THERE IS NO THEORETICAL CURSE OF MULTILINGUALITY FOR EMBEDDING SPACE STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

A central goal of multilingual NLP is to achieve high monolingual performance per language and cross-lingual alignment for large-scale language coverage with a multilingual model. The *curse of multilinguality* describes the phenomenon of degradation in multilingual model performance as we increase language coverage, posing a threat to the above goal. This paper asks whether multilingual embedding spaces are inherently incapable of achieving perfect multilinguality without a prohibitive increase in required capacity. We first formalize the goal of "perfect multilinguality", embodied in two *multilinguality conditions*. We then prove that the minimum dimensionality required for perfect multilinguality grows only logarithmically, as $\Theta(\log L)$, in the number of languages L. That is, we show that there is no theoretical curse of multilinguality for embedding space structure. This suggests that the empirical curse of multilinguality is a result of real world data and training conditions. We back this understanding with a small-scale empirical study. Our paper provides the first theoretical and intrinsic perspective on the curse of multilinguality.

1 INTRODUCTION

The ideal multilingual model exhibits high monolingual performance on individual languages as well as perfect cross-lingual structure, or semantic sharing across languages, for arbitrary language coverage. The well-known *curse of multilinguality* (COM, Conneau et al., 2020), also sometimes termed *negative interference* (Shaham et al., 2023), describes the phenomenon of observed degradation of multilingual performance as we train a fixed-capacity model on increasingly more languages. Folk wisdom generally understands fixed capacity to be an inherent bottleneck (Gurgurov et al., 2024), resulting in "less space" for each language, indicating that perfect multilinguality as described above may not be achievable for high language coverage. Several previous works have demonstrated the COM from the perspective of extrinsic task performance (Conneau et al., 2020; Aharoni et al., 2019), and while other works have explored strategies to mitigate this issue (Wang et al., 2020a; Pfeiffer et al., 2022), it remains a serious challenge. This gives rise to a fundamental question: is it inherently impossible to accommodate increasing language coverage in a limited-capacity model while maintaining desired monolingual and cross-lingual performance on all languages? We provide a theoretical and intrinsic perspective on the COM. Specifically, we study the curse of multilinguality for embedding space structure (COM-EMB), and show that multilingual representations can exhibit intrinsic quality while accommodating more languages without a prohibitive increase in capacity.

First, we formulate two intrinsic requirements for a *perfect multilingual* embedding space, drawing from consensus in the field (Chang et al., 2024; Hämmerl et al., 2024), and term them *multilinguality conditions*: (a) the space exhibits high-quality shared monolingual semantics for each language, and (b) the space exhibits cross-lingual alignment, meaning that translation equivalents across languages are "close" to each other in the space. This is important for applications such as cross-lingual information retrieval (Sasaki et al., 2018; Roy et al., 2020). In the context of embedding spaces, "capacity" can be understood as dimensionality. Then, our theoretical result shows that the minimum dimensionality of a perfect multilingual space encoding L languages grows as $\Theta(\log L)$. Specifically, we assume a base D-dimensional concept space and show that the minimum sufficient extra dimensionality to accommodate L languages while maintaining the multilinguality conditions grows as $O(\log L)$, and that the dimensionality of any such space grows at least as $\Omega(\log L)$. This result shows that multilinguality is not entirely free: it requires an additional capacity cost depending on L.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, given that the number of languages in the world is finite and usually estimated at around $\sim$7000 (Eberhard et al., 2021), the log dependence of the cost indicates only a small multilinguality tax as we increase language coverage, provably as small as 30 extra dimensions for 7000 languages. Thus, we show that there is no theoretical curse of multilinguality for embedding space structure.

The above finding suggests that any empirical curse of multilinguality for embedding space structure is a consequence of real-world conditions such as data and compute constraints, among other things. We characterize the empirical COM-EMB via a controlled study of the empirical degradation of the multilinguality conditions with increasing language coverage in Transformer-based models, in eight different configurations varying compute constraints, sampling conditions, and the set of languages being evaluated. We show that while we do observe "cursed" behaviour in line with general wisdom, this behaviour depends on these configurations and is mitigated in favourable conditions.

Our paper contributes the first theoretical and intrinsic grounding to the curse of multilinguality. In showing that perfect intrinsic multilinguality is theoretically achievable at an arbitrary language scale, we set the bar for empirical multilingual spaces, and promote future work in exploring and closing the gap.

2 BACKGROUND

Dimensionality bounds on embedding spaces Recent work studies the embedding dimension required to express top-k subsets of N documents by single-vector retrieval. Weller et al. (2026) show that, given a fixed margin of separation, the necessary embedding dimension has a logarithmic dependence on the total number of subsets $\binom{N}{k}$, using classical sphere packing arguments (Vershynin, 2019; Conway et al., 1999). Wang et al. (2026) show that $d = 2k + 1$ is a sufficient dimension for expressing all top-k subsets, independent of N, and Okajima & Yoshida (2026) subsequently show that finite-precision quantization introduces a dependence on N for the necessary dimension, which must grow logarithmically with the corpus size for a fixed precision. Our work connects this line of study with fundamental questions in multilingual NLP.

Curse of multilinguality The "curse of multilinguality" or phenomenon of "negative interference" is understood as the observed degradation in multilingual model performance as the number of training languages is increased for a fixed capacity model, commonly described in terms of capacity dilution over languages (Aharoni et al., 2019; Pfeiffer et al., 2022). While multilinguality provides benefits up to a point especially for low-resource languages via cross-lingual transfer from other language training data, high-resource language performance usually suffers; further, after a point, all language performance suffers from including more training languages, constituting a "transfer-dilution" tradeoff (Arivazhagan et al., 2019; Chang et al., 2024; Gurgurov et al., 2024). Shaham et al. (2023) show that the proportion of the target language with respect to total training data is an important factor, and that increasing model size helps, raising the question of the minimum model size required to accommodate a given language coverage. Longpre et al. (2026) develop empirical scaling laws for model and data size while increasing language coverage. There are several works that attempt to mitigate the phenomenon of degradation, including strategies for better data sampling (Wang et al., 2020a; Kreutzer et al., 2021; Foroutan et al., 2025), architectural innovations such as mixture-of-expert models (Pfeiffer et al., 2022), and optimization techniques such as Gradient Vaccine (Wang et al., 2020b). The above works explore the curse of multilinguality from a purely empirical lens. To the best of our knowledge, we are the first to provide a theoretical lens on the curse of multilinguality.

Intrinsic characterization of embedding spaces Previous works in the curse of multilinguality study embedding space quality only indirectly by measuring extrinsic performance on tasks such as XNLI (Conneau et al., 2018) and machine translation (Aharoni et al., 2019). However, there is broad general interest in investigating the structure of multilingual spaces, including assessing cross-lingual overlap and alignment (Xu et al., 2023; Wen-Yi & Mimno, 2023; Shah et al., 2023; Mousi et al., 2024; Hämmerl et al., 2024, inter alia), and the encoding of language versus semantic information (Chang et al., 2022; Xie et al., 2022b). Tasks such as monolingual and cross-lingual retrieval and bitext mining depend directly on representation structure (Roy et al., 2020; Enevoldsen et al., 2025), and inspire the formulation of our multilinguality conditions (Section 3). However, to

the best of our knowledge, no previous work seeks to understand the curse of multilinguality from the intrinsic perspective of its impact on embedding space structure.

3 Multilinguality Conditions for Perfect Multilingual Spaces

In this section, we formulate conditions necessary for perfect multilinguality, given a reference high-quality monolingual concept space.

Preliminaries We begin by defining key structures relevant to our argument. Let $\mathcal{C}$ be a global set of concepts, and let $\mathcal{L}$ be a set of languages. Let Z be the reference concept space, where $c \in \mathcal{C}$ is embedded as z_c with high-quality semantic relationships. For the multilingual embedding space X^L encoding $L = |\mathcal{L}|$ languages, let $x_{c,\ell}$ represent the embedding for concept $c \in \mathcal{C}$ in language $\ell \in \mathcal{L}$. Let $X^L_\ell \subset X^L$ denote the monolingual subspace of language $\ell \in \mathcal{L}$. We use $\cos(\cdot,\cdot)$ to refer to the cosine similarity operator. Finally, we use γ to refer to the resolution scale of Z by which any two distinct points are minimally separated: i.e. we have $\gamma < 1$ such that $\forall c, c' \in \mathcal{C}, c \neq c', \cos(z_c, z_{c'}) < \gamma$.

Monolingual structure A perfect multilingual space should encode high-quality shared monolingual relationships for each encoded language. We use the high-quality concept space Z as reference. Then, perfect monolingual structure requires that for each language ℓ represented in space X^L, its monolingual structure X^L_ℓ exactly mimics Z.[1]

Monolingual structure can be understood in terms of relative similarity judgments between pairs of concepts. Thus, perfect monolingual structure can be understood as the monolingual subspace of each language exhibiting identical similarity rankings as the reference concept space, resulting in high-quality and shared semantics across languages.

Definition 1 (Monolingual Structure)**.** *A multilingual space X^L satisfies the monolingual structure condition if each encoded monolingual subspace in it is identical in terms of similarity rankings ($\sim_R$) to a high-quality concept space Z, i.e. $\forall \ell \in \mathcal{L}, X^L_\ell \sim_R Z$. This condition requires that $\forall \ell \in \mathcal{L}$ and $c, c', d, d' \in \mathcal{C}$, $\cos(z_c, z_{c'}) \leq \cos(z_d, z_{d'}) \iff \cos(x_{c,\ell}, x_{c',\ell}) \leq \cos(x_{d,\ell}, x_{d',\ell})$.*

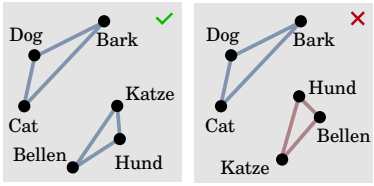


Figure 1: Illustration of Definition 1. *Dog=Hund*, *Cat=Katze*, *Bark=Bellen*. The concept space is depicted in English. The German space on the left exhibits identical ordering of neighbours with the concept space while the German space on the right does not, because the nearest neighbour of *Katze(=Cat)* is *Bellen(=Bark)* and not *Hund(=Dog)*.

