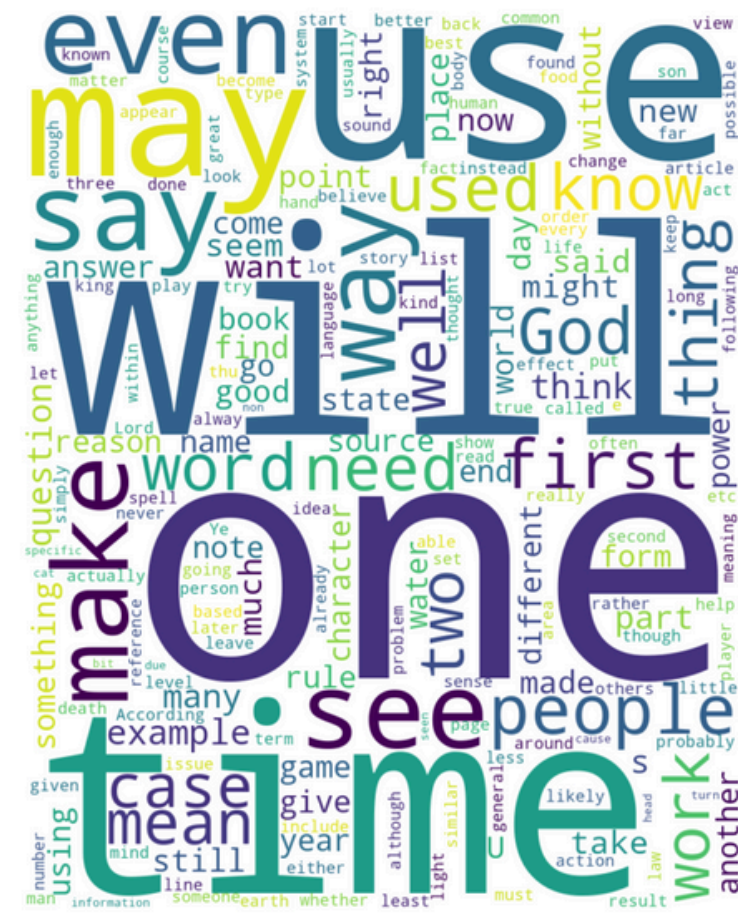# Personalized IR
SE-PQA

Armanni Luca 509085

Ghiotto Alessandro 513944

SE-PQA dataset

Dimension of answers and questions

Word Count - Answers

- - - Average Word Count = 234
- - - Median Word Count = 159

