

Сравнение стратегий подбора обучающих примеров для few-shot оценки языковой компетенции LLM (на примере датасета BLiMP и модели Mistral)



Проектные задачи компьютерной лингвистики
Элеонора Измайлова
07.04.2025

1. Датасет

- **BLiMP: The Benchmark of Linguistic Minimal Pairs for English** [Warstadt et al., TACL 2020]
- создавался для выявления сильных и слабых сторон моделей (*n-gram*, *LSTM*, *Transformer-XL*, *GPT-2*) по сравнению с человеческими оценками
- в отличие от **CoLA** (*Corpus of Linguistic Acceptability*) [Warstadt et al., TACL 2019], **BLiMP** оценивает модели на парах предложений, а не на отдельных предложениях, что позволяет точнее выявлять конкретные лингвистические способности моделей
- был сгенерирован **автоматически** с сохранением структурной аналогии до ключевой позиции контраста, что исключает влияние других факторов на оценку, и использованием словаря с многоуровневой разметкой размером **3000** слов
- состоит из **67** подразделов ('парадигм'), каждый из которых содержит **1 000** минимальных пар предложений на одно грамматическое явление

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	<i>Many girls insulted <u>themselves</u>.</i>	<i>Many girls insulted <u>herself</u>.</i>
ARG. STRUCTURE	9	<i>Rose wasn't <u>disturbing</u> Mark.</i>	<i>Rose wasn't <u>boasting</u> Mark.</i>
BINDING	7	<i>Carlos said that Lori helped <u>him</u>.</i>	<i>Carlos said that Lori helped <u>himself</u>.</i>
CONTROL/RAISING	5	<i>There was <u>bound</u> to be a fish escaping.</i>	<i>There was <u>unable</u> to be a fish escaping.</i>
DET.-NOUN AGR.	8	<i>Rachelle had bought that <u>chair</u>.</i>	<i>Rachelle had bought that <u>chairs</u>.</i>
ELLIPSIS	2	<i>Anne's doctor cleans one <u>important</u> book and Stacey cleans a few.</i>	<i>Anne's doctor cleans one book and Stacey cleans a few <u>important</u>.</i>
FILLER-GAP	7	<i>Brett knew <u>what</u> many waiters find.</i>	<i>Brett knew <u>that</u> many waiters find.</i>
IRREGULAR FORMS	2	<i>Aaron <u>broke</u> the unicycle.</i>	<i>Aaron <u>broken</u> the unicycle.</i>
ISLAND EFFECTS	8	<i>Whose <u>hat</u> should Tonya wear?</i>	<i>Whose should Tonya wear <u>hat</u>?</i>
NPI LICENSING	7	<i>The truck has <u>clearly</u> tipped over.</i>	<i>The truck has <u>ever</u> tipped over.</i>
QUANTIFIERS	4	<i>No boy knew <u>fewer than</u> six guys.</i>	<i>No boy knew <u>at most</u> six guys.</i>
SUBJECT-VERB AGR.	6	<i>These casseroles <u>disgust</u> Kayla.</i>	<i>These casseroles <u>disgusts</u> Kayla.</i>

Table 2: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. *N* is the number of 1,000-example minimal pair paradigms within each broad category.

2. 'Предыдущие' исследования

Тестирование в [Warstadt et al., TACL 2020]: «forced-choice» (модель сравнивает вероятности двух вариантов)

- минимальные пары создавались таким образом, чтобы различия в вероятностях объяснялись исключительно грамматическим контрастом, а не контекстом или длиной предложений

Model	Overall	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Table 3: Percentage accuracy of four baseline models and raw human performance on BLiMP using a forced-choice task. A random guessing baseline would achieve an accuracy of 50%.

3. Данное (квази-) исследование

- тестирование путем **промπτинга** (с прямым обращением к LLM)
- сравнение качества относительно дообученной матрицы проекций
- **Цель:** определить лучшую стратегию подбора обучающих примеров для ***few-shot*** оценки
- размер тестовой выборки: **670** предложений (**10** предложений на каждое из **67** грамматических явлений)
- **МОДЕЛЬ:** `mistral-large-latest`
 - ◆ 123b параметров
 - ◆ доступ через API

3. Как подбирался промпт (на основе zero-shot)

Промпт	Accuracy
Which of the following sentences is grammatically correct ? Please respond with the number of the correct sentence only.	83.58
One of the following sentences contains a grammatical error . Which one is correct? Please respond with the number of the correct sentence only.	82.84
Which of the following sentences has correct {inquiry['linguistic_term']} ? Please respond with the number of the correct sentence only.	82.09
Which of the following sentences is grammatically correct? Please respond with the number of the correct sentence, write a full stop and then explain your reasoning step-by-step in one or two sentences .	85.82
Which of the following sentences is grammatically correct? Please respond with the number of the correct sentence, write a full stop and then explain your reasoning step-by-step .	87.01

4. Few-shot эксперименты

Сетапы:

1. **Без использования NNS по эмбедингам**
 - a. случайный выбор из **всего** датасета
 - b. случайный выбор из **подраздела** датасета на то же грамматическое явление
2. **С использованием NNS (по эмбедингам предложений)**
 - a. поиск по близким к **неграмматичному** предложению в тестовой паре
 - b. поиск по близким к **грамматичному** предложению в тестовой паре
3. **С использованием NNS (по эмбедингам пар предложений)**
 - a. поиск по близким к **разности** эмбедингов тестовой пары
 - b. поиск по близким к **покомпонентному умножению** эмбедингов тестовой пары
 - c. поиск по близким к **усредненному** эмбедингу тестовой пары
 - d. поиск по близким к **конкатенированному** эмбедингу тестовой пары
4. **С использованием NNS (комбинированный подход)**
 - a. поиск по близким к конкатенации разности и **покомпонентного умножения**
 - b. поиск по близким к конкатенации разности и **усреднения**
 - c. поиск по близким к конкатенации разности и **конкатенации**

4. Few-shot эксперименты

Here are some examples of grammatically correct and incorrect sentences:

Example 1:

1: Who would Brenda respect Ellen before alarming?

2: Who would Brenda respect before alarming Ellen?

Correct sentence: 2

Example 2:

1: Who has Suzanne irritated without alarming Homer?

2: Who has Suzanne irritated Homer without alarming?

Correct sentence: 1

Example 3:

1: Who has Homer hated Brad while concealing?

2: Who has Homer hated while concealing Brad?

Correct sentence: 2

Now, please, determine, which of the following sentences is grammatically correct. Please respond with the number of the correct sentence, write a full stop and then explain your reasoning step-by-step.

<<<

1. Who had Andrew cared for while confusing Leslie?

2. Who had Andrew cared for Leslie while confusing?

>>>

4. Few-shot эксперименты

>>> MODEL RESPONSE <<<

1.

Here's the step-by-step reasoning:

- Both sentences are trying to convey that Andrew cared for someone while confusing Leslie.
- In sentence 1, "Who had Andrew cared for while confusing Leslie?", the structure is clear: "cared for" is followed by "while confusing Leslie", which is a dependent clause modifying the verb "cared for".
- In sentence 2, "Who had Andrew cared for Leslie while confusing?", the structure is awkward because "while confusing" is left hanging without a clear object. It should be "while confusing [someone or something]".
- Therefore, sentence 1 is the grammatically correct choice.