Cross-lingual alignment A perfect multilingual space displays cross-lingual alignment, understood as the property that language variants of a given concept are close together, and farther from other concepts. Past literature distinguishes a "strong" and "weak" view of this idea (Roy et al., 2020; Hämmerl et al., 2024). In the "strong" view, we require a concept in language ℓ to be closer to its translation equivalent in language ℓ' than any different concept in *any* language, including other concepts in ℓ, whereas in the weak view, it is only required to be closer than any other ℓ'-language points. We use the strong view, which also trivially implies the weak one.

Definition 2 (Cross-lingual alignment)**.** *A multilingual space X^L satisfies the cross-lingual alignment condition if a concept encoding in a given language is closer to the encoding of the same concept expressed in any other language than to the encoding of any other concept in any language. That is, this condition requires that for all concepts $c, c' \in \mathcal{C}$ where $c \neq c'$ and languages $\ell, m, n \in \mathcal{L}$, $\cos(x_{c,\ell}, x_{c,m}) > \cos(x_{c,\ell}, x_{c',n})$.*

[1] In practice, we do not know what perfect concept relationships are. However, we can consider the monolingual space of a high-resource language (e.g. English) as a reference space, making the monolingual structure condition equivalent to requiring parity for monolingual quality and shared semantics across languages.

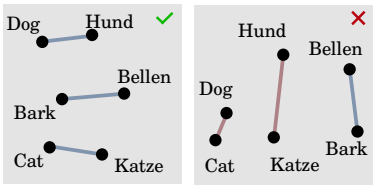


Figure 2: Illustration of Definition 2, with two languages. In the left space, *Hund(=Dog)* is closer to *Dog* than any different-concept word, but in the right space, the different-concept word *Cat* is closer to *Dog* than *Hund=(Dog)*.

Non-degeneracy. A multilingual space can place all concept translation equivalents as co-incident at the same point in the reference high-quality monolingual space, and thus trivially exhibit monolingual structure as well as cross-lingual alignment. However, this violates common practical assumptions of embedding spaces where an embedding space must be capable of distinguishing between any two embedded points with some fixed scale of resolution (Weller et al., 2026). Thus, we require that our multilingual space have a resolution scale, or a maximum cosine similarity $\gamma_X < 1$ such that for any distinct $(c, \ell), (d, m) \in \mathcal{C} \times \mathcal{L}$, $\cos(x_{c,\ell}, x_{d,m}) < \gamma_X$.

4 Theoretical Curse of Multilinguality for Embedding Space Structure

4.1 Defining the curse of multilinguality

The COM is described in the field for empirical spaces as the degradation of desired properties – usually, extrinsic task performance for some set of languages – with an increasing number of encoded languages, given fixed model capacity. For the COM with respect to embedding space structure (COM-EMB), "desired properties" correspond to perfect multilinguality, and "capacity" corresponds to dimensionality. **We formulate the COM-EMB as the inability of a multilingual embedding space to maintain perfect multilinguality given an increasing number of encoded languages L without a prohibitive increase in dimensionality.**

4.2 Theorem

In this section, we answer the following question: *how much dimensionality does a perfect multilingual space accommodating L languages theoretically require in terms of L?* If the required dimensionality grows prohibitively with the number of languages, e.g. linearly or exponentially, then this can be termed a theoretical COM-EMB.

Theorem 1. *The minimum dimensionality of a perfect multilingual space encoding L languages grows as $\Theta(\log L)$.*

We prove this in two parts. First, we upper-bound the minimum *sufficient* dimensionality of a perfect multilingual space in terms of the number of encoded languages and the dimensionality of a fixed target concept space, in line with our setup in Section 3.

Theorem 1.1. *Let D_Z be the dimensionality of the reference concept space Z. Then there exists a perfect multilingual space X^L encoding L languages with minimum sufficient dimensionality $D_{X^L} = D_Z + O(\log L)$.*

Next, we lower-bound the dimensionality of the space, i.e. show the *necessary* dimensionality of a perfect multilingual space.

Theorem 1.2. *Any perfect multilingual space X^L encoding L languages must have dimensionality $D_{X^L} = \Omega(\log L)$.*

Since D_Z is independent of L, Theorem 1.1 and Theorem 1.2 together establish Theorem 1.

The following proofs make a fixed unit norm assumption for Z and X^L. This is standard in the use of text embeddings for applications such as retrieval (Weller et al., 2026). Our proofs are also extendable to the case of variable norm as sketched in Appendix A.4.

4.3 PROOF OF THEOREM 1.1

We proceed constructively. Given Z and L, we describe a multilingual space X^L and show that it maintains the multilinguality conditions, and that its dimensionality follows the required bound.

Multilingual space construction Recall that for a concept c, we have its embedding z_c in the concept space Z, with the resolution scale γ, and dimensionality D_Z.

We now define a space U of *language offsets*. This space uses D_L new dimensions, and is populated with L language offset vectors $u_\ell \in \mathbb{R}^{D_L}$, with a fixed norm $\|u_\ell\| = r$, and with a maximum similarity $\rho < 1$ between any two language offsets. That is, we have $\cos(u_\ell, u_m) < \rho$ for all $\ell, m \in \mathcal{L}$ where $\ell \neq m$.

We then define our multilingual space X^L, by a direct sum construction of the concept space and the language offsets. That is, $\forall c \in \mathcal{C}, \ell \in \mathcal{L}, x_{c,\ell} := \frac{1}{\sqrt{1+r^2}} z_c \oplus u_\ell$, where $\oplus$ refers to vector concatenation. As a result, $x_{c,\ell} \in X^L$ has dimensionality $D_{X^L} = D_Z + D_L$. Note the norm term $\frac{1}{\sqrt{1+r^2}}$. Since $\|z_c\| = 1$ and $\|u_\ell\| = r$, and z_c and u_ℓ are orthogonal when embedded in X^L, we now have X^L as a unit norm space with $\|x_{c,\ell}\| = 1$ for all c, ℓ and dimensionality D_{X^L}.

Similarity bounds on offset vectors We construct the language offset vectors in the D_L-dimensional space U with fixed norm r such that their minimum and maximum cosine similarity lies in the following interval, guaranteed to be non-empty by a choosing $r \geq 1$ and a suitable choice of ρ:

$$1 - \frac{1-\gamma}{r^2} < \cos(u_\ell, u_m) < \rho < 1 \qquad \text{for all } \ell \neq m. \tag{1}$$

Similarity decomposition Given that X^L and Z are unit-norm, we have the following relationship: $\cos(x_{c,\ell}, x_{d,m}) = \langle x_{c,\ell}, x_{d,m} \rangle$; analogously for Z. Since z_c and u_ℓ are orthogonal when embedded in X^L, we make the following useful inner-product decomposition for points in X^L, used throughout our proof:

$$\langle x_{c,\ell}, x_{d,m} \rangle = \left\langle \frac{z_c \oplus u_\ell}{\sqrt{1+r^2}}, \frac{z_d \oplus u_m}{\sqrt{1+r^2}} \right\rangle = \frac{1}{1+r^2} \langle z_c \oplus u_\ell, z_d \oplus u_m \rangle = \frac{\langle z_c, z_d \rangle + \langle u_\ell, u_m \rangle}{1+r^2} \tag{2}$$

Lemma 1. *The space X^L satisfies the monolingual structure condition (Definition 1).*

Proof. We will show that for every language ℓ, $X^L_\ell \sim_R Z$, i.e. each language subspace is identical to Z under point similarity ranking. Using Equation 2 for points in X^L_ℓ, we have

$$\begin{aligned} \cos(x_{c,\ell}, x_{c',\ell}) = \langle x_{c,\ell}, x_{c',\ell} \rangle &= \frac{\langle z_c, z_{c'} \rangle + \langle u_\ell, u_\ell \rangle}{1+r^2} = \frac{\langle z_c, z_{c'} \rangle + r^2}{1+r^2} = \frac{\cos(z_c, z_{c'}) + r^2}{1+r^2} \\ &= a \cos(z_c, z_{c'}) + b \end{aligned}$$

for fixed and global a and b depending only r. Note that this automatically gives us, for any two languages ℓ, m,

$$\cos(x_{c,\ell}, x_{c',\ell}) = a \cdot \cos(z_c, z_{c'}) + b = \cos(x_{c,m}, x_{c',m})$$

Given this affine behavior, and the language agnosticity of the result, it follows that the similarity ranking behavior is identical in all languages $\ell \in \mathcal{L}$ with respect to Z, i.e. $\forall \ell \in \mathcal{L}, X^L_\ell \sim_R Z$. Hence, X^L satisfies the monolingual structure condition. □

Lemma 2. *The space X^L satisfies the cross-lingual alignment condition (Definition 2).*

Proof. We show that for every $c, c' \in \mathcal{C}$ where $c \neq c'$ and all languages $\ell, m, n \in \mathcal{L}$, $\cos(x_{c,\ell}, x_{c,m}) > \cos(x_{c,\ell}, x_{c',n})$. Recall that points in Z have resolution scale γ and points in U have norm r. From the similarity decomposition (Equation 2), we can simplify and bound the two similarity expressions as follows for any $c \neq c', \ell, m, n$:

$$\cos(x_{c,\ell}, x_{c,m}) = \frac{\langle z_c, z_c \rangle + \langle u_\ell, u_m \rangle}{1 + r^2} = \frac{1 + \langle u_\ell, u_m \rangle}{1 + r^2}$$
$$\cos(x_{c,\ell}, x_{c',n}) = \frac{\langle z_c, z_{c'} \rangle + \langle u_\ell, u_n \rangle}{1 + r^2} < \frac{\gamma + r^2}{1 + r^2}.$$

