Universitat Pompeu Fabra Barcelona

Treball de fi de màster Recerca

Exploring Geometric Compression Across Languages in Multilingual Language Models

Autor: Eder Ruiz Moreno

Director: Dr. Corentin Kervadec

Màster: Lingüística Teòrica i Aplicada

Edició: 2023-2024

Any de defensa: 2024

Col·lecció: Treballs de fi de màster

Departament de Traducció i Ciències del Llenguatge

Abstract

This study explores geometric compression of linguistic data across languages in multilingual language models using the Europarl corpus, focusing on three models: BLOOM, XLM-RoBERTa, and Mistral. We estimate the intrinsic dimension (ID) of hidden representations at each layer to quantify geometric compression. In Transformer-based LMs, the last hidden representation arises from a series of intermediate representations computed through a number of identical modules. Our analysis reveals that the ID of these representations is significantly smaller than the ambient dimension, with distinct compression patterns across languages. Languages from the same family exhibit similar ID amplitudes, suggesting that shared linguistic properties impact the dimensionality of model representations. Additionally, we find that the model's performance on a language correlates with the ID amplitude at the first high-dimensionality phase, indicating that the learned linguistic properties influence compression. These findings complement those found in other studies, bringing new insights to our understanding of how state-of-the-art LLMs process and compress linguistic data in different languages.

Keywords: LLMs, Transformers, multilingual LMs, geometric compression, intrinsic dimension.

Table of Contents

Table of Figures

1. Introduction

Language models (LMs) have revolutionized the field of natural language processing (NLP), demonstrating unprecedented capabilities in understanding and generating human language. These state-of-the-art models, which are based on the Transformer architecture, need to distill vast amounts of linguistic information into workable representations. A crucial aspect of this process is the concept of geometric compression, which can be quantified using intrinsic dimension (ID) estimators. Intrinsic dimension can be understood as the minimum number of variables required to faithfully capture the essential features of the data. Understanding this compression can offer insights into the internal workings of language models and their ability to process and represent different types of linguistic data.

Although Transformer-based language models achieve remarkable performance, our understanding of their inner workings and the nature of the information they retain remains an open question. The hidden representation in LLMs arises from a series of intermediate representations processed through a sequence of identical blocks, resulting in a succession of vector spaces that are dimensionally uniform but geometrically diverse.

In particular, recent studies have shown that the ID of the hidden representations is significantly reduced compared to their extrinsic dimension (Valeriani et al., 2023). Typically, hidden layers in large language models exhibit dimensions of approximately 2048 or more, while the intrinsic dimension is usually less than 50. This notable compression demonstrates that large language models are highly effective in extracting the essential features of the input data into a much reduced subspace.

Moreover, this compression has been shown to vary across the layers of the transformer, suggesting that different layers contribute differently to the overall processing dynamics of the model (Valeriani et al., 2023). Cheng et al. (2023) further revealed that the ID profile of LLMs

changes depending on the domain of the text being processed, suggesting that the semantic content has an impact on compression.

Despite these advances, a gap remains in our understanding of how LLMs handle linguistic data from multiple languages, particularly in multilingual settings. Most research has focused on monolingual models or single-language datasets, leaving a significant gap in our knowledge of multilingual LLMs. This study aims to fill this gap by measuring and analyzing the intrinsic dimension of the hidden representations computed by multilingual language models across various languages. We aim to build upon previous work and offer insights into the compression patterns of LLMs on different linguistic data. In particular, if ID profiles vary across discourse domains, given the diversity and complexity of human languages, **what should be expected when the model processes text from a single specific domain but in several languages?**

In order to approach this issue, we use the Europarl corpora (Koehn, 2005), which contains transcriptions of European Parliament proceedings in a number of European languages. This provides a convenient dataset that covers one specific discourse domain: parliamentary debates. We further process[1] the corpus and generate various semantically aligned datasets in several languages. We then use the Maximum Likelihood Estimator (MLE) to quantify the intrinsic dimension of the hidden representations at each layer of three models: the multilingual models BLOOM and XLM-RoBERTa, and the latest-generation Mistral model.

Our findings confirm that the intrinsic dimension of the representations in the analyzed multilingual LLMs is significantly smaller than their ambient dimension, consistent with previous research. We find that Transformer-based multilingual LLMs exhibit distinct levels of compression when processing texts from different languages. The intrinsic dimension

[1] The code used for this study is available at github.com/ederruiz98/compression-multilingual-LMs

profiles exhibit similar patterns across different languages, suggesting that syntax and script do not have an impact on ID. However, we observe that languages within the same family tend to have similar ID amplitudes, which implies that compression is similar when processing text in closely related languages with shared linguistic properties. Additionally, we find that the performance of XLM-R on a language (measured via perplexity) and the ID amplitude at the first high-dimensionality phase are correlated. These results suggest that the linguistic properties learned by the models influence the dimensionality of their representations.

Regarding the structure of this paper, we first present the research question and state two hypotheses, followed by a reasonably comprehensive section devoted to the background and related work. In this section, we devote the *Language Modeling* section purely to the background, covering chronologically from word embeddings to neural language models and, later on, transformers. We then present multilingual models, previous studies and foundations of the geometry of representation in these models, and Indo-European languages. Next, we describe the methodology used in depth, and conclude by presenting and discussing our results, followed by an outline of the limitations of this study and the potential lines of future research.

2. Research Question and Hypotheses

Toward this objective, we formulate the following research question:

RQ: Does geometric compression of LLMs vary across languages? In particular, what are the compression patterns of a LLM when processing texts from various languages?

We quantify geometric compression by estimating the intrinsic dimension of the hidden representations computed by the model. Since training in transformers is usually auto-regressive and they are composed of identical self-attention blocks, the vector representation of the last token of the input sequence is the one used for the prediction of the next token.

Moreover, this last hidden representation is the one that encodes all the contextual information, so here is where all the necessary information to be used for the prediction is gathered. We replicate the experiment with a masked language model (XLM) and also used the last token representation in order to be comparable with the auto-regressive models (BLOOM and Mistral)[2].

Accordingly, two hypotheses are put forward:

H1: LLMs demonstrate distinct compression patterns when processing texts from diverse languages.

H2: The ID amplitude of languages from the same family will share similarities.

3. Background and Related Work

In this section we outline the fundamental concepts underlying current language model architectures and their representations, from neural networks to Transformer-based language models. We also provide a review of previous research on data compression of language models, presenting a coherent outline of recent research and the state of the art.

3.1. Language Modeling

3.1.1. Word Embeddings

In order for a machine learning model to process any type of linguistic data, it must be first transformed into vectors, or *word embeddings*. These are essentially representations of words in a vector space. Word embeddings are based on distributional semantics, which is the area that studies both theories and methods for quantifying semantic similarities between

[2] A possible alternative is to compute the ID on the concatenation of all the tokens of the input sentences. However, this would have required much more computing resources.

linguistic elements, taking into account their distributional properties in large linguistic corpora. The core idea of distributional semantics can be summarized in the so-called distributional hypothesis: "linguistic elements with similar distributions have similar meanings". That is to say, words that are used and appear in the same contexts tend to convey similar meanings. The British linguist John Rupert Firth popularized this underlying idea that a word is defined by its companies.

> The statement of meaning by collocation and various collocabilities does not involve the definition of word-meaning by means of further sentences in shifted terms. Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*. This kind of mutuality may be paralleled in most languages and has resulted in similarities of poetic diction in literatures sharing common classical sources. (Firth, 1957)

This idea can be put into practice in a straightforward way: if we can access meaning through context, then the key issue is to capture information from the context and include it in the representation of the word itself. This may be carried out by means of count-base methods, which involves the manual incorporation of data derived from global corpus statistics, and by means of prediction-based methods. Probably, the most remarkable example of the latter is Word2Vec, where word vectors are learned by training them to predict their contexts (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). It operates iteratively by scrolling through a text. At each stage it focuses on a word and its context words according to the specified context window, and calculates the probabilities of occurrence of each context word, adjusting the vectors each time to approximate those probabilities. Some of the most noteworthy pre-trained vectors are those of GloVe, which resemble count-based methods in that they obtain information from global corpus statistics, but also prediction-based methods, since word vectors are learned by means of gradient descent rather than dimensionality reduction (Pennington et al., 2014).

Despite their usefulness, traditional word embeddings also present challenges. One notable issue is their static nature. For instance, they do not account for polysemy, where the one single word may have more than one meaning based on the context. Recent advancements in contextual embeddings, such as CoVe[3] (McCann et al., 2017) and ELMo[4] (Peters et al., 2018), address this limitation by generating dynamic word representations that vary with context.

3.1.2. Neural Language Models

Language models basically predict the probability of a sequence of tokens. Statistical language models estimate the probability distribution of a sequence of words by computing the probability of a given word occurring based on its preceding words (Dauphin et al., 2017). Thus, the probability of a token sequence is:

$$P(y_1, \dots, y_n) = \prod_{t=1}^{n} P(y_t | y_1, \dots, y_{t-1})$$

Traditionally, *n-gram* language models estimate the probability distribution by extracting global statistics from a corpus first (Kneser & Ney, 1995; Chen & Goodman, 1999). They assume the Markov property, i.e., that the probability of a word depends solely on a fixed number of preceding words (Voita, 2020). Formally, the assumption is as follows:

$$P(y_t | y_1, \dots, y_{t-1}) = P(y_t | y_{t-n+1}, \dots, y_{t-1})$$

[3] CoVe are context vectors that are learned by using the encoder from a model trained for machine translation. Since these encoders first need to process the source sentence to be able to provide its translation into another language, the encoder's vector representations inevitably need to contain some contextual information. As CoVe represents words in context and GloVe represents individual words, McCann et al., (2017) proposed to use both of them together, which resulted in a noticeable improvement in most common NLP tasks.

[4] Instead of using the vector representations from the encoder of a machine translation model, those of a language model could also be used. Such is the case of ELMo (Peters et al., 2018). The architecture is straightforward, comprising two-layer language models in both forward and backward directions, providing each token with contextual information from both left and right.

For instance, in a trigram model (n=3), the probability of a word happening once the previous words have happened is equal to the probability of that word happening once the previous two words have happened:

$$P(y_t|y_1, \dots, y_{t-1}) = P(y_t|y_{t-2}, y_{t-1})$$

When generating text with n-gram language models, one quickly realizes that the model uses a very short context, a small number of tokens. That is precisely the main drawback of these models, their limited context understanding.

Later, it was shown that language models based on neural networks performed much better than classical n-gram models (Bengio et al., 2003; Józefowicz et al., 2016). In contrast to the latter, neural models do not define formulas from global statistics, but rather train a neural network to predict them. Conceptually, neural language models undertake two primary tasks:

1) **Contextual encoding.** Firstly, they encode the preceding context into a vector representation. The manner in which the models handle this is specific to their architecture. Two noteworthy examples are Recurrent Neural Networks or RNN (Mikolov et al., 2010) and Convolutional Neural Networks or CNN (Dauphin et al., 2017).
2) **Token probability prediction.** Subsequently, the model generates a probability distribution for the next token from the vector representation of the context.

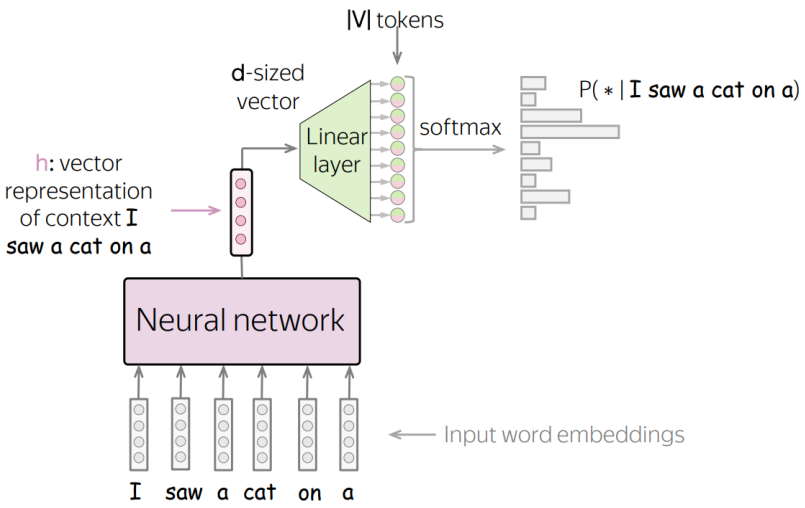

Figure 1: General Pipeline of a Neural Language Model Regardless of the Model Architecture.

Note. From Voita, E. (2020, September). *NLP Course For You*. https://lena-voita.github.io/nlp_course.html

The vector *h* is of size *d*, and it is the contextual representation made by the neural network. This *d*-sized vector is transformed linearly from size *d* to size |V| (vocabulary tokens). Finally, the softmax operation is applied to the |V|-sized vector so the vector representation of the context is transformed into a probability distribution.

A key feature of these neural networks is that the contextual representation is fixed. This presents a major dual problem. On the one hand, conversion of the entire sentence to a single vector by the encoder is highly challenging. Since the potential number of sentences and meanings is infinite, the process of conversion to a single vector is a compression process in which information is lost. On the other hand, at each stage of the text generation process certain parts of the context become more important than others, but the decoder is only provided with a single fixed context representation.

