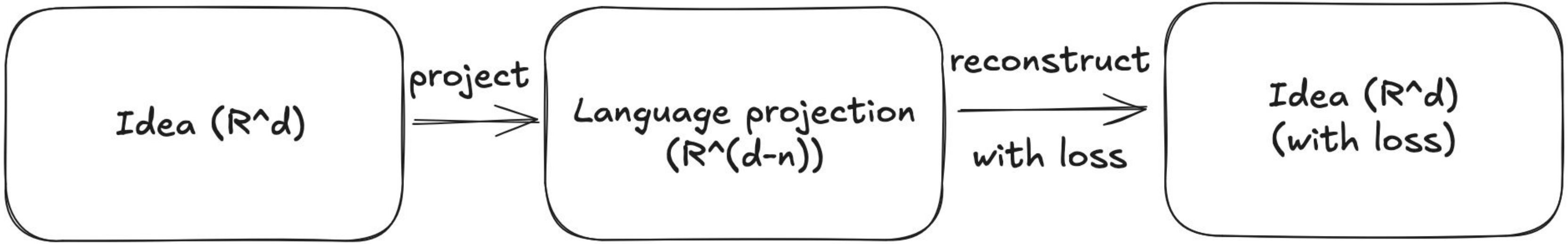
Plato meets LLMs

ideas vs lexical reflections



Idea (R^d)