Now it suffices to show that

$$1 + \langle u_\ell, u_m \rangle > \gamma + r^2.$$

Since $\langle u_\ell, u_m \rangle = r^2 \cos(u_\ell, u_m)$, the above expression is equivalent to

$$\cos(u_\ell, u_m) > 1 - \frac{1 - \gamma}{r^2}.$$

which is guaranteed by our construction in Equation 1. Thus, X^L satisfies the cross-lingual alignment condition. □

Lemma 3. *The space X^L is non-denegerate.*

Proof. The cosine similarity between any two distinct points in X^L is given by Equation 2. We bound the maximum similarity between any two points based on the three cases: same concept with different language, different concept with same language, and different concept with different language. If $c = c'$ and $\ell \neq m$, then $\cos(x_{c,\ell}, x_{c,m}) < \frac{1+r^2\rho}{1+r^2}$. If $c \neq c'$ and $\ell = m$, then $\cos(x_{c,\ell}, x_{c',\ell}) < \frac{\gamma+r^2}{1+r^2}$. If both $c \neq c'$ and $\ell \neq m$, then

$$\cos(x_{c,\ell}, x_{c',m}) = \frac{\langle z_c, z_{c'} \rangle + \langle u_\ell, u_m \rangle}{1 + r^2} < \frac{\gamma + r^2\rho}{1 + r^2} < \gamma_X := \max\left\{ \frac{\gamma + r^2}{1 + r^2}, \frac{1 + r^2\rho}{1 + r^2} \right\}.$$

Note that $\gamma_X < 1$, since $\gamma < 1$ and $\rho < 1$. γ_X is independent of L and only depends on the constants γ, ρ and r; thus, X^L maintains a fixed cosine scale independent of the number of languages encoded. Thus, X^L satisfies the non-degeneracy condition at fixed resolution scale $\gamma_X < 1$. □

Lemma 4. *The minimum sufficient dimensionality of X^L is $O(\log L)$.*

Proof. Recall that $D_{X^L} = D_Z + D_L$. We will bound D_L, the dimensionality of the language offset space U.

U is constructed by placing L fixed-norm language offsets in U such that each point pair has cosine similarity in the non-empty interval defined in Equation 1. Bounding D_L therefore reduces to the following question: how many dimensions D_L are sufficient to pack L points on a sphere in a D_L-dimensional with specified maximum and minimum cosine similarity?

Using standard spherical cap packing bounds, we can show that the number of points that can be placed on a sphere in a D_L-dimensional space following these constraints grows exponentially in D_L.[2] This is shown in detail in Appendix A.1. Briefly, the maximum similarity constraint imposes an angular region within which all points must lie, and for a maximal packing, caps with angular radius equal to the minimum angular separation centered at the packed points must cover this containing region. This allows us to lower-bound the number of points in a maximal packing by a ratio of spherical cap measures, which grows exponentially with the dimensionality of the sphere. Therefore, to represent L language offsets, it is possible to choose some constant $C > 0$ independent of L such that a sphere in a D_L-dimensional space with $D_L = \left\lceil \frac{1}{C} \log L \right\rceil$ fits $\exp(CD_L) \geq L$ such points. Therefore, the minimum sufficient dimension to pack L points maintaining the required constraints is $D_L = O(\log L)$. Since D_Z is constant in terms of L, X^L has the minimum sufficient dimension:

$$D_{X^L} = D_Z + D_L = D_Z + O(\log L) \implies D_{X^L} = O(\log L)$$

Note that it suffices to exhibit a single perfect multilingual space with the above scaling to upper-bound the minimum sufficient dimensionality for perfect multilinguality. By the existence of X^L, we establish the minimum sufficient dimension for perfect multilinguality grows as $O(\log L)$. □

[2] This comes with the constraint on a choice of ρ. For $r = 1$, this is $\arccos(\rho) < \arccos(\gamma)/2$.

Fitting 7000 languages in 30 dimensions For practical purposes, the constant involved in the $O(\log L)$ bound may be of concern. In Appendix A.3, we prove that it is theoretically possible to fit $L = 7000$ language offset unit vectors into $D_L = 30$ dimensions while satisfying all required constraints on minimum and maximum similarity required for our constructive solution, with a realistic choice of γ. Briefly, we show that a random construction of L points in a D_L-dimensional space has a non-zero probability of satisfying our required constraints, proving existence. Thus, our theoretical scaling factor is of minor practical concern given the typical dimensionality scale of representation spaces.

4.4 Proof of Theorem 1.2

Proof. Fix any concept c. Then the L language variants of this concept, $x_{c,1}, \ldots, x_{c,L}$, are all distinct fixed-norm vectors, or distinct points on a sphere. Since X^L is non-degenerate as we assume it is perfectly multilingual, any two distinct points in it must remain separated at some fixed cosine resolution scale $\gamma_X < 1$, equivalent to a fixed angular separation $\theta_X := \arccos(\gamma_X)$, independent of L.[3] Thus, X^L contains at least L points on a multidimensional sphere with a fixed angular separation, and its necessary dimensionality is lower bounded by the number of dimensions needed to place points in such a way.

We now apply a standard spherical cap bound. For any fixed angular separation $\theta_X > 0$, the maximum number of vectors that can be placed on a D_{X_L}-dimensional sphere in with pairwise angular separation at least θ_X grows at most exponentially in D_{X_L}. We show this in detail in Appendix A.2. Briefly, the minimum angular separation means that if we place a spherical cap with half the required angular separation around each point then these caps must be disjoint. Their total measure cannot exceed that of the sphere. This upper-bounds the number of points by the reciprocal of a spherical cap measure, which grows at most exponentially with the dimensionality of the sphere. That is, there exists a constant $A_{\theta_X} > 0$, depending only on θ_X, such that

$$L \leq \exp(A_{\theta_X} D_{X_L}) \Rightarrow \log L \leq A_{\theta_X} D_{X_L} \Rightarrow D_{X_L} \geq \frac{1}{A_{\theta_X}} \log L \quad \Rightarrow D_{X_L} = \Omega(\log L)$$

Thus, the dimensionality of any non-degenerate multilingual space encoding L languages has a logarithmic dependence on L. Note that if the original monolingual concept space "fully" requires D_Z dimensions to represent concept semantics, then these $\Omega(\log L)$ dimensions required to construct language directions are additional to these D_Z dimensions. Under this minimality assumption on the base concept space, the additional cost can be understood as a multilinguality tax:[4]

$$D_{X^L} = D_Z + \Omega(\log L)$$

In general, the dimension D_{X^L} of any perfect multilingual space X^L encoding L languages must grow at least as $\Omega(\log L)$. □

5 Empirical curse of multilinguality for embedding space structure

While our theorem shows that a multilingual space is theoretically capable of exhibiting perfect multilinguality under language scaling without a prohibitive cost, there may still exist an empirical `COM-EMB` for real-world multilingual models. In this section, we formulate metrics for monolingual structure and cross-lingual alignment, and provide a small-scale controlled study showing the empirical curse of multilinguality for embedding space structure under eight configurations mimicking various real-world conditions. Specifically, we train multilingual models from scratch on an increasing number of languages and evaluate embedding space structure under language scaling, in different training and evaluation configurations.[5] We provide precise metric definitions and extra experimental details and results in Appendix B.

[3] Fixing the resolution scale across the family of spaces X^L allows us to study the asymptotic behaviour of the necessary dimensionality of X^L with increasing L in a meaningful manner, and follows the assumptions in related work studying dimensionality scaling (Weller et al., 2026; Okajima & Yoshida, 2026).

[4] In general, real embedding spaces may not use their ambient dimensionality (Ansuini et al., 2019). Instead, this can be understood as placing pressure on true or effective dimensionality of such a space.

[5] Our characterization of the theoretical curse of multilinguality for embedding space structure allows for a small growth based on number of languages, while in our experiments, we keep the dimensionality fixed with

Note on previous work Previous works studying the COM and related concepts generally differ with respect to their experimental conditions and assumptions. Differences include (i) whether token compute is fixed (Shaham et al., 2023) or increases with more training languages (Aharoni et al., 2019), (ii) whether the training data distribution is uniform over languages (Shaham et al., 2023) or follows a real-world distribution (Aharoni et al., 2019; Conneau et al., 2020), and (iii) whether performance metrics are tracked over all training languages (Conneau et al., 2020) or a fixed group such as high-resource languages (Arivazhagan et al., 2019), among other variables.

Configurations We study the empirical COM-EMB under eight configurations consisting of all combinations of the three above dimensions: (i) {fixed-compute, increasing-compute}, (ii) {uniform-sampling, realistic-sampling}, and (iii) {aggregate-all, aggregate-fixed}.

Metrics Our metrics are inspired from the multilinguality conditions. We consider a "concept" to be a text, and compute the following using a fixed concept set across all target languages obtained from multiway parallel data. We evaluate **monolingual structure** against a strong reference space, by measuring, for each concept, the overlap of the monolingual neighbourhood of that concept in each target language with the gold neighbourhood of that concept in a strong monolingual reference space. We report this metric (*k-nearest-neighbour overlap; MS-NNO*) as a percentage of $k = 20$, averaged over concepts and languages. We also look at language mean over *token-normalized masked language modeling loss* on a monolingual test set per language (MS-MLM) as an indirect measure of monolingual structure quality. We evaluate **cross-lingual alignment** by checking for each language ℓ_1, whether, for each concept and each language ℓ_2, the ℓ_2 equivalent of that concept is closer to the ℓ_1 equivalent than other ℓ_2 concepts (weak view; similar trends for strong view shown in Appendix B.3). This metric (CLA-WV) is reported as a percentage of the concept set, averaged over language pairs. We also check for non-degeneracy by checking for looking at the minimum, maximum, and mean cosine similarity over all point pairs in the space as a function of increasing language coverage. We do not see evidence of degeneracy in any configuration (Appendix B.3).