Word Count - Question Text

- - - Average Text Word Count: 118
- - - Median Text Word Count: 89
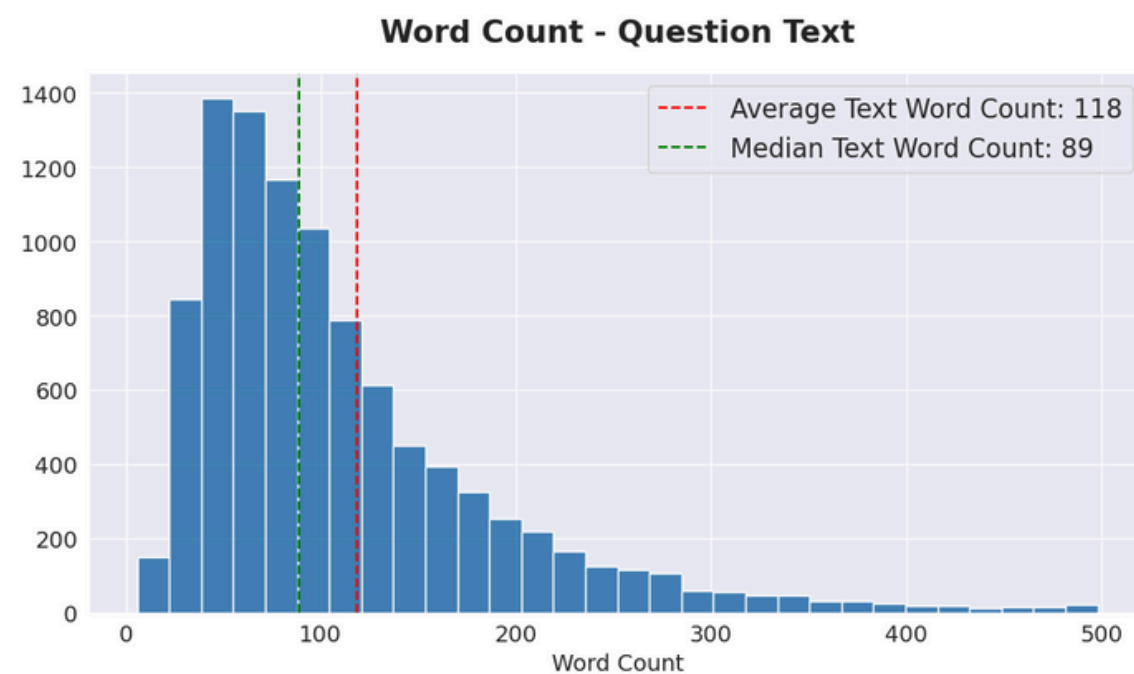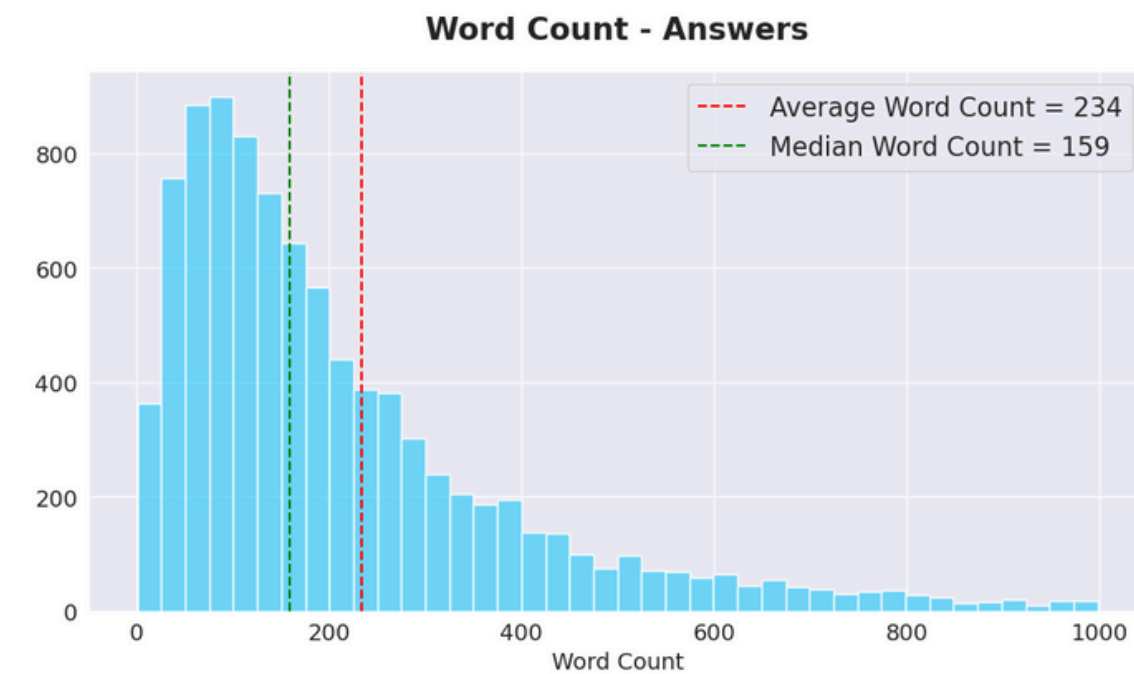
Topic of our documents

Sparsity of user profiles...