Therefore, the solution is somewhat straightforward: allowing the model's decoder to assign different importance or weights to each part of the context at each stage of the process. In line with this idea, the attention mechanism is presented in 2016 (Bahdanau et al., 2016). Now, the encoder provides representations for each and every token in the context, rather than a single compressed representation of the whole context. There are several functions to compute the attention score at every stage. Two of the most remarkable ones are the bilinear function or Luong attention (Luong et al., 2015), and the multi-layer perceptron or Bahdanau attention, presented in the original paper (Bahdanau et al., 2016).

3.1.3. Transformers

Once the attention mechanism was introduced, it was only a year later that the Transformer was presented, a model whose architecture is based purely and exclusively on attention mechanisms (Vaswani et al., 2017). Now, the interaction process between the encoder and the decoder no longer relies on recurrences or convolutions, but only on attention mechanisms. These models became the state of the art and remain so to this day (including some modifications), not only because they perform considerably better than previous recurrent neural networks, but also because they are enormously more efficient and, consequently, faster (Radford et al., 2018; Al-Rfou et al., 2018; Kaplan et al., 2020).

In classical neural networks, intuitively, the encoder processes word by word, which means that sometimes the exact meaning of a word cannot be known until a larger part of the context has been encoded. On the contrary, in a Transformer encoder the tokens are interacting with one another all at the same time, continuously updating representations, which greatly speeds up this process. The decoder works in a similar way, with a self-attention mechanism[5],

[5] Self-attention takes place between representations of the same nature (for instance, between tokens at the same state, or between all the states of the encoder at one particular layer).

allowing tokens to interact with each other, and also making use of the states of the encoder. When tokens interact with each other via an attention mechanism, extra information of the context is collected, thus updating their own representation. These processes take place simultaneously, not one by one. However, today, encoder-decoder architectures are no longer used in favor of decoder-only architectures.

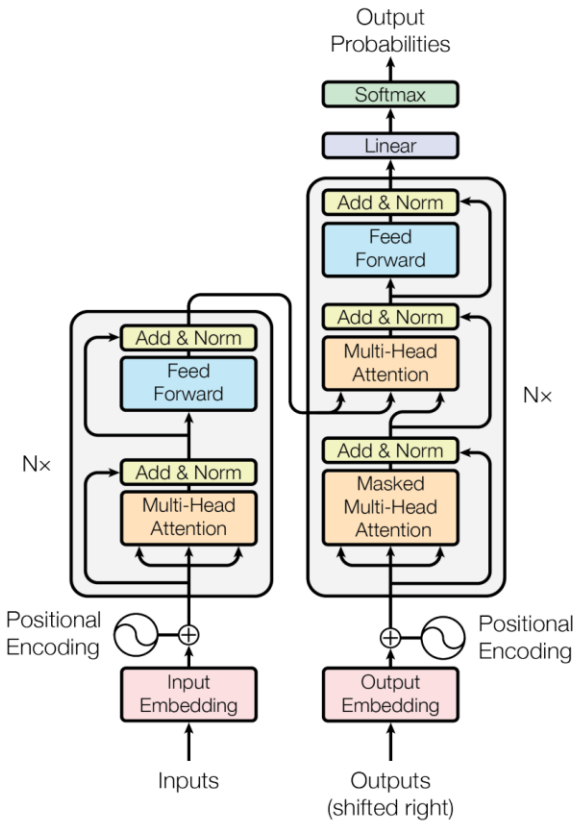

Figure 2: Overview of a Transformer's Model Architecture.

Note. From Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

By way of synthesis of the overall process, tokens interact with each other in the encoder, thus updating their representations, whereas in the decoder tokens update their representations by interacting with previously generated tokens and the source. Apart from the attention and self-attention mechanisms, two other main elements can be highlighted: the feed-forward network and residual connections. The feed-forward network takes the information collected by the attention mechanism and processes it. After this, residual connections simply add the input[6] to the output, a very simple mechanism but one that facilitates flow and allows the simultaneous use of several layers.

Previous architectures were strongly dependent on the task for which they were used, i.e., the architecture of each model was different depending on which task they were intended for. On the contrary, one of the advantages of Transformers is that they are independently pre-trained and their architecture remains intact during fine-tuning. This framework is known as *transfer learning*, as one single architecture can be fine-tuned on different tasks.

For instance, GPT is a Transformer-based left-to-right language model composed of a Transformer decoder with 12 layers (Radford et al., 2018). As it can be seen in Figure 3, the architecture of the model remains the same at the fine-tuning stage, while the input data is fed differently depending on the objective task. Later, successive generations of GPT were released, with different model sizes for each generation, and modifications to the hyperparameters and training dataset (Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023).

[6] It is important to note that the vector representation of each input token is the result of the sum of the representation of the token itself and its position. Previously, only the representation of the token was needed. This is due to the fact that the Transformer is unaware of the positions of the input tokens, since it does not have recurrence or convolution mechanisms, but only attention mechanisms.

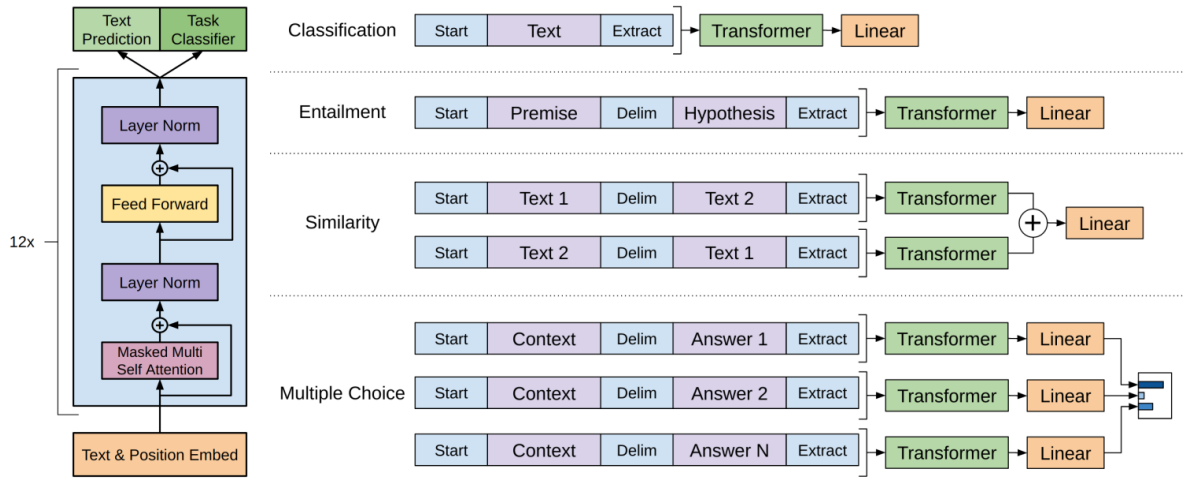

Figure 3: GPT's Transformer Architecture and Input Transformations for Fine-Tuning on Several Tasks.

Note. From Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. https://api.semanticscholar.org/CorpusID:49313245

Another Transformer-based language model that is worth noting is BERT, which uses a Transformer's encoder (Devlin et al., 2019). If GPT was a left-to-right language model, BERT is bidirectional. In GPT's Transformer decoder, each token can interact only with previous tokens, so the final representations from the final layer only encode preceding context. In contrast, BERT uses another pre-training approach: Masked Language Modeling (MLM). This target is different in that the model is able to look at the entire context, left and right[7], but some tokens are masked out (usually replaced with a special token), and the goal is to predict the original tokens.

[7] ELMo representations also know about both left and right contexts, but this required concatenating the representations of two unidirectional models, a forward and a backward one.

3.2. Multilingual Models

For all the aforementioned reasons, transformers became, and are nowadays, the state-of-the-art architecture for language models. The introduction of ELMo, and later on GPT and BERT, resulted in the popularization of pre-trained language models for initial setup for further fine-tuning on downstream tasks. Thereafter, newer and improved models have been developed, often employing different pre-training strategies (Lewis et al., 2019; Liu et al., 2019; Raffel et al., 2019; Zhang et al., 2019).

It has been empirically shown that the size of a model and its performance are positively correlated (Hestness et al., 2017; Kaplan et al., 2020). Models tend to perform better the larger they are. This is referred to as *scaling* language models, and it involves scaling both the model parameters and the computational load during training. This positive tendency is usually maintained under different variations in the hyperparameters.

This finding has led to scaling models up, and new models have become progressively larger. Consequently, the increasing costs of training large language models have meant that only large companies and organizations with substantial capital investment can afford to do so, which in turn translates into a certain reticence to release open-source models. These circumstances and the fact that most of the research community has been somewhat deprived of participating in the development of state-of-the-art models has had major consequences. Among the most notable ones, and the one that is relevant here, is the fact that the vast majority of LLMs are trained predominantly in English (Scao et al., 2022).

Thus, most language models are only functional in English. Others have been trained with more languages, yet LLMs are usually trained on a relatively low number of languages, especially neglecting low-resource languages. To overcome this problem and with the goal of democratizing LLMs, hundreds of researchers collaborated and released the BigScience Large

Open-science Open-access Multilingual Language Model (BLOOM) (Scao et al., 2022). Another noteworthy multilingual model is XLM-RoBERTa (Conneau et al., 2019), which significantly outperforms the multilingual mBERT.

The latest generation of language models is still primarily trained in English, but also starts to reach decent performances in other languages. Two examples of these models are ChatGPT, which is a sister model to InstructGPT (Ouyang et al., 2022), and Mistral (Jiang et al., 2023). However, we can only hypothesize so to a certain extent, because neither model is open-source and we do not know what dataset they have been trained on.

In this experiment we use the multilingual models BLOOM and XLM-RoBERTa, and the latest-generation model Mistral (Jiang et al., 2023). The specific features of these models are detailed in the methodology section.

3.3. The Geometry of Representation in Transformers

3.3.1. Intrinsic Dimension of the Representation Space

With the proliferation of high dimensional datasets from areas like speech analysis, image processing, and genetic research, there is a need for effective analysis methods. Many of these datasets have numerous features, which in the area of machine learning was once believed to be the *curse of dimensionality*. However, real-world datasets present regularities and certain patterns that give rise to strong correlations between their features (Valeriani et al., 2023). These concentrations of measure phenomena were discovered at the end of the nineteenth century, but it was not until the beginning of the twenty-first century that it was evidenced that the correct understanding and handling of these phenomena would turn the previous curse into what some authors call the *blessing of dimensionality*, especially in machine learning (Gorban & Tyukin, 2018).

One approach, termed *Manifold Learning*, assumes this hypothesis, i.e. it assumes that these high dimensional datasets typically exist close to a low dimensional manifold. To put it another way, most high dimensional data from the real world lie along the vicinity of a low dimensional manifold, although everything occurs within a high dimensional space. This underlying hypothesis is often referred to as the *Manifold Learning Hypothesis* or simply *Manifold Hypothesis* (Narayanan & Mitter, 2010; Fefferman et al., 2013).

While data are typically represented by high-dimensional vectors, they can often be effectively represented in spaces with fewer dimensions without sacrificing information integrity. Such simplification, known as *dimensionality reduction*, is often crucial for algorithms to function effectively. Consequently, numerous techniques for reducing dimensionality have been introduced recently (Campadelli et al., 2015). Generally, this is the manner of quantifying the so-called *geometric compression* of a data representation in a neural network (Cheng et al., 2023). In this context, a key issue is determining the least number of variables required to faithfully capture a system's significant features. This minimal number is referred to as the *Intrinsic Dimension* (ID) of the data manifold. "A dataset's intrinsic dimension (ID) is the dimensionality of the manifold *approximating* the data and can be described as the minimum number of coordinates that allow specifying a data point approximately without information loss" (Valeriani et al., 2023).

However, estimating the ID may be challenging from a statistical point of view. In particular, there are two known problems that ID estimators usually face: the curvature of the manifold and the local variations of the density of the data points (Ansuini et al., 2019). Traditionally, the methods for estimating ID are classified into global and local methods (Campadelli et al., 2015). Global ones operate under the premise that the dataset is distributed across a single geometric object, determining its dimensionality by considering the dataset as

a whole. Conversely, local ID estimators focus on calculating the ID for individual data points within their respective neighborhoods and subsequently formulate a global estimate by combining these local measurements (Cheng et al., 2023).

Two well-known examples of global estimators are TwoNN (Facco et al., 2017) and those based on Principal Component Analysis or PCA (Jolliffe, 2002). Local estimators, however, have been shown to obtain a better global estimate of the intrinsic dimension (Carter et al., 2010), as they work better at high dimensions. Moreover, Cheng et al., (2023) tested 12 different ID estimators, both global and local. Based on their results, we use the *Maximum Likelihood Estimator* (MLE) (Levina & Bickel, 2004) due to its optimal performance. MLE is a local estimator based on the application of the maximum likelihood principle between close neighbor distances.

3.3.2. Compression in Transformer-Based Language Models

As discussed earlier, language models based on the Transformer architecture have become the standard choice for current language models for a broad range of NLP tasks. Despite their outstanding performance, little do we know about their underlying functioning and the type of information they retain. The input representation in large transformer models is processed through a series of identical modules, thereby generating a succession of vector spaces that, although dimensionally consistent, differ significantly in their geometric structure.

In particular, Valeriani et al. (2023) discovered that the ID of the hidden representations is significantly lower than the extrinsic dimension of the model's hidden layers. For instance, while the hidden layers typically have dimensions in the order of 2048 or more, the ID is often around 50. This substantial compression suggests that LLMs are highly efficient in distilling the essential features of the input data into a much smaller subspace. Moreover, further insights from Cheng et al. (2023) have revealed that different models with different extrinsic hidden

dimensions compress data into similar ranges of ID. They also found that the ID of the representations is 2 orders of magnitude smaller than their ambient dimension, in line with the results of other studies.