Experimental setup We study a 4-layer Transformer-based BERT encoder architecture with a masked-language modeling training objective (Devlin et al., 2019). We use the MADLAD-400 corpus (Kudugunta et al., 2023) for monolingual training data, and FLORES+ (Goyal et al., 2022; Team et al., 2024), BOUQuET (Andrews et al., 2025; Omnilingual MT Team et al., 2026), and WMT24++ (Deutsch et al., 2025) as multiway parallel data for evaluation. We use the English concept space from e5-large-v2 (Wang et al., 2024) as a strong reference space for MS-NNO. We create nine language groups with incrementally more languages added in descending order of resourcedness, with the largest language group containing 100 languages. The token count per language depends on the token-compute and sampling settings: fixed-compute experiments are trained on a total budget of 500M tokens, and increasing-compute with uniform-sampling uses 50M tokens per language. The realistic-sampling setting uses 500M as the maximum language token count, with other language token count proportionate to resource level in our training corpus. Each model run consists of 5 seeds, with each seed sampling different data up to the budget, and random initialization of model. The tokenizer is trained on a subset of the training corpus. For the aggregate-all setting, we evaluate all training languages covered by each evaluation dataset. For aggregate-fixed, the smallest language group is 10, and we evaluate these 10 fixed languages across all language groups. We use the [CLS] token embedding as a text embedding.

Results and discussion If all metrics exhibit deterioration for a configuration, we term it a "cursed" configuration. Figure 3 shows results for all eight configurations with BOUQuET, separating cursed from non-cursed configurations (similar trends for other evaluation datasets in Appendix B.3). We observe four cursed configurations of eight, showing consistent degradation of monolingual structure and cross-lingual alignment. **Therefore, we observe an empirical curse of multilinguality for embedding space structure, parallelling previous work in the COM.**

In the other four non-cursed configurations, we observe stability or improvements in one or more metrics. We find that these are the configurations that generally maintain high token count for

language scale for simplicity. In practice, trained embedding spaces use very little of the ambient dimensionality (Ethayarajh, 2019; Ansuini et al., 2019), and the permissible numerical inflation for around a hundred languages is small as discussed in Appendix A.3.

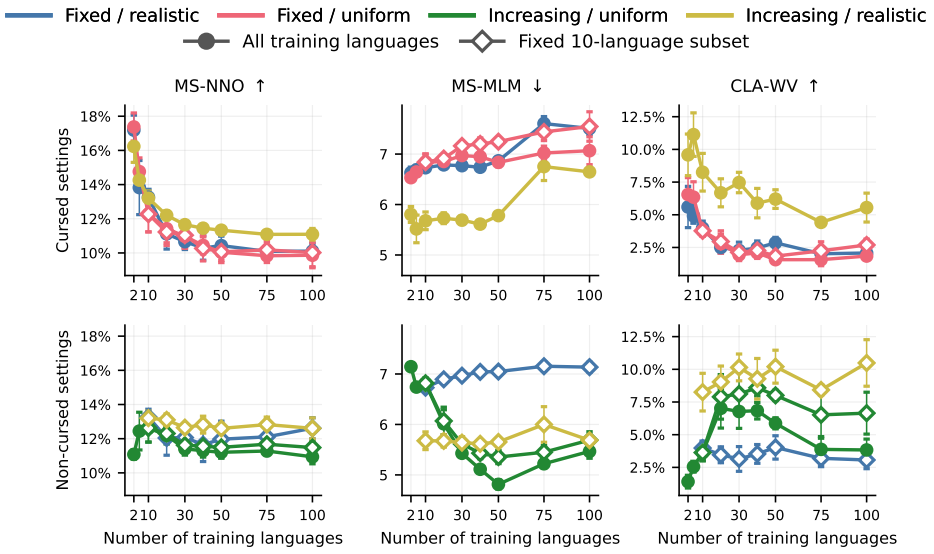


Figure 3: `MS-NNO`, `MS-MLM`, and `CLA-WV` on `BOUQuET` for encoder models with as the number of training languages increases, in eight configurations. The top and bottom rows show cursed and non-cursed configurations respectively. Arrows indicate whether higher or lower metric values are better. Error bars show standard deviation over five model seeds.

the target languages of interest as a consequence of compute and sampling choices. For example, with `increasing-compute`, `realistic-sampling`, and `aggregate-fixed` (—◇—), token count for the 10 high-resource target languages remains high and unchanged with more training languages, and metrics stay roughly stable. We also see that metrics may actually show some improvements with more languages up to a point, e.g. with `increasing-compute` and `uniform-sampling` (—), which indicates that total token count and cross-lingual transfer may also play a role. **In general, our findings suggest that the empirical curse of multilinguality for embedding space structure is sensitive to training and evaluation configurations and is mitigated in favourable conditions.**

Recommendation regarding empirical studies The above finding regarding the `COM-EMB` parallels the fact that various previous empirical studies of the extrinsic `COM` differ in their characterization of the phenomenon. These often operate in different configurations based on their application of interest, overloading the catchall term "curse of multilinguality" and related concepts. Given our findings, we recommend that future work make their configuration of interest explicit to promote scientific clarity regarding the `COM`.

6 Conclusion

Our work provides theoretical and intrinsic grounding to the curse of multilinguality, which describes the phenomenon of performance degradation in fixed capacity multilingual models with increasing language scale. Specifically, we look at whether multilingual representation spaces can theoretically can be scaled in language coverage while maintaining their quality. We formulate the notion of *perfect multilinguality* for embedding spaces, embodied by two *multilinguality conditions*. Our main contribution is a theoretical result that shows it is necessary and sufficient to scale representation space capacity only logarithmically in the number of languages while maintaining perfect multilinguality. This shows that there is no theoretical curse of multilinguality for embedding space structure. We also demonstrate the empirical curse of multilinguality for embedding space structure for the first time, and show that it is sensitive to data, compute, and evaluation conditions. Our findings contribute to better formulating and understanding the curse of multilinguality, and raise questions regarding bridging the gap between theory and practice.

AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims, assist in the writing of proofs, provide feedback on research methodology or experiments, and implement methods. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, support qualitative and thematic data analysis, interpret results, or formulate mathematical claims, and the remaining tasks of generating synthetic data sets, assisting with translation, and cleaning and reformatting datasets are not applicable to this work. Additionally, we used generative AI tools to create scientific figures or images, create and edit software code, source and search for information, and identify relevant literature. We have reviewed all AI-assisted work. All proofs and content in this paper are written and cross-checked by the authors. LLM-generated code was manually reviewed by authors. Our literature search used AI-assisted tools in combination with manual search and all papers were manually reviewed for relevance. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The details of our proofs are provided in the attached appendix. The code for our experiments is provided in this anonymized repository.

REFERENCES

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively Multilingual Neural Machine Translation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 3874–3884, Minneapolis, Minnesota, jun 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388/.

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. Bouquet : dataset, benchmark and open initiative for universal quality evaluation in translation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27515–27535, Suzhou, China, nov 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1400. URL https://aclanthology.org/2025.emnlp-main.1400/.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges, jul 2019. URL http://arxiv.org/abs/1907.05019.

Sambartha Ray Barman, Andrey Starenky, Sophia Bodnar, Nikhil Narasimhan, and Ashwin Gopinath. The Geometry of Forgetting, mar 2026. URL http://arxiv.org/abs/2604.06222.

Károly Böröczky Jr and Gergely Wintsche. Covering the sphere by equal spherical balls. In *Discrete and Computational Geometry: The Goodman-Pollack Festschrift*, pp. 235–251. Springer, 2003. URL https://link.springer.com/content/pdf/10.1007/978-3-642-55566-4.pdf#page=242.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The Geometry of Multilingual Language Model Representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.9. URL https://aclanthology.org/2022.emnlp-main.9.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.236. URL https://aclanthology.org/2024.emnlp-main.236/.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2475–2485, 2018. URL https://aclanthology.org/D18-1269/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, jul 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

John H Conway, Chaim Goodman-Strauss, and N Sloane. Recent progress in sphere packing. *Current Developments in Mathematics*, 1999(1):37–76, 1999.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. Wmt24++: Expanding the language coverage of wmt24 to 55 languages and dialects, 2025. URL https://arxiv.org/abs/2502.12404.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, 2019. URL https://aclanthology.org/N19-1423/.

David M Eberhard, Gary F Simons, and Charles D Fennig. Ethnologue: Languages of the world, sil international. 2021. *Online version*, 2021. URL http://www.ethnologue.com.

Kenneth C. Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çagatan, Akash Kundu, and Orion Weller. Mmteb: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fc0e3f908a2116ba529ad0a1530a3675-Abstract-Conference.html.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, nov 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Negar Foroutan, Paul Teiletche, Ayush Kumar Tarun, and Antoine Bosselut. Revisiting Multilingual Data Mixtures in Language Model Pretraining, 2025. URL http://arxiv.org/abs/2510.25947.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. URL https://direct.mit.edu/tacl/article-abstract/doi/10.1162/tacl_a_00474/110993.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daniil Gurgurov, Tanja Bäumel, and Tatiana Anikina. Multilingual Large Language Models and Curse of Multilinguality. 2024. URL http://arxiv.org/abs/2406.10602.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. Understanding Cross-Lingual Alignment—A Survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10922–10943, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.649. URL https://aclanthology.org/2024.findings-acl.649/.