4. Few-shot эксперименты

Результаты:

1. Без использования NNS по эмбедингам

- a. случайный выбор из **всего** датасета [Асс. 86.87]
- b. случайный выбор из **подраздела** датасета на то же грамматическое явление [Асс. 92.69]

2. С использованием NNS (по эмбедингам предложений)

- a. поиск по близким к **неграмматичному** предложению в тестовой паре [Асс. 91.04]
- b. поиск по близким к **грамматичному** предложению в тестовой паре [Асс. 93.28]

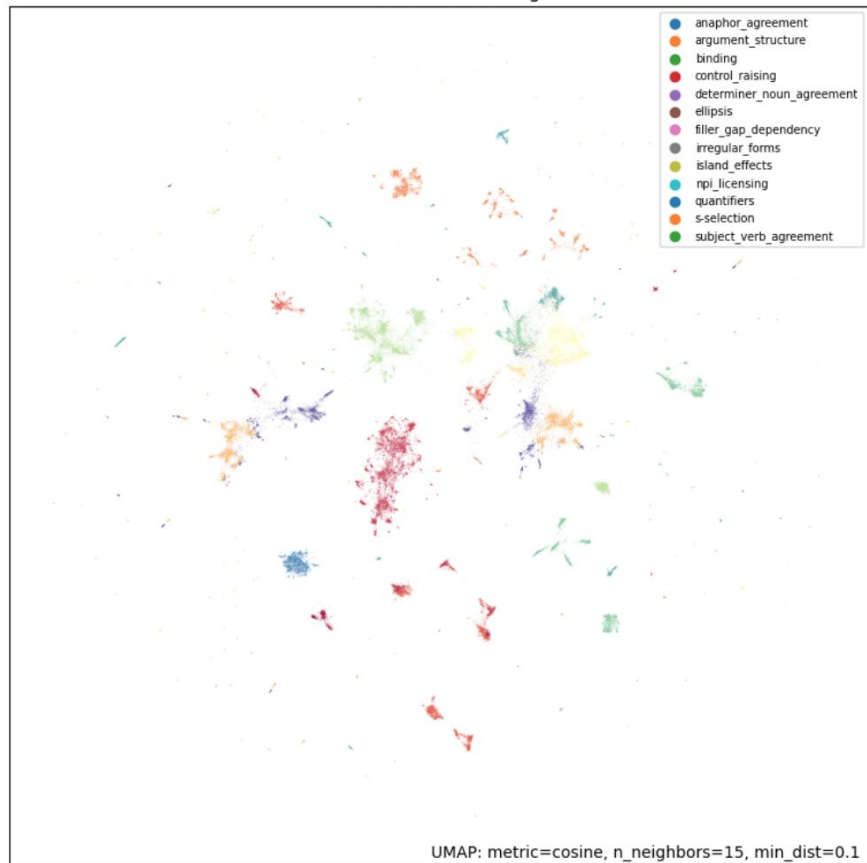
3. С использованием NNS (по эмбедингам пар предложений)

- a. поиск по близким к **разности** эмбедингов тестовой пары [Асс. 93.73]
- b. поиск по близким к **покомпонентному умножению** эмбедингов тестовой пары [Асс. 92.24]
- c. поиск по близким к **усредненному** эмбедингу тестовой пары [Асс. 90.09]
- d. поиск по близким к **конкатенированному** эмбедингу тестовой пары [Асс. 92.09]

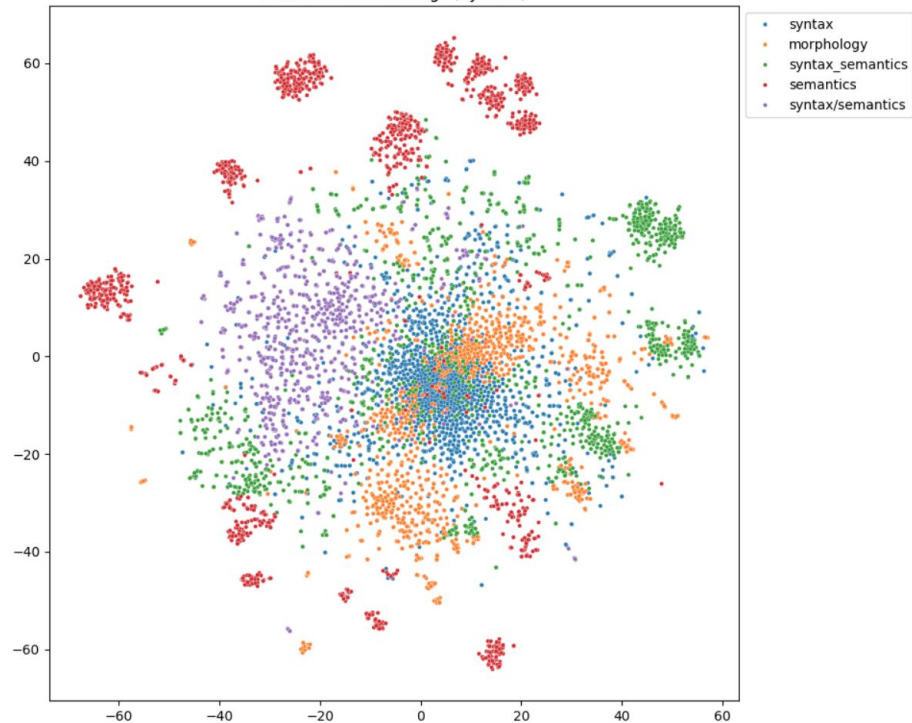
4. С использованием NNS (комбинированный подход)

- a. поиск по близким к конкатенации разности и **покомпонентного умножения** [Асс. 89.94]
- b. поиск по близким к конкатенации разности и **усреднения** [Асс. 93.58]
- c. поиск по близким к конкатенации разности и **конкатенации** [Асс. 91.94]

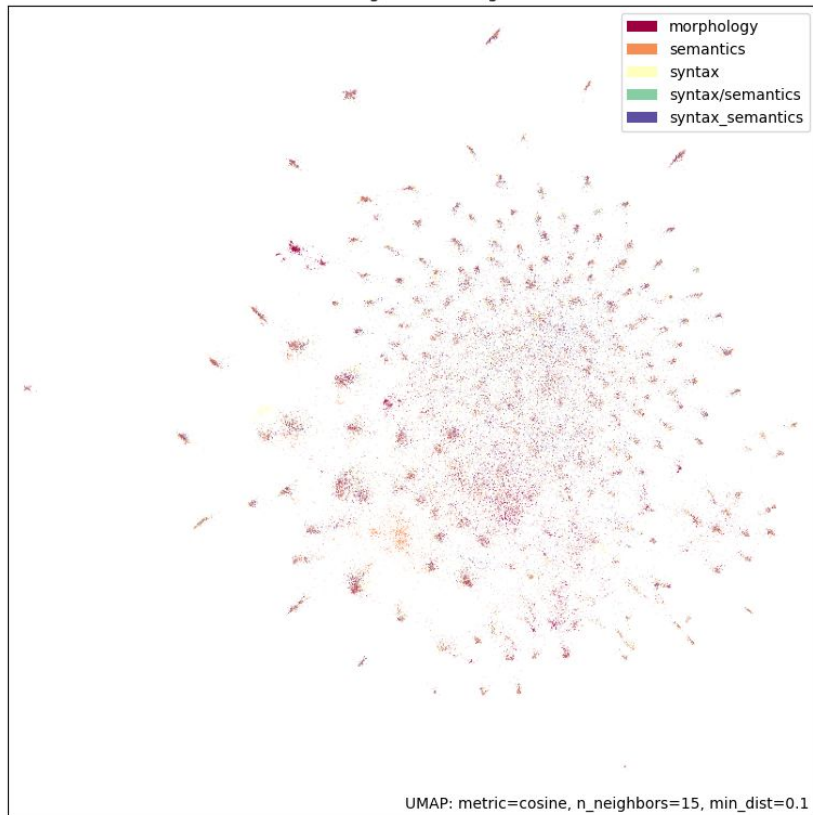
Difference Embeddings



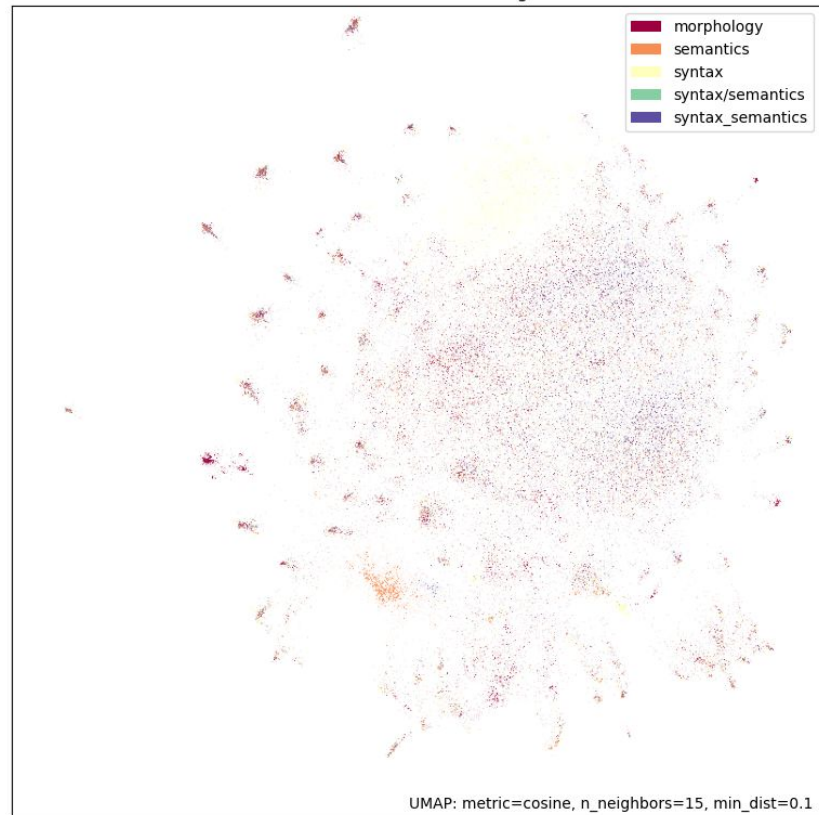
Difference Embeddings (by Field)