Just like humans, the generation of coherent text does not have to do with the prior memorization of all possible combinations of words or tokens, but rather with the extraction of a finite number of certain rules and vocabulary that allow the production of that potentially infinite number of utterances (Cheng et al., 2023). This is directly related to the previously described manifold hypothesis: language models exhibit a high dimensionality, but it can be reduced to a minimum number of parameters that allows to span the entire space. There have even been attempts to apply this same concept to languages themselves, offering an approach to compute the dimensionality of discourse (Doxas et al., 2010).

Thus, one of the cornerstones of the performance of language models is their ability to compress information. In recent years, much research has been devoted to this process of data compression. For instance, the geometric properties of the representations across the layers of a transformer have been studied, and it has been found that these representations evolve similarly whether the transformer has been pre-trained with a dataset of protein sequences, for tasks regarding protein language, or one of images, for tasks regarding image reconstruction (Valeriani et al., 2023). In the results they provide, some changes in compression are also observed across layers and phases, suggesting that the representations evolve through the layers in different phases. Hence, this compression is not uniform across layers: each layer contributes differently to the overall compression.

Also, the form and extent of data compression has been linked to adaptability to NLP tasks (Cheng et al., 2023). In particular, they show that higher compression is positively correlated with ease of adaptation to different tasks. Their results also demonstrate that the ID

profile changes depending on the domain of the text being processed. Moreover, they find that the removal of linguistic structure (e.g., by permuting words or constructing sequences with the same words but in random order) increases the intrinsic dimension and perplexity of data representations. This suggests that it is precisely the linguistic structure that allows such compression. In its absence, compression diminishes.

Further insights from Cheng et al. (2024) reveal that the first high-ID phase corresponds to the first full linguistic abstraction, and that an earlier occurrence of this phase predicts a better model's performance. They also find that this first peak is notably reduced when the model processes random text.

3.4. The Indo-European Family of Languages and Linguistic Typology

In this study we aim to explore geometric compression in multilingual LMs applied on languages primarily from the Indo-European family. The term *Indo-European* is normally used to refer to the Indo-European family of languages, which exhibit similarities that are often attributed to the fact that there was an earlier shared progenitor language called Proto-Indo-European (Anthony & Ringe, 2015). The discovery of phonetic similarities among words with identical meanings across various Indo-European languages suggests the presence of a Proto-Indo-European antecedent (Mallory & Adams, 2006).

The 21 languages of the Europarl dataset (Koehn, 2005), which are the ones analyzed here, are divided into language families and branches as follows:

- **Indo-European family:**
    - **Romance:** French, Italian, Spanish, Portuguese, Romanian.
    - **Germanic:** English, Dutch, German, Danish, Swedish.
    - **Slavic:** Bulgarian, Czech, Polish, Slovak, Slovene.

    - **Baltic:** Latvian, Lithuanian.
    - **Hellenic:** Greek.
- **Finno-Ugric family:** Finnish, Hungarian, Estonian.

Almost all languages belong to the Indo-European family of languages, but they fall into 5 different branches. On the other hand, Finnish, Hungarian and Estonian are the only languages that do not have an Indo-European origin, but instead belong to the Finno-Ugric family of languages.

Languages are grouped into branches and sub-branches by their history, studying their origins, their common antecedents, etc. In short, this classification results in the genealogical tree of languages. However, languages belonging to the same group not only are related historically, but this very fact means that they are languages with many linguistic properties in common. Indo-European languages exhibit shared elements in their basic lexicon and grammatical affixes, where the forms of these elements across languages can be systematically explained through specific phonetic rules.

Linguistic typology is concerned with analyzing, comparing and classifying languages according to their common characteristics of structure and form, regardless of their history. Fruitful classifications based on a structural characteristic are those that result in groups and subgroups of languages that share other characteristics as well (Trask, 2007). Probably the best-known classification is that of word order. It was proposed by Joseph Greenberg in 1963, and is based on the order of the basic syntactic constituents (Velupillai, 2012). It mainly deals with the relative order of subject, object, and verb.

For instance, languages structured with a Subject-Object-Verb (SOV) order typically exhibit certain syntactic characteristics: modifiers tend to appear before their corresponding head nouns, auxiliary verbs follow the main verbs, postpositions are used instead of

prepositions, and nouns often have an extensive case system. Conversely, languages with a Verb-Subject-Object (VSO) order generally display the opposite features: modifiers come after nouns, auxiliary verbs precede the main verbs, prepositions are used instead of postpositions, and nouns lack a case system (Trask, 2007).

Consequently, the Indo-European languages, besides having a common ancestor, share linguistic features such as word order. Almost all Indo-European languages exhibit a SVO pattern. In terms of word order, the 21 analyzed languages are classified as follows (Dryer, 2013):

- **SVO:** French, Italian, Spanish, Portuguese, Romanian, English, Danish, Swedish, Bulgarian, Czech, Polish, Slovak, Slovene, Latvian, Lithuanian, Finnish, Estonian.
- **No dominant order:** Dutch, German, Greek, Hungarian.

It can be seen how the only Indo-European languages that do not follow the SVO scheme are Dutch, German and Greek, and this is not even because they have a different word order, but because they have no dominant word order at all. On the other hand, among the three languages belonging to the Finno-Ugric family, Hungarian does not have any dominant order either, but Finnish and Estonian also follow an SVO pattern.

For the purpose of this research, and without going into too much detail, it makes sense to study the compression of language models in various languages by considering the families and subfamilies of these languages, since these groups bring together languages with linguistic characteristics in common.

4. Methodology

Our objective is to explore geometric compression of linguistic data across languages in multilingual language models. To do so, we use MLE to quantify the intrinsic dimension of the hidden representations at each layer of three different models on multiple datasets in various languages.

4.1. Models

The selected models and versions are the multilingual models BLOOM 3B (Scao et al., 2022) and XLM-R_{XL} 3.5B (Goyal et al., 2021), and Mistral 7B (Jiang et al., 2023). The differences they present and which are detailed below make this a broader experiment so that it is not limited to a very specific type of model.

As said before, BLOOM is an open-access LLM released in 2022. It is trained to output text based on a prompt; therefore, it is a *causal language model*. "BLOOM is a 176 billion parameter language model trained on 46 natural languages and 13 programming languages that was developed and released by a collaboration of hundreds of researchers" (Scao et al., 2022). This is indeed the essential feature of BLOOM, the fact that it is a multilingual model at its core, already from the pre-training phase, making use of a carefully curated dataset in a wide variety of languages. Data curation was taken as a fundamental part of the process and therefore a multitude of researchers with linguistic knowledge of these languages collaborated in order to avoid making mistakes previously made in automated data curation processes. Here we use the variant with 3B parameters, not the full 176B one.

XLM-RoBERTa, or simply XLM-R, is a Transformer-based masked language model pre-trained on 100 languages. In particular, we use the XL variant. Unlike BLOOM, XLM-R is not an autoregressive model, as it was pre-trained with the MLM objective. This introduces

a variation, as this allows the model to learn representations in a bidirectional manner. Also, masked language models have been shown to be effective at cross-lingual transfer (Conneau et al., 2019). Regarding the training dataset, they built it using several dumps of CommonCrawl in 100 languages. The variant we use here, XLM-R_{XL} 3.5B, is pre-trained on the same dataset, and has 3.5 billion parameters (Goyal et al., 2021).

Finally, Mistral is not an open-source model, so we do not have all the details about how it was trained, namely the training data or the optimization process. It certainly works for more languages than only English, but there is a huge difference in performance in English with respect to other languages. Technically, it is not a multilingual model, or at least it is not specifically designed for a multilingual setting. It is a fairly well-balanced model regarding performance and efficiency. In fact, when it was released at the end of 2023, this model with 7 billion parameters outperformed "the best open 13B model (Llama 2) across all evaluated benchmarks, and the best released 34B model (Llama 1) in reasoning, mathematics, and code generation" (Jiang et al., 2023).

4.2. Datasets

Previous research has shown that LLMs demonstrate distinct compression patterns when processing texts from diverse domains, such as Wikipedia articles, tweets, or code snippets (Cheng et al., 2023). As we are trying to see if LLMs also exhibits distinct ID profiles when processing texts from various languages, it is important that the different datasets processed by the model are semantically aligned, i.e. their domain is the same and, ideally, their content is also identical.

For this purpose, the Europarl corpus is a convenient dataset (Koehn, 2005). This compilation contains the records of proceedings from the European Parliament, beginning in 1996. Originally, the corpus was conceived as training data for statistical machine translation.

For this reason, the corpus comprises subcorpora in 21 languages of European Union member countries. Furthermore, these subcorpora are structured into pairs of languages. They also performed document alignment, extracting and mapping parallel chunks of text. Finally, for each of the languages the text was split into sentences.

Because of these reasons, the Europarl corpus is perfectly suited for our purpose. First, the domain in all subsets of the corpus is the same: European parliamentary debates. Moreover, it also meets the second requirement, as the subsets, which correspond to each language pair, are semantically aligned.

4.3. Datasets Preprocessing

The selected and extracted subsets from the Europarl corpus were subjected to a preprocessing procedure. The purpose of this preprocessing is to structure the datasets in a convenient way to be processed by the model and to be able to properly estimate the ID at each layer. First, all language pairs were extracted from the corpus, with the objective of creating one dataset per pair, as we need the dataset to be semantically aligned. From this point on and for each language pair, the preprocessing phase consisted of the following main steps:

1) **Randomly extract 20,000 sentences from the dataset.**

Some languages in the dataset have more than 2 million sentences. Cheng et al. (2023) also studied the convergence of ID estimators, and their results show that all ID estimates start to converge at around 10,000 sequences. For computational efficiency, we chose to randomly extract 20,000 sentences. This way we are also above the convergence threshold of the ID estimates. In addition, as some languages tend to have longer words (such as German or Finnish), this translates into fewer words per sentence. Thus, by doing this random extraction we make sure there is a certain margin while ensuring the validity of the results. Moreover, by

performing the extraction of sentences from the original language-pair dataset, we ensure the dataset is semantically aligned within the language pair.

2) Preprocess sentences.

First, we add a period at the end of the sentences that lack one[8]. Once this is done, we concatenate all the sentences into a single string. Finally, we split the string by words.

3) Process sentences and convert them into 20-word lines.

Once the sentences are concatenated and divided by words, the objective is to transform them into lines of equal length. These lines with the same number of words are the input sequences that will be processed by the model. Thus, from the extracted sentences, we created lines of 20 words each. Cheng et al. (2023) found that short sequences give inconsistent results when estimating ID, often resulting in lower dimension, but very long sentences have no impact whatsoever on ID estimation.

Finally, it is important to keep spaces at correct places. For instance, the model encodes differently "the" and "[space]the": "the" may be a fragment of a word (e.g. "ma-the-matics"), while "[space]the" is either the word "the" or the first fragment of a word (e.g. "the-sis"). Bearing this in mind, we make sure that the lines never end with a space, but rather with an actual token. If the line is already 20 words long but the sentence is not over yet, we start the next line where we left off in the sentence but adding a space at the beginning. Naturally, if the end of the sentence coincides with the end of the line, the next line starts normally, straight with the first token of the sentence.

[8] In the original dataset, some sentences lack a period. This raises a problem for our purpose because the period indicates the end of a sentence, and its absence interferes with the way the model processes sentences (a sentence without a period will merge with the beginning of the next sentence). Therefore, we first add the period in case any of the randomly extracted sentences do not have it.

By way of illustration, the reader can refer to Appendix A, where 5 sample pairs of sentences extracted from the original *en-es* (English-Spanish) Europarl subcorpus are displayed, just as they appear originally. Below them are the same sentences in their final form after the described preprocessing, exactly as they appear in our final datasets for the *en-es* subcorpus (one for English and one for Spanish). Moreover, the algorithm used for preprocessing the datasets is provided as a single function in pseudo-code in Appendix B. The whole Python file, as well as the ones used for extracting the hidden representations and for computing the intrinsic dimension are available at github.com/ederruiz98/compression-multilingual-LMs.

5. Results and Discussion

We find that the intrinsic dimension of the representations of linguistic data is notably smaller than the ambient dimension. This is consistent with the results of previous research, such as with pre-trained models on protein sequences and images (Valeriani et al., 2023), and also on linguistic data (Cheng et al., 2023). In particular, the ambient dimension of representations in the models BLOOM 3B and XLM-R_{XL} 3.5B is $D = 2560$ for both of them, while for Mistral 7B it is $D = 4096$. In contrast, our ID estimate of representations at any given layer does not surpass 30 at the maximum peak in BLOOM, 40 in XLM-R, and 50 in Mistral.

Our main results show that Transformer-based multilingual LLMs exhibit distinct compression patterns when processing texts from different languages. Although the ID profile does not seem to change depending on the language, the ID amplitude does. In particular, we find that the ID amplitude is similar in languages that belong to the same family or branch, suggesting that the linguistic properties learned by the model and shared between related languages affect the dimensionality of the representations. Additionally, we find that the

performance of XLM-R on a language and the ID amplitude at the first high-dimensionality phase are correlated.

