CAN LLM AS ENCONDER GOOD?

Task: Entity Matching

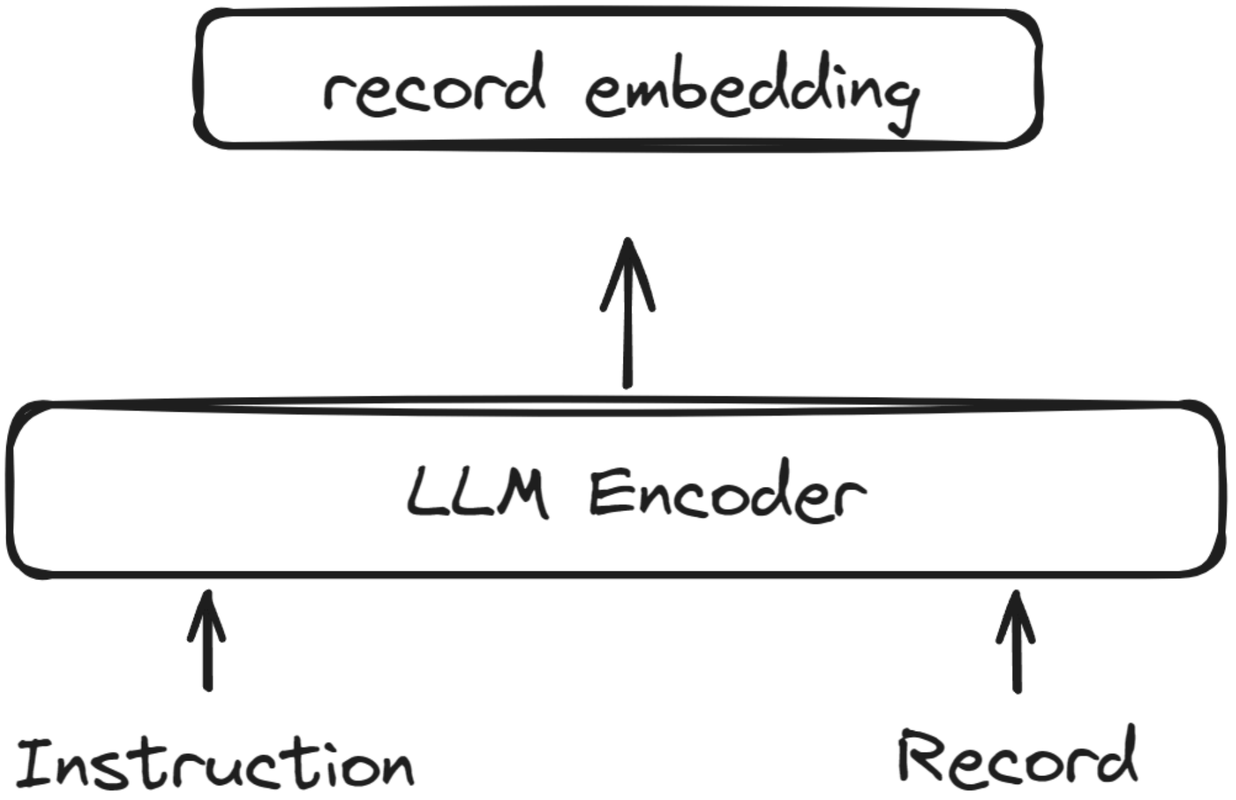
Starting problem

LLM achieve state of the art in the tast of Entity Matching in the majority of the dataset, but have several problem like:

- opacity;
- unable to provide numeric accuracy

Idea

Using the idea of **LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders**, we want to embed a record and use numerical vector instead of generative answer for evaluation.



