using both the Pearson correlation (r) and Graph-based Correlation (ρ_G) metrics.

6.1 Constructing Language Graphs

To use our graph-based correlation to analyze memorization in MLLMs, we construct language similarity-based graphs, at varying thresholds θ based on equation 1, which specifies the minimum similarity required for two languages to be considered meaningfully related. Thus, θ directly controls the sparsity of the resulting language graph. Fig. 3 illustrates this effect using a subset of MGPT-101 pre-training languages, showing how edge density and connectivity increase as θ increases. As expected, higher language similarity thresholds θ, the fewer connected graphs. By varying θ, we adjust the granularity of the language similarity topology, enabling analysis under different levels of relational strictness.

6.2 Data Availability in Memorization

We evaluate the relationship between per-language memorization rates and the token counts in training data in a MLLM, for example, MGPT-101. As shown in Table 1, our graph-based metric ρ_G, by incorporating language similarity, largely *accentuates* the negative correlation between language-wise memorization and token count, in comparison

Mem. Metric	r	ρ_G
EM	-0.13	-0.24
PM	-0.36	-0.56
RM (BLEU)	-0.23	-0.36
RM (Rouge-L)	-0.06	-0.30

Table 1: Correlations between Memorization Rates and Training Data in Token Counts of MGPT-101. The ρ_G with graph-based metric threshold $\theta = 0.41$. **Takeaway:** the proposed ρ_G accentuates the correlation.

to the Pearson correlation coefficient. This negative trend suggests that, among similar languages, those with fewer training tokens tend to exhibit higher memorization, which further corroborates our hypothesis that memorization in MLLMs cannot be explained by training data volume alone.

6.3 Cross-lingual Transferability vs. Memorization

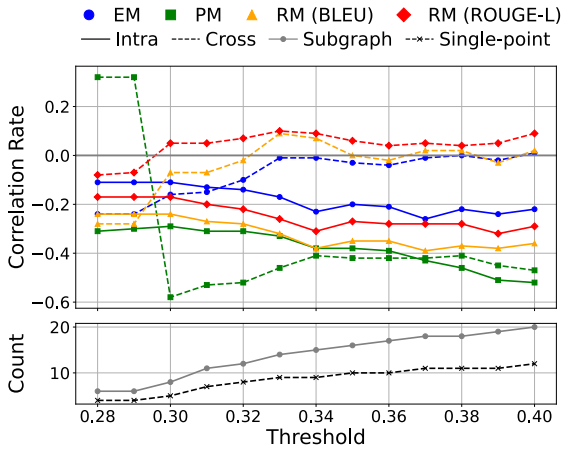

Figure 4: Intra-Topology and Cross-Topology Correlation Coefficients (ρ_G) across varying thresholds θ. **Top**: Memorization Rates across Thresholds. **Bottom**: Topology graph information via subgraph and singleton counts at varying threshold (x-axis), from 6 to 20 language groups (y-axis), with a total of 95 languages. **Takeaway**: Cross-lingual transferability among similar languages impact memorization.

Leveraging the constructed language graph, we measure the topology-based correlation for both intra-topo and cross-topo at various θ. As shown in Fig. 4, among cross-topo languages, EM and RM become largely uncorrelated with token counts, spanning from -0.2 to 0.05 with the growing number of language groups. While PM has a stronger negative correlation, the correlation becomes generally weaker as more cross-topo language groups are created, from -0.6 to -0.4. This highlights that, *across distinctive language groups, the correlation between memorization and data volume becomes weaker*. In contrast, consistent with previous findings (cf. Section 6.2), intra-topo ρ_G values grow increasing negative (down to -0.6) across memorization metrics, as more similar languages are grouped together (as θ becomes higher), indicating *an inverse relationship between training data and memorization within similar languages*.

From both cross-topo and intra-topo perspectives, our results show that as MLLMs are trained with richer data from similar languages, memorization decreases —— evidence that *cross-lingual transferability* among similar languages plays an essential role in memorization in MLLMs.

Mem. Metric	EM		PM	
Model	r	ρ_G	r	ρ_G
MT5-SMALL	0.05	0.12	0.10	-0.53
MT5-BASE	-0.12	-0.15	0.48	0.03
MT5-LARGE	0.01	0.08	0.47	0.19
MGPT-1.3B	0.22	**-0.49**	-0.39	-0.63
MGPT-13B	0.18	-0.13	-0.39	**-0.78**

Table 2: Correlations between Memorization Rates and Training Data across Models and Scales, with specific θ for Intra-Topology Correlation. **Takeaway:** In contrast to r, ρ_G presents stronger correlations and more consistent alignment with prior memorization analyses. **Bold** values indicate the highest-magnitude correlation.

Mem. Metric	RM (BLEU)		RM (Rouge-L)	
Model	r	ρ_G	r	ρ_G
MGPT-1.3B	-0.18	-0.53	0.42	-0.42
MGPT-13B	-0.21	-0.31	0.4	-0.04

Table 3: Relaxed Memorization Rates across MGPT models, with specific θ for Intra-Topology Correlation.

6.4 Memorization across Model Architectures & Scales

Since language similarity is model-specific, its scores exhibit different distributions across models. We select the specific threshold θ to better incorporate structural patterns based on language similarity, while confirming that the observed trends hold across a range of thresholds (cf. Appendix B.2).