For the analysis of the results, it is necessary to take into account the languages present in the datasets with which the models were pre-trained on. In this case, among the 21 languages of the Europarl dataset, BLOOM includes English, French, Portuguese, and Spanish, XLM-R all 21, and we do not know with certainty about Mistral. For this reason, the focus is on XLM-R and Mistral in cases where the analysis involves several languages from different families.

A. ID profile is similar across languages.

The way ID changes across different layers illustrates the sequential process by which Transformers compress and decompress linguistic representations. Although the ID profiles differ from one model to another, we can see the same three phases found in previous studies with models pre-trained on protein sequences and images (Valeriani et al., 2023) and with models pre-trained on linguistic data (Cheng et al., 2023). This can be seen in Figure 4 and in the line plots with languages divided by families (Appendix C). In particular, we can distinguish a first peak that reaches a local maximum, a second plateau or relative minimum phase, and one last (slight) ascent phase that usually ends with a reconstruction phase with a short descent of the ID.

The ID profile of Mistral representations show a more drastic increase than the other models in the first peak. Moreover, it reaches the highest maximum ID value among the three. These facts make sense because, compared to the other analyzed models, the ambient dimension of Mistral representations is almost twice as large.

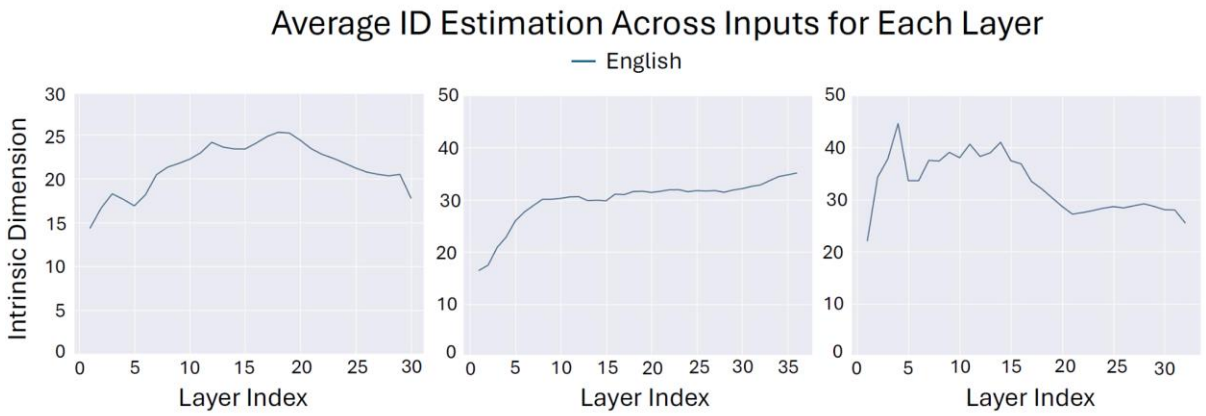

Figure 4: Average ID estimation for each layer on the English dataset with BLOOM 3B (left), XLM-R_{XL} 3.5B (middle), and Mistral 7B (right).

B. ID amplitude is similar among languages from the same family.

We observe that languages from the same family or branch exhibit a similar range of ID values (see Appendix C). This suggests that the model's learned linguistic properties, common among related languages, have an impact on the dimensionality of the representations.

In Figures 5 and 6, we observe that the Slavic language branch and the Finno-Ugric language family are particularly well distinguished. One aspect to note about these is that the ID profile is flatter than with other languages, and this holds for both XLM-R and Mistral. The model representations for both language groups reach a lower dimensionality. In addition, the ID profile of the Finno-Ugric family languages is slightly flatter than the Slavic languages one.

In contrast, although the Germanic and Romance language branches also form two distinct groups, the range of ID values is greater than for the previous cases. This is especially observed with Mistral, where the Germanic languages are more dispersed, mostly explained by the fact that the ID in English reaches much higher values. As Mistral has been (presumably) mostly trained on English, this result suggests that the ID on one language is impacted by the LM's performance on that language (as we will show later).

Moreover, results from Valeriani et al. (2023) revealed that the semantic information is best expressed at the end of the first peak, and Cheng et al. (2024) found that the end of the first peak is where the language model reaches a full representation of the semantics and the syntax. We can see more clearly the relation between the language families and the ID amplitude at that point, which for XLM-R is around layer 10 and for Mistral around layer 14.

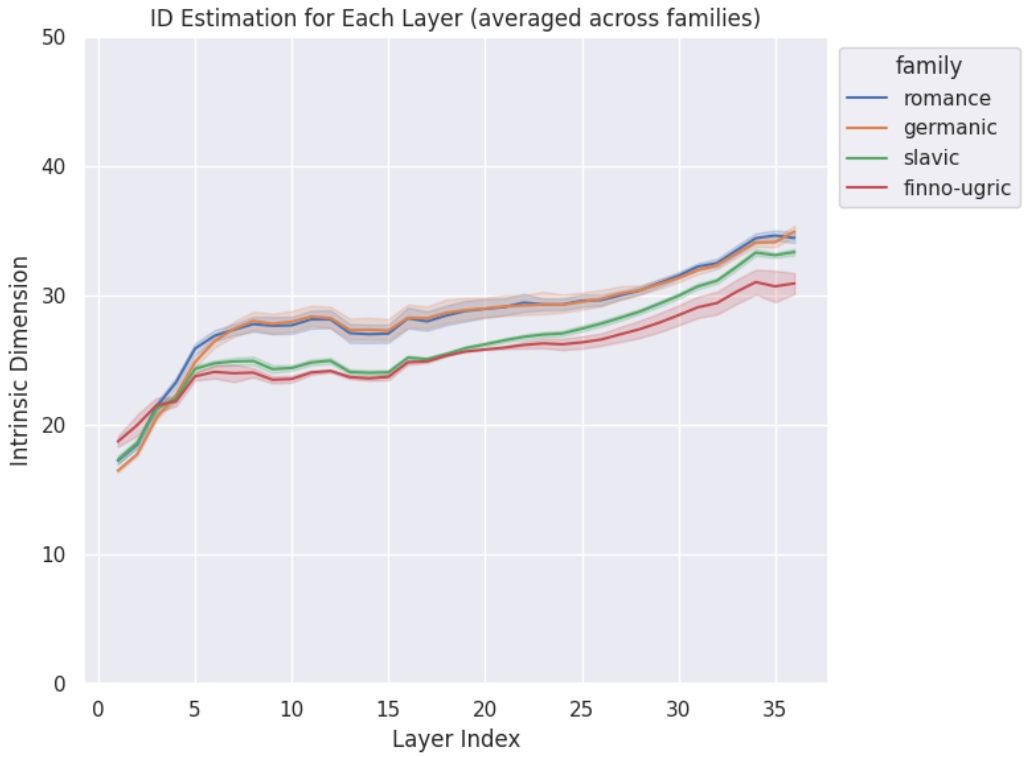

Figure 5: Average ID estimation for each layer on each family of languages (averaged across languages within each family) with XLM-R_{XL} 3.5B.

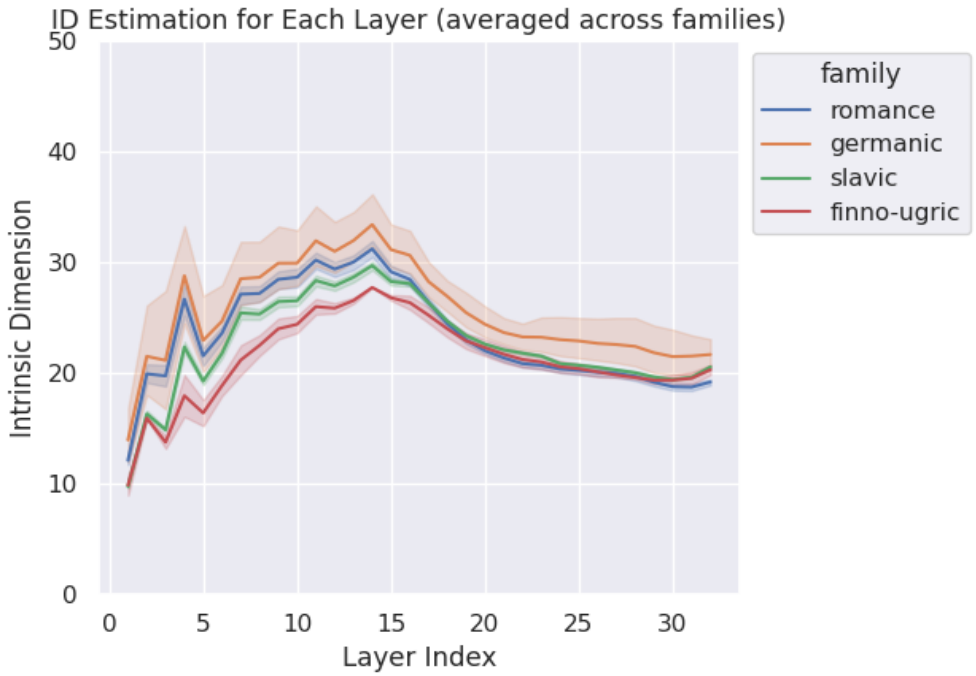

Figure 6: Average ID estimation for each layer on each family of languages (averaged across languages within each family) with Mistral 7B.

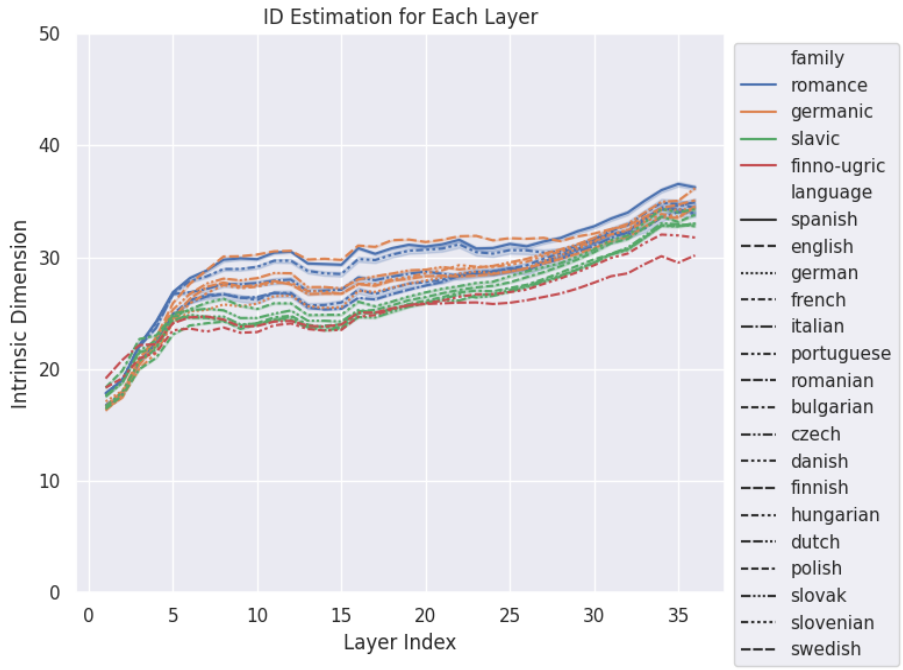

Figure 7: Average ID estimation for each layer on all languages (grouped by family) with XLM-R_{XL} 3.5B.

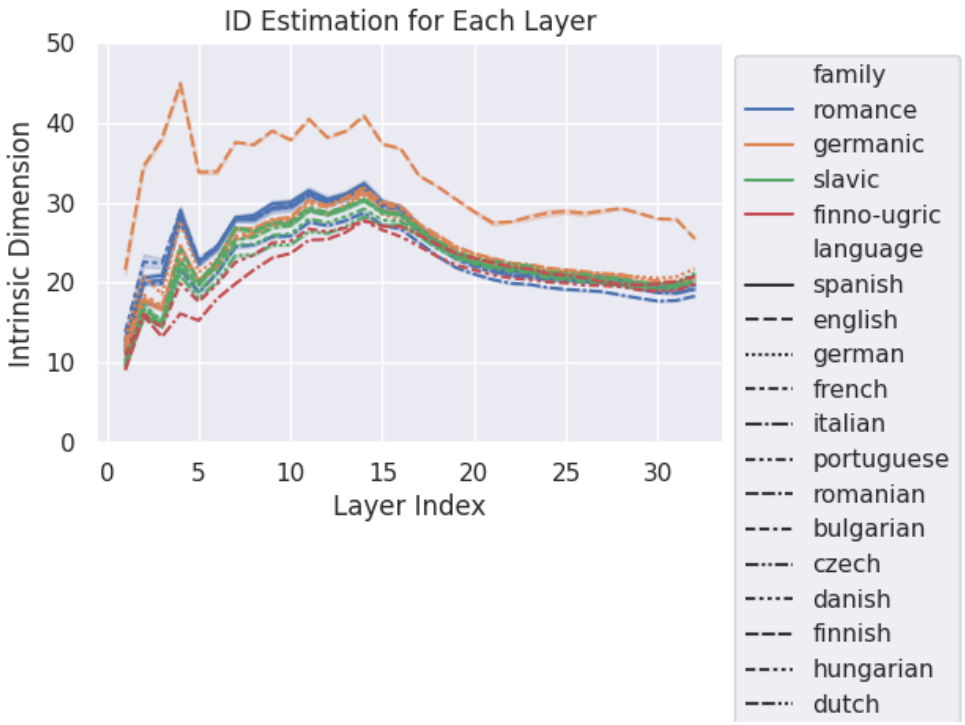

Figure 8: Average ID estimation for each layer on all languages (grouped by family) with Mistral 7B.