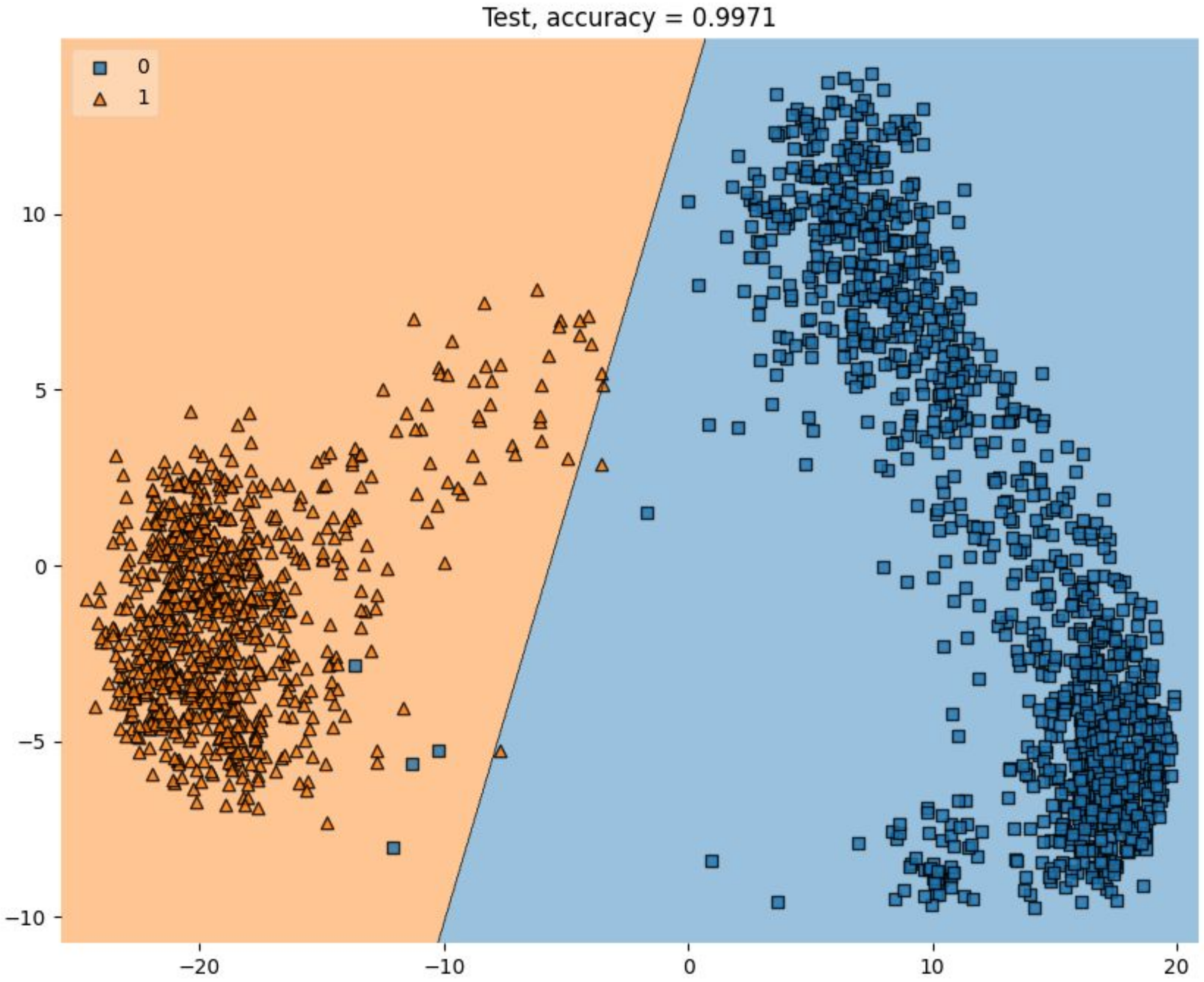
project

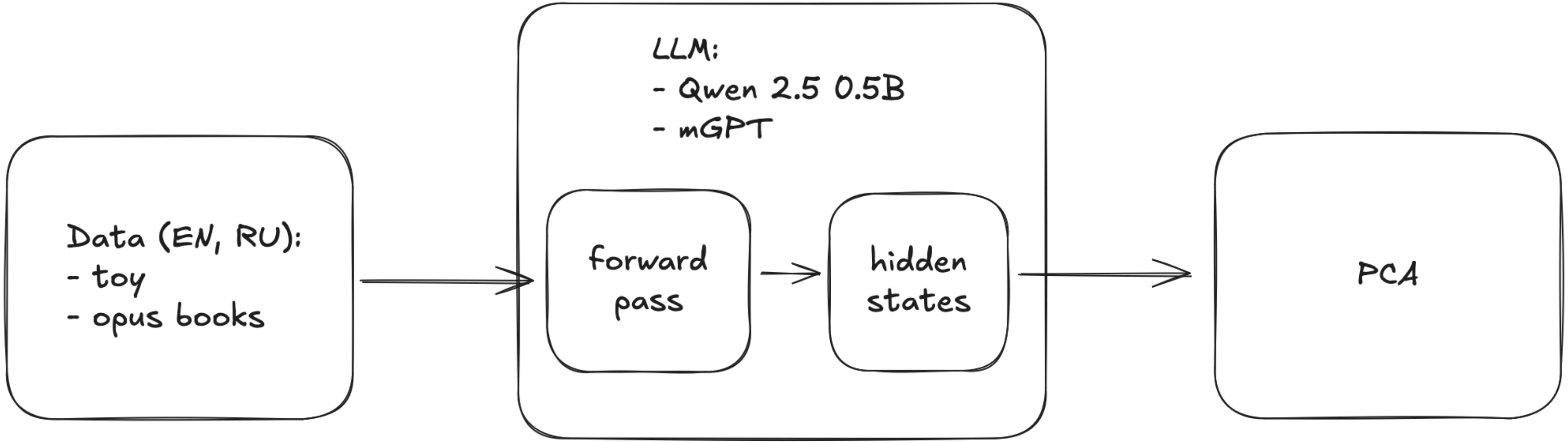
Language projection (R^(d-n))

reconstruct

with loss

Idea (R^d) (with loss)