Julia Kreutzer, David Vilar, and Artem Sokolov. Bandits Don't Follow Rules: Balancing Multi-Facet Machine Translation with Multi-Armed Bandits. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3190–3204, Punta Cana, Dominican Republic, nov 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.274. URL https://aclanthology.org/2021.findings-emnlp.274/.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d49042a5d49818711c401d34172f9900-Abstract-Datasets_and_Benchmarks.html.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I.-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan Arik, Chen-Yu Lee, and Sayna Ebrahimi. ATLAS: Adaptive Transfer Scaling Laws for Multilingual Pretraining, Finetuning, and Decoding the Curse of Multilinguality. 2026. doi: 10.48550/arXiv.2510.22037. URL http://arxiv.org/abs/2510.22037.

Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. Exploring Alignment in Shared Cross-lingual Spaces. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6326–6348, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.344. URL https://aclanthology.org/2024.acl-long.344/.

Márton Naszódi. On some covering problems in geometry. *Proceedings of the American Mathematical Society*, 144(8):3555–3562, 2016. URL https://pubs.ams.org/journals/proc/2016-144-08/S0002-9939-2016-12992-4.

Koki Okajima and Tsukasa Yoshida. What limits does quantization place on dense top-k retrieval? a theoretical study. 2026. URL https://arxiv.org/abs/2606.11780.

Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Xiang Cao, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, Jean Maillard, Gabriel Mejia Gonzalez, Holger Schwenk, Edan Toledo, Arina Turkatenko, Albert Ventayol-Boada, Rashel Moritz, Alexandre Mourachko, Surya Parimi, Mary Williamson, Shireen Yates, David Dale, and Marta R. Costa-jussà. Omnilingual MT: Machine translation for 1,600 languages. 2026. doi: 10.48550/arXiv.2603.16309. URL https://arxiv.org/abs/2603.16309.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023. URL https://arxiv.org/abs/2311.03658.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3479–3495, Seattle, United States, jul 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.255. URL https://aclanthology.org/2022.naacl-main.255/.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. LAReQA: Language-Agnostic Answer Retrieval from a Multilingual Pool. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5919–5930, Online, nov 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.477. URL https://aclanthology.org/2020.emnlp-main.477/.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. Cross-lingual learning-to-rank with shared representations. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 458–463, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2073. URL https://aclanthology.org/N18-2073/.

Cheril Shah, Yashashree Chandak, and Manan Suri. The Geometry of Multilingual Language Models: An Equality Lens. 2023. doi: 10.48550/arXiv.2305.07839. URL http://arxiv.org/abs/2305.07839.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. Causes and cures for interference in multilingual translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15849–15863, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.883. URL https://aclanthology.org/2023.acl-long.883/.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024. URL https://doi.org/10.1038/s41586-024-07335-x.

Roman Vershynin. High-dimensional probability. *Cambridge Series in Statistical and Probabilistic Mathematics*, 47, 2019.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024. URL https://arxiv.org/abs/2212.03533.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. Balancing Training for Multilingual Neural Machine Translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8526–8537, Online, jul 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.754. URL https://aclanthology.org/2020.acl-main.754/.

Zihao Wang, Hang Yin, Lihui Liu, Hanghang Tong, Yangqiu Song, Ginny Wong, and Simon See. $\mathbb{R}^{2k}$ is theoretically large enough for embedding-based top-k retrieval. 2026. URL https://arxiv.org/abs/2601.20844.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient Vaccine: Investigating and Improving Multi-task Optimization in Massively Multilingual Models. 2020b. doi: 10.48550/arXiv.2010.05874. URL http://arxiv.org/abs/2010.05874.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=k9CzIvzfaA.

Andrea W Wen-Yi and David Mimno. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1124–1131, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.71. URL https://aclanthology.org/2023.emnlp-main.71/.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. Discovering Low-rank Subspaces for Language-agnostic Multilingual Representations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5617–5633, Abu Dhabi, United Arab Emirates, dec 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.379. URL https://aclanthology.org/2022.emnlp-main.379/.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5617–5633, 2022b. URL https://aclanthology.org/2022.emnlp-main.379/.

Ningyu Xu, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. Are structural concepts universal in transformer language models? towards interpretable cross-lingual generalization. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13951–13976, Singapore, dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.931. URL https://aclanthology.org/2023.findings-emnlp.931/.

A PROOF DETAILS

A.1 SPHERICAL CAP BOUND FOR MINIMUM SUFFICIENT DIMENSIONALITY

In this section, we describe and prove the result necessary for the proof in Section 4.3. Specifically, we will show that the number of points that can be placed on a D_L-dimensional sphere with radius r given a minimum and maximum cosine similarity following Equation 1, and given a suitable choice of ρ, grows exponentially in D_L.

Recall that $\gamma < 1$ is the maximum allowed similarity in the non-degenerate space Z, given to us. We can assume that $\gamma > 0$; if observed maximum cosine similarity is negative, the non-degeneracy condition would still hold for a larger γ.

We will use $r = 1$ in our construction for simplicity and without loss of generality. Equation 1 simplifies to

$$\gamma < \cos(u_\ell, u_m) < \rho < 1$$

We will choose ρ such that $\gamma < \rho < 1$ and $\arccos(\rho) < \arccos(\gamma)/2$.

Now, we will show the following.

Lemma 5. *Fix constants $0 < \gamma < \rho < 1$, such that $\arccos(\rho) < \arccos(\gamma)/2$. Then d dimensions are sufficient to pack $\exp(\Omega(d))$ points on the unit sphere S^{d-1} with maximum cosine similarity ρ and minimum cosine similarity γ. That is, there exists a set $\mathcal{M} \subset S^{d-1}$ such that for all distinct $u, u' \in \mathcal{M}$, $\gamma < \langle u, u' \rangle < \rho$, and $|\mathcal{M}| = \exp(\Omega(d))$.*

Proof sketch We will place points on a cap on a multi-dimensional sphere bounded with some angular radius. Since all points are within this cap, the maximum cosine similarity is bounded, letting us satisfy the maximum similarity constraint. Now we will fit points on this cap such that any two points have a minimum angular radius from each other. Imagine each such point making a small cap around itself with angular radius as the minimum required angular separation. Now, consider the maximal set of fitting points in the big cap. The union of the area of the small caps must cover the area of the big cap, since, if there was an uncovered point, that would mean that it was not within the minimum required separation of any other point, and we would be able to add in to our maximal set. Since we can express the union of the area of the small caps in terms of the number of points, we therefore get a lower bound on the maximum number of fitting points in terms of the areas of the two caps. By standard spherical cap results, the area of a spherical cap grows exponentially in the dimension of the sphere. Thus, as long as the angular radius of the big cap is large enough relative to that of the small cap, guaranteed by our choice of ρ, the maximum number of points we can fit that satisfy the above constraints also grow exponentially in the dimension of the sphere, giving us the required result that the sufficient dimension to fit L points is logarithmic in L.

Proof. Let $\theta_{\min} := \arccos(\rho)$ and $\theta_{\max} := \arccos(\gamma)$. Since $\arccos(\rho) < \arccos(\gamma)/2$, and $\gamma > 0$, we have $\theta_{\min} < \theta_{\max}/2 < \pi/4$.

Choose β such that $\theta_{\min} < \beta < \frac{\theta_{\max}}{2} < \pi/4$. Fix a direction $v \in S^{d-1}$ and define the spherical cap around v:

$$\mathcal{C}_\beta(v) := \{u \in S^{d-1} : \angle(u, v) < \beta\}.$$

This means that any two points in this cap u and u' can have angular distance at most $2\beta < \theta_{\max}$ (since each has angular distance at most β from v).

Now consider a set of points $\mathcal{Q}$ inside $\mathcal{C}_\beta(v)$ such that every pair in this set has angular distance greater than $\theta_{\min}$. Equivalently, $\langle u, u' \rangle < \rho$. Thus, for any distinct $u, u' \in \mathcal{Q}$, we will have

$$\begin{aligned} &\theta_{\min} < \angle(u, u') < 2\beta < \theta_{\max} \\ &\implies \cos(\theta_{\max}) < \langle u, u' \rangle < \cos(\theta_{\min}) \\ &\implies \gamma < \langle u, u' \rangle < \rho \end{aligned}$$

Thus, the set $\mathcal{Q}$ satisfies our required angular constraints. In order to prove our lemma, it is enough to show that the maximal size of $\mathcal{Q}$ grows exponentially with d; in other words, that d dimensions are enough to pack some set of points $\mathcal{M}$ with size growing exponentially in d.

Let $\mathcal{M} = \{u_1, \ldots, u_M\} \subset \mathcal{C}_\beta(v)$ with size $M = |\mathcal{M}|$ be a maximal set satisfying $\angle(u_i, u_j) > \theta_{\min}$ for all $i \neq j$.