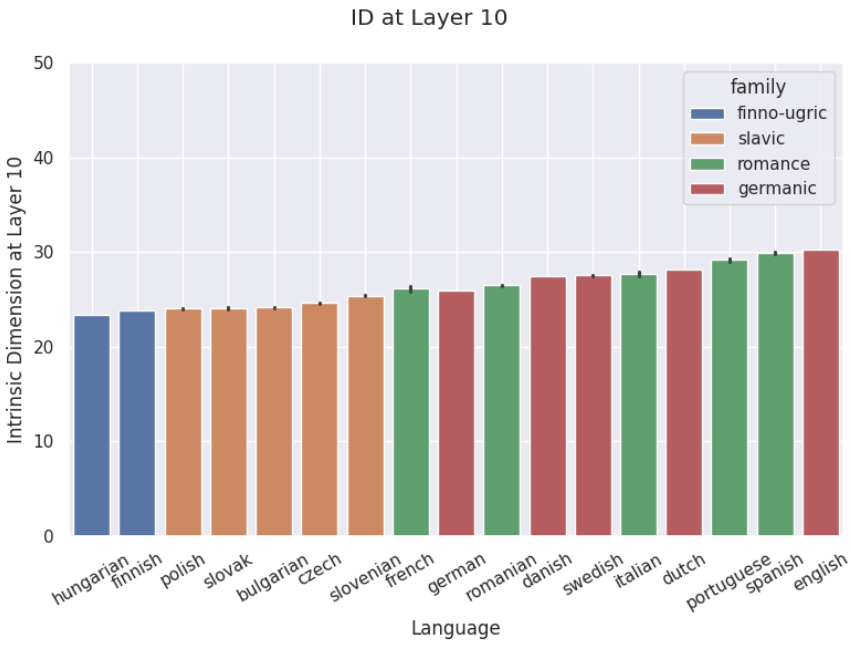

Figure 9: Average ID estimation at the first high-dimensionality phase (layer 10) on all languages (grouped by family) with XLM-R_{XL} 3.5B.

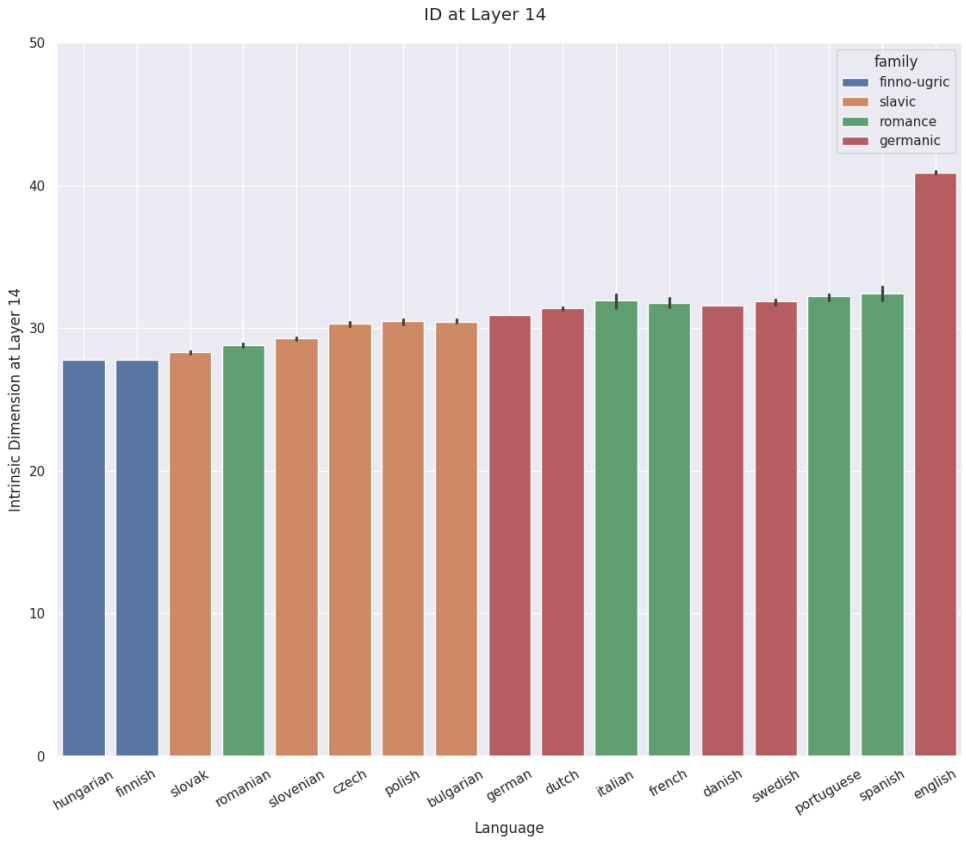

Figure 10: Average ID estimation at the first high-dimensionality phase (layer 14) on all languages (grouped by family) with Mistral 7B.

C. ID profile is flat on languages not present in the training dataset.

Besides these results, it is interesting to see what happens when the model processes text from a language that was not present in its training dataset. Since we do not know the Mistral training dataset and the XLM-R training dataset includes all the languages we analyzed, we can examine the situation when BLOOM processes text from languages in which it was not trained. We get a similar ID profile for the 4 languages that were in its training dataset (English, French, Portuguese, and Spanish). However, the ID profile is much flatter and the first peak is severely reduced in those languages the model was not trained on (see Appendix D). This is coherent with Cheng et al. (2024) results, where they show that the first peak (corresponding to the first high dimensionality phase) is notably reduced when the model processes random

text. In these terms, text in an unknown language is similarly processed by the model as random text (permutations of words, altered word order, etc.).

D. The ID profile is not or very little impacted by the syntax nor the script.

As it can be seen in Figures 5 to 8, the ID profile remains almost the same for all languages: they present the same phases with similar lengths. This occurs both with languages of different but related branches (such as Indo-European languages) and with distant languages (such as Finnish and Spanish, for example). Moreover, the ID profile for Bulgarian is similar to that of other Slavic languages, although Bulgarian is the only language that uses Cyrillic script. The rest of the analyzed languages use Latin script.

For these reasons, the syntax and the script of the linguistic data do not have a significant impact on the ID profile, suggesting that these do not play a significant role in the dimensionality of the hidden representations computed by the model. In contrast, as the datasets are semantically aligned, we can conclude that the semantic content is one of the factors that determine the ID profile, which is consistent with the results of Cheng et al. (2023), who found different ID profiles for different datasets under the same model.

E. Dataset size and ID are not correlated.

Looking at Figures 9 and 10, it appears that languages associated to the lowest IDs are the ones that could be described as low resource languages (e.g. Hungarian, Finnish, etc.). It is fair to expect that the size of a language in the training dataset is directly correlated with the ID amplitude at the end of the first peak. It is intuitive to think that if a model has been more heavily trained in one language than in another, the number of minimum parameters needed to span the vector space of representations in that language will be greater, which could also predict performance in that language. In fact, this is usually the case in rather extreme cases.

For instance, it is likely that in the Mistral training dataset there was much more English than any other language, and this is precisely reflected in Figure 8, where it can be seen how the ID reaches much higher values, attaining a higher overall dimensionality. It also happens with XLM-R, where the English training dataset is by far the largest (55,608M tokens) and it is the language that gets the highest ID at the first peak.

That being said, we do not have information about Mistral training data, but we do have the XLM-R training dataset sizes divided by languages. We obtain that, for XLM-R, p-value is 0.88, quite above the standard threshold of 0.05, concluding that there is no statistically significant correlation between dataset size and ID amplitude at layer 10 (cf. Figure 11). However, transferability across languages could be affecting the results. XLM has been trained on 100 different languages. Hence, it is possible that, beyond the dataset size of individual languages, interactions between languages have an impact on the ID.

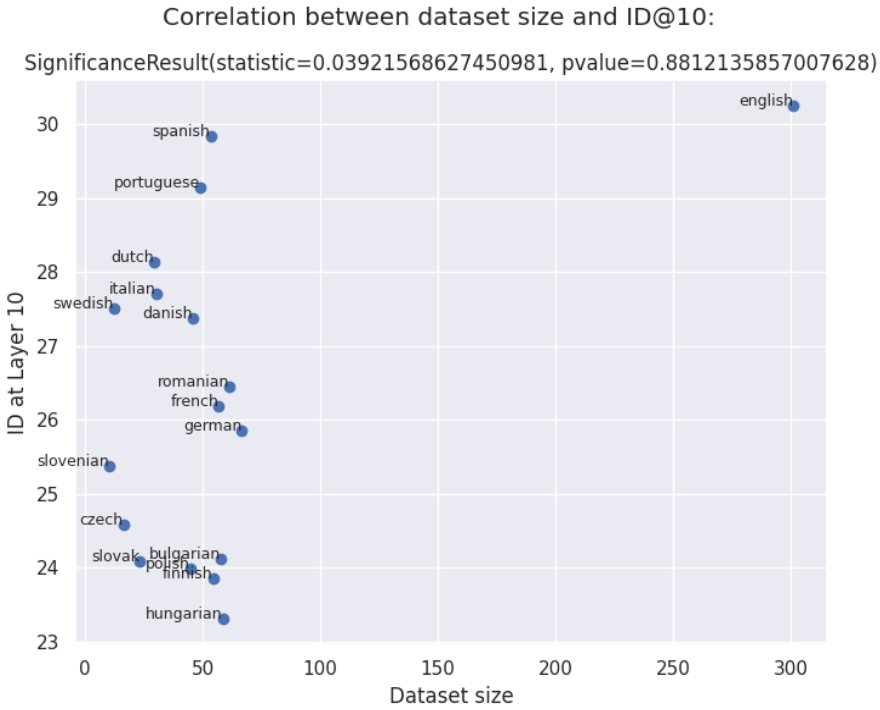

Figure 11: Correlation between dataset size (in GiB) and ID at layer 10 with XLM-R_{XL} 3.5B. We find no statistically significant correlation between the two.

F. Performance and ID of the first peak are correlated.

Perplexity (PPL) is a measure that indicates how good a language model is at next-token prediction. A low PPL means that the model is good at next-token-prediction on the dataset. We do find a statistically significant negative correlation between PPL and ID at the end of the first peak in XLM-R_{XL} (cf. Figure 12)[9], meaning that when the model performs well on a language (indicated by low PPL), the ID the first peak will be higher. On the contrary, if the PPL is high on a language, meaning that the model does not perform well on that language, the ID at the first peak will be lower and the whole ID profile will be flatter[10].

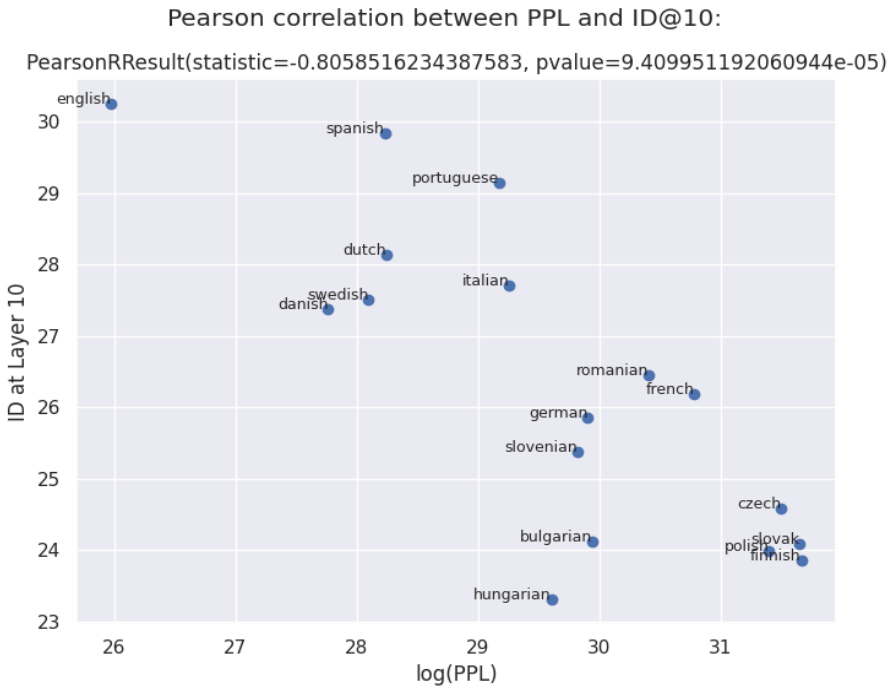

Figure 12: Pearson correlation between PPL and ID at layer 10 with XLM-R_{XL} 3.5B. We find PPL and ID at layer 10 to be correlated.

[9] XLM-R is a masked language model; therefore, it is not possible to use the standard measure of perplexity. Instead, we used a variant called *pseudo-perplexity* (Salazar et al., 2020)

[10] We observe the opposite trend with Mistral: we get that PPL and ID at layer 14 are correlated. This is probably due to the fact that Mistral has not been trained on multiple languages, making cross-languages comparisons not trivial. See results in Appendix E.