Number of Questions per User

Average: 2.83
Median: 1
Users with 1 question: 75.15%
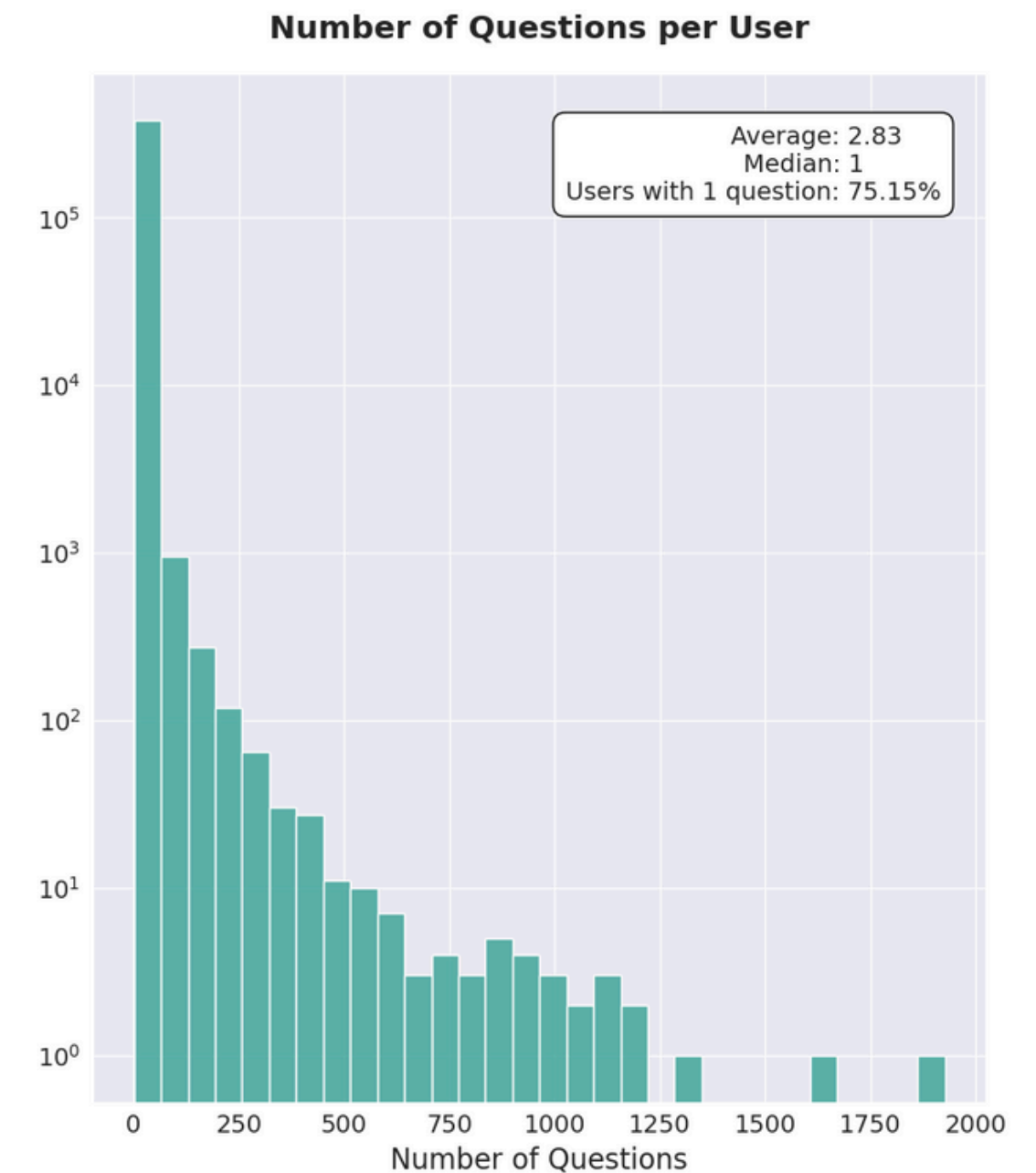
Roadmap