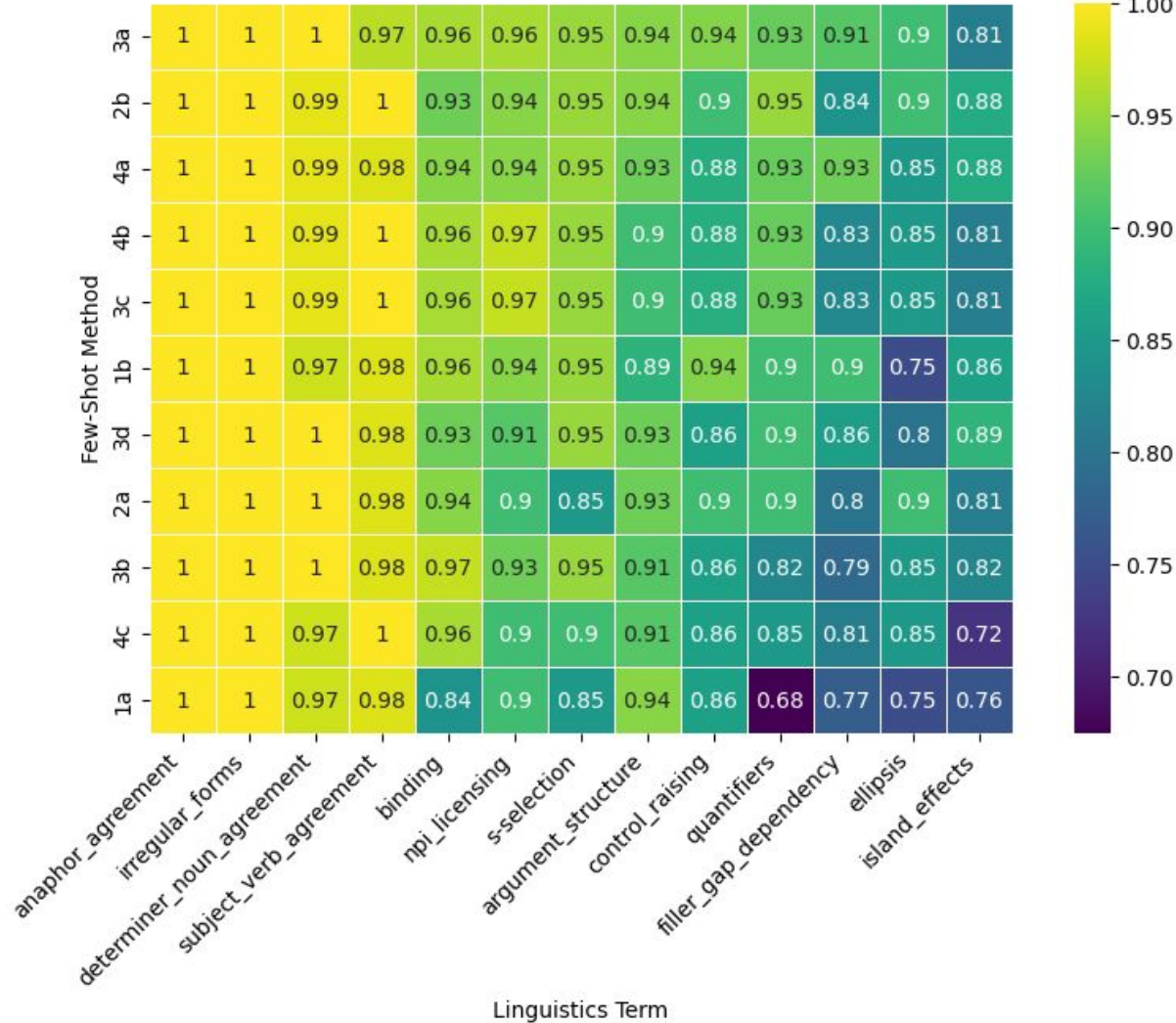
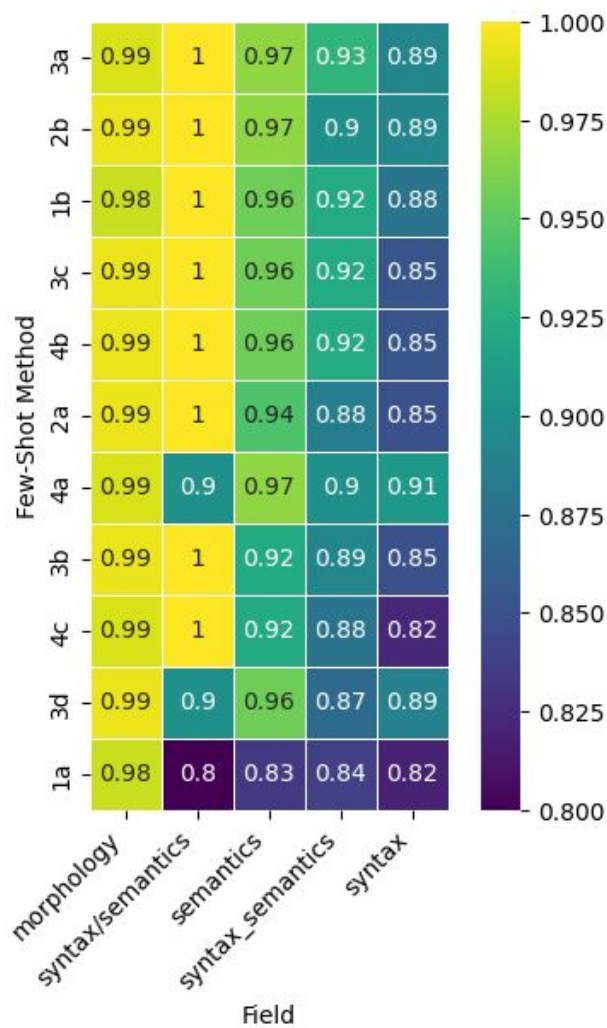


Average Embeddings



Concatenated Embeddings





1	0.8	0.9	0.8	0.9	0.8	0.9	0.7	0.6	0.9	0.5
0.9	0.9	0.8	1	0.8	0.8	0.7	0.6	0.9	0.7	0.6
0.9	0.5	0.8	0.6	0.8	0.8	0.7	0.8	0.6	0.5	0.6
0.8	0.8	1	0.9	0.6	0.7	0.8	0.8	0.7	0.8	0.6
0.9	0.8	0.6	0.8	0.8	0.8	0.8	0.5	0.7	0.4	0.8
0.6	0.8	0.9	0.8	0.8	0.6	0.6	0.9	0.4	0.7	0.5
0.6	0.8	0.9	0.8	0.8	0.6	0.6	0.9	0.4	0.7	0.5
0.6	0.6	1	0.8	0.8	0.7	0.6	0.6	0.5	0.5	0.6
0.9	0.9	0.6	0.6	0.7	0.8	0.6	0.7	0.7	0.5	0.6
0.8	1	0.6	0.8	0.8	0.7	0.8	0.5	0.7	0.6	0.4
0.9	0.8	0.6	0.7	0.7	0.8	0.8	0.7	0.7	0.3	0.6

wh_vs_that_with_gap
 principle_A_reconstruction
 existential_there_quantifiers_2
 wh_questions_subject_gap_long_distance
 ellipsis_n_bar_2
 drop_argument
 tough_vs_raising_1
 only_npi_scope
 complex_NP_island
 wh_vs_that_with_gap_long_distance
 sentential_subject_island

Sentential Subject Island (*Island Effects*)

"Who would the men's appreciating Sara disturb."