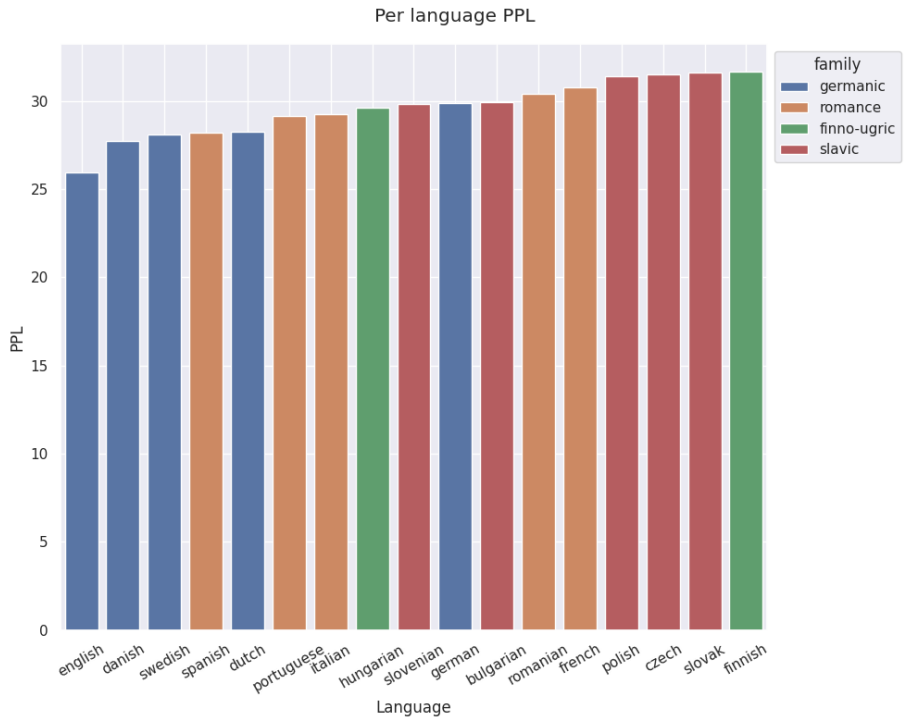

Figure 13: Perplexity computed on our dataset per each language with XLM-R_{XL} 3.5B.

G. Multilingual LMs demonstrate strong cross-lingual transfer abilities.

At a closer look, a number of insights into XLM-R transfer abilities can be drawn from Figures 12 and 13. Interestingly, Swedish is one of the languages in which the model performs best, according to the PPL measures on our dataset. However, the Swedish training dataset is only 77.8M tokens, being one of the languages with the lowest representation in the XLM-R training dataset[11]. Such behavior may probably be attributable to the model's transfer abilities. In fact, XLM-R is a multilingual masked language model (MLM), and models trained with the MLM objective have been shown to be effective at cross-lingual transfer (Conneau et al., 2019).

[11] Note that this size is very small. Other typically low-resource languages have much more representation in the training dataset. For instance, Georgian, Esperanto, and Icelandic datasets are, respectively, 469M, 157M and 505M tokens. Swedish is indeed minimally represented in XLM-R training dataset.

This assumption is reinforced in this case because even though Swedish is significantly underrepresented in the training dataset, other Germanic languages do comprise a large part of the tokens of the training dataset. Namely, English, German, and Danish datasets are, respectively, 55608M, 10297M, and 7823M tokens. In this scenario, the model could infer certain rules and linguistic properties from other Germanic languages, learning features shared with Swedish from closely related languages. This fact would explain why Swedish perplexity is so low even though it comprises a minimal portion of the training dataset.

Moreover, according to our PPL measures, Finnish is the worst performing language. However, XLM-R training dataset has 6730M Finnish tokens, a rather high number compared to the rest of the languages in the training dataset, and especially to languages which the model performs better on. Finnish belongs to the Finno-Ugric language family, which also includes Estonian and a few minority languages from the Baltic Sea area, so it is a rather isolated branch of languages. In the same scenario as before, this could explain the model's poor performance in Finnish.

6. Limitations and Further Research

This study provides some valuable insights into the impact that languages have on the geometric compression of representations in transformer-based multilingual language models. Moreover, some of our results complement those obtained in previous studies on the topic (Valeriani et al., 2023; Cheng et al., 2023; Cheng et al., 2024), since we aim to build upon these novel findings and provide a more fine-grained analysis based on them. However, this study also presents certain limitations that should be taken into account in order to correctly contextualize the results obtained and to define future lines of research. The following primary limitations can be highlighted:

A. Model diversity.

Our study is limited to three specific models: BLOOM 3B, XLM-R_{XL} 3.5B, and Mistral 7B. Each of these models has distinct characteristics and training datasets, which may influence their behavior and performance. BLOOM and XLM-R are explicitly designed for multilingual contexts, whereas Mistral is not specifically optimized for multiple languages. In addition, the pre-training objective is also different, as BLOOM is a causal LM and XLM-R is a masked LM. This lack of uniformity in model design and training can affect the comparability of results across models. However, this was also our intention, since introducing this diversity meant that we avoided making the scope of the research too narrowly focused, which could prevent the results from being generalizable.

Future research should include a broader range of multilingual language models, including those specifically designed for multilingual settings and others that are not. This would provide a more comprehensive understanding of how different model architectures and training objectives affect geometric compression across languages. In addition, it would be ideal if situations like Mistral's, where the training dataset is unknown, could be avoided in the future. However, this is not always possible, as the best performing LMs are usually not open-source.

B. Limited range of languages.

Although the Europarl corpus is well-suited for our study due to its aligned multilingual data, it only covers European languages. For this reason, the majority of the languages analyzed belong to the Indo-European family of languages, with the exception of Finnish, Hungarian and Estonian, which belong to the Finno-Ugric family.

Expanding the dataset to include non-European languages, especially distant languages, would test the robustness and generalizability of the findings. This could also include languages with non-Latin scripts, since all the analyzed languages use Latin script, with the exception of Hungarian, which uses Cyrillic script. This is also a limitation, as we find that Cyrillic script in Hungarian does not introduce any ID variability, but it would be interesting to check if this holds true for other scripts and for a wider range of languages.

C. Domain diversity.

The Europarl corpus contains only European parliamentary proceedings. This is not necessarily a drawback, since we wanted a semantically aligned dataset on one specific discourse domain. However, replicating the experiment on other specific domains would provide robustness to the generalizability of the results.

D. Fine-grained linguistic analysis.

Finally, although we have tried to keep the interference of uncontrolled parameters to a minimum, it would be valuable to carry out very specific research that further reduces uncontrolled parameters. For instance, by comparing closely related languages and making datasets that introduce little variability in terms of linguistic properties (numerous probes to test specific linguistic features have already been proposed), it would be possible to study in more detail which linguistic features are actually learned by the model and have an impact on geometric compression.

7. Conclusion

In this study, we explored the geometric compression of linguistic data across multiple languages in multilingual language models. In particular, we did so by estimating the intrinsic dimension (ID) of hidden representations across the layers of three different models: BLOOM, XLM-RoBERTa, and Mistral. For this purpose, we needed a semantically aligned dataset, in order to avoid interference and study only the language variable. In order to obtain a dataset in the correct form for our analysis, we performed a specific preprocessing on the Europarl corpus, which was already a convenient corpus for our needs. This way, we were able to analyze the ID profiles for various languages, providing insights into how these models compress information and handle the same semantic content in different languages.

Our results are aligned with those obtained in previous studies. Specifically, we get that the ID of the representations computed by the models are 2 orders of magnitude inferior to the extrinsic dimension, which reinforces the idea that language models compress large amounts of linguistic data into their basic parameters in much smaller vector spaces. However, we extend these insights into the field of multilingual LLMs, revealing distinct compression patterns when these models process texts in different languages.

One key finding is that while the ID profile remains largely consistent across languages, the ID amplitude varies, particularly among distant languages. This suggests that (1) the syntax and the script do not have an impact on the ID profile, but semantic content does, and (2) the model infers certain linguistic properties that results in similar ID amplitude among closely related languages. For instance, languages from the Slavic and Finno-Ugric families exhibit flatter ID profiles with lower dimensionality, whereas Germanic and Romance languages display greater ID values.

Another significant finding is the correlation between model performance, as measured by perplexity (PPL), and ID amplitude at the first peak. For XLM-R, higher ID at the first high-ID phase correlates with better performance on next-token prediction, indicating that the model performs better on languages with higher. This relationship, however, is not present in Mistral, likely due to differences in training data and model architecture (causal vs. masked LMs). We also find strong cross-lingual transfer abilities in XLM-R. In this case, we find that languages poorly represented in the training dataset perform well, as long as there are similar languages (of the same branch or family) that are also present in the training dataset.

In summary, our study sheds light on the compression process of multilingual LLMs, building upon previous work and offering some insights into how they handle linguistic data. Our findings emphasize the relevance of linguistic properties and features in shaping the geometric compression patterns of Transformer-based multilingual LMs. Finally, future research may address the limitations of this study and help with the potential generalizability of these results.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Corentin Kervadec. Your guidance and support were instrumental in the successful completion of this thesis. I am also grateful to the COLT investigators, as this study builds on their foundational work. Special thanks go to Marco Baroni for providing essential code, and to Iuri Macocco and Emily Cheng for their invaluable tools and ideas. Finally, I extend my heartfelt thanks to my professor, Thomas Brochhagen, for the insightful lessons that have significantly influenced my research. This thesis is a testament to the knowledge and skills I have gained from you all.

References

Achiam, O. J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., … Zoph, B. (2023). *GPT-4 Technical Report*. https://api.semanticscholar.org/CorpusID:257532815

Al-Rfou, R., Choe, D., Constant, N., Guo, M., & Jones, L. (2018). Character-Level Language Modeling with Deeper Self-Attention. *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:52004855

Ansuini, A., Laio, A., Macke, J. H., & Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:170079070

Anthony, D. W., & Ringe, D. (2015). The Indo-European Homeland from Linguistic and Archaeological Perspectives. In *Annual Review of Linguistics* (Vol. 1, Issue Volume 1, 2015, pp. 199–219). https://doi.org/10.1146/annurev-linguist-030514-124812

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. http://arxiv.org/abs/1409.0473

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*(Feb), 1137–1155.

BigScience, *BigScience Language Open-science Open-access Multilingual (BLOOM) Language Model*. International, May 2021-May 2022

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv*, *abs/2005.14165*. https://api.semanticscholar.org/CorpusID:218971783

Campadelli, P., Casiraghi, E., Ceruti, C., & Rozza, A. (2015). Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. *Mathematical Problems in Engineering*, *2015*, 1–21.

Carter, K. M., Raich, R., & Hero, A. O. (2010). On Local Intrinsic Dimension Estimation and Its Applications. *IEEE Transactions on Signal Processing*, *58*, 650–663.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–394. https://doi.org/10.1006/csla.1999.0128

Cheng, E., Doimo, D., Kervadec, C., Macocco, I., Yu, J., Laio, A., & Baroni, M. (2024). Emergence of a High-Dimensional Abstraction Phase in Language Transformers. *ArXiv*, *abs/2405.15471*. https://api.semanticscholar.org/CorpusID:270045386

Cheng, E., Kervadec, C., & Baroni, M. (2023). Bridging Information-Theoretic and Geometric Compression in Language Models. *ArXiv*, *abs/2310.13620*. https://api.semanticscholar.org/CorpusID:264406051

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:207880568

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). Language Modeling with Gated Convolutional Networks. *Proceedings of the 34th International Conference on Machine Learning*, 933–941. https://proceedings.mlr.press/v70/dauphin17a.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:52967399

Doxas, I., Dennis, S., & Oliver, W. L. (2010). The dimensionality of discourse. *Proceedings of the National Academy of Sciences*, *107*(11), 4866–4871. https://doi.org/10.1073/pnas.0908315107

Dryer, M. S. (2013). Order of Subject, Object and Verb (v2020.3). In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Zenodo. https://doi.org/10.5281/zenodo.7385533

Facco, E., d'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, *7*. https://api.semanticscholar.org/CorpusID:3991422

Fefferman, C., Mitter, S. K., & Narayanan, H. (2013). Testing the Manifold Hypothesis. *arXiv: Statistics Theory*. https://api.semanticscholar.org/CorpusID:50258911

Firth, J. R. (1957). Modes of Meaning. In *Papers in Linguistics, 1934-1951*. Oxford University Press.

Gorban, A. N., & Tyukin, I. Y. (2018). Blessing of dimensionality: Mathematical foundations of the statistical physics of data. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *376*(2118). https://doi.org/10.1098/rsta.2017.0237

Goyal, N., Du, J., Ott, M., Anantharaman, G., & Conneau, A. (2021). Larger-Scale Transformers for Multilingual Masked Language Modeling. *ArXiv, abs/2105.00572*. https://api.semanticscholar.org/CorpusID:233481097

Hestness, J., Narang, S., Ardalani, N., Diamos, G. F., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. *ArXiv, abs/1712.00409*. https://api.semanticscholar.org/CorpusID:2222076

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de L., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A.,

Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *ArXiv*, *abs/2310.06825*. https://api.semanticscholar.org/CorpusID:263830494

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the Limits of Language Modeling. *CoRR*, *abs/1602.02410*. http://arxiv.org/abs/1602.02410

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *ArXiv*, *abs/2001.08361*. https://api.semanticscholar.org/CorpusID:210861095

Kneser, R., & Ney, H. (1995). Improved backing-off for M-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing*, *1*, 181–184 vol.1.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of Machine Translation Summit X: Papers*, 79–86. https://aclanthology.org/2005.mtsummit-papers.11

Levina, E., & Bickel, P. J. (2004). Maximum Likelihood Estimation of Intrinsic Dimension. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:14865278

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:204960716

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, *abs/1907.11692*. https://api.semanticscholar.org/CorpusID:198953378

Luong, T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d15-1166

Mallory, J. P., & Adams, D. Q. (2006). *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. OUP Oxford.