Table 2 presents the intra-topology correlations with model-specific thresholds. For MGPT-1.3B and MGPT-13B - trained on a corpus with a less pronounced long-tailed distribution, r appears positive, seemingly contradictory to previous findings (cf. Sections 6.2; 6.4). However, leveraging language similarity and filtering out noisy language

pairs, ρ_G shows negative correlations, consistent with prior findings. Notably, with PM, MGPT-13B presents the strongest negative correlation, suggesting that larger models trained on a more balanced corpus reveal the strongest inverse link between memorization and data availability in similar languages. In contrary, MT5's EM results exhibit a different trend compared to MGPT models, which might be attributed to its encoder-decoder architecture. As RM is not applicable to MT5-based models (cf. Section 3.2), we show the relaxed memorization metrics for MGPT models in Table 3. We observe a consistent trend aligns with our earlier findings: memorization is negatively correlated with training data quantity among similar languages.

In summary, our analysis and results support the claim that memorization in MLLMs is not shaped solely by training data volume - as commonly observed in computer vision task - but also by intricacies among languages. Specifically, when language similarity is incorporated via a topology-based metric, we show that languages with fewer training tokens tend to exhibit higher memorization — a pattern that only becomes evident when language relations are explicitly modeled.

6.5 Effect of Prompt Length & Model Scale on Memorization

Model	Prompt. Len.	EM (%)	PM	RM (B)	RM (R)
GPT2 Decoder-only: MGPT-101					
MGPT-101	50	0.22	-44.4	3.2	9.8
	100	0.42	-41.9	3.6	10.1
	150	**0.56**	**−40.9**	**3.9**	**10.1**
GPT3 Decoder-only: MGPT-*1.3B / 13B*					
MGPT-1.3B	50	0.31	-33.7	**4.1**	**5.7**
	100	0.29	-32.0	3.7	5.7
	150	**0.32**	**−31.1**	3.5	4.8
MGPT-13B	50	1.01	-32.2	7.1	7.6
	100	1.38	-30.2	8.1	8.2
	150	**1.56**	**−29.5**	**8.6**	**8.4**
Encoder-Decoder: MT5 *family*					
MT5-SMALL	50	0.02	-66.1	–	–
	100	0.15	**−56.9**	–	–
	150	**0.25**	-61.3	–	–
MT5-BASE	50	0.07	-45.7	–	–
	100	0.50	-35.0	–	–
	150	**0.90**	**−31.4**	–	–
MT5-LARGE	50	0.02	-78.4	–	–
	100	0.23	-52.6	–	–
	150	**0.49**	**−39.0**	–	–

Table 4: Memorization Rates across various prompt lengths $(35, 85, 135)$, model architectures and scales. The predicted token length is fixed at 15. The highest memorization rates for each model are **bold**. **Takeaway**: Overall, the memorization rates increase with the increasing prompt lengths, with a few exceptions.

To investigate the effects of experimental setup on memorization, we measure memorization across models of different architectures and scales at varying prompt-length $(35, 85, 135)$, with the fixed output token length of 15. The prompt-length refers to prefix-length in the context of decoder-only models. As shown in Table 4, across all model types, we observe a consistent trend: *longer prompts lead to higher memorization*. This pattern holds across the memorization metrics, with a few exceptions, as underlined, and aligns with prior findings on memorization in monolingual LMs, indicating that longer contexts offer more cues for memorization (Carlini et al., 2021, 2022).

In GPT-3–based decoder-only models, we also observe a clear **scaling effect**: *larger models exhibit stronger memorization*, particularly in exact memorization. For example, EM increases from 0.32% in MGPT-1.3B to 1.56% in MGPT-13B with the prefix of length 135. Results in other metrics (e.g., PM, RM) follow this trend with few exceptions. In comparison, the encoder-decoder models tells a different story. While memorization generally increases with growing scale (e.g., MT5-SMALL to MT5-BASE), the largest model (MT5-LARGE) exhibits lower memorization when compared to MT5-BASE.

In addition, we observe that MT5-LARGE — without downstream finetuning — produces more broken completions for masked tokens. We hypothesize that this instability may lead to reduced memorization rates in MT5-LARGE, especially in a masked language modeling context. We provide a random example of such unstable generation in Appendix B.6.

6.6 Language-Level Memorization across Prompt Lengths & Model Scales

We analyze how language-level memorization varies across different prompt lengths and model scales by computing Pearson correlations of per-language memorization rates under each condition. Across all models, language-level memorization distributions at different prompt lengths remain strongly correlated. For decoder-only models, Pearson correlations consistently exceed 0.9 in all memorization metrics, while for the MT5 models, they are generally above 0.8, with the lowest still above 0.66. These results indicate that languages with high memorization tend to remain highly memorized regardless of prompt length. See Table 12 and 13 for detailed results.

A similar trend holds across model scales. Across all metrics and model scales, the Pearson correlation is consistently shows a strong positive correlation, with the lowest value being 0.71. These results suggest that memorization tendencies are stable, intrinsic language-level characteristics that generalize across both prompt length and model scale. We observe a "the poorer get poorer" phenomenon, where languages with high memorization consistently remain high across settings. See Table 14 for full results.