"Who would the men's appreciating disturb Sara."

wh vs. that with gap long distance (*Filler-Gap Dependency*)

"Anne wasn't finding out **who** a lot of drivers that won't love Tanya weren't concealing."

"Anne wasn't finding out **that** a lot of drivers that won't love Tanya weren't concealing."

Complex NP Island (*Island Effects*)

"Who isn't Karen's friend who couldn't confuse Elaine listening to?"

"Who isn't Elaine listening to Karen's friend who couldn't confuse?"

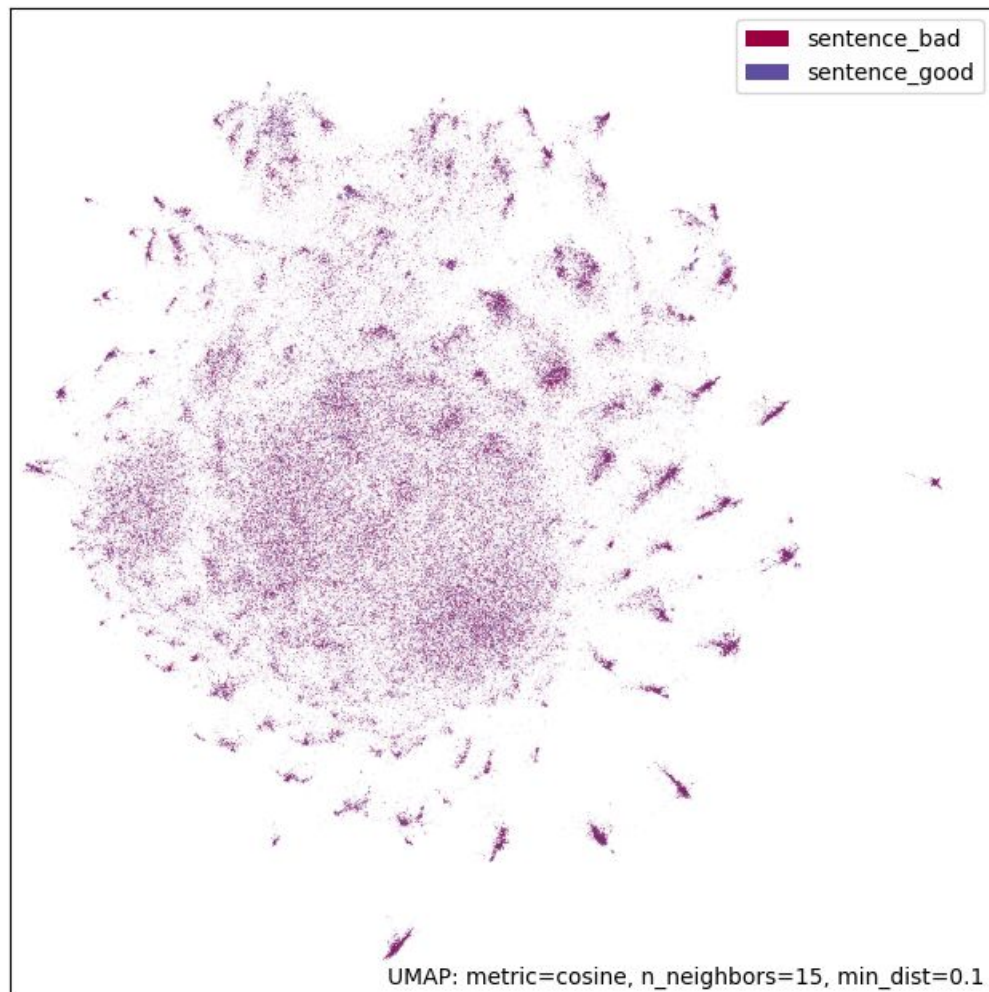
Only NPI Scope (*NPI Licensing*)

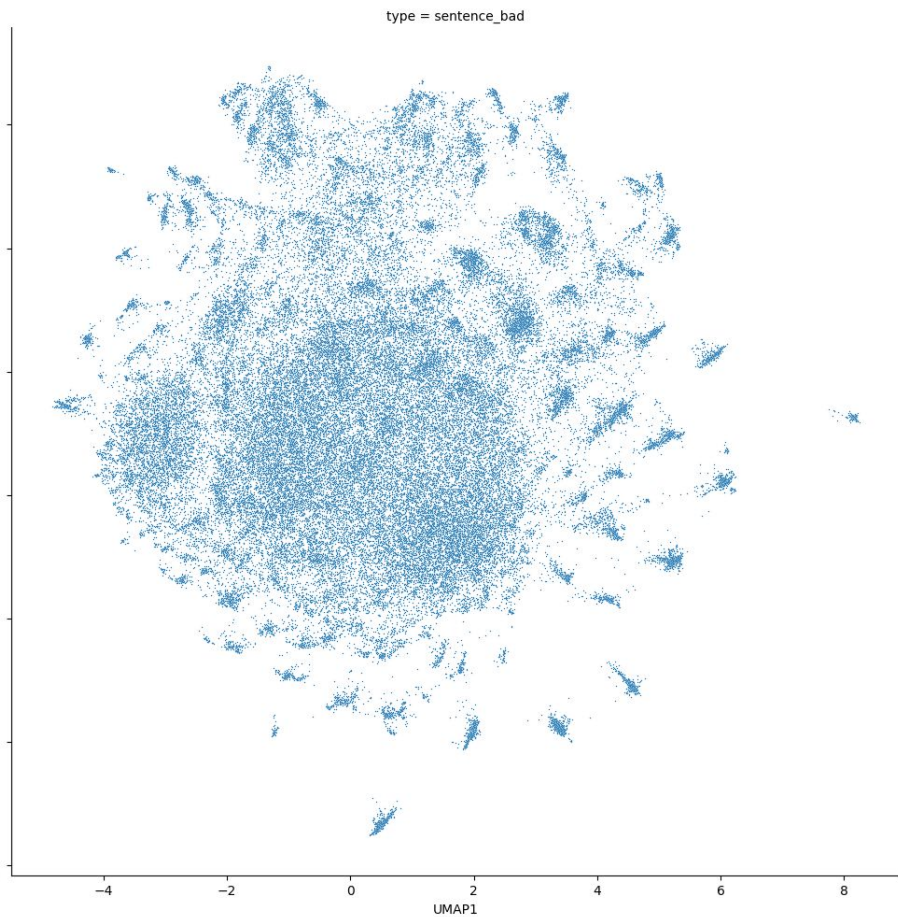
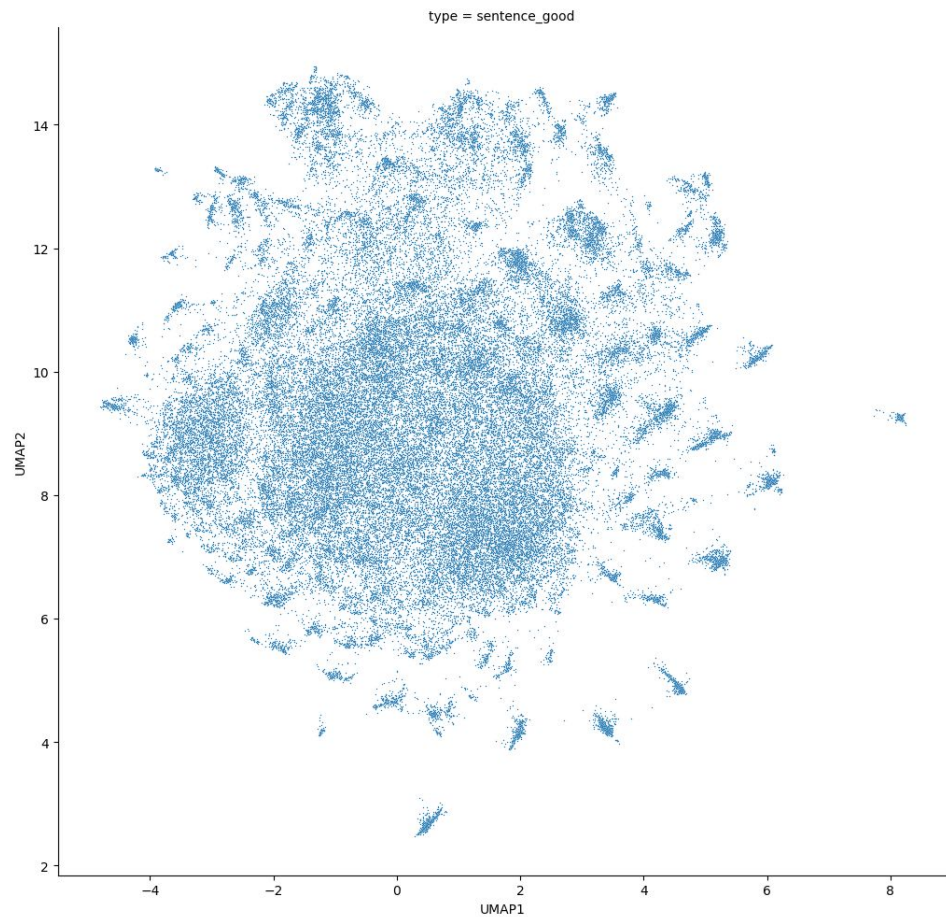
"Only Carlos's tutor who most adults insult ever performed."

