

Исследование внутренних представлений моделей

Иванова Анастасия, Лебедева Анна, Курдюков Никита, Гришина Екатерина

30 августа 2024 г.

1 Цели и пайплайн

Цель проекта - исследовать геометрические характеристики эмбеддингов моделей для разных языков.

Пайплайн:

- Делаем forward на наборе предложений на одном языке.
- Сохраняем эмбеддинги разных слоев.
- Считаем геометрические характеристики эмбеддингов.

Мы используем датасет FLORES – набор параллельных переводов на >100 языках. Эксперименты проводились с декодерами Phi3.5, Phi3 и энкодером Bert.

2 Внутренняя размерность

Внутреннюю размерность данных можно рассматривать как количество переменных, необходимых для минимального представления данных. Мы использовали алгоритм подсчёта внутренней размерности, реализованный в LLM-Microscope.

Наши эксперименты показали, что ID для языков из типологической выборки имеет схожий тренд вне зависимости от языка 1. Для языков, использующих две письменности ID в среднем выше, если язык записан латиницей, а не арабицей 2. Еще мы измерили ID для текстов на английском с правильной и ошибочной грамматикой из [датасета](#). Внутренняя размерность неграмматичных текстов оказалась выше 3.

Также было сделано сравнение внутренней размерности языков с разной морфологической сложностью (среднее количество морфем на слово) и малоресурсных и высокоресурсных языков (по парам близкородственных языков), однако корреляции ID с этими характеристиками не было найдено 4.

Мы исследовали зависимость между качеством модели и внутренней размерностью эмбеддингов. На графике 5 видно, что ID эмбеддингов языков линейно связана с качеством модели. Чем меньше ID, тем выше качество на MMLU.

Если сравнивать внутреннюю размерность энкодерных и декодерных моделей, то Encoder (BERT) имеет меньшую ID в сравнении с Decoder (Phi-3.5) 6.

3 KL дивергенция

KL дивергенция – показывает расстояние между двумя распределениями.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Ungrammatical	Standard
The students studies for the exam.	The students study for the exam.
The car need to be repaired.	The car needs to be repaired.