7 Conclusion and Future Work

We present the first large-scale study of memorization in MLLMs, grounding observed memorization patterns through language similarity and revealing *cross-linguality* as a key factor shaping memorization in MLLMs. To this end, we define memorization metrics tailored to language models and propose a graph-based correlation measure that incorporates language similarity, uncovering patterns that linear metrics fail to capture. Notably, the tendency for languages with fewer training tokens to exhibit higher memorization, a trend that only becomes apparent when language relationships are explicitly modeled. We experiment on a range of language models, across architectures, scales and 95 languages, showing consistent memorization trends. Our findings urge a paradigm shift toward language-aware memorization audits in MLLMs, particularly for under-resourced languages vulnerable to cross-lingual leakage. We encourage further work at the intersection of multilingualism and memorization to develop effective strategies to mitigate memorization in MLLMs.

Limitations

Our proposed memorization metric relies on a manually selected similarity threshold to construct the graph, making it sensitive to this parameter and limiting its applicability to languages with low similarity to others, which often become isolated nodes and reduce interpretability. A more robust approach could involve adaptive threshold optimization or the development of threshold-free methods that fully leverage language similarity without requiring manual intervention. While our work provides the first large-scale analysis of memorization in MLLMs, we primarily examine models in their pre-trained state and do not explore how fine-tuning or instruction tuning may alter memorization behavior, particularly in task-specific or alignment-sensitive contexts. Nonetheless, we believe our study offers a principled and extensible foundation for understanding memorization through the lens of language similarity in multilingual models.

Ethics Statement

We comply with the ACL Ethics Policy. This work aims to improve understanding of memorization risks in multilingual language models, with the broader goal of enabling safer and more privacy-preserving NLP systems. All experiments are conducted on publicly available pre-trained models and benchmark datasets. We do not train on, extract from, or attempt to infer sensitive personal information from proprietary or private data.

Acknowledgements

XL, YC and JB are funded by the Carlsberg Foundation, under the Semper Ardens: Accelerate programme (project nr. CF21-0454). XL is additionally supported by the EU ChipsJU and the Innovation Fund Denmark through the project CLEVER (no. 101097560). We further acknowledge the support of the AAU AI Cloud and express our gratitude to DeiC for providing computing resources on the LUMI cluster (project nr. DeiC-AAU-N5-2024085-H2-2024-28). Finally, we thank the Aalborg University AI:X initiative for enabling this work via the AI:SECURITY lab.

References

Compact language detector v3 (cld3). https://github.com/google/cld3. Accessed: 2025-04-30.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, and 26 others. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Verena Blaschke, Masha Fedzechkina, and Maartje ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. *arXiv preprint arXiv:2501.14491*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Paulo Cavalin, Pedro Henrique Domingues, Claudio Pinhanez, and Julio Nogima. 2024. Fixing rogue memorization in many-to-one multilingual translators of extremely-low-resource languages by rephrasing training samples. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4503–4514, Mexico City, Mexico. Association for Computational Linguistics.

Hongyan Chang and Reza Shokri. 2021. On the privacy risks of algorithmic fairness. pages 292–303. IEEE.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to chatgpt/gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327.

Bowen Chen, Namgi Han, and Yusuke Miyao. 2024. A multi-perspective analysis of memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11190–11209.

Yiyi Chen, Qiongxiu Li, Russa Biswas, and Johannes Bjerva. 2025. Large language models are easily confused: A quantitative metric, security implications and typological analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3810–3827, Albuquerque, New Mexico. Association for Computational Linguistics.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during lm fine-tuning. *arXiv preprint arXiv:2305.13286*.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672.

Hyunjin Choi, Judong Kim, Seongho Joe, Seungjai Min, and Youngjune Gwon. 2021. Analyzing zero-shot cross-lingual transfer in supervised nlp tasks. *arXiv preprint arXiv:2101.10649*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. 33:2881–2891.

Yinghua Gao, Yiming Li, Linghui Zhu, Dongxian Wu, Yong Jiang, and Shu-Tao Xia. 2023. Not all samples are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition*, 139:109512.

Isha Garg, Deepak Ravikumar, and Kaushik Roy. 2023. Memorization through the lens of curvature of loss function around samples. *arXiv preprint arXiv:2307.05831*.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2022. Understanding transformer memorization recall through idioms. *arXiv preprint arXiv:2210.03588*.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. 2020. Exploring the memorization-generalization continuum in deep learning. *arXiv preprint arXiv:2002.03206*.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36:20750–20762.

Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Conference on Empirical Methods in Natural Language Processing*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Bo Li, Qinghua Zhao, and Lijie Wen. 2024a. Rome: Memorization insights from text, logits and representation. *arXiv preprint arXiv:2403.00510*.

Qiongxiu Li, Xiaoyu Luo, Yiyi Chen, and Johannes Bjerva. 2025. Trustworthy machine learning via memorization and the granular long-tail: A survey on interactions, tradeoffs, and beyond. *arXiv preprint arXiv:2503.07501*.

Xiao Li, Qiongxiu Li, Zhanhao Hu, and Xiaolin Hu. 2024b. On the privacy effect of data enhancement via the lens of memorization.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, Andre Martins, and Hinrich Schuetze. 2024. mPLM-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 276–310, St. Julian's, Malta. Association for Computational Linguistics.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

Xinyu Liu, Runsong Zhao, Pengcheng Huang, Chunyang Xiao, Bei Li, Jingang Wang, Tong Xiao, and Jingbo Zhu. 2024. Forgetting curve: A reliable method for evaluating memorization capability for long-context models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4667–4682.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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. *arXiv preprint arXiv:2311.17035*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. *arXiv preprint arXiv:2305.16768*.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *arXiv preprint arXiv:2403.19851*.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hasti Toossi, Guo Qing Huai, Jinyu Liu, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2024. A reproducibility study on quantifying language similarity: The impact of missing values in the uriel knowledge base. In *North American Chapter of the Association for Computational Linguistics*.

Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. Truth serum: Poisoning machine learning models to reveal their secrets. pages 2779–2792.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Søren Wichmann, Taraka Rama, and Eric W Holman. 2011. Phonological diversity, word length, and population sizes across languages: The asjp evidence.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. 2024. Discovering low-rank subspaces for language-agnostic multilingual representations. *arXiv preprint arXiv:2401.05792*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Qinghua Zhao, Vinit Ravishankar, Nicolas Garneau, and Anders Søgaard. 2024. Word order and world knowledge. *arXiv preprint arXiv:2403.00876*.

Dengyong Zhou and Bernhard Schölkopf. 2004. A regularization framework for learning from graph data. In *ICML 2004 workshop on statistical relational learning and its connections to other fields (SRL 2004)*, pages 132–137.

A Language-specific Subspaces

The algorithm for identifying language-specific subspace is as in Algorithm 1, refer to Xie et al. (2024) for more details.

Algorithm 1: Language-specific Subspace Identification

Input: Languages' mean Embeddings $\boldsymbol{M}$, rank of subspace r.
Output: Language-agnostic component μ, language-specific subspace $\boldsymbol{M}_s$, coordinates Γ.
/* 1) Approximate $\boldsymbol{M}$ in low rank */
$\mu' \leftarrow \frac{1}{d}\boldsymbol{M}\mathbb{1}^\intercal$
$\boldsymbol{M}'_{s,\_,\Gamma} \leftarrow \text{Top-}r\text{SVD}(\boldsymbol{M} - \mu'\mathbb{1}^\intercal)$;
$\boldsymbol{M}' \leftarrow \mu'\mathbb{1}^\intercal + \boldsymbol{M}'_s\Gamma'^\intercal$;
/* 2) Force orthogonality */
$\mu \leftarrow (1/||\boldsymbol{M}' + \mathbb{1}||^2)\boldsymbol{M}'^{+}\mathbb{1}$
$\boldsymbol{M}_s, \_, \Gamma \leftarrow \text{Top-}r\text{SVD}(\boldsymbol{M}' - \mu\mathbb{1}^\intercal)$

B Appendix

B.1 Models detail

Model	#Params	#Langs. (used.)	Architecture	Layers
MGPT-101	560M	101 (95)	GPT-2 based	24
MGPT-61	1.3B	61 (48)	GPT-3 based	24
MGPT-61	13B	61 (48)	GPT-3 based	40
MT5-SMALL	300M	101 (95)	Encoder-Decoder	8
MT5-BASE	580M	101 (95)	Encoder-Decoder	12
MT5-BASE	1.2B	101 (95)	Encoder-Decoder	24

Table 5: MLLMs and their Scale, Datasets, Languages (analyzed), Architectures.

B.2 Cross-lingual correlation

	Threshold θ							
MGPT-101	0.31	0.33	0.35	0.37	0.39	0.41	0.43	0.45
# **Subgraph**	11	14	16	18	19	25	26	35
# **Single Point**	7	9	10	11	11	18	18	24
EM Intra	-0.13	-0.17	-0.20	-0.26	-0.24	-0.24	-0.19	-0.17
EM Cross	-0.15	-0.01	-0.03	-0.01	-0.02	0.04	0.04	-0.09
PM Intra	-0.31	-0.33	-0.38	-0.43	-0.51	-0.56	-0.54	-0.57
PM Cross	-0.53	-0.46	-0.42	-0.42	-0.45	-0.35	-0.36	-0.36
RM (B) Intra	-0.27	-0.32	-0.35	-0.39	-0.38	-0.36	-0.33	-0.31
RM (B) Cross	-0.07	0.09	-0.00	0.02	-0.03	0.04	0.04	-0.12
RM (R) Intra	-0.20	-0.26	-0.27	-0.28	-0.32	-0.30	-0.26	-0.24
RM (R) Cross	0.05	0.10	0.06	0.05	0.05	0.41	0.42	0.18

Table 6: Cross-topo vs. intra-topo correlation at low thresholds for mGPT-101.

	Threshold θ							
MGPT-1.3B	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89
# **Subgraph**	8	11	12	13	22	28	31	33
# **Single Point**	5	7	8	9	17	23	26	27
EM Intra	-0.04	-0.09	-0.12	-0.11	-0.16	-0.26	-0.49	-0.43
EM Cross	0.74	0.76	0.56	0.55	0.25	0.19	0.20	0.19
PM Intra	-0.45	-0.46	-0.49	-0.51	-0.59	-0.50	-0.60	-0.63
PM Cross	0.05	0.18	0.18	0.14	-0.06	-0.13	-0.20	-0.22
RM (B) Intra	-0.20	-0.22	-0.27	-0.29	-0.35	-0.40	-0.53	-0.50
RM (B) Cross	0.34	0.43	0.40	0.34	0.04	-0.06	-0.09	-0.11
RM (R) Intra	0.24	0.19	0.09	0.05	-0.14	-0.27	-0.42	-0.35
RM (R) Cross	-0.03	-0.17	-0.16	-0.14	0.00	0.25	0.28	0.32

Table 7: Cross-topo vs. intra-topo correlation at high thresholds for mGPT-1.3B.