"Carlos's tutor who only most adults insult ever performed."

4. Дообучение матрицы проекций эмбеддингов

- попытаться разделить пространство эмбеддингов по грамматичности
- посмотреть, как качество соотносится с результатами промптинга
- Конфигурация обучения (для обеих моделей)
 1. **Фреймворк:** PyTorch
 2. **Оптимизатор:** Adam
 - Скорость обучения: 0,001 (1e-3)
 3. **Функция потерь:** Contrastive Loss
 - Порог (*margin*): 0,5
 4. **Протокол обучения:**
 - Эпохи: 100 (максимум)
 - Ранняя остановка:
 - *patience*: 5 эпох
 - Дельта (минимальное улучшение): 0,0075
 - Порог сходства: 0,25

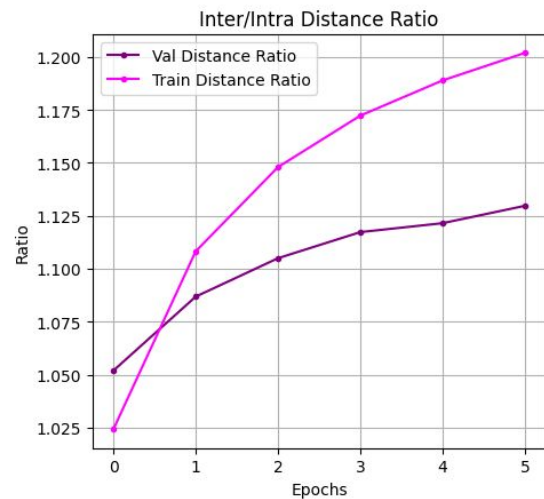
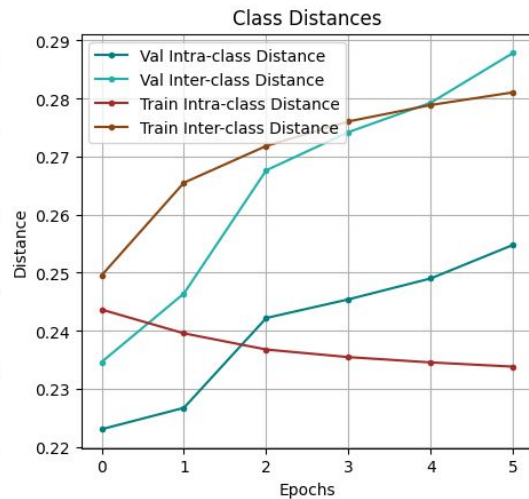
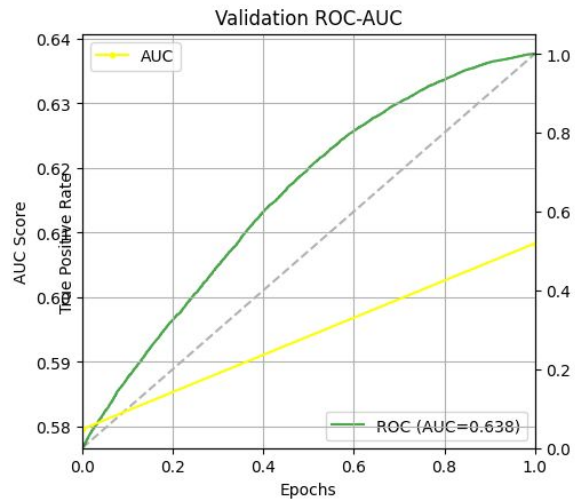
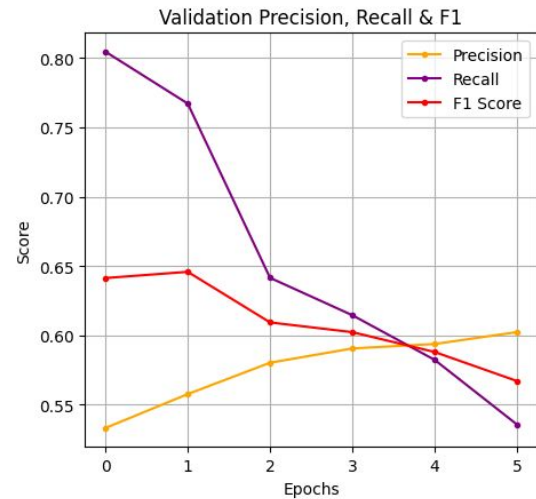
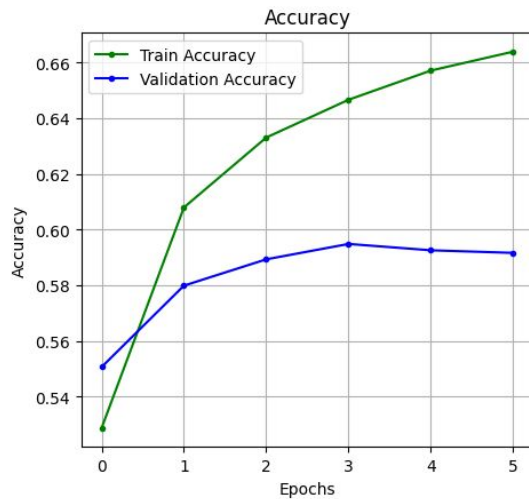
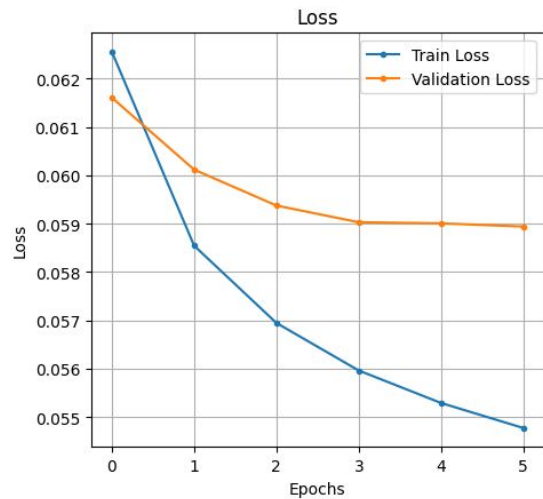