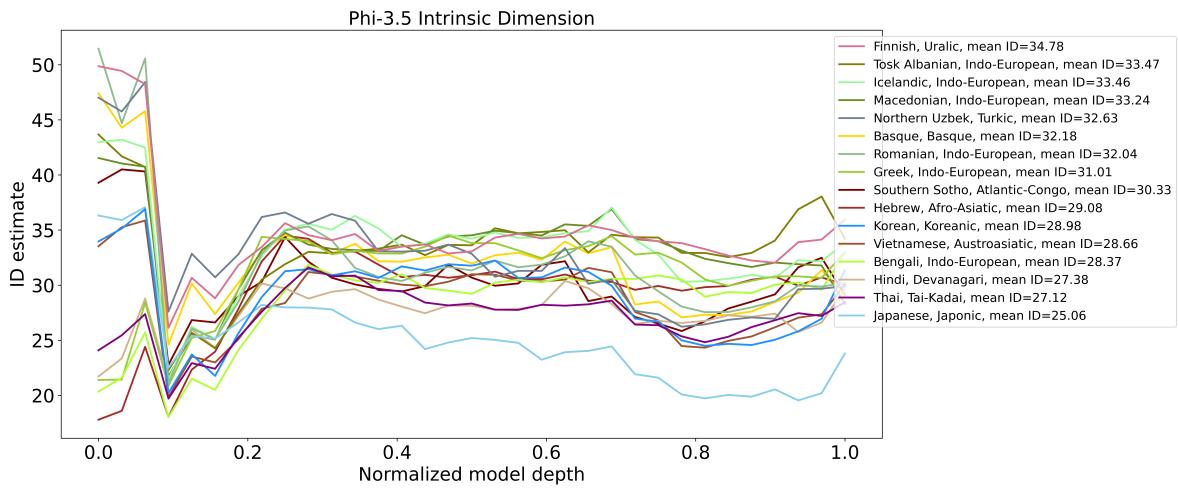


Рис. 1

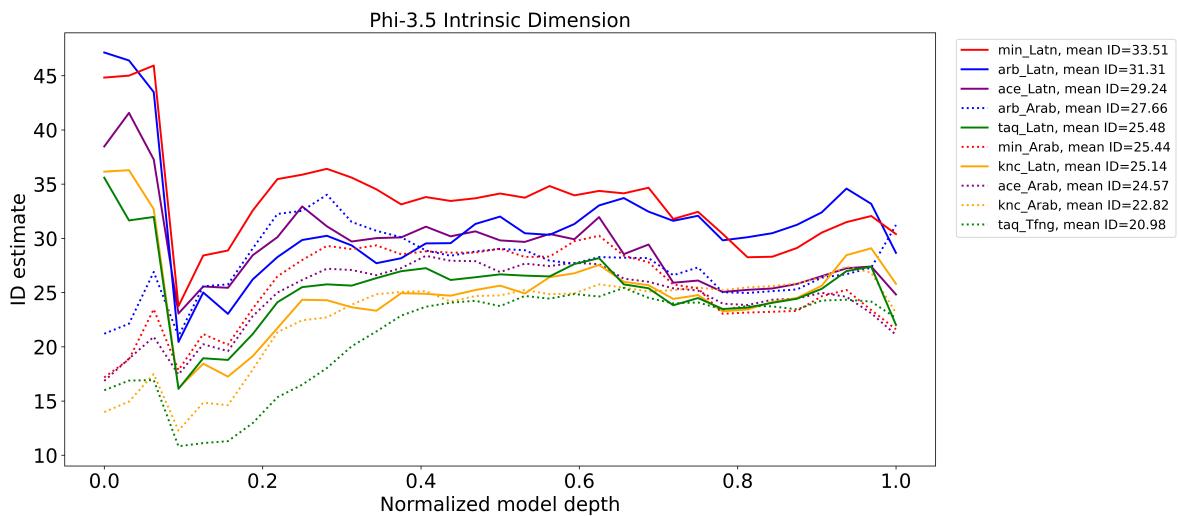


Рис. 2

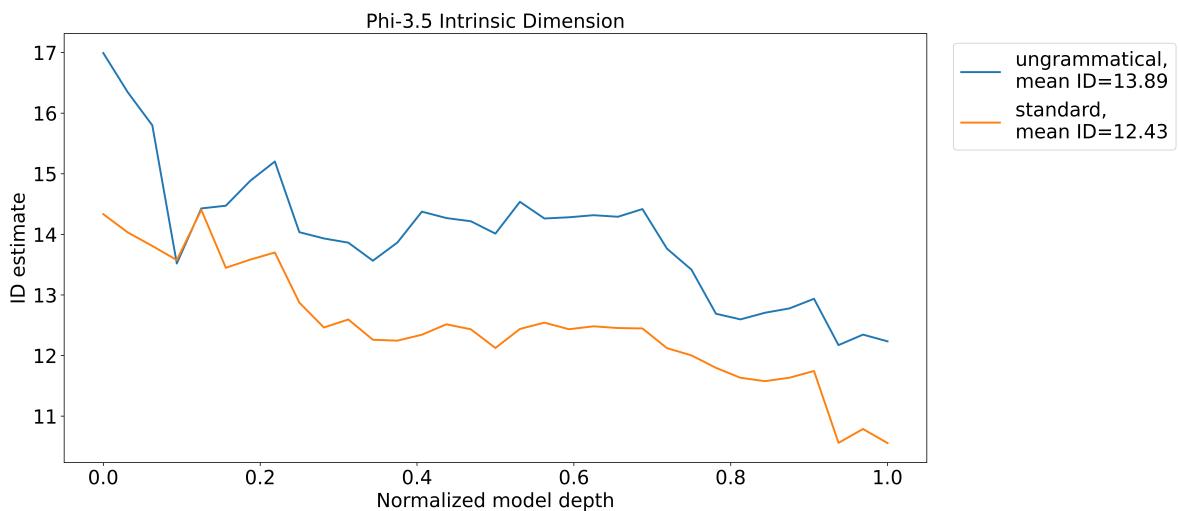


Рис. 3

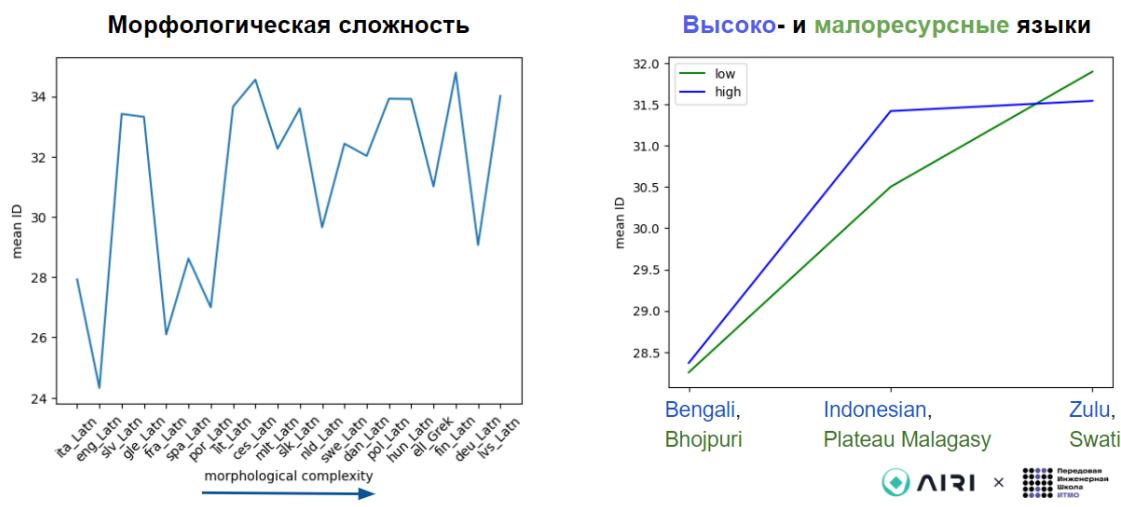


Рис. 4

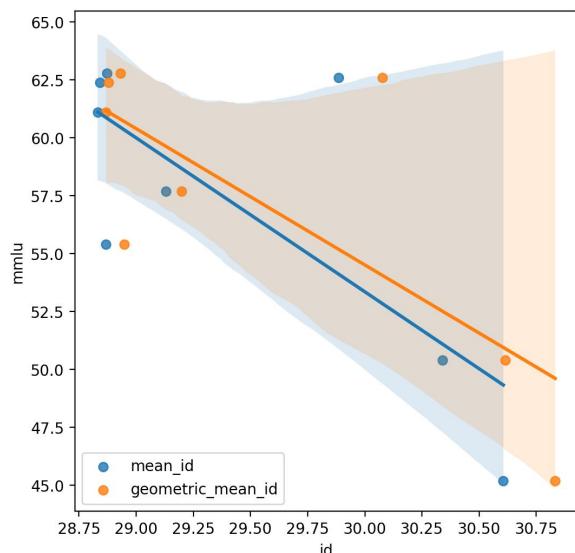


Рис. 5

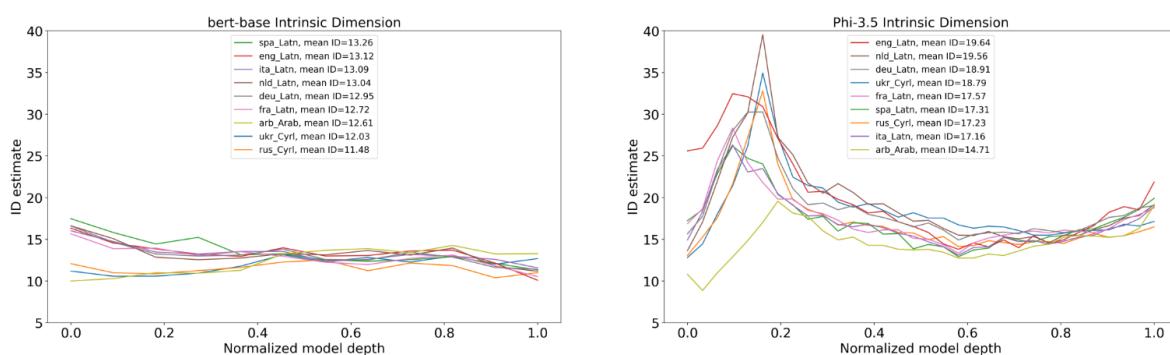


Рис. 6

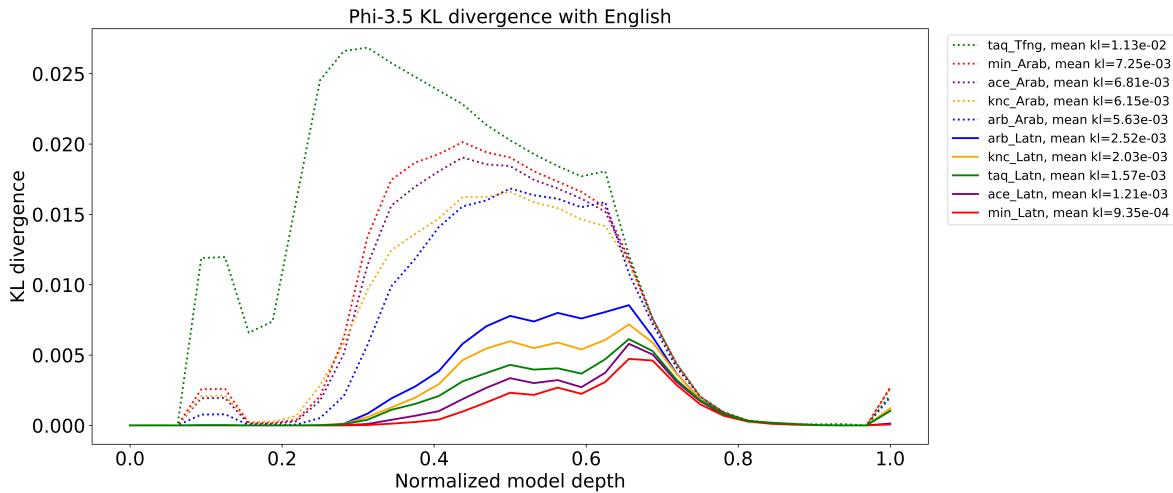


Рис. 7

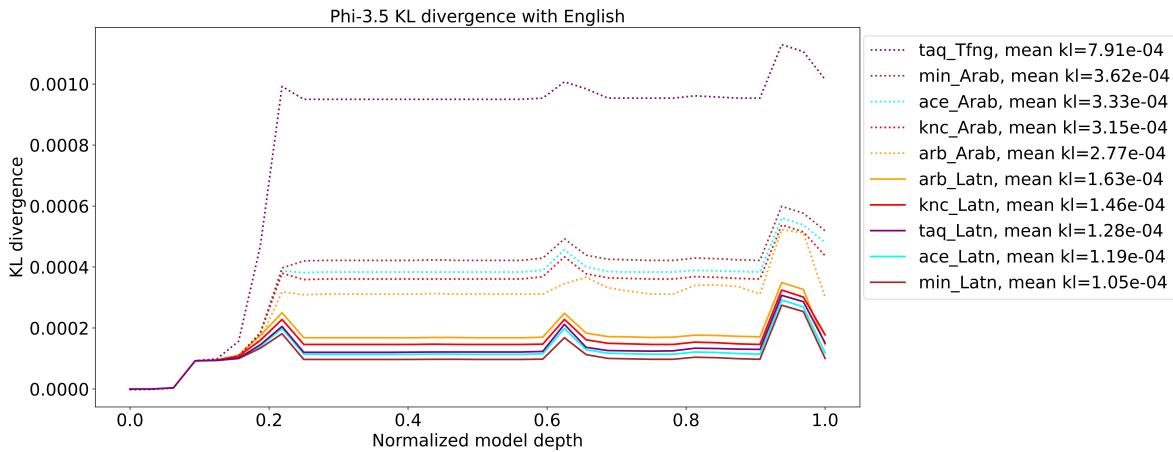


Рис. 8