Now, any point in $\mathcal{C}_\beta(v)$ must be within angular distance $\theta_{\min}$ of some point in $\mathcal{M}$; otherwise this point could be added to $\mathcal{M}$. Therefore, $C_\beta(v)$ is a subset of the union of spherical caps created by defining a minimum angle radius around each point in $\mathcal{M}$. In other words:

$$\mathcal{C}_\beta(v) \subseteq \bigcup_{i=1}^{M} \mathcal{C}_{\theta_{\min}}(u_i).$$

Now we will consider the surface area on the sphere. Taking normalized spherical measure σ, and since σ only depends on angular radius,

$$\sigma(\mathcal{C}_\beta(v)) \leq \sum_{i=1}^{M} \sigma(\mathcal{C}_{\theta_{\min}}(u_i)) = M\sigma(\mathcal{C}_{\theta_{\min}})$$
$$\implies M \geq \frac{\sigma(\mathcal{C}_\beta)}{\sigma(\mathcal{C}_{\theta_{\min}})}.$$

For fixed angular radii, spherical cap measure scales exponentially in d as per standard cap measure results (Naszódi, 2016). Specifically, for $0 < \alpha < \pi/2$, the normalized measure of a spherical cap of angular radius α in S^{d-1} given by Naszódi (2016, Lemma 3.4, Eq. (7)), attributed to Böröczky Jr & Wintsche (2003), satisfies

$$\sigma(\mathcal{C}_\alpha) = \Theta_\alpha\left(\frac{\sin^{d-1}(\alpha)}{\sqrt{d}}\right)$$

Applying this with $\alpha = \beta < \pi/2$ and $\alpha = \theta_{\min} < \pi/2$ gives

$$M \geq \frac{\sigma(\mathcal{C}_\beta)}{\sigma(\mathcal{C}_{\theta_{\min}})} = \Omega\left(\left(\frac{\sin\beta}{\sin\theta_{\min}}\right)^{d-1}\right),$$

Since $\beta > \theta_{\min}$, we have $\sin\beta > \sin\theta_{\min}$. Since β and $\theta_{\min}$ are fixed, there exists a constant $C_M > 0$, independent of d, such that

$$M \geq \exp(C_M d) \implies M = \exp(\Omega(d))$$

That is, there exists some set $\mathcal{M}$ containing $\exp(\Omega(d))$ points whose pairwise cosine similarities lie between γ and ρ, on a d-dimensional sphere.

In other words, the minimum sufficient dimension to fit L points whose pairwise cosine similarities lie between γ and ρ on $S^{d-1} \subset \mathbb{R}^d$ is $d = O(\log L)$. □

A.2 Spherical cap bound for necessary dimensionality

In this section, we describe and prove the result necessary for the proof in Section 4.4. Specifically, we will show that the dimensionality required to place L fixed-norm points on a D_{X^L}-dimensional sphere with some minimum angular separation grows at least logarithmically in L.

Given $\gamma < 1$ as the maximum allowed cosine similarity, we will use minimum angular separation $\theta = \arccos(\gamma)$, with $0 < \theta < \pi$. Now, we will show the following:

Lemma 6. *Fix a constant $0 < \theta < \pi$. Suppose $\mathcal{M} \subset S^{d-1}$ is a set of $M = |\mathcal{M}|$ points on a unit sphere with pairwise angular separation greater than θ. That is, for all distinct $u, u' \in \mathcal{M}$, $\angle(u, u') > \theta$. Then there exists a constant $A_\theta > 0$, independent of M and d, such that $M \leq \exp(A_\theta d)$. Equivalently, any such packing of M points requires $d = \Omega(\log M)$.*

Proof sketch We will place M points on a multidimensional sphere. Think of each point in $\mathcal{M}$ forming a spherical cap on the sphere with angular radius as half of the minimum required angular separation. Since each pair of points must have the required angular separation, these spherical caps

must be disjoint: if a point on the sphere belonged to caps associated with two points in $\mathcal{M}$, those two points would not satisfy the criterion by triangle inequality. Thus, the surface area of the whole sphere must be larger than the union of the surface area of the spherical caps. Since we have M spherical caps and can express the union of the surface area of the spherical caps in terms of M, we get an upper bound on M in terms of an inverse spherical cap measure, which scales exponentially in the dimension of the sphere. Thus, we get that the minimum dimension that can fit M points is at least logarithmic in M.

Proof. Let $\mathcal{M} = \{u_1, \ldots, u_M\} \subset S^{d-1}$ satisfy $\angle(u_i, u_j) > \theta \qquad$ for all $i \neq j$.

Around each point u_i, place a spherical cap of angular radius $\theta/2$:

$$\mathcal{C}_{\theta/2}(u_i) := \{u \in S^{d-1} : \angle(u, u_i) \leq \theta/2\}.$$

These caps are disjoint, since if some point w belonged to both $\mathcal{C}_{\theta/2}(u_i)$ and $\mathcal{C}_{\theta/2}(u_j)$, then by the triangle inequality, $\angle(u_i, u_j) \leq \angle(u_i, w) + \angle(w, u_j) \leq \theta$, contradicting the separation constraint.

Taking normalized spherical measure σ, we have

$$\sum_{i=1}^{M} \sigma(\mathcal{C}_{\theta/2}(u_i)) \leq \sigma(S^{d-1}) = 1$$

Since σ only depends on angular radius,

$$M\sigma(\mathcal{C}_{\theta/2}) \leq 1 \implies M \leq \frac{1}{\sigma(\mathcal{C}_{\theta/2})}.$$

For fixed angular radius, spherical cap measures are at least exponentially small in d. More concretely, for $0 < \alpha < \pi/2$, the normalized measure of a spherical cap of angular radius α in S^{d-1}, as given by Naszódi (2016, Lemma 3.4, Eq. (7)), attributed to Böröczky Jr & Wintsche (2003) satisfies

$$\sigma(\mathcal{C}_\alpha) \geq \frac{1}{\sqrt{2\pi d}} \sin^{d-1}(\alpha).$$

Applying this with $\alpha = \theta/2 < \pi/2$ gives

$$M \leq \frac{1}{\sigma(\mathcal{C}_{\theta/2})} \leq \sqrt{2\pi d} \left(\frac{1}{\sin(\theta/2)}\right)^{d-1}$$
$$\implies \log M \leq \frac{1}{2}\log(2\pi d) + (d-1)\log\left(\frac{1}{\sin(\theta/2)}\right)$$

Since θ is fixed, $\sin(\theta/2)$ is a fixed constant in $(0, 1)$. Since $\log(2\pi d) = o(d)$, there exists a constant $A_\theta > 0$ depending only on θ, such that

$$\log M \leq A_\theta d \implies d \geq \frac{1}{A_\theta} \log M \implies d = \Omega(\log M)$$

Therefore any set of L points on $S^{d-1} \subset \mathbb{R}^d$ with fixed pairwise cosine separation γ requires $d = \Omega(\log L)$. □

A.3 Fitting 7000 languages in 30 extra dimensions

Our constructive proof in Section 4 and Appendix A.1 constructs a language offset space U with dimensionality D_L, and shows that that $D_L = O(\log L)$ dimensions are sufficient to fit L points on a D_L-dimensional unit sphere given minimum cosine similarity γ and maximum cosine similarity ρ.

However, this bounds hides a constant which may be of concern for practical considerations. In this section, we prove the following:

Lemma 7. *It is possible to fit $L = 7000$ points on a 30-dimensional sphere with minimum pairwise cosine similarity $\gamma = 0.9$ and maximum cosine similarity $\rho = 0.99$.*

γ is set based on the choice of 0.1 as a standard scale of cosine separation for text embedding models (Weller et al., 2026). Note that our proof requires $\arccos(\rho) < \arccos(\gamma)/2$ which holds for this choice of ρ.

Sketch of proof We will construct each point having 1 shared dimension with an appropriate length, which will guarantee minimum similarity. We will use the remaining 29 dimensions to guarantee difference, i.e. ensure that any two points have atmost the maximum similarity. In particular, we will prove this statement using an existence argument: if we choose 7000 points at random on a sphere in 29 dimensions, the probability that no pair of points crosses a particular threshold of similarity is non-zero. The existence of such a configuration gives us the above result.

Proof. Consider the space $\mathbb{R}^{30} = \text{span}(e_0) \oplus \mathbb{R}^{29}$, where e_0 is a unit vector orthogonal to $\mathbb{R}^{29}$. Now, construct

$$x_i = \sqrt{0.945}\, e_0 + \sqrt{0.055}\, y_i, \qquad y_i \in S^{28} \subset \mathbb{R}^{29}.$$

Then $\|x_i\|_2 = 1$ and since all vectors have unit norm, cosine similarity is equal to the inner product. Now, we consider the similarity between any two vectors x_i and x_j:

$$\langle x_i, x_j \rangle = 0.945 + 0.055 \langle y_i, y_j \rangle.$$

Therefore, if $|\langle y_i, y_j \rangle| \leq \frac{9}{11}$, for all $i \neq j$, then

$$\begin{aligned} \langle x_i, x_j \rangle &\in \left[0.945 - 0.055 \cdot \frac{9}{11},\, 0.945 + 0.055 \cdot \frac{9}{11} \right] \\ &= [0.9, 0.99]. \end{aligned}$$

It remains to show that such $y_1, \ldots, y_{7000}$ exist. Suppose that

$$y_1, \ldots, y_{7000} \overset{\text{i.i.d.}}{\sim} \text{Unif}(S^{28}).$$

Fix any pair $i \neq j$. Since the distribution on the sphere is rotationally invariant, we may fix one point $Y = e_1 = [1, 0, 0, ...]$ and choose only the second point $Z \sim \text{Unif}(S^{28})$. The distribution of the inner product is unchanged. Thus, the dot product of Y and Z is distributed identically to just the first coordinate of Z:

$$\langle Y, Z \rangle \overset{d}{=} Z_1,$$

Z_1 is the first coordinate of a uniformly random point on S^{28}. The distribution of the coordinate's value can be expressed in terms of a Beta distribution, as follows,[6]

$$\langle Y, Z \rangle^2 \sim \text{Beta}\left(\frac{1}{2}, 14 \right).$$

Therefore, for any fixed pair,

$$p := \Pr\left(|\langle Y, Z \rangle|^2 > \left(\frac{9}{11} \right)^2 \right) \tag{3}$$

$$= I_{40/121}\left(14, \frac{1}{2} \right) \approx 3.3465 \times 10^{-8}, \tag{4}$$

where $I_x(a, b)$ denotes the regularized incomplete beta function.

[6] Briefly, this is because of the following: we generate a random point by choosing 29 dimensions independently and normalizing all coordinates to result in unit norm. The first coordinate can then be expressed as $Z_1^2 \sim \frac{A}{A+B}$ where $A \sim \chi_1^2$ and $B \sim \chi_{28}^2$ and A and B are independent. This standardly gives us $Z_1^2 \sim \text{Beta}\left(\frac{1}{2}, 14\right)$.