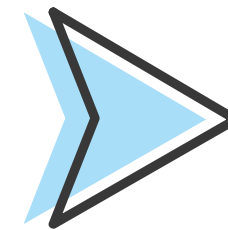**1** Baseline retrieval

2 Neural-reranking

3 Query expansion

4 Tags score

5 Learn to rank

6 Best model

1. Baseline retrieval

Input: question **text**

Best preprocessing

remove emojis

remove links

remove special characters

lowercase

remove whitespaces

remove stopwords

stemming
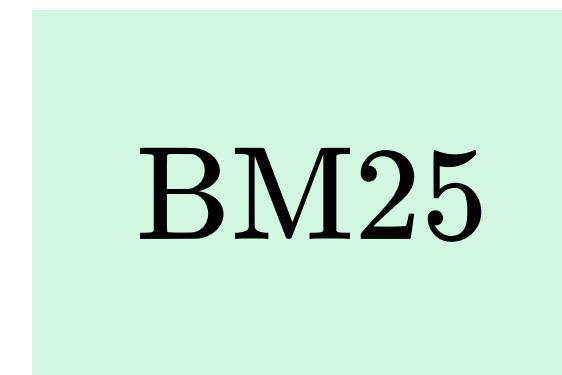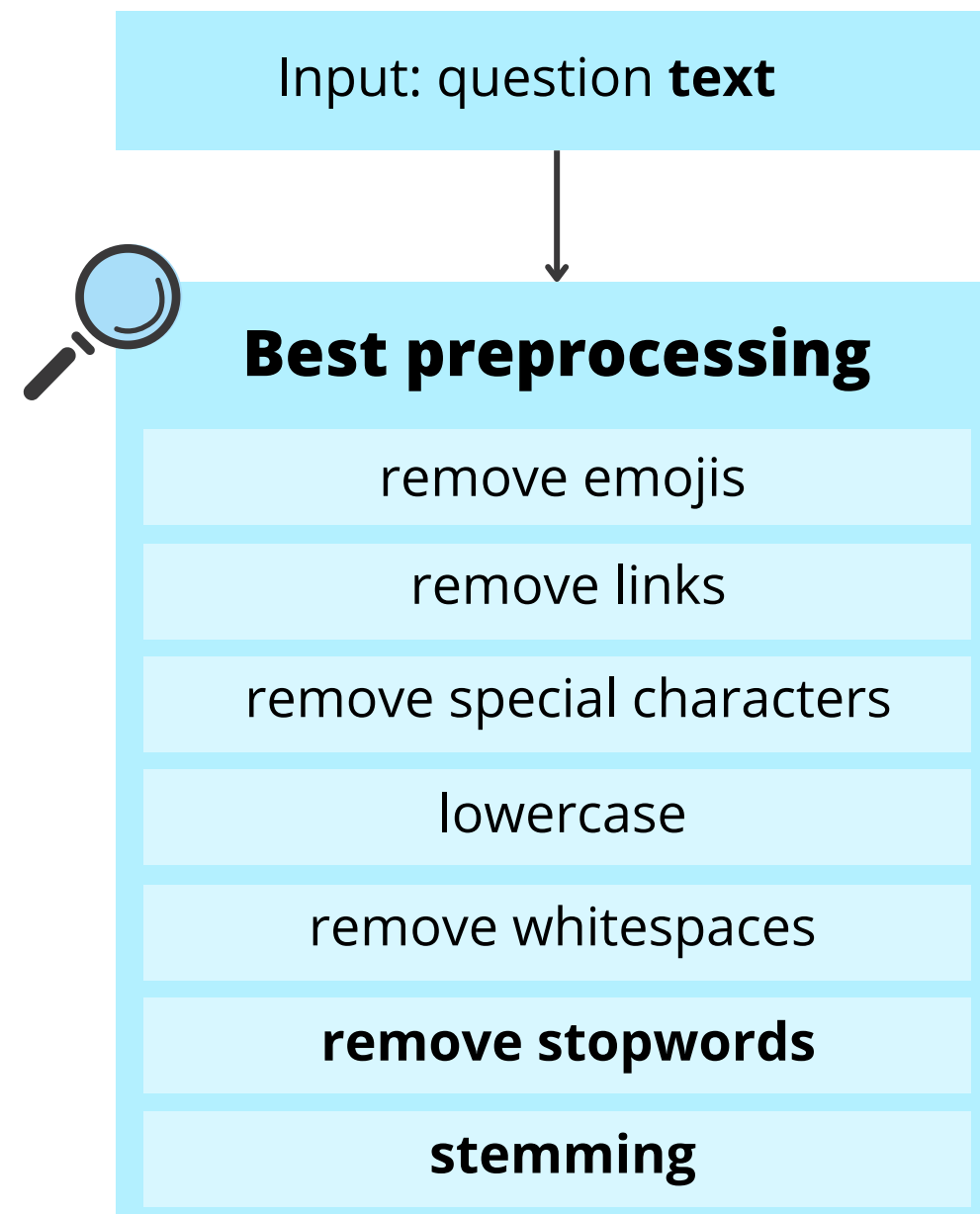
BM25
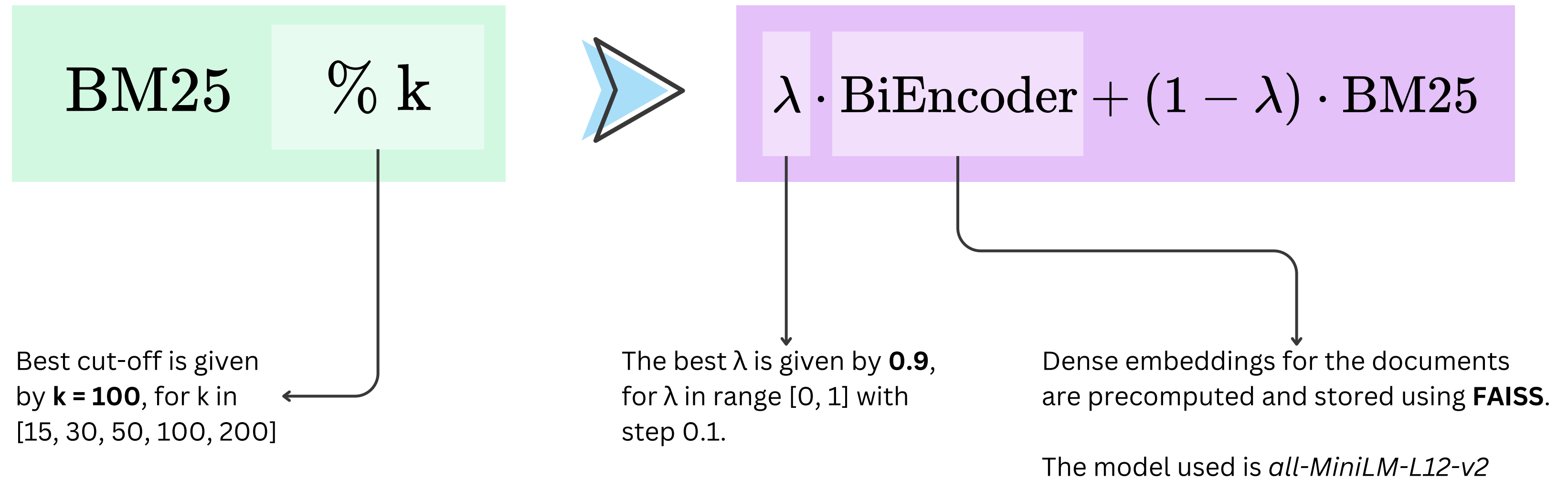
Best weighting parameters:
- c = 1.0, k_1 = 1.2 for precision
- c = 1.0, k_1 = 2.5 for recall