Мы измеряем KL между разными языками и английским. Можно вычислять KL, получая средний эмбеддинг по предложению (7) или дополняя нулями предложения до нужной длины и не агрегируя эмбеддинги (8). Графики отличаются, но в любом случае языки одинаково располагаются относительно друг друга. Мы можем сделать вывод, что для языков, использующих две письменности KL дивергенция с английским выше, если текст записывается латиницей.

4 Анизотропия

Анизотропия - свойство векторного пространства. Геометрически эмбеддинги в таком пространстве будут лежать в узком конусе. Мы хотим оценить, насколько узок этот конус. Это можно сделать через главную компоненту - оценить сколько дисперсии она объясняет:

$$anisotropy(X) = \frac{\sigma_{max}^2(X)}{\sum_i \sigma_i^2(X)}$$

Чем больше анизотропия, тем более похожи все эмбеддинги между собой.

Для вычисления анизотропии в библиотеке llm-microscope использовалось SVD, асимптотика которого $O(n^3)$, n - размер матрицы X . Однако можно вычислять анизотропию быстрее, используя степенной метод для вычисления старшего сингулярного числа и фробениусову норму для вычисления знаменателя. Например, на матрице размера 1000×3000 мы получили ускорение в 15 раз (6 ms работает степенной метод и 92 ms SVD).

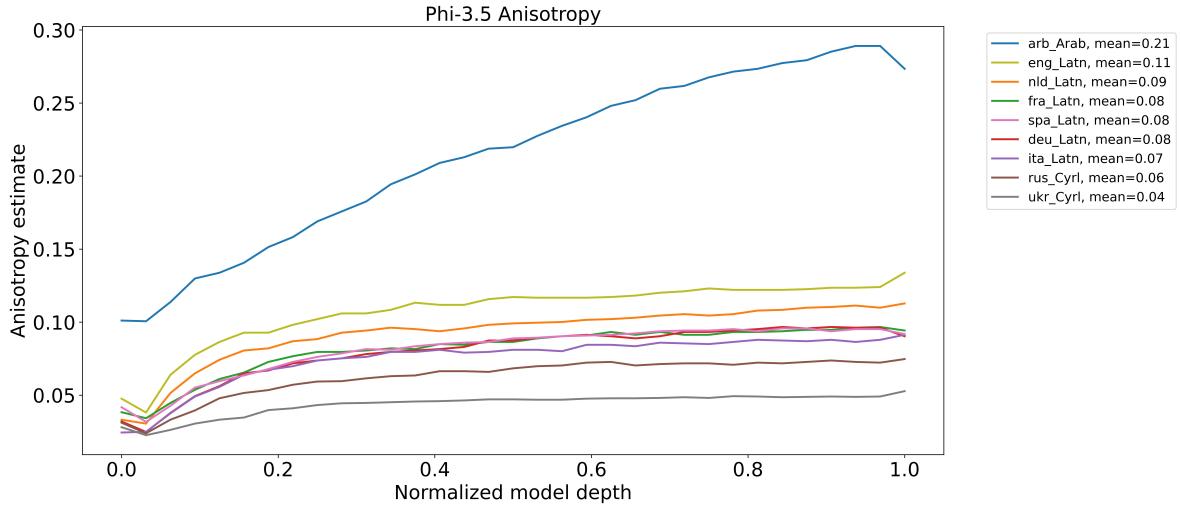


Рис. 9: Анизотропия необученной модели.

$$anisotropy(X) = \frac{\sigma_{max}^2(X)}{\sum_i \sigma_i^2(X)} = \frac{\sigma_{max}^2(X)}{\|X\|_F^2}$$