	Threshold θ							
MGPT-13B	0.28	0.30	0.32	0.34	0.36	0.38	0.40	0.42
# **Subgraph**	4	8	10	14	22	26	31	31
# **Single Point**	2	5	7	10	13	17	21	21
EM Intra	-0.07	-0.12	-0.10	-0.10	0.18	0.19	0.17	0.17
EM Cross	0.15	0.29	0.23	0.11	0.37	0.31	0.30	0.30
PM Intra	-0.35	-0.42	-0.46	-0.55	-0.57	-0.65	-0.78	-0.78
PM Cross	-0.85	0.28	0.28	0.17	-0.00	-0.12	-0.25	-0.25
RM (B) Intra	-0.21	-0.27	-0.27	-0.31	-0.18	-0.21	-0.33	-0.33
RM (B) Cross	-0.96	0.26	0.27	0.14	0.19	0.08	-0.07	-0.07
RM (R) Intra	0.08	0.10	-0.01	-0.04	0.12	0.23	0.20	0.20
RM (R) Cross	0.56	0.02	0.01	0.08	0.32	0.34	0.44	0.44

Table 8: Cross-topology vs. intra-topology Pearson correlation at varying thresholds for MGPT-13B.

	Threshold θ							
MT5-SMALL	0.54	0.56	0.58	0.60	0.62	0.64	0.66	0.68
# **Subgraph**	30	38	46	52	56	66	72	77
# **Single Point**	16	23	29	36	41	57	61	70
EM Intra	0.27	0.27	0.26	0.24	0.20	0.22	0.16	0.12
EM Cross	-0.14	-0.06	-0.03	-0.04	-0.02	-0.04	-0.01	0.01
PM Intra	-0.13	-0.13	-0.11	-0.12	-0.19	-0.32	-0.38	-0.53
PM Cross	0.33	0.15	0.17	0.10	0.11	0.16	0.14	0.12

Table 9: Cross-topology vs. intra-topology Pearson correlation at varying thresholds for MT5-SMALL.

	Threshold θ							
MT5-BASE	0.72	0.74	0.76	0.78	0.80	0.82	0.84	0.86
# **Subgraph**	1	2	7	14	29	48	62	74
# **Single Point**	0	0	6	9	21	39	50	64
EM Intra	-0.15	-0.13	-0.10	-0.12	0.04	0.04	-0.02	-0.14
EM Cross	0.00	-1.00	-0.24	-0.27	-0.07	-0.14	-0.20	-0.16
PM Intra	0.20	0.13	0.07	0.07	0.15	0.22	0.13	0.03
PM Cross	0.00	-1.00	-0.44	-0.16	0.04	0.21	0.27	0.28

Table 10: Cross-topology vs. intra-topology Pearson correlation at varying thresholds for MT5-BASE.

	Threshold θ							
MT5-LARGE	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92
# **Subgraph**	7	9	21	27	51	65	79	86
# **Single Point**	6	8	18	23	48	60	74	83
EM Intra	0.18	0.24	0.19	0.17	0.27	0.27	0.18	0.07
EM Cross	-0.31	-0.22	-0.11	-0.14	0.08	-0.06	-0.00	0.02
PM Intra	0.30	0.27	0.21	0.19	0.20	0.23	0.12	-0.13
PM Cross	0.43	0.52	0.52	0.52	0.51	0.57	0.52	0.50

Table 11: Cross-topology vs. intra-topology Pearson correlation at varying thresholds for MT5-LARGE.

B.3 Prompt length impact

Model	EM		PM	
	50 vs. 100	100 vs. 150	50 vs. 100	100 vs. 150
GPT2 Decoder-only: MGPT-101				
MGPT-101	0.97	0.98	0.99	0.99
GPT3 Decoder-only: MGPT-1.3B / 13B				
MGPT-1.3B	0.90	0.98	0.99	0.99
MGPT-13B	0.96	0.99	0.99	0.99
Encoder-Decoder: MT5 *family*				
MT5-SMALL	0.81	0.96	0.84	0.88
MT5-BASE	0.86	0.97	0.99	0.98
MT5-LARGE	0.66	0.94	0.94	0.97

Table 12: Correlation of memorization metrics (exact and family vs probability) between prompt lengths 50 vs 100 and 100 vs 150 across model families. EM = Exact Memorization, PM = Probability Memorization.

Model	RM BLEU		RM ROUGE-L	
	50 vs. 100	100 vs. 150	50 vs. 100	100 vs. 150
GPT2 Decoder-only: MGPT-101				
MGPT-101	0.92	0.97	0.99	0.99
GPT3 Decoder-only: MGPT-1.3B / 13B				
MGPT-1.3B	0.95	0.99	0.99	0.99
MGPT-13B	0.99	0.99	0.99	0.99

Table 13: Correlation of relaxed memorization metrics between different prompt lengths.

Model Pair	Mem. Metric	r
MT5-SMALL vs. MT5-BASE	EM	0.71
	PM	0.76
MT5-BASE vs. MT5-LARGE	EM	0.81
	PM	0.72
MGPT-1.3B vs. MGPT-13B	EM	0.92
	PM	0.99
	RM (BLEU)	0.97
	RM (ROUGE-L)	0.99

Table 14: Pairwise memorization correlation (r) between adjacent model scales for exact memorization (EM), probability memorization (PM), and reference match metrics (RM).