Test, accuracy = 0.9971
0
1
10
5
0
−5
−10
−20
−10
0
10
20



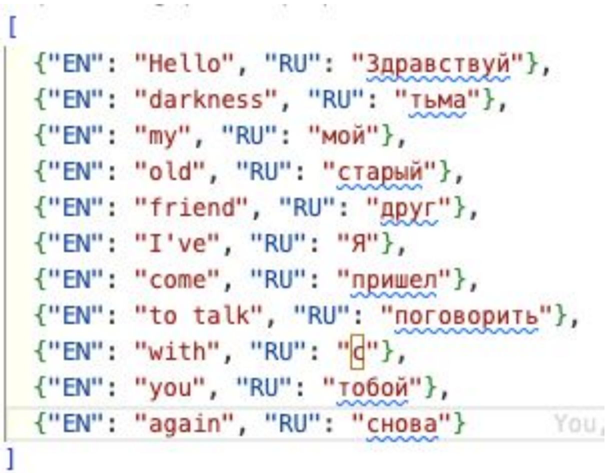
Data (EN, RU):
- toy
- opus books

LLM:
- Qwen 2.5 0.5B
- mGPT

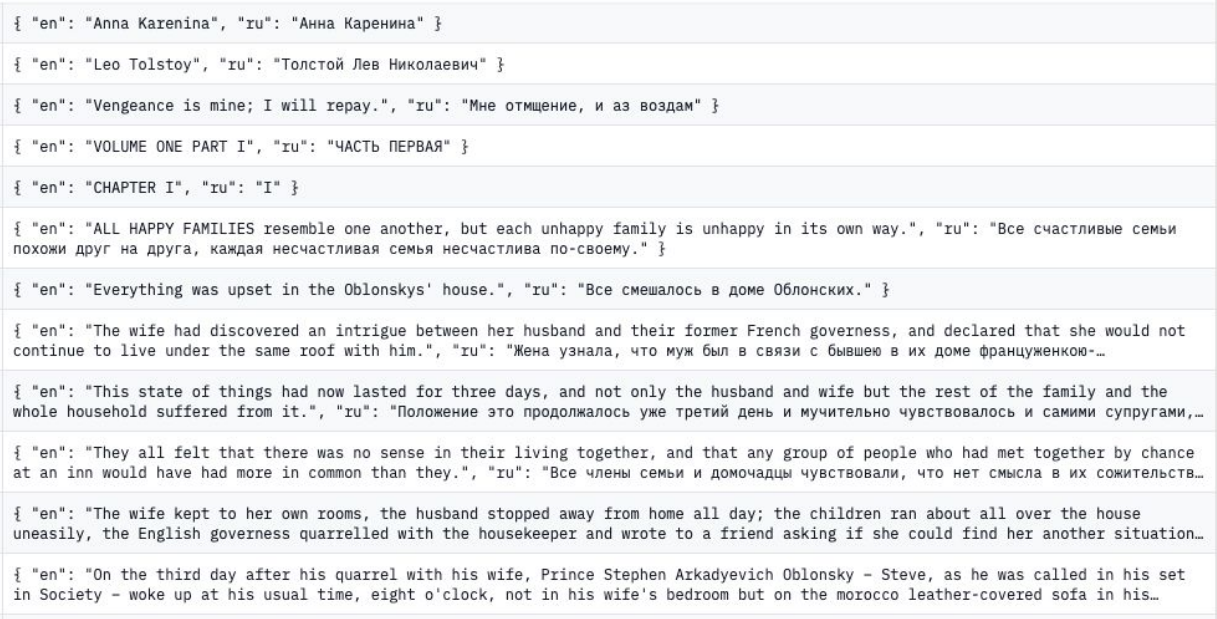
forward pass

hidden states

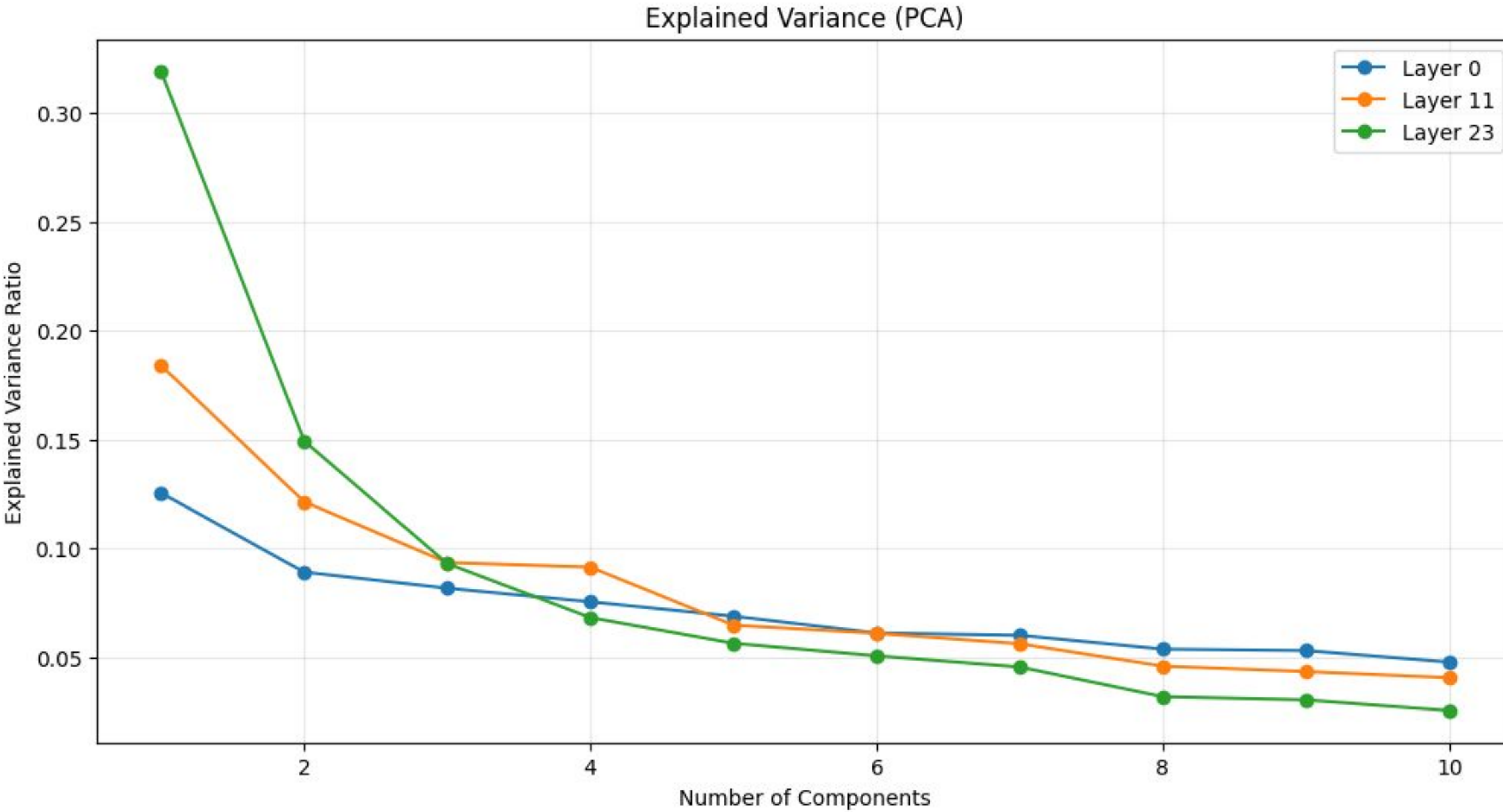
PCA

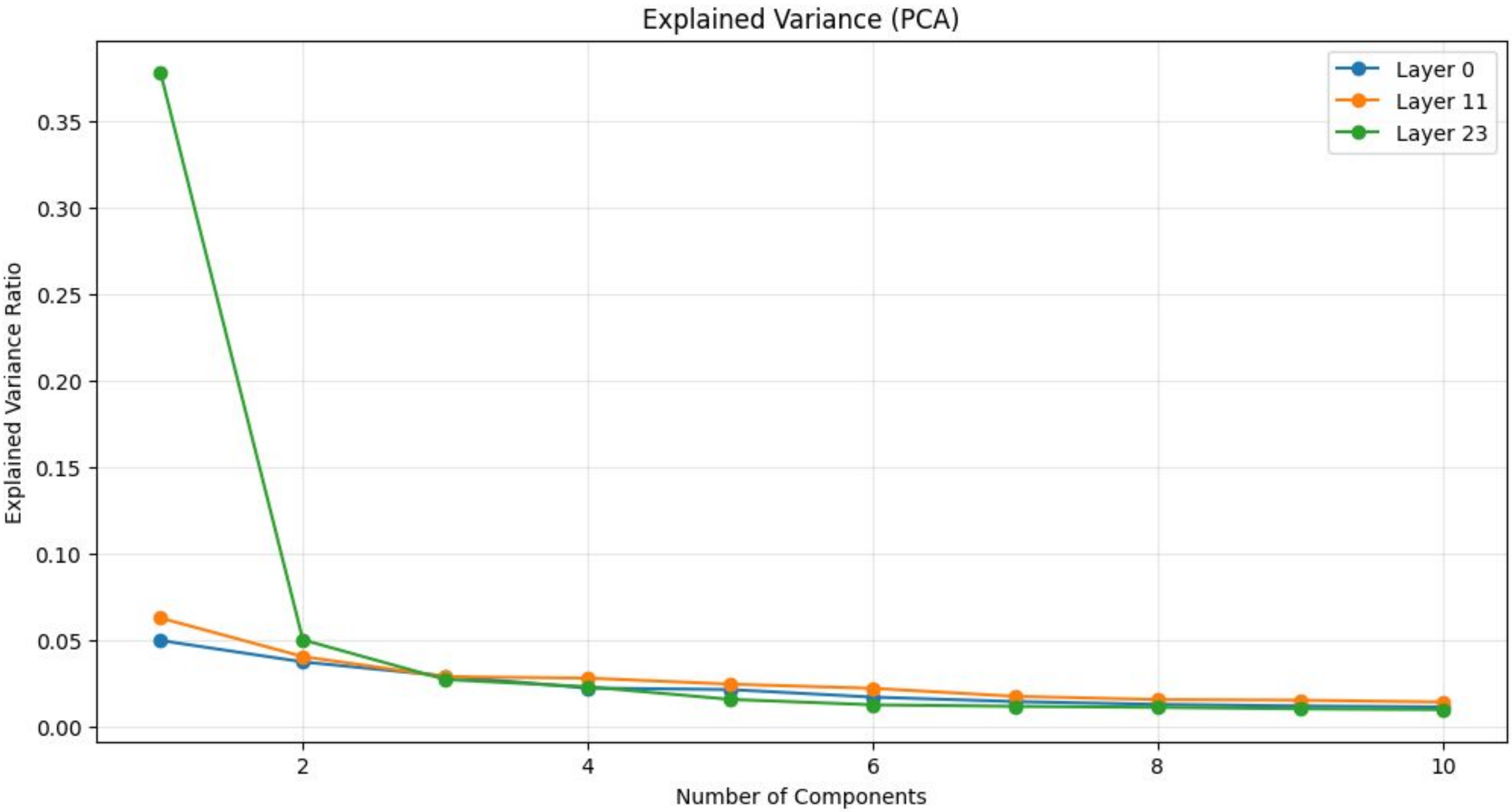


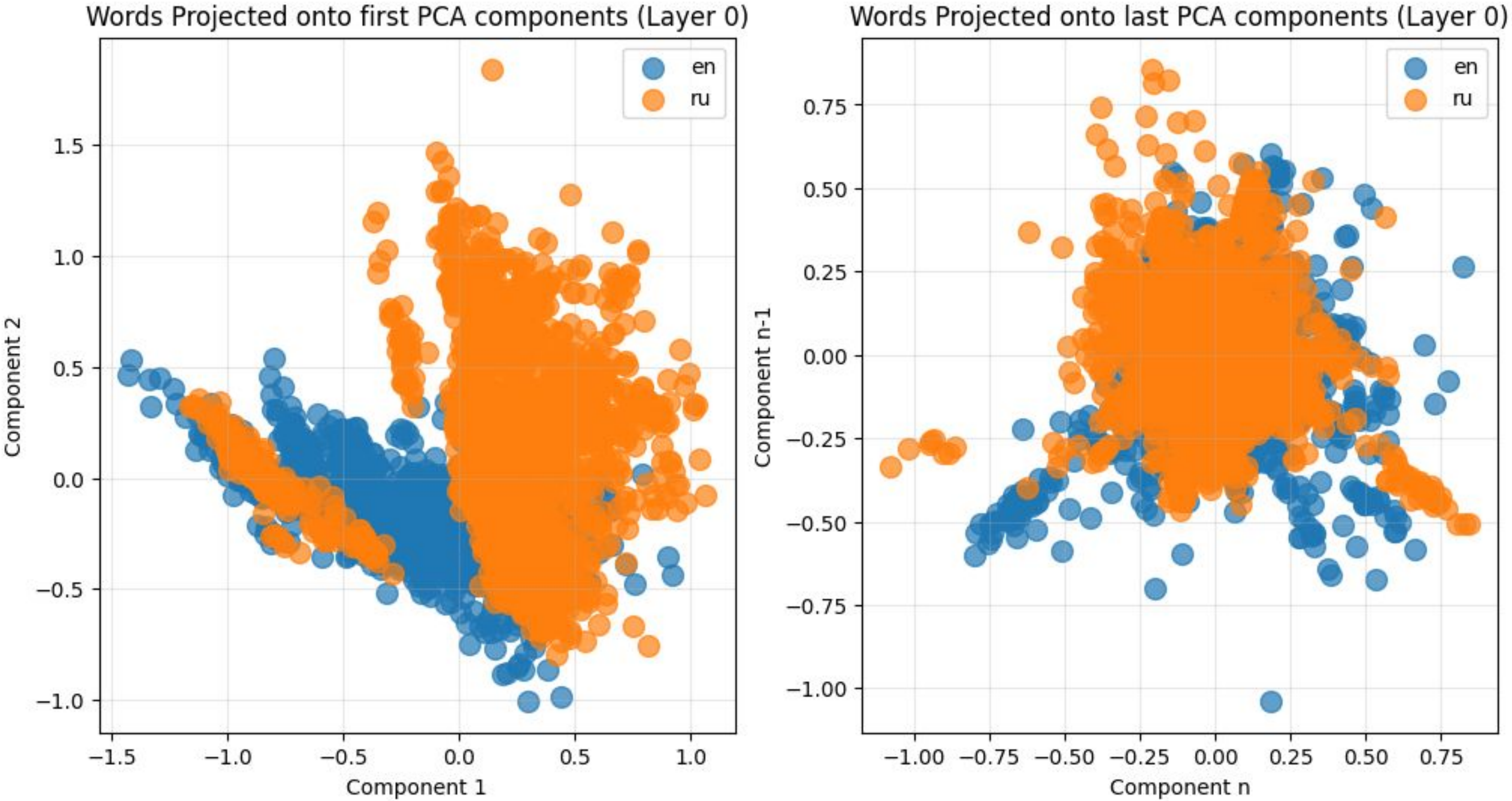
```
[
  {"EN": "Hello", "RU": "Здравствуй"},
  {"EN": "darkness", "RU": "тьма"},
  {"EN": "my", "RU": "мой"},
  {"EN": "old", "RU": "старый"},
  {"EN": "friend", "RU": "друг"},
  {"EN": "I've", "RU": "Я"},
  {"EN": "come", "RU": "пришел"},
  {"EN": "to talk", "RU": "поговорить"},
  {"EN": "with", "RU": "с"},