Power method

```

 $u = randn(n)$ 
for 1 ... n_iters do
     $u = Xu / \|Xu\|_2$ 
     $u = u^t X / \|u^t X\|_2$ 
end for
 $\sigma_{max} = \|Xu\|_2$ 
```

Анизотропия обученной и необученной декодерной модели Phi-3.5 заметно отличается 9 и 10. Из чего можно сделать вывод, что высокая анизотропия - свойство не архитектуры, а обученной модели. График для обученных декодеров всегда имеет схожую форму с высокой анизотропией, начиная с ~ 3 слоя, у энкодеров (Bert), наоборот, анизотропия почти нулевая на всех слоях 13, это наблюдение согласуется с [1, 2].

Для языков, использующих две письменности: анизотропия меньше, если текст записан арабицей 11.

Также была гипотеза, что может отличаться внутренняя размерность и анизотропия моделей, обученных на естественных и синтетических текстах (Phi-3 и Phi-3.5), но она не подтвердилась, паттерны совпадают 12.

Мы также проверили связь ВРС с качеством на MMLU и внутренней размерностью, но не нашли корреляции 14.

5 Выводы

- Эмбеддинги текстов в разных письменностях имеют разные распределения, что показывает ID, KL дивергенция и анизотропия.
- Связь внутренней размерности с морфологической сложностью языков и малоресурсностью не прослеживается.
- Неграмматичные тексты имеют большую внутреннюю размерность.
- Внутренняя размерность представлений линейно связана с качеством модели.

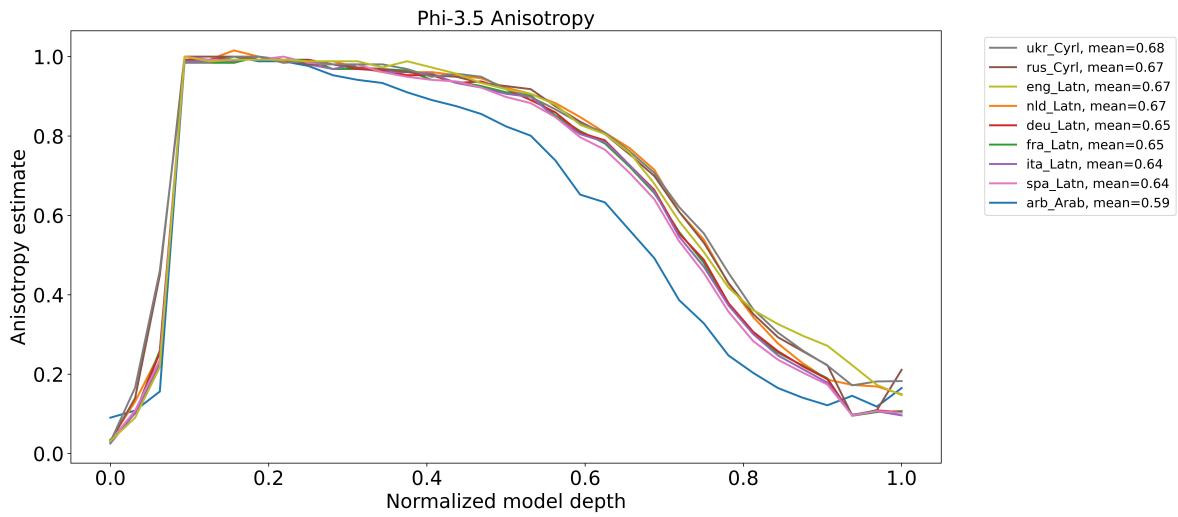


Рис. 10: Анизотропия обученной модели.