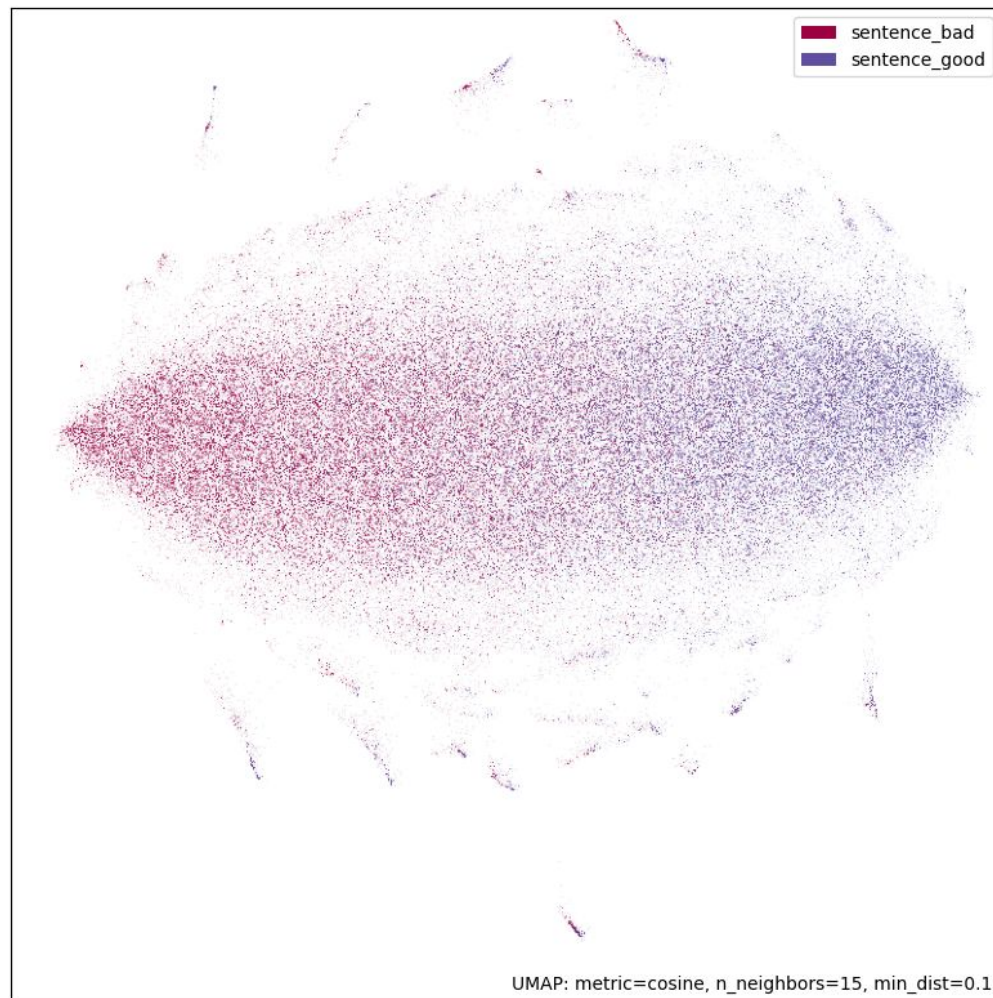


4. Дообучение матрицы проекций эмбеддингов

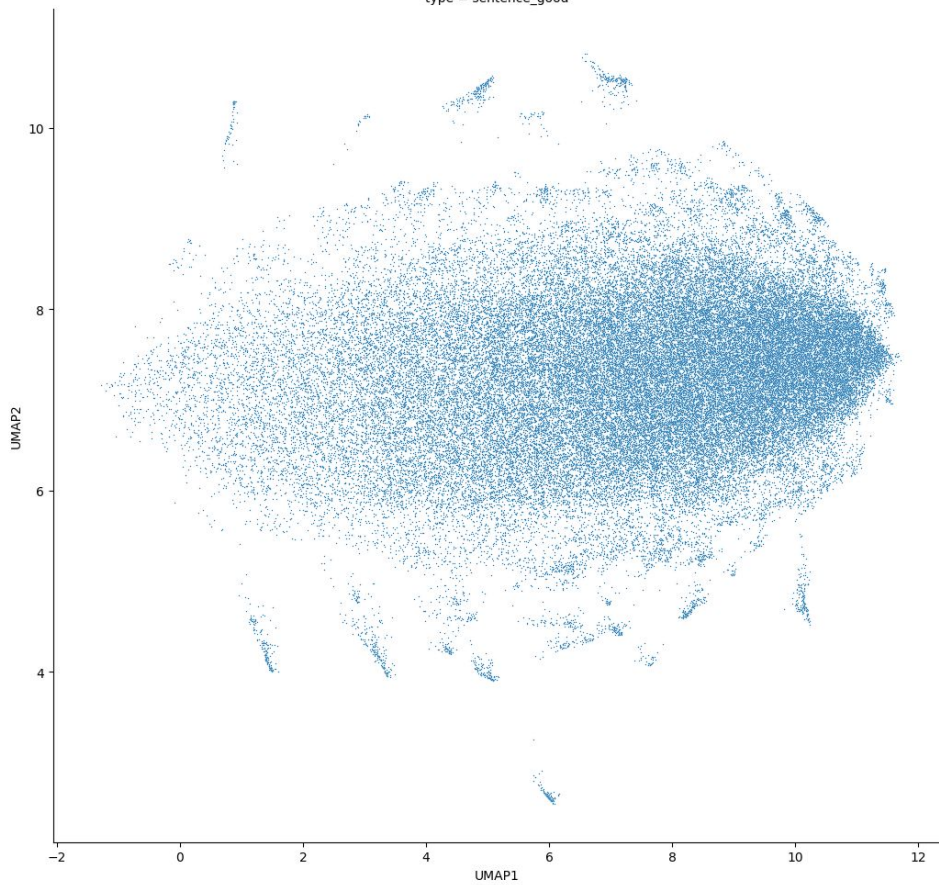
Модель 1: ProjectionMatrix

1. **Тип архитектуры:** Однослойная линейная проекционная модель
2. **Входная размерность:** 1024
3. **Выходная размерность:** 512
4. **Параметры:**
 - Одна матрица весов и вектор смещения
 - Количество параметров: ~524 тыс. ($1024 \times 512 + 512$)
5. **Метрики качества:**
 - accuracy: 0.59
 - precision: 0.6
 - recall: 0.54
 - f1_score: 0.57
 - auc: 0.64
 - mean_intra_class_dist: 0.25
 - mean_inter_class_dist: 0.29
 - distance_ratio: 1.13