  {"EN": "you", "RU": "тобой"},
  {"EN": "again", "RU": "снова"}        You,
]
```



```
{ "en": "Anna Karenina", "ru": "Анна Каренина" }
{ "en": "Leo Tolstoy", "ru": "Толстой Лев Николаевич" }
{ "en": "Vengeance is mine; I will repay.", "ru": "Мне отмщение, и аз воздам" }
{ "en": "VOLUME ONE PART I", "ru": "ЧАСТЬ ПЕРВАЯ" }
{ "en": "CHAPTER I", "ru": "I" }
{ "en": "ALL HAPPY FAMILIES resemble one another, but each unhappy family is unhappy in its own way.", "ru": "Все счастливые семьи
похожи друг на друга, каждая несчастливая семья несчастлива по-своему." }
{ "en": "Everything was upset in the Oblonskys' house.", "ru": "Все смешалось в доме Облонских." }
{ "en": "The wife had discovered an intrigue between her husband and their former French governess, and declared that she would not
continue to live under the same roof with him.", "ru": "Жена узнала, что муж был в связи с бывшею в их доме француженкою-…
{ "en": "This state of things had now lasted for three days, and not only the husband and wife but the rest of the family and the
whole household suffered from it.", "ru": "Положение это продолжалось уже третий день и мучительно чувствовалось и самими супругами,…
{ "en": "They all felt that there was no sense in their living together, and that any group of people who had met together by chance
at an inn would have had more in common than they.", "ru": "Все члены семьи и домочадцы чувствовали, что нет смысла в их сожительств…
{ "en": "The wife kept to her own rooms, the husband stopped away from home all day; the children ran about all over the house
uneasily, the English governess quarrelled with the housekeeper and wrote to a friend asking if she could find her another situation…
{ "en": "On the third day after his quarrel with his wife, Prince Stephen Arkadyevich Oblonsky – Steve, as he was called in his set
in Society – woke up at his usual time, eight o'clock, not in his wife's bedroom but on the morocco leather-covered sofa in his…
```