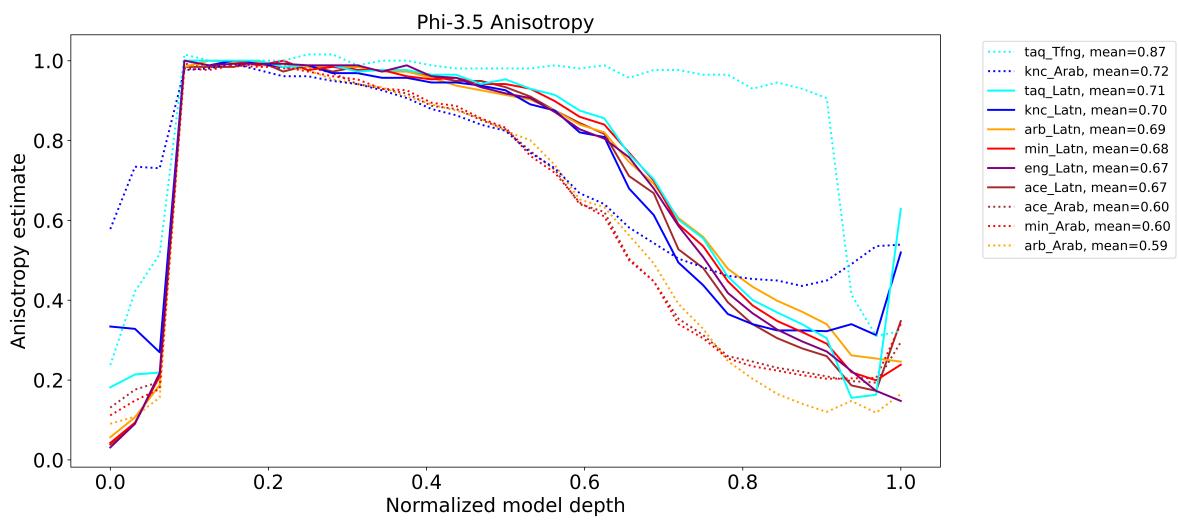


Рис. 11: Слева анизотропия необученной модели, справа - обученной модели.

Phi3 vs Phi3.5: ID и анизотропия

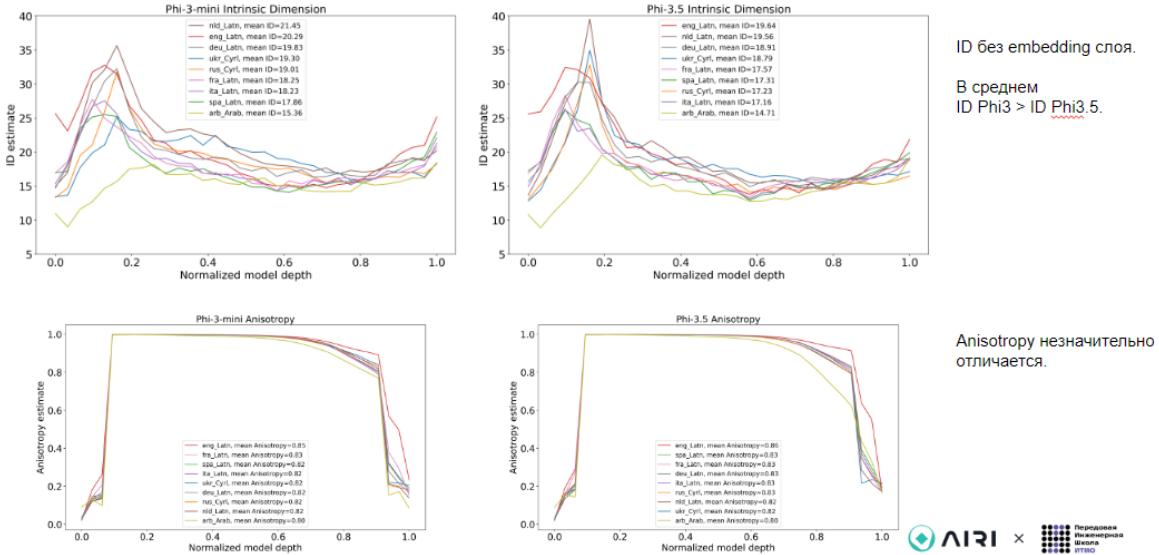


Рис. 12: Слева анизотропия необученной модели, справа - обученной модели.