There are $\binom{7000}{2}$ pairs. By the union bound,

$$\Pr\left(\exists\, i<j : |\langle y_i, y_j\rangle| > \frac{9}{11}\right) \leq \binom{7000}{2} p \approx 0.8198 < 1.$$

Hence,

$$\Pr\left(|\langle y_i, y_j\rangle| \leq \frac{9}{11} \text{ for all } i \neq j\right) > 0.$$

Therefore, there exists a realization $y_1, \ldots, y_{7000}$ satisfying the required bound, and the corresponding vectors $x_1, \ldots, x_{7000} \in \mathbb{R}^{30}$ have pairwise cosine similarities in $[0.9, 0.99]$. That is, it is possible to fit 7000 points, or language offsets, satisfying the constraints from our proof such that the resulting space would maintain the multilinguality conditions, in 30 additional dimensions. □

A.4 Fixed versus variable norm assumptions

Our proof of Theorem 1 assumes a fixed-norm concept space Z and multilingual space X^L. This proof is extendable to the variable norm case. We provide a brief sketch of the extension below.

Extending Theorem 1.1 We will still use fixed-norm U in our construction, assuming variable-norm Z, resulting in variable-norm X^L constructed as $x_{c,\ell} = z_c \oplus \|z_c\| u_\ell$. Since $\|u_\ell\| = r$, we have $\|x_{c,\ell}\| = \|z_c\|\sqrt{1+r^2}$, and therefore the similarity decomposition in Equation 2 still works out as

$$\begin{aligned}\cos(x_{c,\ell}, x_{d,m}) &= \frac{\langle z_c, z_d\rangle + \|z_c\| \cdot \|z_d\| \langle u_\ell, u_m\rangle}{\|z_c\| \cdot \|z_d\|(1+r^2)} \\ &= \frac{\cos(z_c, z_d) + \langle u_\ell, u_m\rangle}{1+r^2}\end{aligned}$$

Now, our proofs showing that X^L upholds the multilinguality conditions and is non-degenerate follow as before. The proof of the upper-bound on the sufficient dimensionality of U in Appendix A.1 also works as-is, since we used fixed-norm U.

Extending Theorem 1.2 Given that X^L is variable-norm, we can simply normalize each point in it and obtain unit-norm space $\widehat{X}^L$. This operation does not change the dimension of the space and maintains cosine similarities. Now, we can lower-bound the dimension of $\widehat{X}^L$ exactly as in Section 4.4 and this bound holds for variable norm X^L.

B Additional details and results for the empirical curse

B.1 Metrics for multilinguality conditions

We provide more details here on metrics used in Section 5. We use multiparallel data, and so the concept set is identical across languages.

Monolingual structure As per Definition 1, the monolingual structure condition is concerned with the quality of monolingual relationships for a given language, and uses a theoretical ideal concept space as reference. Given a reference space Z and target space X, we use ***k-nearest-neighbour overlap (MS−NNO)*** to measure the quality of monolingual neighbourhoods in the subspace $X_t \subset X$, for each language t. Given $\mathcal{T}$ is the set of target languages, $\mathcal{C}$ is the set of language-agnostic concepts, $\mathcal{N}_X^t(c,k)$ is the neighbourhood of k closest concepts for c in language t in space X, and $\mathcal{N}_Z(c,k)$ is the reference concept neighbourhood,

$$\mathrm{NNO}(X, Z, k) = \frac{1}{|\mathcal{T}| \cdot |\mathcal{C}| \cdot k} \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} \sum_{n^t \in \mathcal{N}_Z(c,k)} \mathbf{1}[n^t \in \mathcal{N}_X^t(c,k)]$$

We consider the English concept space from a strong reference model as our theoretical ideal space Z and use mappings from multiparallel data to check set membership as in above.

Note that monolingual text embedding quality in the field of retrieval is measured by `recall@k` among other metrics, given query and relevance judgments. Our evaluation is an adjustment given the lack of multiparallel retrieval datasets and relevance judgments at our required language scale, and therefore treating monolingual neighbourhoods from a strong model as the gold target set.

We use $k = 20$ for reporting results, and observe similar trends for varying $k = 5, 10, 20, 50$.

We also look at language mean over *normalized masked language modeling loss* on a monolingual test set per language (`MS-MLM`) as an indirect measure of monolingual structure quality.

Cross-lingual alignment Drawing from Definition 2, and similar to previous work (Roy et al., 2020; Hämmerl et al., 2024), we measure cross-lingual alignment in embedding space structure by checking for each language ℓ_1, whether, for each concept and each language ℓ_2, the ℓ_2 equivalent of that concept is closer to the ℓ_1 equivalent than other ℓ_2 concepts (weak view; `CLA-WV`) or any any other concepts regardless of language (strong view; `CLA-SV`). Given the set of target languages $\mathcal{T}$, $\ell_s, \ell_t \in \mathcal{T}$, and $\mathcal{G}$ as the comparison group depending on strong or weak view,

$$\text{CLA} = \frac{1}{|\mathcal{C}| \cdot |\mathcal{T}|(|\mathcal{T}| - 1)} \sum_{\ell_s} \sum_{\ell_t \neq \ell_s} \sum_{c} \mathbf{1}[\text{argmax}_{x' \in G}(\cos(x_{c,\ell_s}, x')) = x_{c,\ell_t}] \,.$$

where $\mathcal{G} = \{x_{c',\ell_t} | c' \neq c\} \cup \{x_{c,\ell_t}\}$ in the weak view and $\mathcal{G} = \{x_{c',\ell_m} | c' \neq c, \ell_m \in \mathcal{T}\} \cup \{x_{c,\ell_t}\}$ for the strong view. We study both the weak view and strong view in our experiments and find similar trends. We report the former. Note that this evaluation is equivalent to measuring performance on a parallel text mining task.

B.2 Further experimental details

Models For the reference model used for monolingual structure metric, we use a strong English text embedding model: `e5-large-v2`[7] (Wang et al., 2024), and `Llama-3.1-8B-Instruct`[8] (Grattafiori et al., 2024) with last token pooling. We obtain consistent trends. Reported results use the former.

Languages Training languages were chosen in priority of token availability in the `MADLAD-400` corpus, as well as availability in evaluation datasets. We evaluate on all languages covered in the training language group: `FLORES+` (90 languages) (Goyal et al., 2022; Team et al., 2024), `BOUQuET` (90 languages) (Andrews et al., 2025; Omnilingual MT Team et al., 2026), and `WMT24++` (55 languages) (Deutsch et al., 2025).

B.3 Additional results

Other datasets We show results for two other multiparallel datasets in Figure 4 and Figure 5. We also show `CLA-SV`, which shows similar trends for `WMT24++`, but is generally too low for `FLORES+` and `BOUQuET` to show meaningful trends.

These show similar trends as shown in Figure 3.

Non-degeneracy We plot the mean and standard deviation of pairwise cosine similarity across models for `BOUQuET` in Figure 6 (similar for other datasets). This is stable, although generally high, which is evidence of high anisotropy in the space.

Maintaining token count In Section 5, we discussed that the four settings in which metrics stay stable or improve are generally the settings that maintain **token count for the target languages of interest**. We elaborate on this here.

With `fixed-compute`, the token count for each language decreases as we add more languages; thus, mean conditions deteriorate with `aggregate-all` for both types of sampling. However, when we have `fixed-compute` and only look at high-resource languages in

[7] `huggingface.co/intfloat/e5-large-v2`

[8] `huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

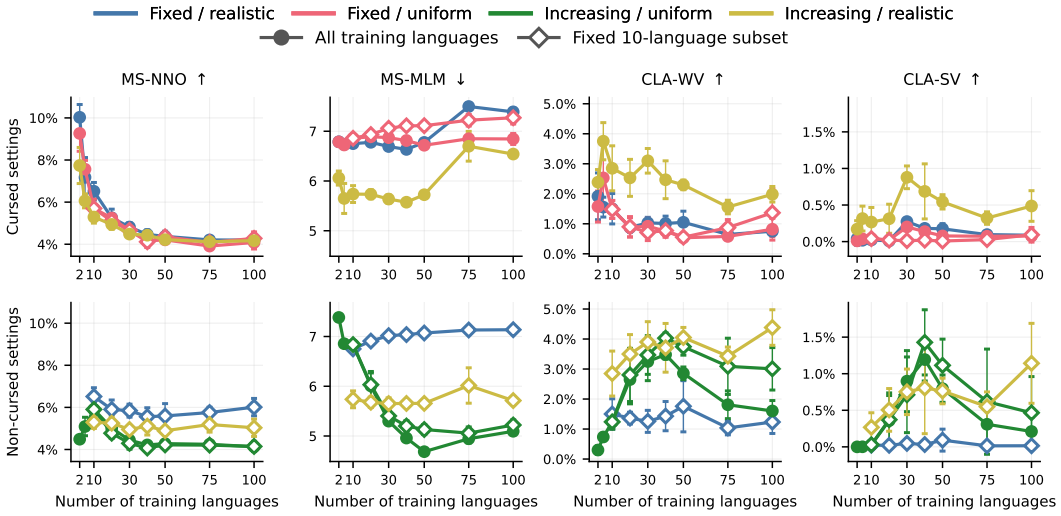


Figure 4: `MS-NNO`, `MS-MLM`, and `CLA` on `FLORES+` for encoder models with as the number of training languages increases, in eight configurations. The top and bottom rows show cursed and non-cursed configurations respectively. Arrows indicate whether higher or lower metric values are better. Error bars show standard deviation over five model seeds.