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in Translation: Contextualized Word Vectors. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:9447219

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:5959482

Mikolov, T., Karafiát, M., Burget, L., Honza \vCernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech*. https://api.semanticscholar.org/CorpusID:17048224

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:16447573

Narayanan, H., & Mitter, S. K. (2010). Sample Complexity of Testing the Manifold Hypothesis. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:7645673

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L. E., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. J. (2022). Training language models to follow instructions with human feedback. *ArXiv*, *abs/2203.02155*. https://api.semanticscholar.org/CorpusID:246426909

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *ArXiv*, *abs/1802.05365*. https://api.semanticscholar.org/CorpusID:3626819

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. https://api.semanticscholar.org/CorpusID:49313245

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://api.semanticscholar.org/CorpusID:160025533

Raffel, C., Shazeer, N. M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, *21*, 140:1-140:67.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked Language Model Scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.240

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ili’c, S., Hesslow, D., Castagn’e, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., Moral, A. V. del, … Wolf, T. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *ArXiv*, *abs/2211.05100*. https://api.semanticscholar.org/CorpusID:253420279

Trask, R. L. (2007). *Language and Linguistics: The Key Concepts* (2nd ed.). Routledge. https://doi.org/10.4324/9780203961131

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., & Cazzaniga, A. (2023). The geometry of hidden representations of large transformer models. *ArXiv*, *abs/2302.00294*. https://api.semanticscholar.org/CorpusID:256459698

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Velupillai, V. (2012). *An Introduction to Linguistic Typology*. John Benjamins Publishing Company.

Voita, E. (2020, September). *NLP Course For You*. https://lena-voita.github.io/nlp_course.html

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced Language Representation with Informative Entities. *ArXiv*, *abs/1905.07129*. https://api.semanticscholar.org/CorpusID:158046772

Appendix A

Datasets Preprocessing: Examples From the Original and the Final Datasets

5 elements from the original *en-es* Europarl dataset:

{ “en”: “My intention with these proposals to adjust agricultural prices is to send a clear message to the European farming community. I want to assure them that this House intends to defend their legitimate interests at a particularly difficult time when doubt is creeping in among farmers, particularly the younger ones.”, “es”: “Con las propuestas de adaptación de los precios agrícolas he querido enviar una señal fuerte al campesinado europeo para confirmarle que este Parlamento defiende sus intereses legítimos en una época especialmente difícil en la que las dudas asaltan a los profesionales, principalmente a los jóvenes.” }

{ “en”: “As this directive is implemented in Greece, if someone is treated in a psychiatric clinic even for a single day, the clinic is obliged to inform the police and the police the Ministry for Traffic, with the consequence that the licence is revoked immediately.”, “es”: “En aplicación de esta directiva, en Grecia sucede que, cuando alguien está en tratamiento, aunque sea por solo un día, en una clínica psiquiátrica, ésta tiene la obligación de informar a la policía, y la policía, a su vez, al Ministerio de Tráfico, lo cual lleva a la privación inmediata del permiso.” }

{ “en”: “And look what is happening: the number is rising all the time.”, “es”: “Y hete aquí: la cifra sigue aumentando.” }

{ “en”: “The final version of the draft agenda for the present part-session as drawn up by the Conference of Presidents at its meeting of 17 February pursuant to Rules 130 and 131 of the Rules of Procedure has been distributed.”, “es”: “Se ha distrubuido el proyecto definitivo del orden del día del presente período parcial de sesiones, de acuerdo con lo aprobado por la Conferencia de Presidentes, en su reunión del jueves 17 de febrero pasado, y de conformidad con los artículos 130 y 131 del Reglamento.” }

{ “en”: “The characteristic picture in Greek families with students finishing high-school is that in the evening, when the father or mother get home from work, the 17- or 18-year-old child sits there reciting the book to them from memory, because self-testing is not good enough.”, “es”: “La imagen característica de los estudiantes que finalizan el instituto en las familias griegas es por la noche, cuando el padre y la madre regresan del trabajo y el chico de 17 o 18 años se sienta y les recita la lección de memoria. Porque él mismo no debe controlarse.” }

Same 5 elements as they appear in our English *en-es* dataset after preprocessing:

My intention with these proposals to adjust agricultural prices is to send a clear message to the European farming community.

I want to assure them that this House intends to defend their legitimate interests at a particularly difficult time when

doubt is creeping in among farmers, particularly the younger ones. As this directive is implemented in Greece, if someone is

treated in a psychiatric clinic even for a single day, the clinic is obliged to inform the police and the

police the Ministry for Traffic, with the consequence that the licence is revoked immediately. And look what is happening: the

number is rising all the time. The final version of the draft agenda for the present part-session as drawn up

by the Conference of Presidents at its meeting of 17 February pursuant to Rules 130 and 131 of the Rules

of Procedure has been distributed. The characteristic picture in Greek families with students finishing high-school is that in the evening,

when the father or mother get home from work, the 17- or 18-year-old child sits there reciting the book to

them from memory, because self-testing is not good enough. […]

Same 5 elements as they appear in our Spanish *en-es* dataset after preprocessing:

Con las propuestas de adaptación de los precios agrícolas he querido enviar una señal fuerte al campesinado europeo para confirmarle

que este Parlamento defiende sus intereses legítimos en una época especialmente difícil en la que las dudas asaltan a los

profesionales, principalmente a los jóvenes. En aplicación de esta directiva, en Grecia sucede que, cuando alguien está en tratamiento, aunque

sea por solo un día, en una clínica psiquiátrica, ésta tiene la obligación de informar a la policía, y la

policía, a su vez, al Ministerio de Tráfico, lo cual lleva a la privación inmediata del permiso. Y hete aquí:

la cifra sigue aumentando. Se ha distrubuido el proyecto definitivo del orden del día del presente período parcial de sesiones,

de acuerdo con lo aprobado por la Conferencia de Presidentes, en su reunión del jueves 17 de febrero pasado, y

de conformidad con los artículos 130 y 131 del Reglamento. La imagen característica de los estudiantes que finalizan el instituto

en las familias griegas es por la noche, cuando el padre y la madre regresan del trabajo y el chico

de 17 o 18 años se sienta y les recita la lección de memoria. Porque él mismo no debe controlarse.

Appendix B

Datasets Preprocessing: Algorithm in Pseudo-Code

```
FUNCTION process_sentences(sentences)
    # Add a dot at the end of sentences if there isn't one already
    FOR EACH sentence IN sentences
    IF sentence DOES NOT END WITH punctuation THEN
            APPEND '.' TO sentence
    ENDIF
    ENDFOR

    # Concatenate all sentences into a single string
    concatenated_sentences = CONCATENATE sentences WITH " "

    # Split concatenated sentences into words
    words = SPLIT concatenated_sentences BY " "

    # Initialize variables
    lines = EMPTY LIST
    line = ""

    # Create lines with 20 words each
    FOR EACH word IN words
    IF COUNT OF words IN line < 20 THEN
            APPEND word TO line
    ELSE
            REMOVE trailing space FROM line
            APPEND line TO lines
            line = word + " "
    ENDIF
    ENDFOR

    # Append the last line if it is not empty
    IF line IS NOT EMPTY THEN
    APPEND line TO lines
    ENDIF

    RETURN lines

END FUNCTION
```

Appendix C

Figures: ID at Each Layer on All Languages (Grouped by Family) With XLM-R and Mistral

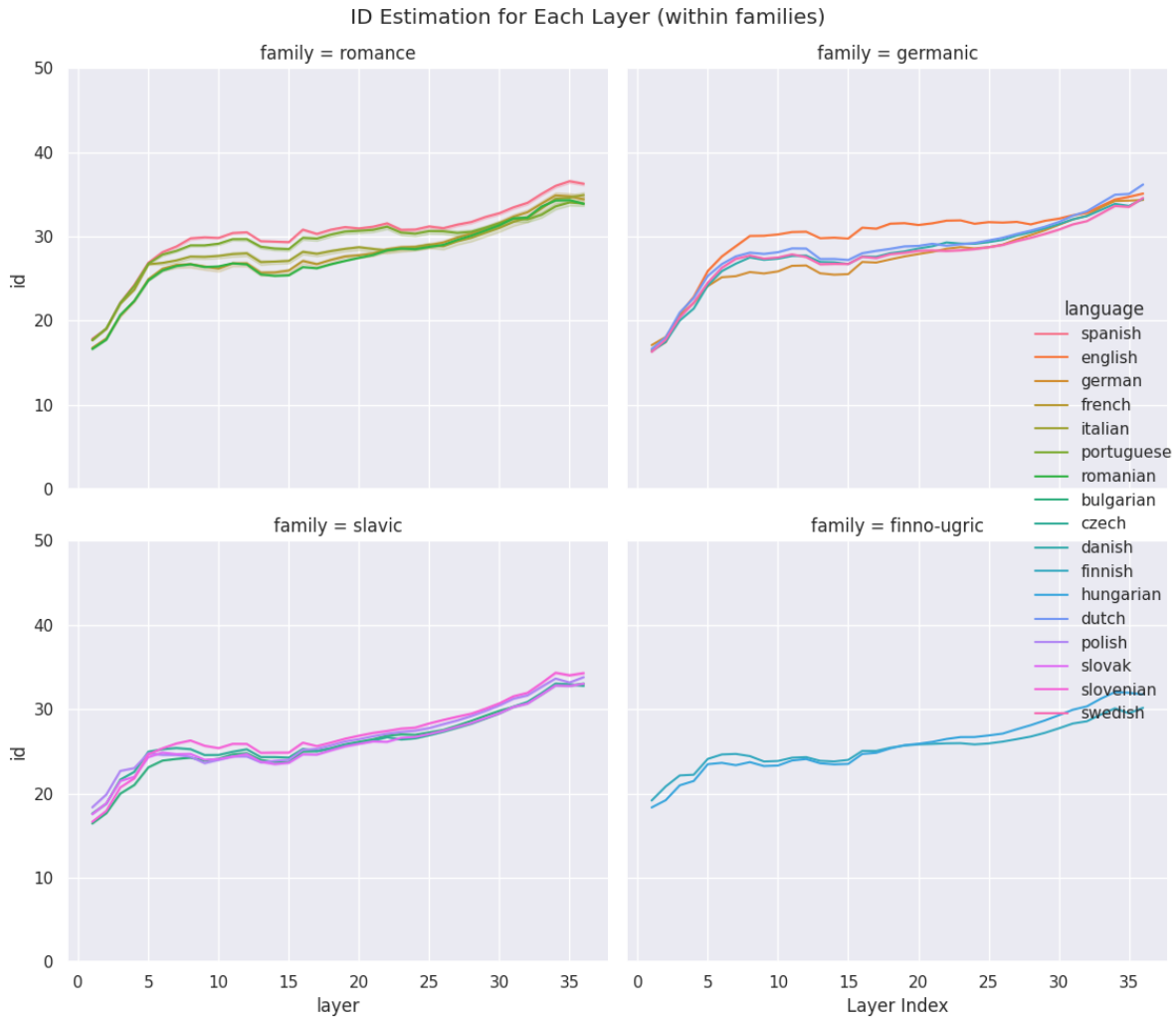

Figure C.1: Average ID estimation for each layer on all languages (grouped by family) with XLM-R_{XL} 3.5B.

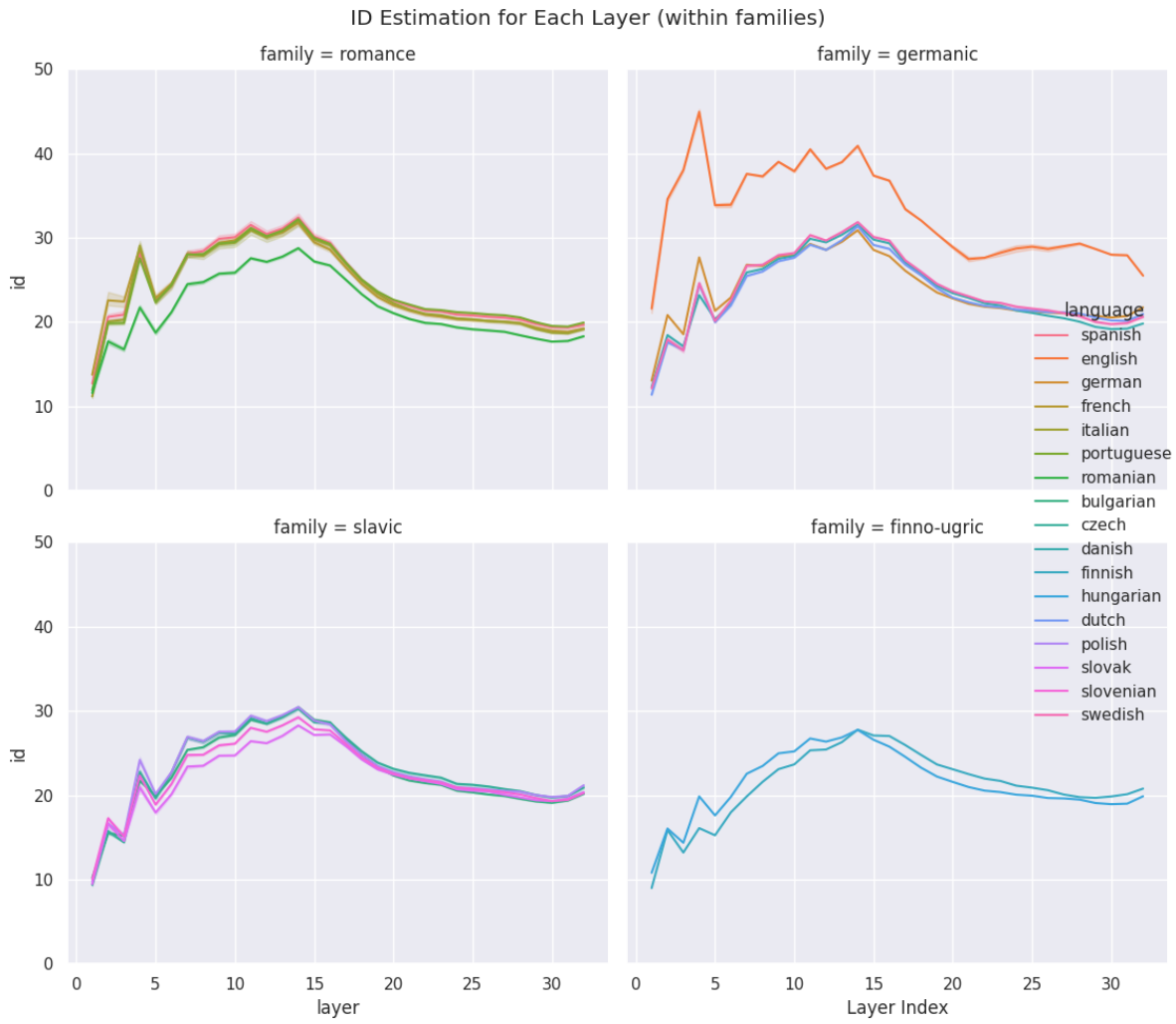

Figure C.2: Average ID estimation for each layer on all languages (grouped by family) with Mistral 7B.