B.4 Layer-wise Lang2Vec correlation

We include supplementary visualizations showing how various linguistic feature correlations evolve across layers for different multilingual models.

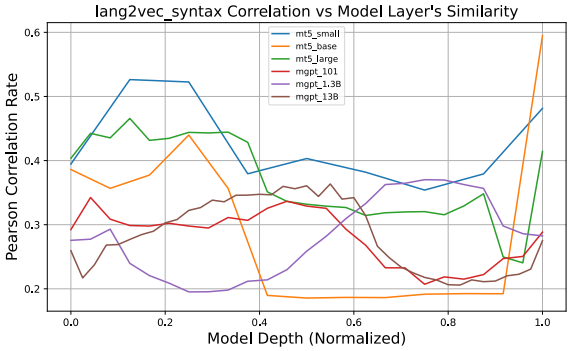

Figure 5: Layer-wise trend for `Lang2Vec (Syntax)`.

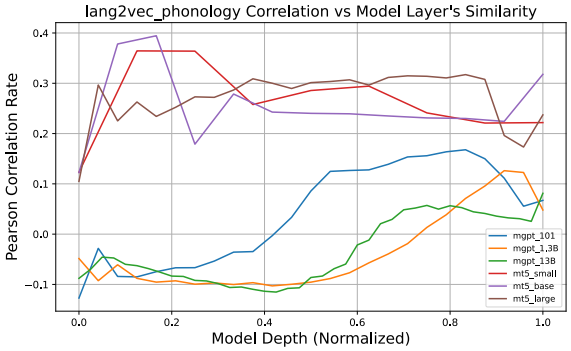

Figure 6: Layer-wise trend for `Lang2Vec (Phonology)`.

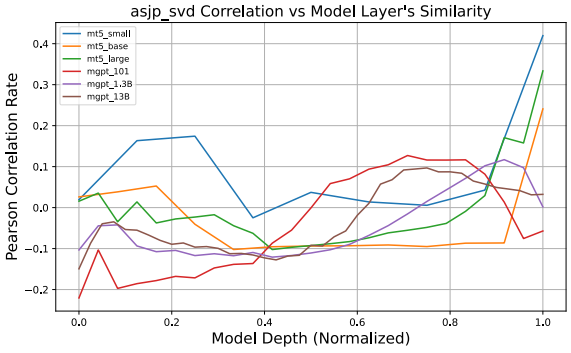

Figure 7: Layer-wise trend for `ASJP (SVD)`.

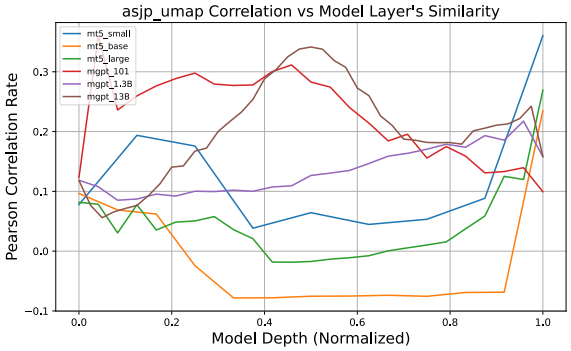

Figure 8: Layer-wise trend for `ASJP (UMAP)`.

B.5 Examples of Exact Memorization

Danish MGPT-101

Prompt:
Americas Best Value Inn Santa Rosa tilbyder også mange faciliteter der vil berige dit ophold i Santa Rosa (CA). Hotellet tilbyder sine gæster adgang til et stort udvalg af servicetilbud, som trådløst internet i fællesområder, parkering, familieværelse. Hotellets bekvemmeligheder er særligt udvalgt for at sikre den højeste komfort. På nogle af værelserne kan gæsterne finde internetadgang - trådløst,

Reference:
ikke-rygerværelser, aircondition, skrivebord

Prediction:
ikke-rygerværelser, aircondition, skrivebord

Danish MT5-BASE

Prompt:
- Se på kort Mere om Pensjonat Mi **<extra_id_0>** Miłosna er indrettet til **<extra_id_1>**- og forretningsrejse **<extra_id_2>**er idéelt i Kwidzyn; én af byens mest populære beliggenheder. Herfra har gæster glæde af nem adgang til alt, hvad denne livlige by kan tilbyde. Med sin praktisk **<extra_id_3>** b **<extra_id_4>** dette hotel nem adgang til byens vigtigste sev

Reference:
<extra_id_0>łosna Pensjonat **<extra_id_1>** både ferie **<extra_id_2>**nde og ligg **<extra_id_3>**e **<extra_id_4>**eliggenhed tilbyder

Prediction:
<extra_id_0>łosna Pensjonat **<extra_id_1>** både ferie **<extra_id_2>**nde og ligg **<extra_id_3>**e **<extra_id_4>**eliggenhed tilbyder

German MGPT-101

Prompt:

Wir denken ebenfalls, dass solcherlei akzeptabel recherchierte Tests, überaus hilfreich sind. Trotzdem wollen wir du jene Gattung von Produktvorstellungen nicht anbieten, weil der Markt außerordentlich schnelllebig und dynamisch ist und zum wiederholten Male neumodische Produktkette dazukommen und die "alten" Produktmodelle uninteressant werden, egal um welches Produkt es geht. Deswegen bieten wir auf unserer Seite ausschließlich eine Darstellung von den jetzigen 5 Produkte an. Somit kann