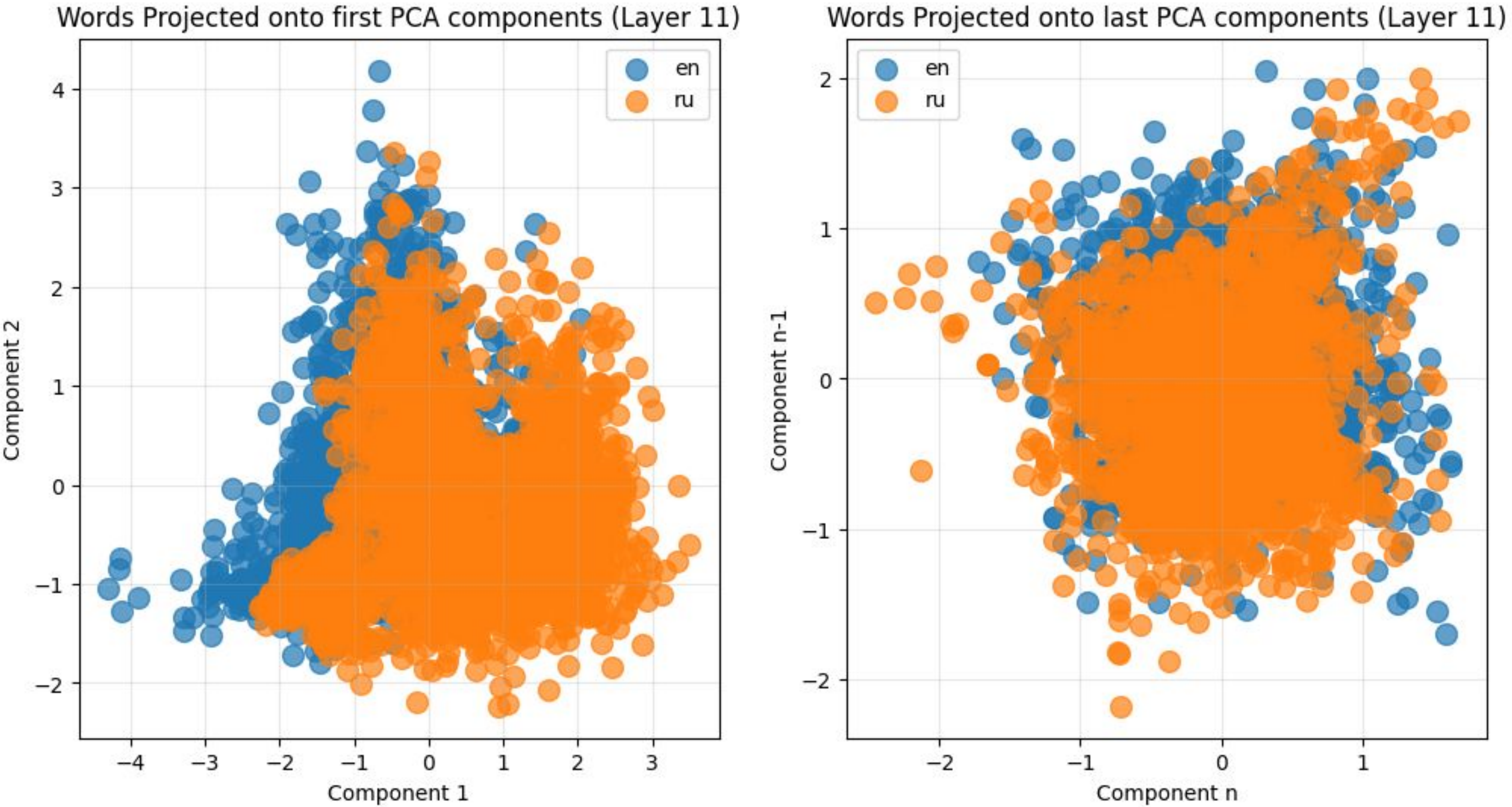

Explained Variance (PCA)
Explained Variance Ratio
0.30
0.25
0.20
0.15
0.10
0.05
2
4
6
8
10
Number of Components
Layer 0
Layer 11
Layer 23


Explained Variance (PCA)
Layer 0
Layer 11
Layer 23
Explained Variance Ratio
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
4
6
8
10
Number of Components