Appendix D

Figures: ID Estimation for Each Layer on Indo-European Languages Divided by Branch With BLOOM

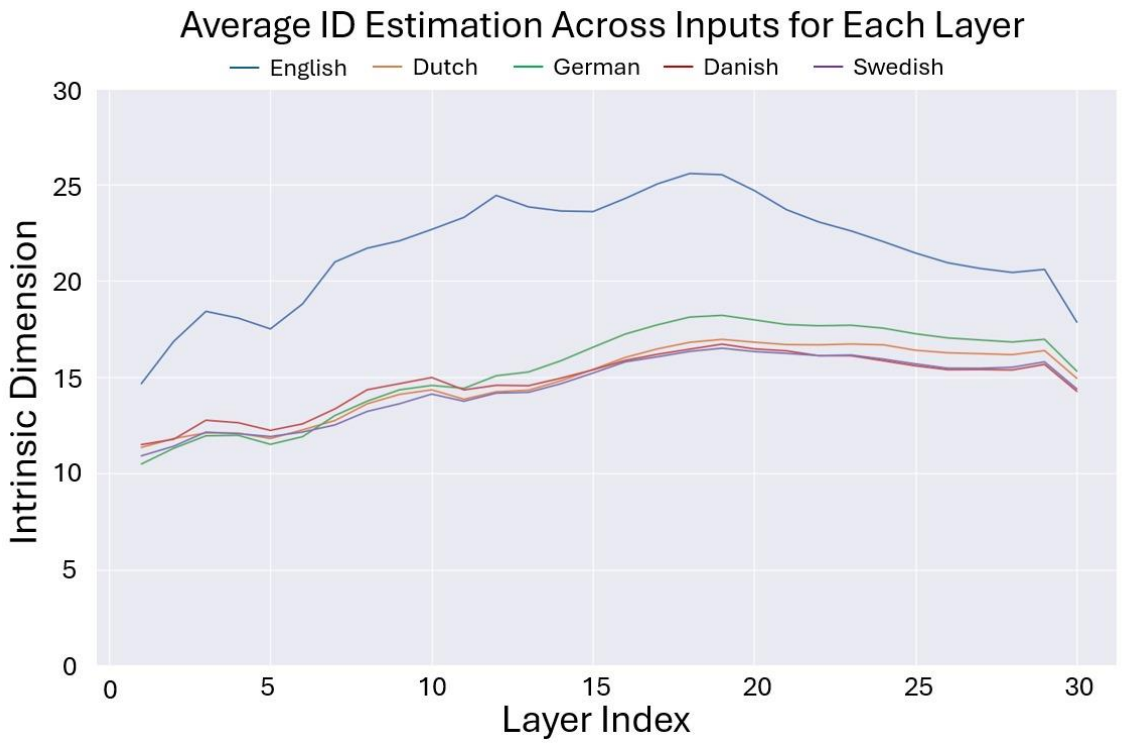

Figure D.1: Average ID estimation for each layer on Germanic languages with BLOOM 3B.

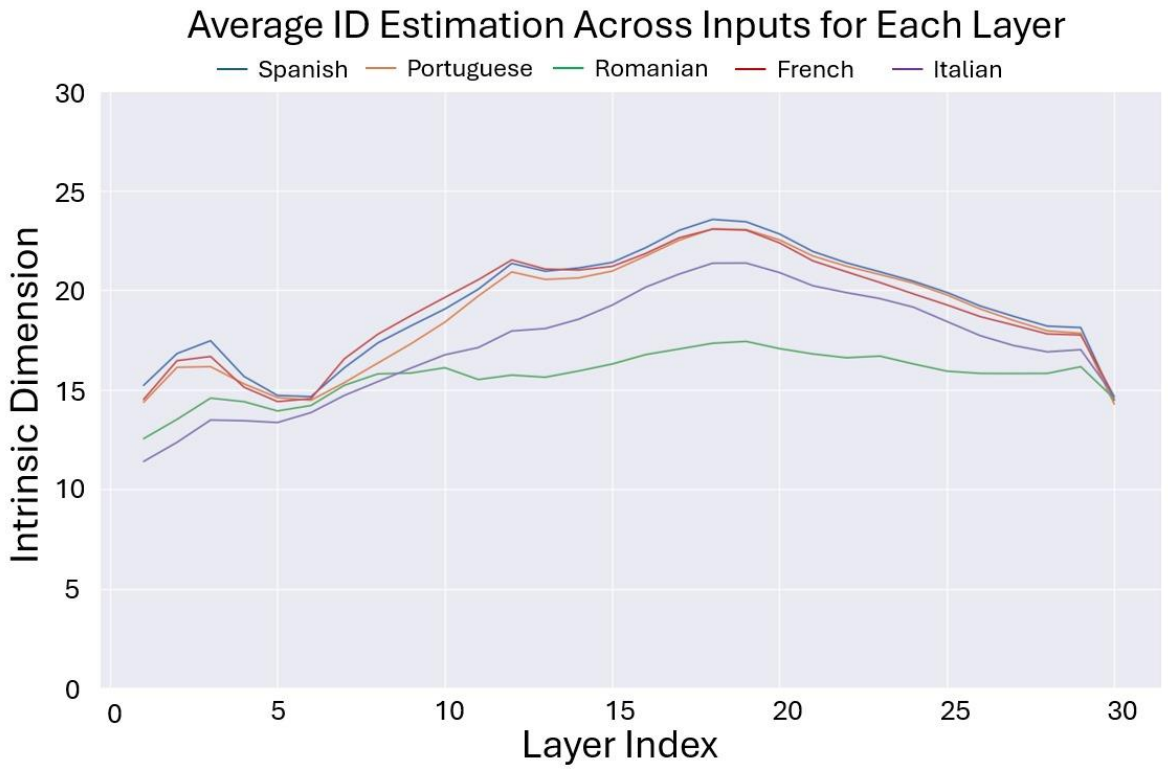

Figure D.2: Average ID estimation for each layer on Romance languages with BLOOM 3B.

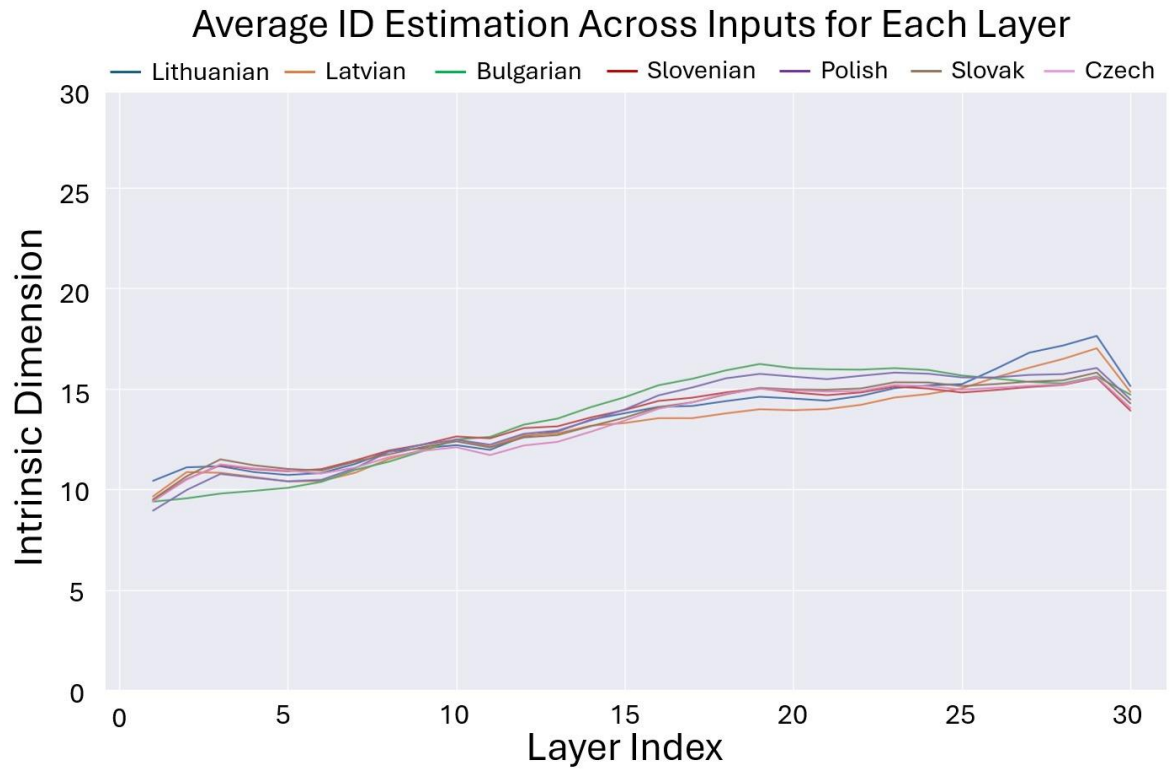

Figure D.3: Average ID estimation for each layer on Balto-Slavic languages with BLOOM 3B.

Appendix E

Figure: Pearson Correlation Between PPL and ID at Layer 14 With Mistral

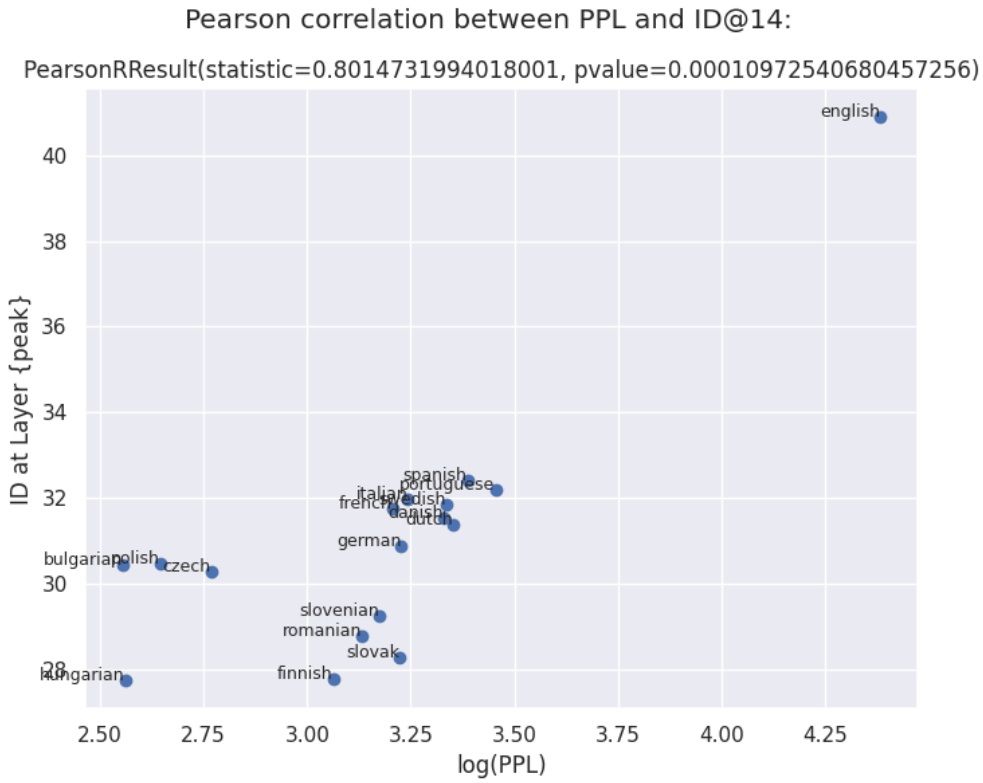

Figure E.1: Pearson correlation between PPL and ID at layer 14 with Mistral 7B. We find PPL and ID at layer 14 to be negatively correlated.