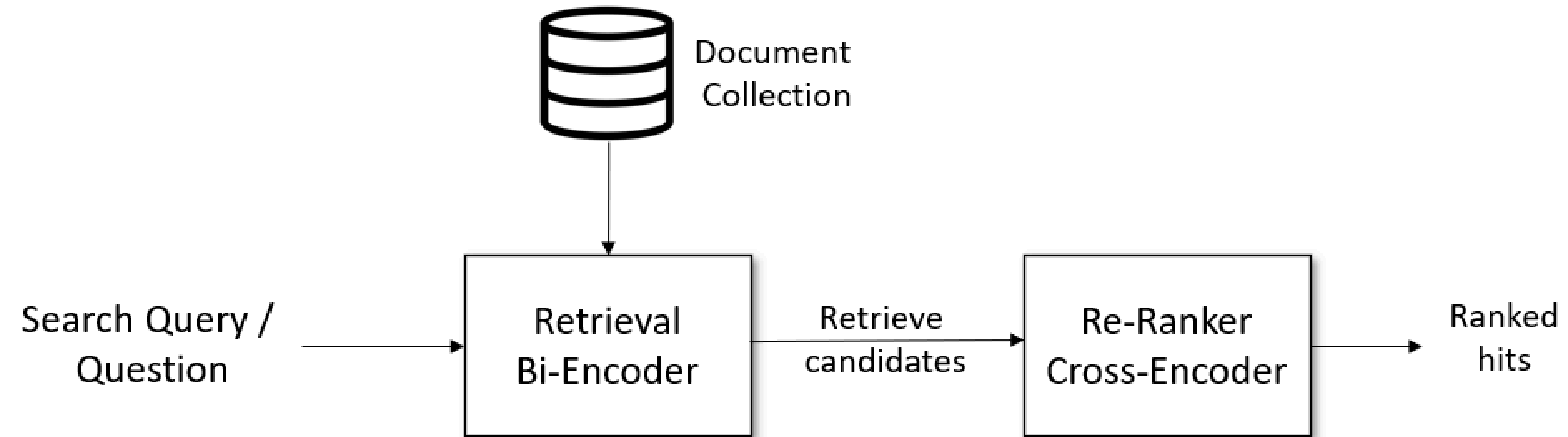


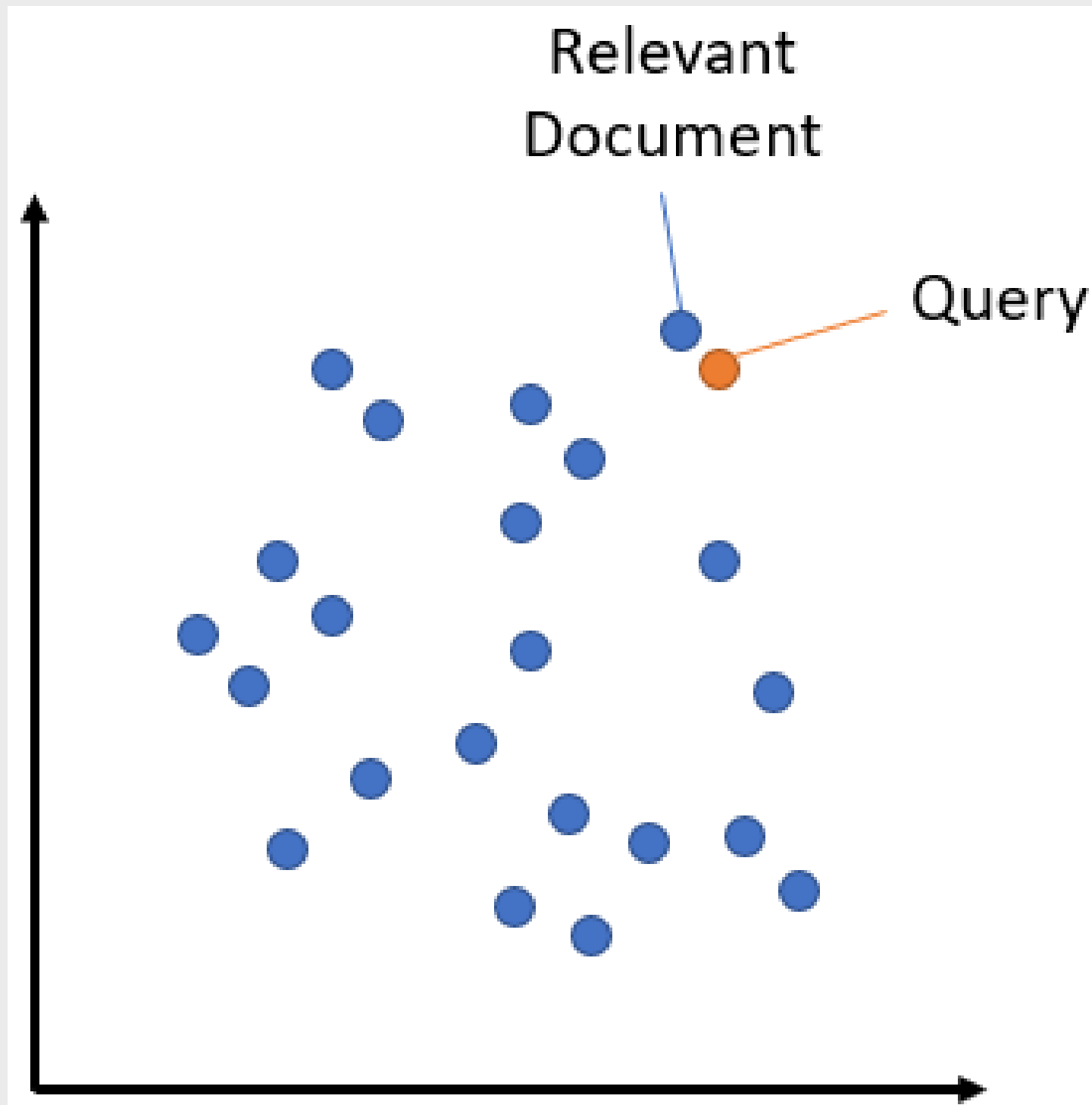
RUSSIAN CROSS- ENCODER

МАКАРОВА ЕЛЕНА
23.Б10

Retrieve & Re-Rank Pipeline

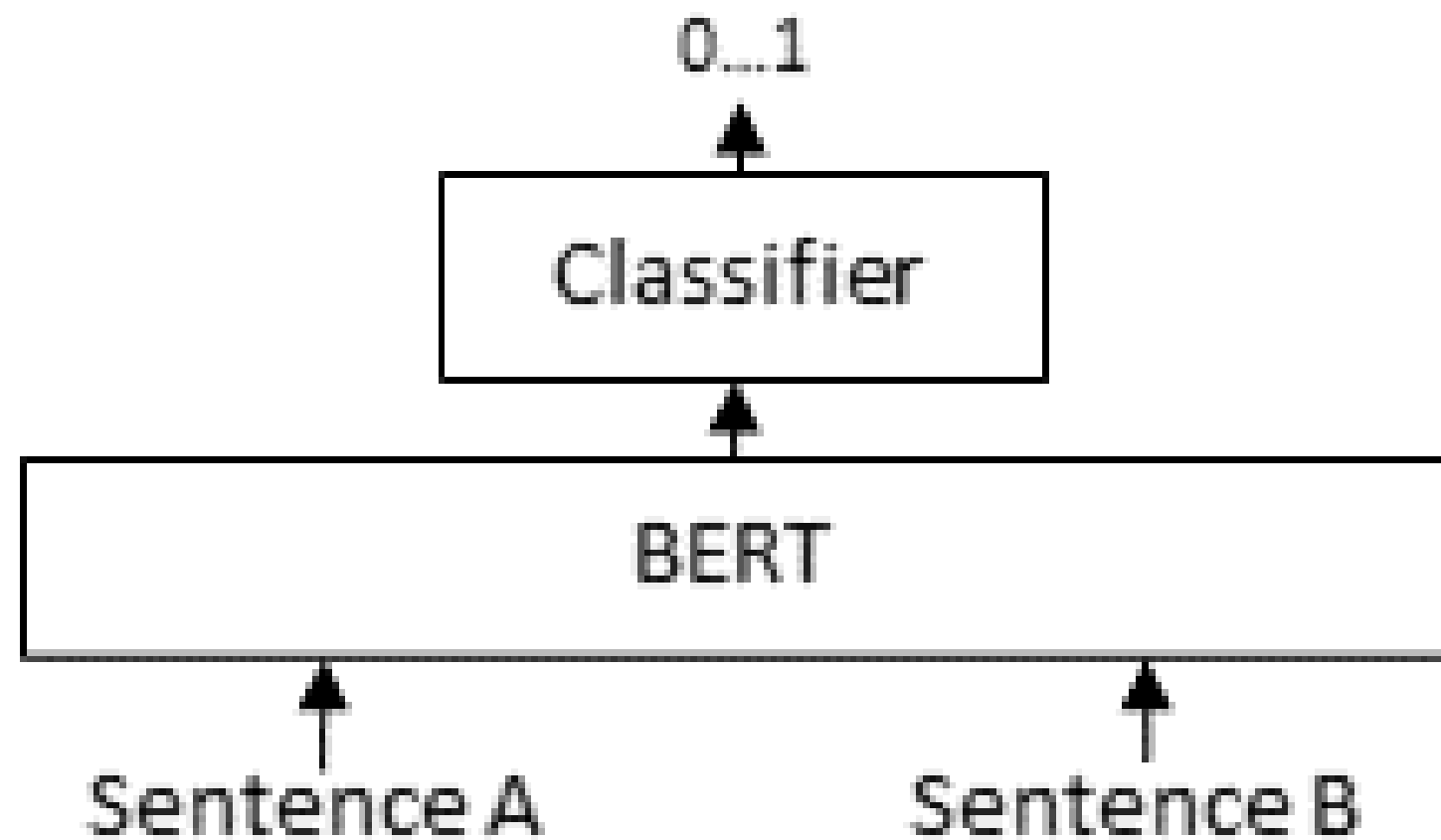


Retrieval: Bi-Encoder



Лексический поиск выполняет поиск буквальных соответствий запрашиваемых слов. Он не распознает синонимы, сокращения или орфографические вариации. В отличие от этого, семантический поиск (или плотный поиск) кодирует поисковый запрос в векторном пространстве и извлекает эмбединги близких по вектору документов.

Retrieval: Bi-Encoder



Запрос и возможный документ передаются одновременно в трансформер, который затем выдает единичную оценку от 0 до 1, указывающую, насколько документ соответствует данному запросу.

Базовая модель

Дообучалась на мультиязычной BERT для адаптации к русскому языку.

Данные для обучения:

- русскоязычная Википедия
- дампы новостных статей
- общий объем: ~10 ГБ текста

Базовая модель

Модель следует архитектуре BERT-Base:

- $L = 12$ (*количество слоев*)
- $H = 768$ (*размер скрытого состояния*)
- $A = 12$ (*голов внимания*)
- общее число параметров: ~180 млн

Cross-Encoder Russian MS-MARCO

- finetuning на MS-MARCO Russian passage ranking dataset
- ~0.2B параметров

Модель может быть использована для поиска информации на русском языке: задаётся запрос, сравнивается со всеми возможными ответами, затем ответы сортируются в порядке убывания.

Cross-Encoder Russian MS-MARCO

Эксперимент:

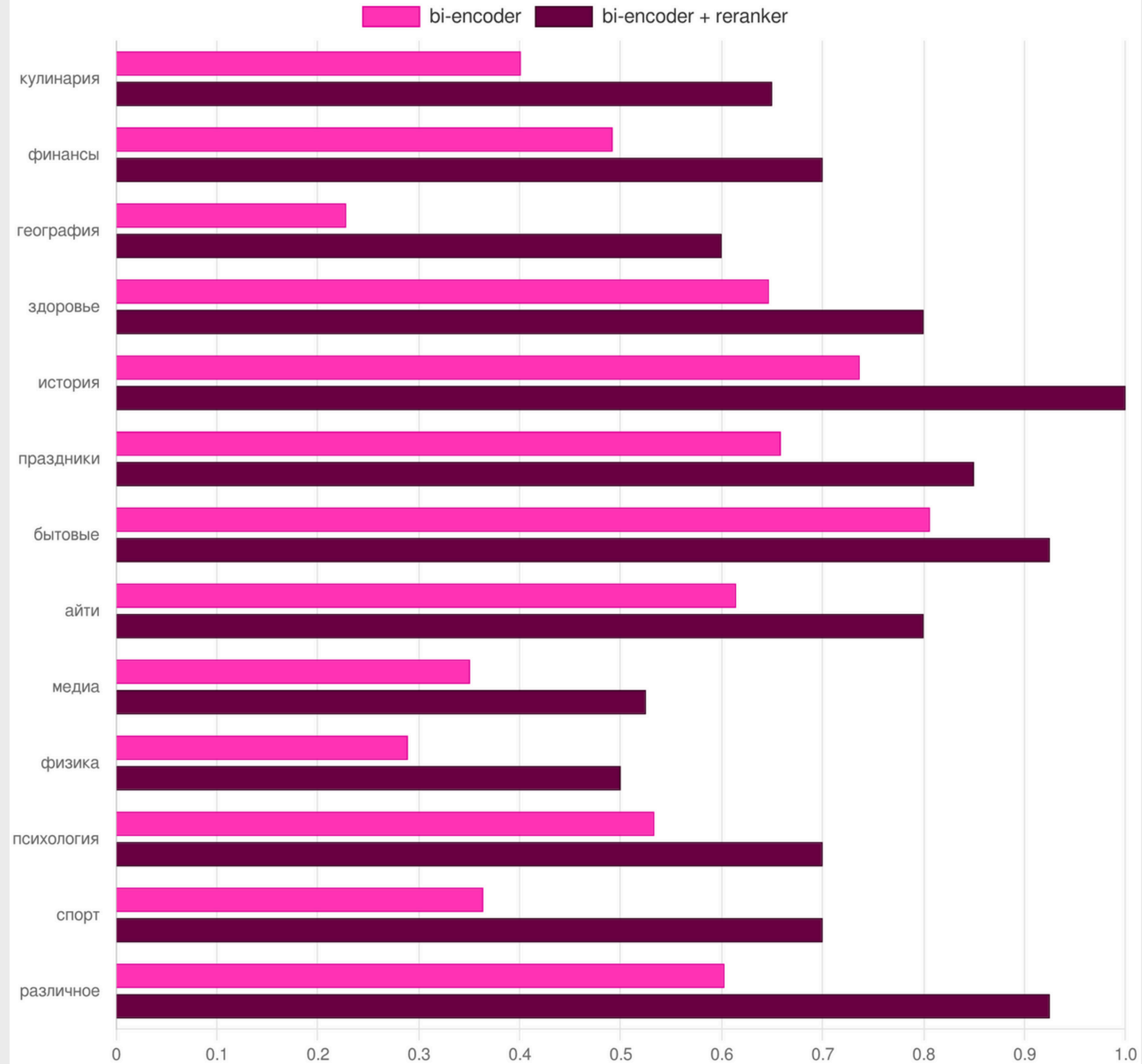
- корпус из ~1200 коротких текстов
- по 10 запросов из нескольких категорий: спорт, еда, здоровье, история, финансы и др.
- сравнение метрики MRR@k до и после использования reranker на отобранных при помощи bi-encoder запросов

Cross-Encoder Russian MS-MARCO

Пример на одном из запросов

```
Query: основой чего является фотоэлектрический эффект (gold idx: 843)
bi-encoder top20 indices: [843, 50, 146, 342, 70, 80, 680, 261, 392, 408, 376, 753, 258, 137, 853, 151, 53, 265, 818, 676]
Time: bi-encoder retrieval 0.052s
Reranked (index : score) :
 843 : 0.9202 --> Фотоэлектрический эффект – явление испускания электронов веществом под действием электромагнитного излучения (фотонов). Лежит в основе работы солнечных батарей.
 853 : 0.2068 --> Фотоэлектронная спектроскопия – метод исследования вещества, основанный на анализе энергии электронов, вылетающих при фотоэффекте.
 680 : 0.0677 --> Фототропизм – рост или движение организма (например, растения) в ответ на световой стимул.
 265 : 0.0331 --> Парниковый эффект – удержание тепла в атмосфере Земли из-за газов, таких как CO2 и метан.
 258 : 0.0289 --> Солнечные панели преобразуют энергию солнечного света в электричество с помощью фотоэлектрического эффекта.
Time: rerank 0.887s (batch_size=8)
Total time (bi + rerank): 0.939s
```

Сравнение моделей



Где используется

- RAG-системы
- классические поисковые системы
- рекомендательные системы
- детекция дубликатов и плагиата
- чат-боты, модерация контента и др.