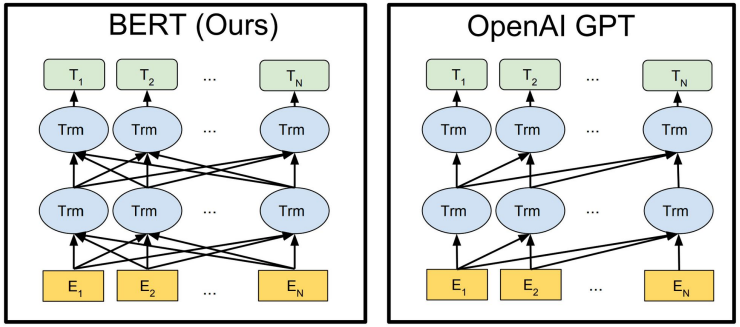
Investigation of approaches for evaluating semantic similarity of texts in the task of matching vacancies and resumes

Main idea:

Obtain vector representations of texts and measure text proximity using **cosine similarity**

Method:

1. Dataset: expertly tagged and synthetic positive pairs
2. Ranking resumes by vacancy, quality metric: **MAP@K**
3. Use **Transformer-based models** for text vectorization: BERT & GPT
4. Additional training of embedders on positive pairs: contrastive loss



$$p@K = \frac{\sum_{k=1}^{K} r^{true}(\pi^{-1}(k))}{K} = \frac{\text{релевантных элементов}}{K}$$

model	Random	BERT multilingual	BERT Russian	DeepPavlov ruBERT
MAP@10	0.7	0.013	0.015	0.017
MAP@20	0.005	0.015	0.018	0.018