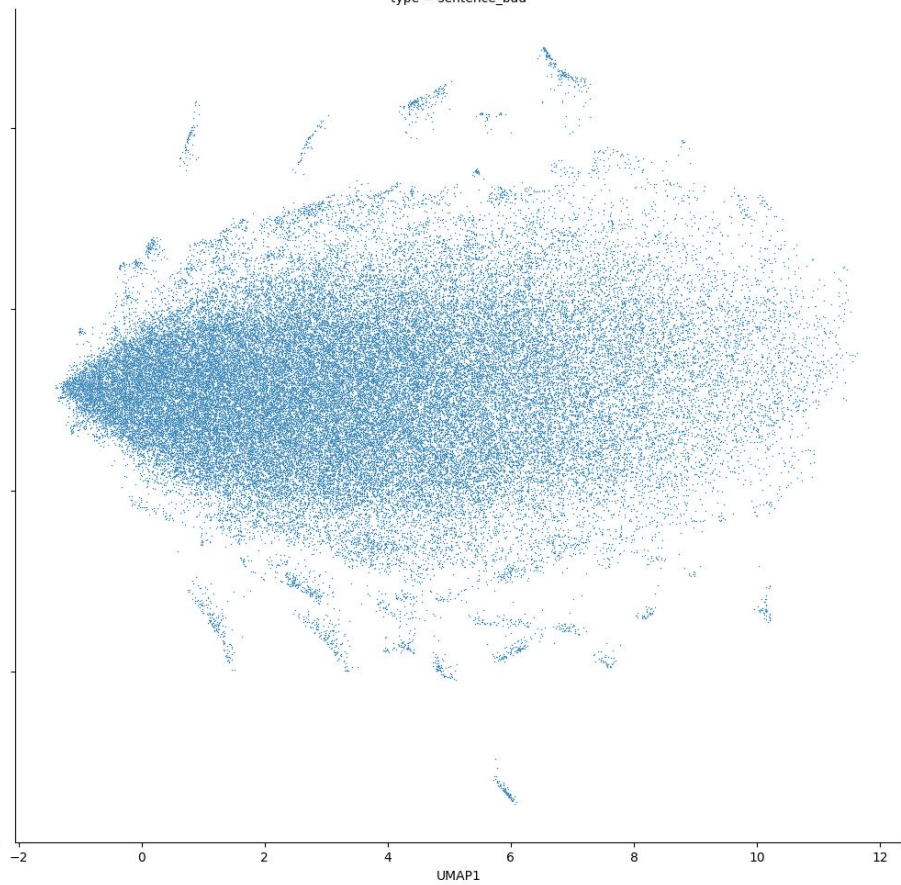




type = sentence_good



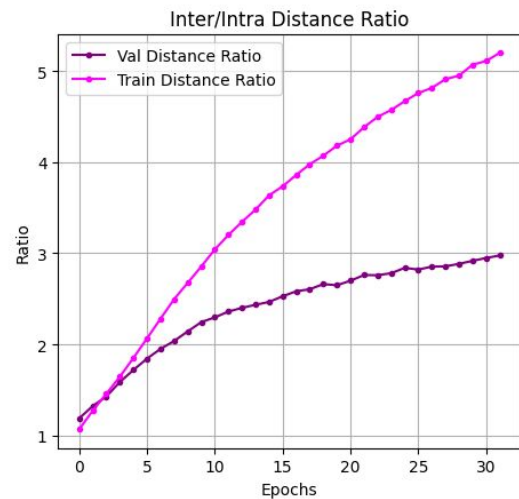
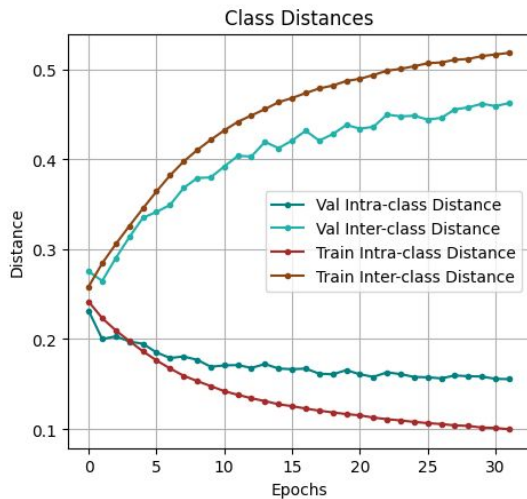
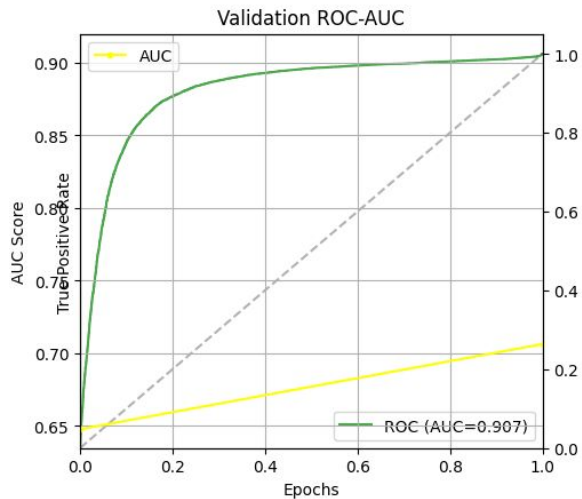
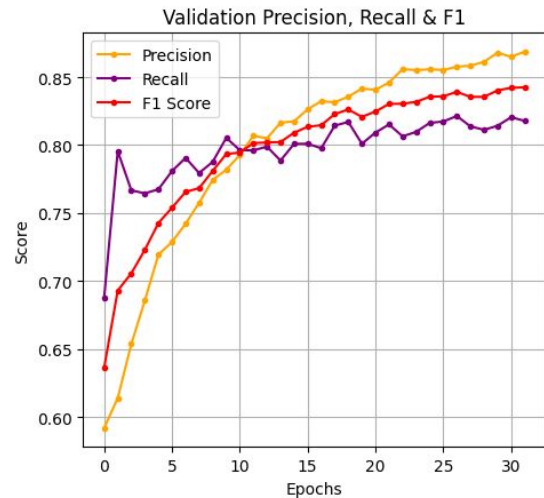
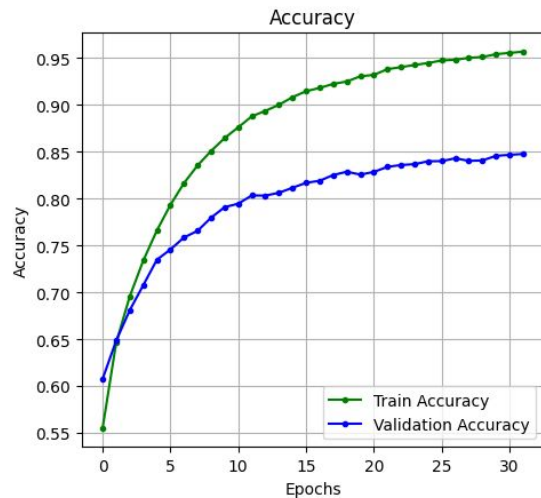
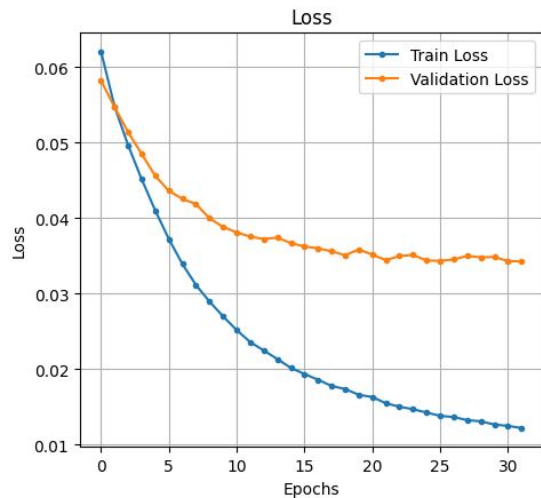
type = sentence_bad