Reference:

man sich selbsttätig seine Favoriten intuitiv raussuchen

Prediction:

man sich selbsttätig seine Favoriten intuitiv raussuchen

German MT5-BASE

Prompt:

die Versandkosten ungeachtet dessen überaus nie **<extra_id_0>**halten werden oder keineswegs erst anfallen. Zu diesem Zweck gehören die Leistung, die getrennten Einstellungen, die Größe des Körpers und der genaue Einsatzbereich. Das **<extra_id_1>** ein außergewöhnlich breites Angebot von Erzeugnissen fix **<extra_id_2>**roduzenten ak **<extra_id_3>**. Häufig werden lediglich wenige be **<extra_id_4>**t, weil die

Reference:

<extra_id_0>drig ge **<extra_id_1>** Kaufportal offeriert **<extra_id_2>** vom P **<extra_id_3>**kurat wie von Händlern **<extra_id_4>**rücksichtig

Prediction:

<extra_id_0>drig ge **<extra_id_1>** Kaufportal offeriert **<extra_id_2>** vom P **<extra_id_3>**kurat wie von Händlern **<extra_id_4>**rücksichtig

English MGPT-101

Prompt:

exactly dimension of Modern Ideas Sports Wallpapers Backgrounds Hd On The App Store was 246x246 pixels. You can also look for some pictures that related to Modern Ideas Sports Wallpapers Backgrounds Hd On The App Store by scroll down to collection on below this picture. If you want to find the other picture or article about Sports Wallpapers just push the next button or previous button; or if you are interested in similar pictures of Modern Ideas Sports Wallpapers Backgrounds Hd On

Reference:

The App Store, you are free to browse through search feature that

Prediction:

The App Store, you are free to browse through search feature that

English MT5-BASE

Prompt:

the administration announced a $6 million investment over two years for provider education and outreach. Expand support **<extra_id_0>** with Alzheimer

<extra_id_1>their families: <extra_id_2> with Alzheimer's disease and their families and care <extra_id_3>requires giving them the tools that they need, helping to plan for future needs, and ensuring that safety and dignity are <extra_id_4>ed," the report says. The announcement proposes an investment

Reference:

<extra_id_0> for people <extra_id_1>'s disease and <extra_id_2> "Supporting people <extra_id_3>givers <extra_id_4>maintain

Prediction:

<extra_id_0> for people <extra_id_1>'s disease and <extra_id_2> "Supporting people <extra_id_3>givers <extra_id_4>maintain

Chinese MGPT-101

Prompt:

大。 新宝gg游戏平台网页版第88届奥斯卡颁奖礼已经落下帷幕,与其有关的话题还在持续。获奖的近20部影片中有不少改编自小说,单是入围"最佳影片"角逐的9部影片就有5部改编自小说。其中,像《荒野猎人》《房间》等获奖影片的原著小说都出版了中文版。此外,获提名的《火星救援》《卡罗尔》等四部电影的原著小说也有了中文版。看过电影后,不妨去读读这些原著小说。 昨日早上5时

Reference:

许,在距离爆炸现场南侧不到400米处的天津港进口

Prediction:

许,在距离爆炸现场南侧不到400米处的天津港进口

Chinese MT5-BASE

Prompt:

也反映了国内垂直电商的困境 <extra_id_0>平台型电商,垂直电商的 <extra_id_1>了,很难形成核心壁垒。"他说到。途棋牌在外贸方面,广东全年进出口顺差为1.54万亿元,出口增速快于进口4.5个百分点;一般贸易 <extra_id_2>比重为49.0%,比上年提高2.0个百分点。从区域看 <extra_id_3>一带一路"沿线国家进出口总额增长6.3%。 途 <extra_id_4> 傲头傲脑

Reference:

<extra_id_0>。"相对 <extra_id_1>获客成本太高 <extra_id_2>占进出口总额的 <extra_id_3>,对" <extra_id_4>棋牌

Prediction:

<extra_id_0>。"相对 <extra_id_1>获客成本太高 <extra_id_2>占进出口总额的 <extra_id_3>,对" <extra_id_4>棋牌

Japanese MGPT-101

Prompt:

最高です。 義実家の姑・義姉は良い人なのですが、クーポンの服には出費を惜しまないためおすすめしていないと大変です。自分が惚れ込んだ物は用品が合わなくたって「いつか着れる」と買ってしまうので、用品がドンピシャの頃には収納に埋もれていたり、出してもアウトドアテーブル 120 80だって着たがらないんですよね。オーセンティックな感じの商品の服だと品質さえ良ければクーポンのことは考えなくて済むのに、カードの趣味

Reference:

や私の反対意見などには耳も貸さずに購入するため、

Prediction:

や私の反対意見などには耳も貸さずに購入するため、

Japanese MT5-BASE

Promp t

の無料を聞いていない <extra_id_0>。用品が話しているときは夢中になるくせに、用品が念を押したことや予約 <extra_id_1>てしまうようです。アウトドアテーブル 120 80だって仕事だってひと通りこなしてきて、クーポンがないわけではないのですが、ポイントもない様子で、パソコンがいまいち噛み合わないのです。クーポンが <extra_id_2>言いませんが、サービスの妻はその傾向が強いです。 夏日になる日も増えてきましたが、私は昔からモバイル <extra_id_3>ダメで <extra_id_4>。この用品