2. Neural re-ranking

$$\text{BM25} \quad \% \text{ k}$$

$$\lambda \cdot \text{BiEncoder} + (1 - \lambda) \cdot \text{BM25}$$

Best cut-off is given by **k = 100**, for k in [15, 30, 50, 100, 200]

The best λ is given by **0.9**, for λ in range [0, 1] with step 0.1.

Dense embeddings for the documents are precomputed and stored using **FAISS**.

The model used is *all-MiniLM-L12-v2*

3. Query expansion

BM25 % 100 → expand query → BiEncoder

zoom in

Original query ⟶
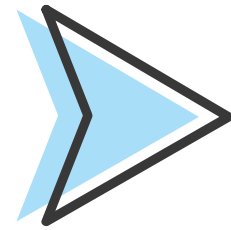Set of tags ⟶ LLM (*Phi-3-mini-4k-instruct*) ⟶ Expanded query
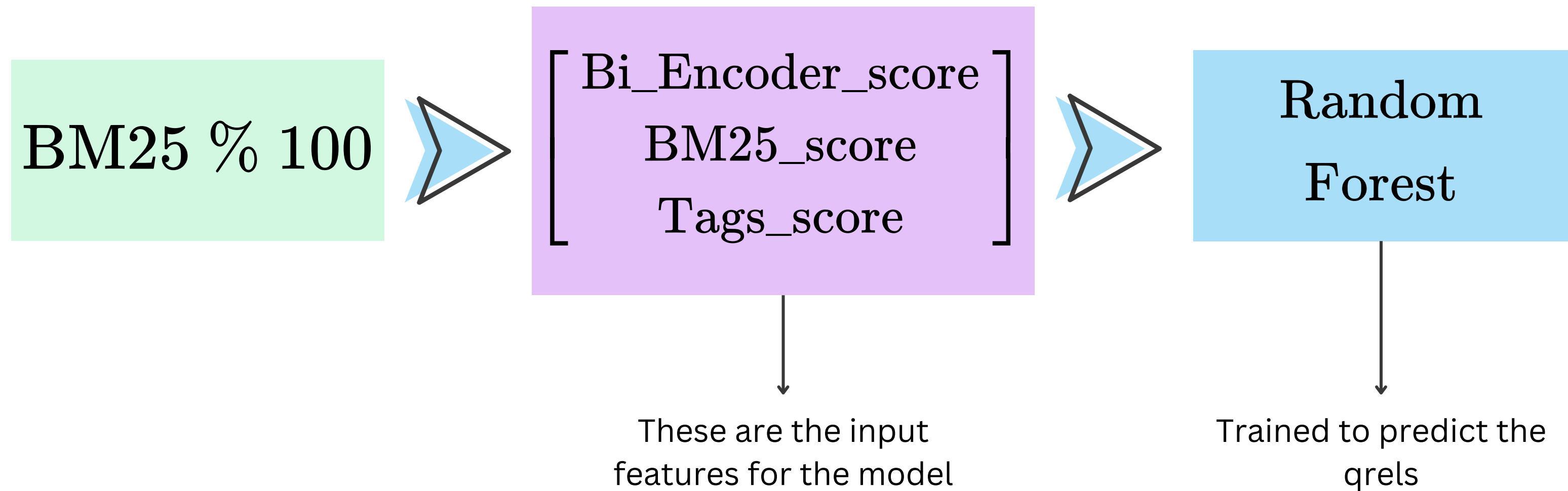
4. Tags score

BM25 % 100

$$\lambda_1 \cdot \text{BiEncoder\_score} + \lambda_2 \cdot \text{BM25\_score} + \lambda_3 \cdot \text{tags\_score}$$

best config:
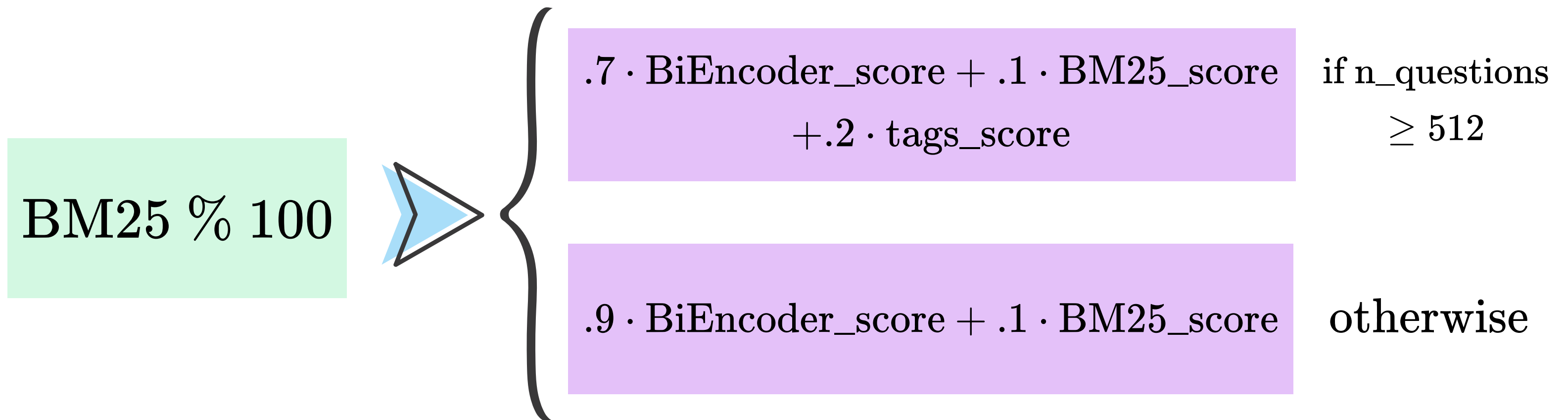$$(\lambda_1, \lambda_2, \lambda_3) = (.7, .1, .2)$$

$$\frac{len(intersection(Tags(u_q, t), Tags(u_a, t))}{len(Tags(u_q, t)) + 1}$$

5. Learn to Rank

BM25 % 100

$$\begin{bmatrix} Bi\_Encoder\_score \\ BM25\_score \\ Tags\_score \end{bmatrix}$$