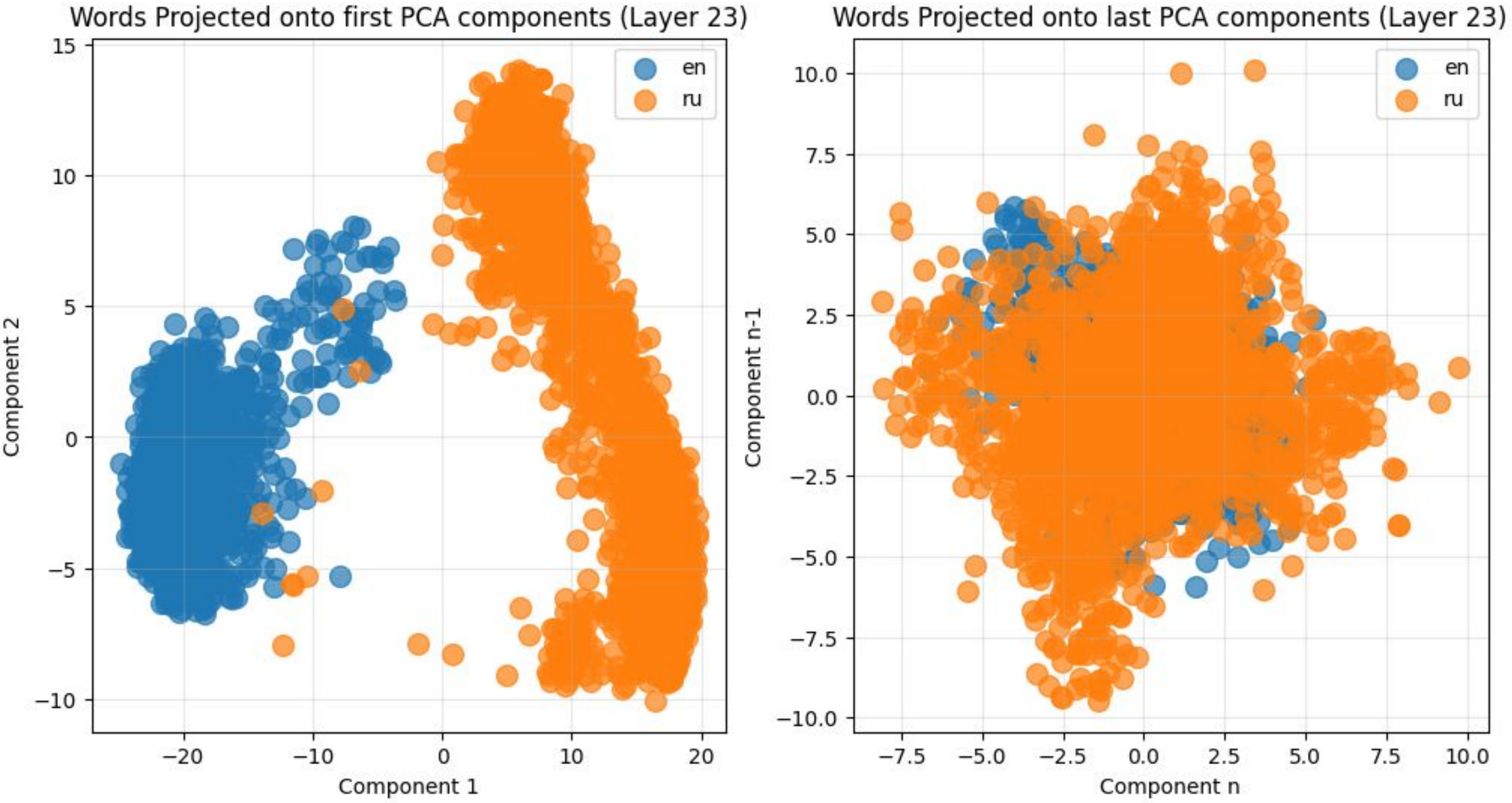


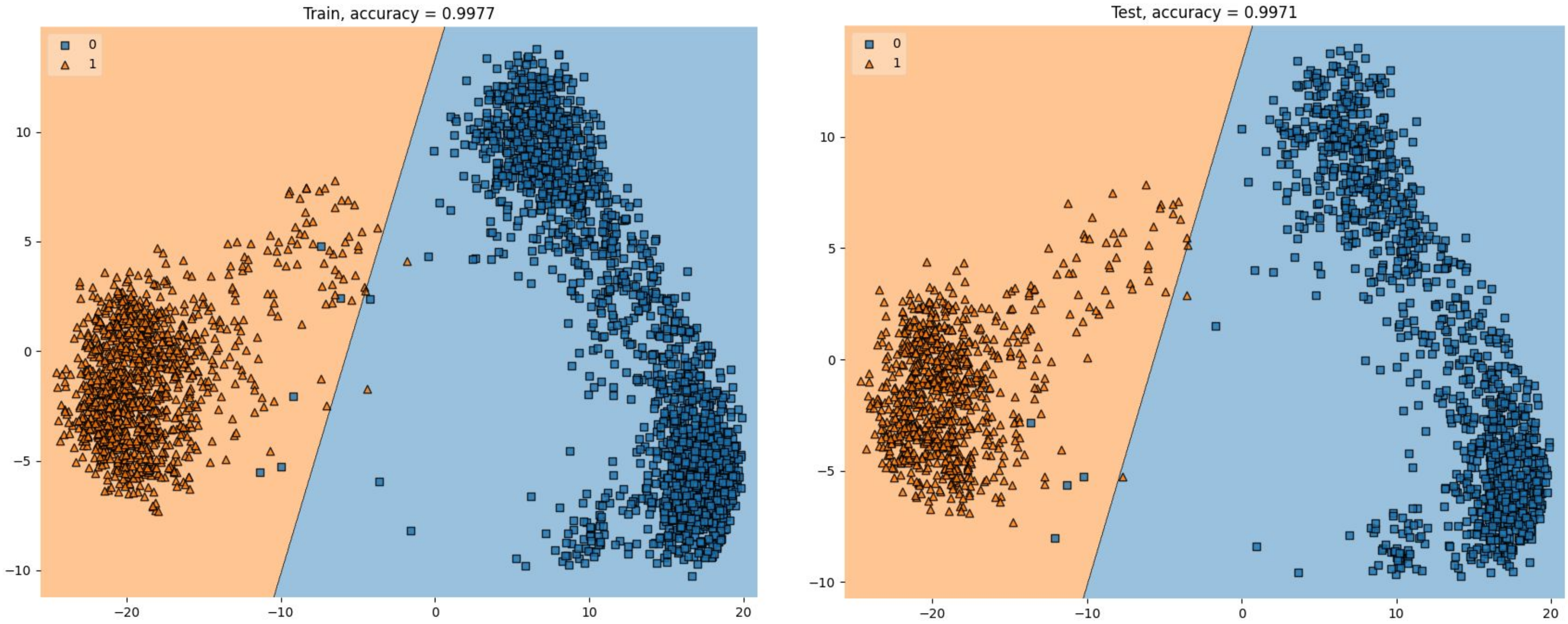
Words Projected onto first PCA components (Layer 0)

Words Projected onto last PCA components (Layer 0)



Words Projected onto first PCA components (Layer 11)
Words Projected onto last PCA components (Layer 11)
en
ru
Component 1
Component 2
Component n
Component n-1
−4
−3
−2
−1
0
1
2
3
4


Words Projected onto first PCA components (Layer 23)
Component 1
Component 2
−20
−10
0
10
20
15
10
5
0
−5
−10
en
ru
Words Projected onto last PCA components (Layer 23)
Component n
Component n-1
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
en
ru



Train, accuracy = 0.9977

Test, accuracy = 0.9971

Next 🧑‍🔬

- Train a classifier on hidden space directly
- Fine-tune LLM to forget one of the languages
- Proceed to lexical subspace separation