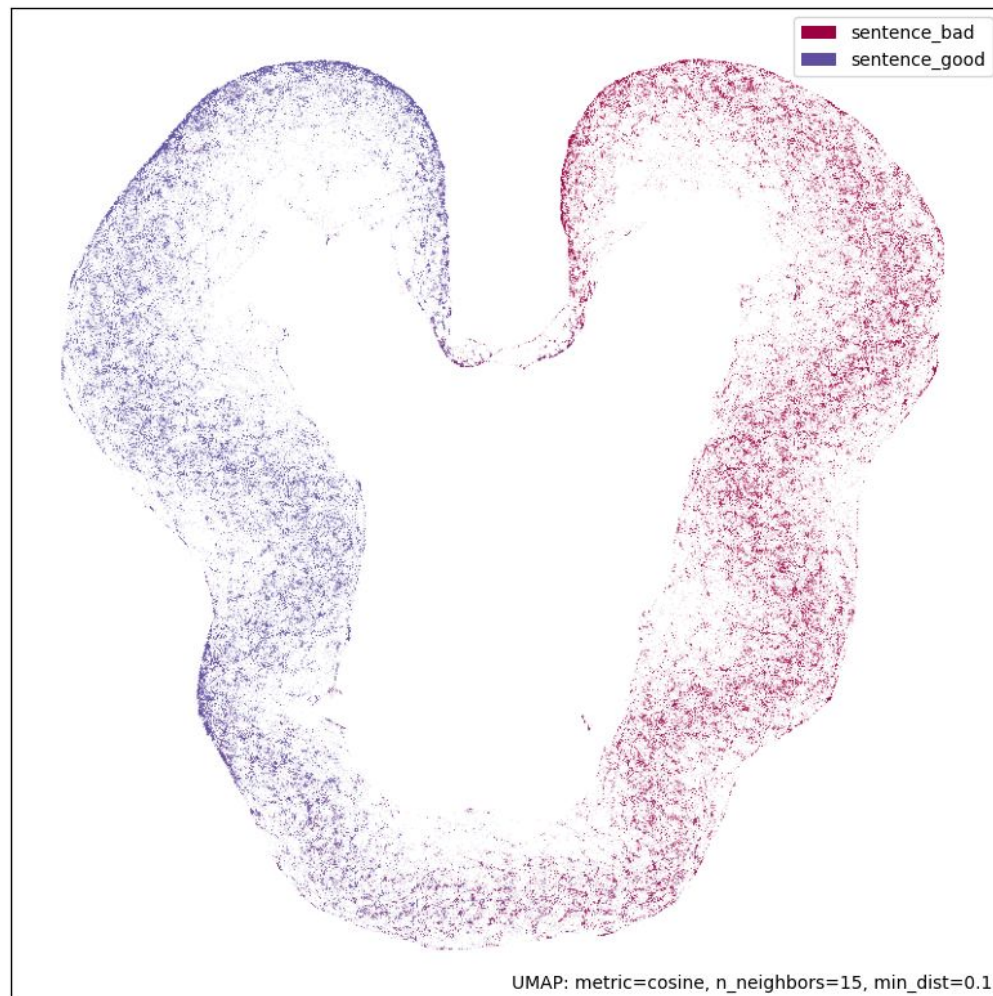


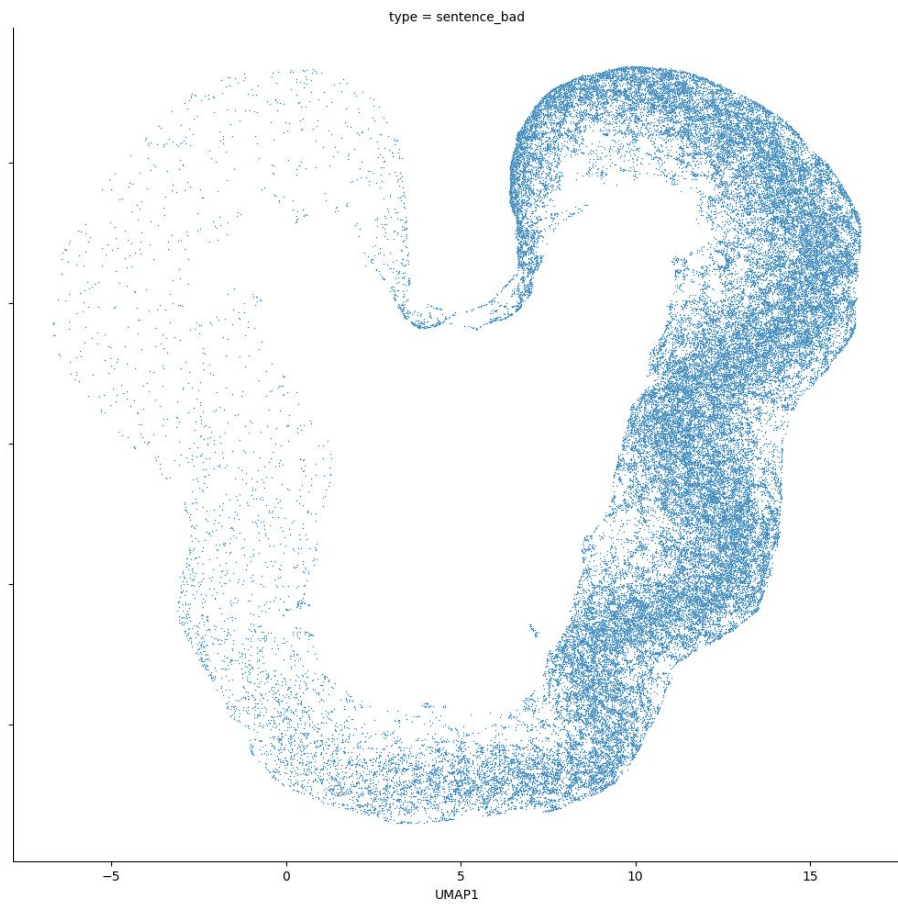
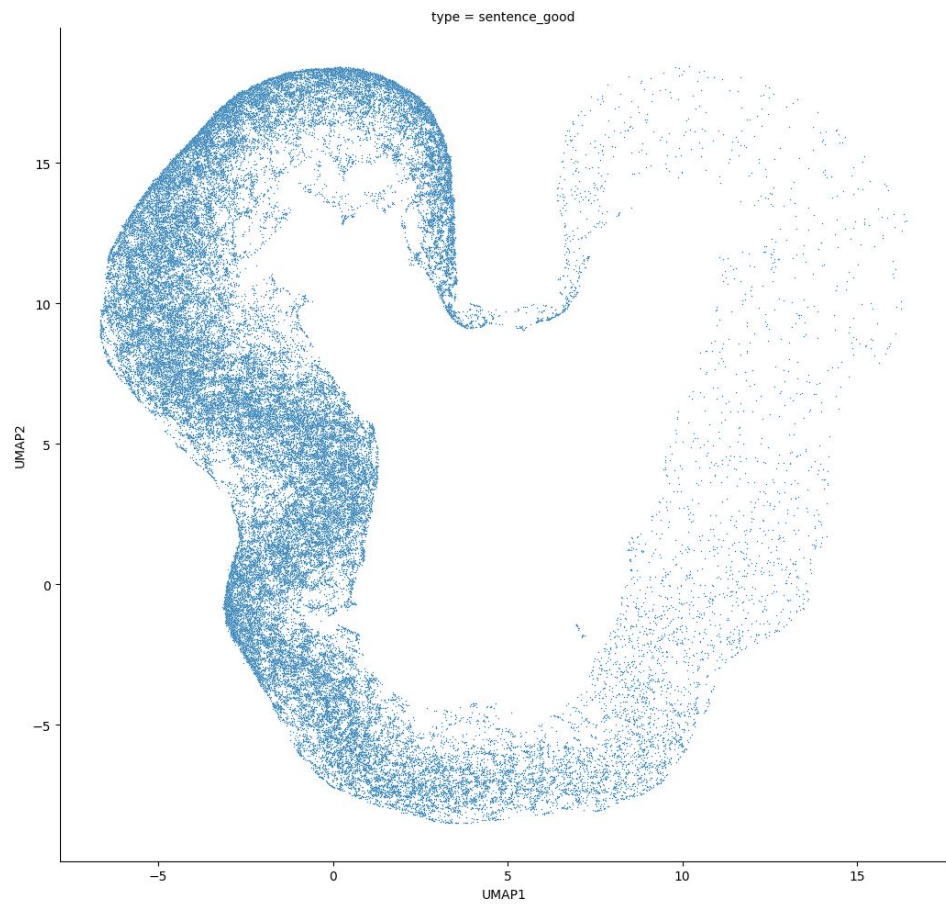
4. Дообучение матрицы проекций эмбеддингов

Модель 2: ProjectionMultiLayer

1. **Тип архитектуры:** Многослойная нейронная сеть
2. **Конфигурация слоёв:**
 - Входная размерность: 1024
 - Скрытый слой: 768 нейронов
 - Выходная размерность: 512
 - Коэффициент *dropout*: 0.1
3. **Параметры:**
 - Первый слой: ~786 тыс. параметров ($1024 \times 768 + 768$)
 - Второй слой: ~394 тыс. параметров ($768 \times 512 + 512$)
 - Всего параметров: ~1,18 млн
1. **Метрики качества:**
 - accuracy: 0.85
 - precision: 0.87
 - recall: 0.82
 - f1 score: 0.84
 - auc: 0.91
 - mean intra class dist: 0.16
 - mean inter class dist: 0.46
 - distance_ratio: 2.98







5. Выводы

1. **Chain-of-Thought** эффективнее простых вопросов:
 - Но и влияние формулировки вопроса имеет вес
 - Запрос на пошаговое объяснение повышает *accuracy* на **3-4%**
 - ☆ Интересно расхождение с поведением людей, для которых подобный эффект не характерен
2. **Few-shot** превосходит **zero-shot** (но есть нюанс):
 - Случайный выбор примеров [Асс. **86.87**] незначительно лучше *zero-shot* [Асс. **87.01**]
 - Целенаправленный подбор примеров повышает *accuracy* до **93.73**
3. **Наибольшую эффективность показали методы с использованием эмбеддингов:**
 - Поиск по разности эмбеддингов тестовой пары: [Асс. **93.73**]
 - Поиск по конкатенации разности и усреднения: [Асс. **93.58**]
 - Выбор примеров из того же грамматического явления: [Асс. **92.69**]
4. **Дообучение проекционной матрицы:**
 - Исходные эмбеддинги не содержат явного разделения по грамматичности
 - Простая линейная модель показывает результаты [Асс. **59**], сопоставимые с *random baseline*
 - Многослойная модель достигает [Асс. **85**], что сравнимо с *zero-shot* промптингом

Библиография

- ★ Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The Benchmark of Linguistic Minimal Pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- ★ Alex Warstadt, Amanpreet Singh, Samuel R. Bowman. 2019. [Neural Network Acceptability Judgments](#). *Transactions of the Association for Computational Linguistics*, 7: 625–641.

Спасибо за внимание!