Random Forest

These are the input features for the model

Trained to predict the qrels

6. Best model

$$\text{BM25 \% 100} \Rightarrow \begin{cases} .7 \cdot \text{BiEncoder\_score} + .1 \cdot \text{BM25\_score} \\ \qquad + .2 \cdot \text{tags\_score} & \text{if n\_questions} \geq 512 \\ \\ .9 \cdot \text{BiEncoder\_score} + .1 \cdot \text{BM25\_score} & \text{otherwise} \end{cases}$$

We use the personalized model if we have enough data about the user

Results

Results on val set

Configurations
- 1. BM25
- 2. Re-Ranking
- 3. Query expansion
- 4. Tags Score
- 5. Random Forest
- 6. Best model
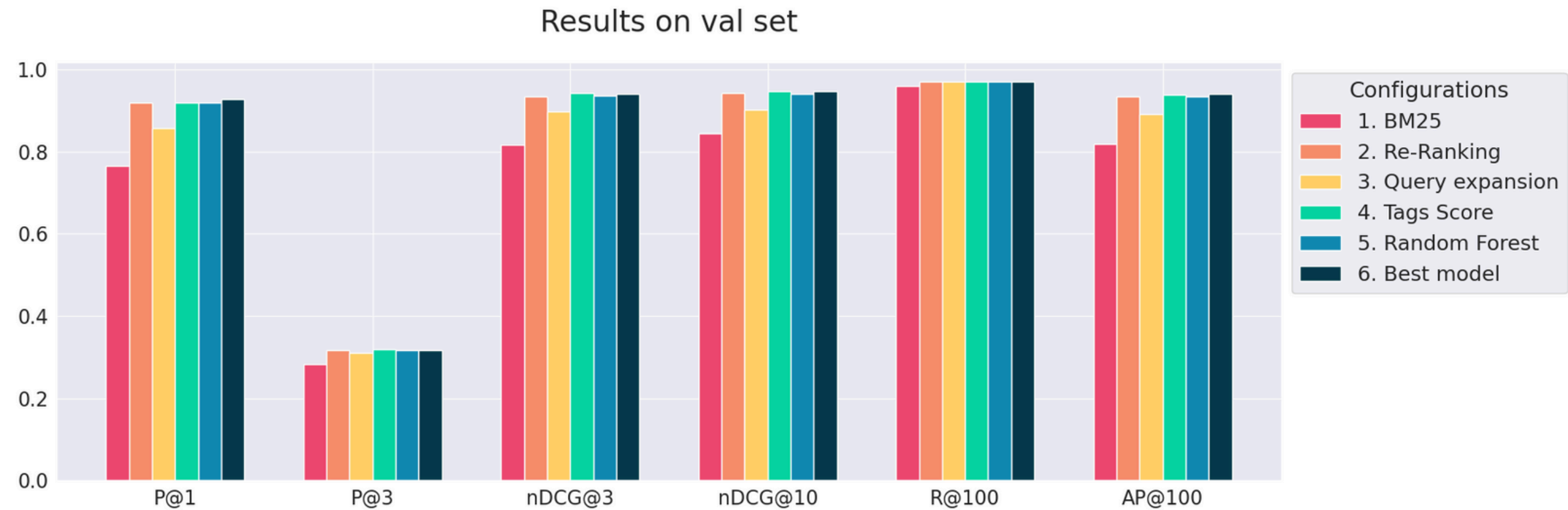
Best model on test set

P@1	P@3	nDCG@3	nDCG@10	R@100	AP@100
0.878	**0.306**	0.901	0.915	0.969	0.902

Thanks for your attention!