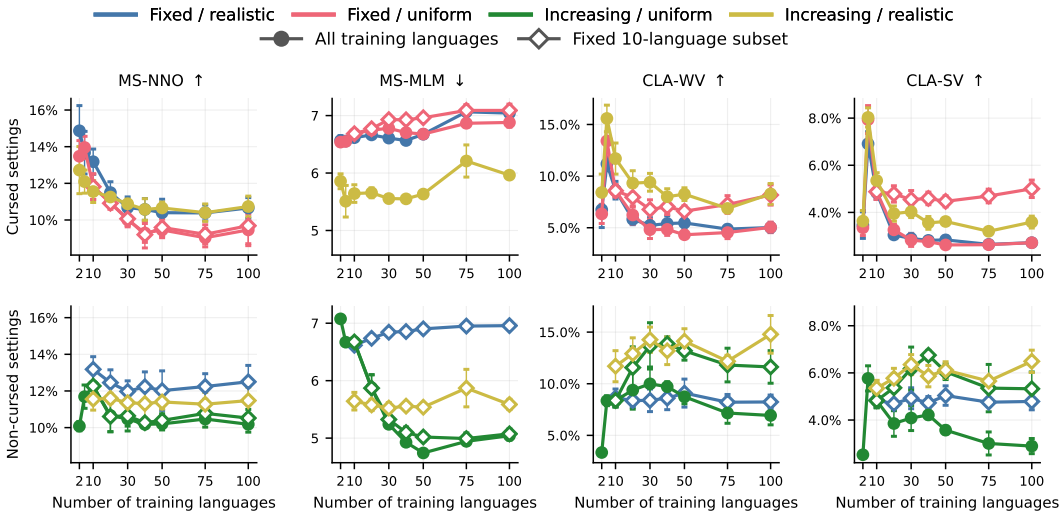


Figure 5: `MS-NNO`, `MS-MLM`, and `CLA` on `WMT24++` for encoder models with as the number of training languages increases, in eight configurations. The top and bottom rows show cursed and non-cursed configurations respectively. Arrows indicate whether higher or lower metric values are better. Error bars show standard deviation over five model seeds.

`aggregate-fixed` with `realistic-sampling`, this effect is somewhat softened since subsequent LRLs consume a smaller part of token budget. With `increasing-compute`, token count for target languages remains unchanged with more training languages with `uniform-sampling`, and metrics stay stable for both types of aggregation. With `increasing-compute` and `realistic-sampling`, similar to above, HRLs are able to maintain token count, allowing metrics to stay stable for `aggregate-fixed`, whereas `aggregate-all` over `realistic-sampling` means that that performance means are driven down by low-resource languages. `COM-EMB` trends may potentially also be affected by other things such as model scale, optimization strategies, and data quality, the impact of which we leave to future work to investigate.

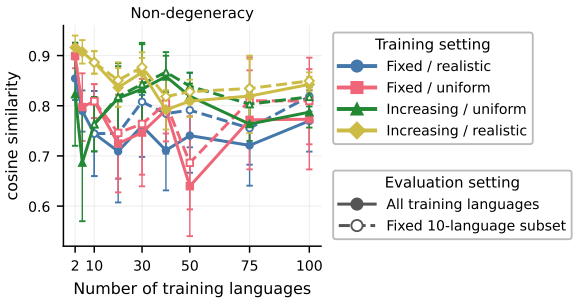


Figure 6: Mean pairwise similarity on BOUQuET across training and evaluation settings. Error bars show standard deviation over five seeds.

B.4 Additional exploration of language subspace use and dimensionality

Our construction achieving perfect multilinguality in Section 4.3 uses compact language offsets that are reused across concepts. This idea of low-dimensional and reusable language feature representations is of interest to several works (Park et al., 2023; Xie et al., 2022a; Chang et al., 2022)

We want to investigate whether, similarly to our constructive solution, language identity is compactly and reusably represented in our trained models that maintain the multilinguality conditions, and whether this differs by language scale.

Defining language subspace efficiency We quantify the efficiency of the model regarding language information as the effective dimensionality of the *language subspace*, or the space of the language offsets or language-expressing directions.

Constructing the language subspace The language subspace is constructed as the space spanned by the difference vectors between pairs of translation equivalents, i.e. points expressing an identical concept in two different languages Formally,

$$U := \text{span}\{x_{c,l} - x_{c,l'} | c \in \mathcal{C}, l, l' \in \mathcal{L}, l \neq l'\}$$

Here, $x_{c,l}$ is unit-normalized. We are now interested in the effective dimensionality of U.

Computing effective dimensionality We use participation ratio as a measure of effective dimensionality, in line with previous work (Barman et al., 2026). Let M stack the language offsets $x_{c,l} - x_{c,l'}$ for all c and unordered pairs (l, l'), and let s_i be the singular values of M after row-centering. We use

$$d_{\text{lang}} = \frac{(\sum_i s_i)^2}{\sum_i s_i^2}.$$

This behaves like a soft dimension count. Intuitively, if all singular values of active directions are equal, it returns the number of active directions. Instead, if the spectrum is dominated by a few directions, it returns a smaller value. We then report $\text{effdim}_{\text{lang}} := \frac{d_{\text{lang}}}{D} \in (0, 1]$, where D is the ambient dimensionality of the space.

Trends We are interested in $\text{effdim}_{\text{lang}}$ for models that maintain the multilinguality conditions. We construct the language subspace over all training languages, and so we only look at `aggregate-all` non-cursed configurations. There is only one such configuration. See Figure 7 for trend of $\text{effdim}_{\text{lang}}$ on BOUQuET (similar for other datasets) on this configuration.

We find that the model consumes considerably more effective dimensionality to represent language information with higher language count, going up to 50% of ambient dimensionality.

Interaction between language and concept subspace The above experiment raises a natural next point of comparison with our constructive solution. Namely: is the language subspace orthogonal

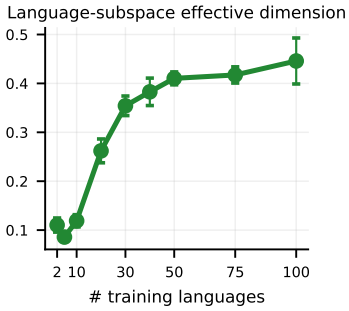


Figure 7: Effective dimensionality of the language subspace of all training languages on BOUQuET for non-cursed training setting. Error bars show standard deviation over five seeds.

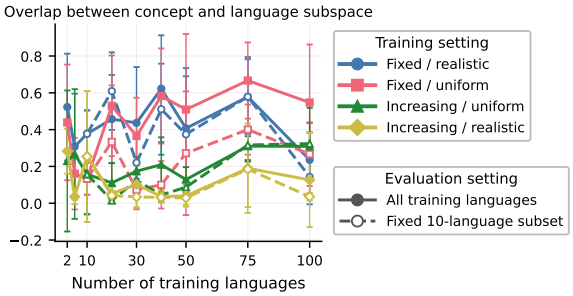


Figure 8: Overlap between concept and language subspace, using BOUQuET.

to the concept subspace, or are semantic and language directions entangled? Note that in our constructive solution, the concept and language subspaces are orthogonal by construction.

We construct the language subspace as above, and analogously construct a concept subspace using same-language different-concept pairs. We then consider the minimal orthonormal bases of both subspaces that capture $k = 20\%$ of energy with dimensionality d_c and d_l, and quantify the overlap between these directions using principal angle between the subspaces, adjusted for random overlap. Given $m = \min(d_c, d_l)$, we formulate this as:

$$O_{\text{CL}} = \frac{\frac{1}{m}\sum_{i=1}^{m}\sigma_i^2 - \frac{\max(d_c,d_l)}{D}}{1 - \frac{\max(d_c,d_l)}{D}}.$$

where $O_{\text{CL}} \leq 1$ and higher values mean more overlap. Motivating this briefly: it can be shown that if $V_{\text{concept}} \in \mathbb{R}^{D\times d_c}$ and $V_{\text{lang}} \in \mathbb{R}^{D\times d_l}$ are the orthonormal bases capturing k proportion of energy, then, $\sigma_i = \text{svd}(V_{\text{concept}}^{\top} V_{\text{lang}})_i = \cos\theta_i$, where θ_i are principal angles between V_{concept} and V_{lang}. It can also be shown that the expected overlap between two subspaces with randomly chosen dimensions given a total dimensionality D is:

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\cos^2\theta_i\right] = \frac{\max(d_c, d_l)}{D}$$

So our metric O_{CL} is a measure of the squared cosines of principal angles between the two subspaces of interest, adjusted for random.

We plot this metric for all settings, shown in Figure 8. We observe that the overlap is generally high across settings and number of training languages, including non-cursed settings. This is in contrast to the ideal constructive solution, where the concept and language subspace are orthogonal, with zero overlap.

In general, our preliminary experiments indicate that empirical embedding spaces at our small scale do not behave similar to our ideal constructive solution even in non-cursed conditions. We leave it to future work to explore the underlying mechanisms of the (non-)curse of multilinguality for embedding space structure in conducive settings and at scale, for various architectures and training configurations.

C Limitations

Perfect multilinguality Our formulation of perfect multilinguality for multilingual representation spaces is most directly relevant for applications such as embedding-based retrieval. Different multilingual representation spaces, such as intermediate representations of decoder models, may require

adaptation based on needs from different downstream application scenarios. Further, this conception uses a simplified scenario consisting of a global concept set with global semantics across languages. This allows the cost of concept semantics to stay constant as language coverage grows in theory. In practice, different languages may introduce language-specific concepts, or have differing semantic relationships between concepts, introducing language-specific needs and additional dimensionality costs. We leave this to future work to incorporate into a theoretical model.

Model scale Our empirical studies are for small encoder models given the controlled nature and language scale of our study. We leave it to future work to study the impact of model scaling and different architectures on the curse of multilinguality for embedding space structure.