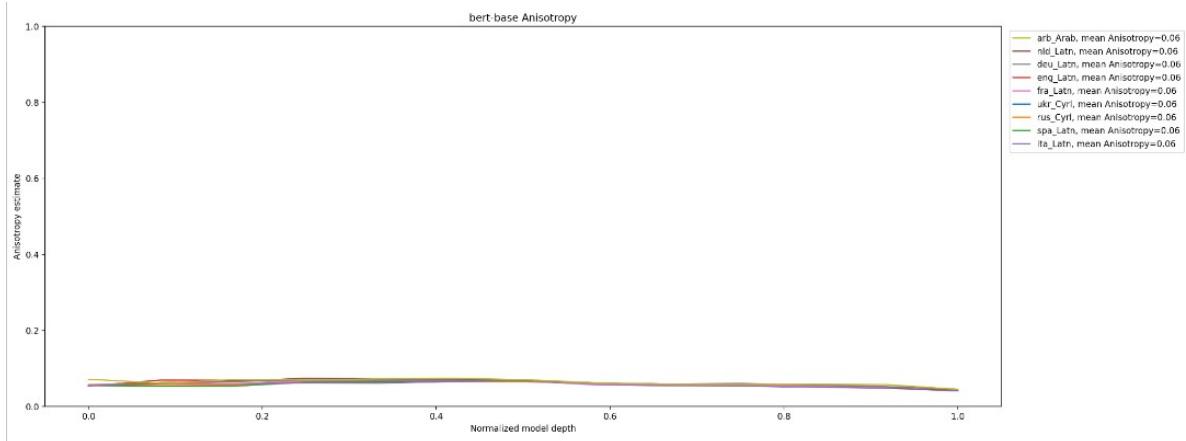


Рис. 13: Слева анизотропия необученной модели, справа - обученной модели.

BPC не коррелирует с результатами на MMLU и с внутренней размерностью (Phi-3.5)

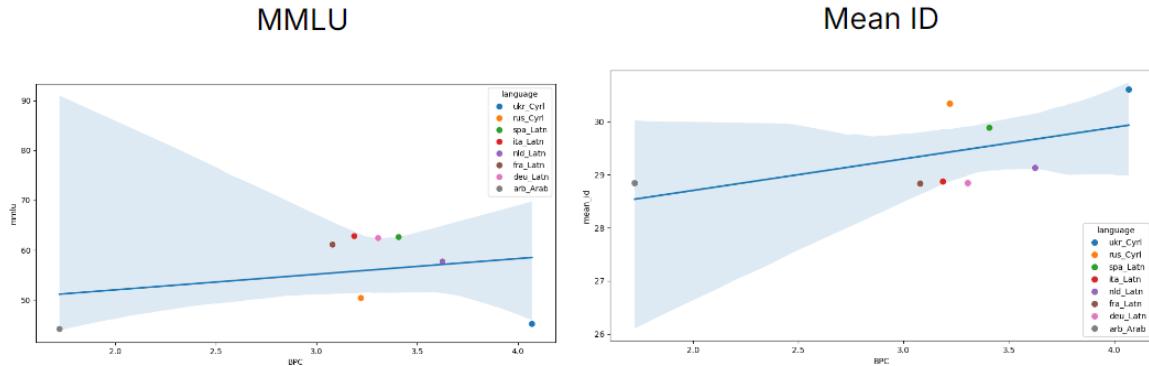


Рис. 14: Слева анизотропия необученной модели, справа - обученной модели.

Список литературы

- [1] Anton Razzhigaev, Matvey Mikhalkuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. Your transformer is secretly linear. *arXiv preprint arXiv:2405.12250*, 2024.
- [2] Kirill Tyshchuk, Polina Karpikova, Andrew Spiridonov, Anastasiia Prutianova, Anton Razzhigaev, and Alexander Panchenko. On isotropy of multimodal embeddings. *Information*, 14(7):392, 2023.