Reference:

<extra_id_0>と感じることが多いです <extra_id_1>はなぜか記憶から落ち <extra_id_2>みんなそうだとは <extra_id_3>が <extra_id_4>湿疹が出てしまいます

Prediction:

<extra_id_0>と感じることが多いです <extra_id_1>はなぜか記憶から落ち <extra_id_2>みんなそうだとは <extra_id_3>が <extra_id_4>湿疹が出てしまいます

B.6 Example of Unstable generation

Unstable Generation Example (MT5-LARGE)

Reference:
`<extra_id_0>` beneficiaries of `<extra_id_1>` the `<extra_id_2>` the bond `<extra_id_3>`, agreeing to invest `<extra_id_4>` $56.6 million in

Predicted:
`<extra_id_0>` . public bond.. mill for parents school students vote mill projects

B.7 Corpus Distribution

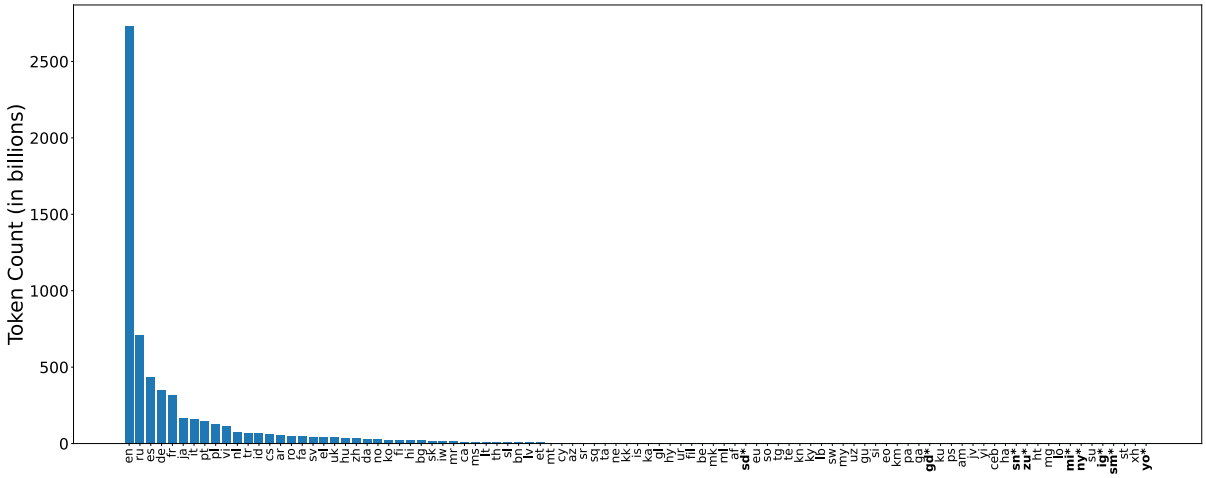

Figure 9: MGPT-101 & MT5 family analyzed language tokens distribution. The Languages marked with * have fewer than 50,000 sampled examples, averaging 33,960 examples per language.

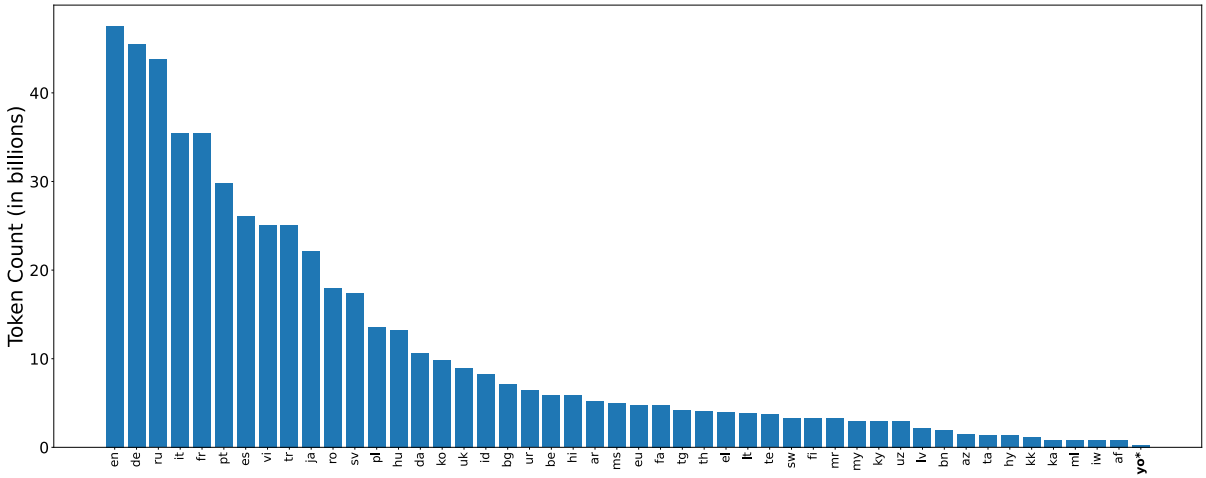

Figure 10: MGPT-61 (1.3B & 13 B) family analyzed language tokens distribution. The language marked with * has fewer than 50,000 sampled examples, with a total of 17,339 examples.