LLM2Vec

briefly

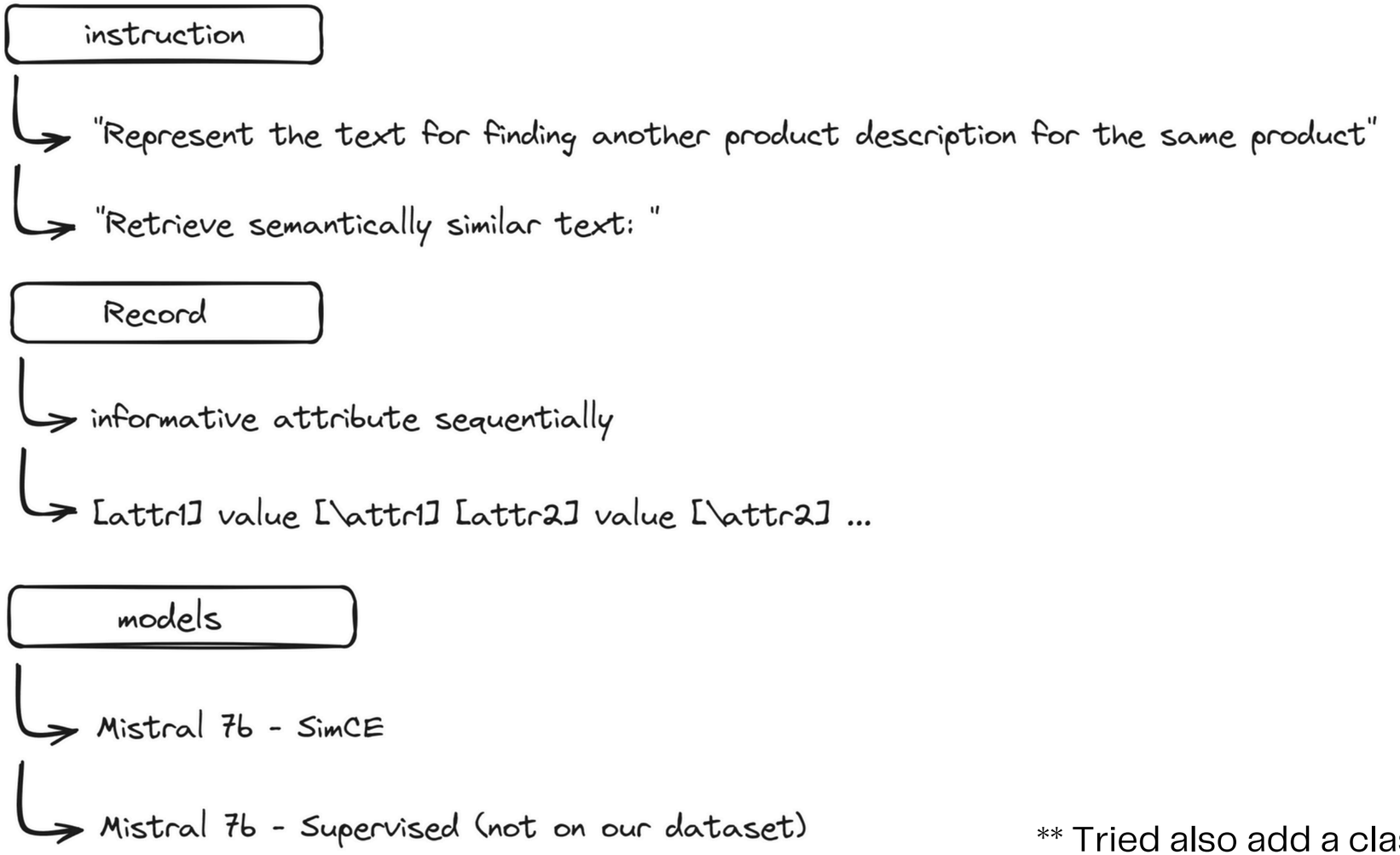
1. **Enabling bidirectional attention (Bi):** The first step of the LLM2Vec approach is to replace the causal attention mask of decoder-only LLMs by an all-ones matrix. This gives each token access to every other token in the sequence, converting it into a bidirectional LLM.
2. **Masked next token prediction (MNTP):** training objective that combines next token prediction with masked language modeling. Used to adapt a model to make use of its bidirectional attention.
3. **Unsupervised contrastive learning (SimCSE):** The model is trained to maximize the similarity between two representations while minimizing the similarity with representations of other sequence in the batch.

Dataset

Standard dataset taken by github Deepmatcher for benchmarking.
The first is clean, the others are dirty by inserting noise in the records.

	A_size	B_size	testing_pairs	num_pos_match	num_neg_match	attributes
abt_buy	1081	1092	1916	206	1710	3
dirty_dblp_acm	2616	2294	2473	444	2029	4
dirty_dblp_scholar	2616	64263	5742	1070	4672	4
dirty_itunes_amazon	6907	55923	109	27	82	8
dirty_walmart_amazon	2554	22074	2049	193	1856	5

Approches



** Tried also add a classifier on top with null result

Results

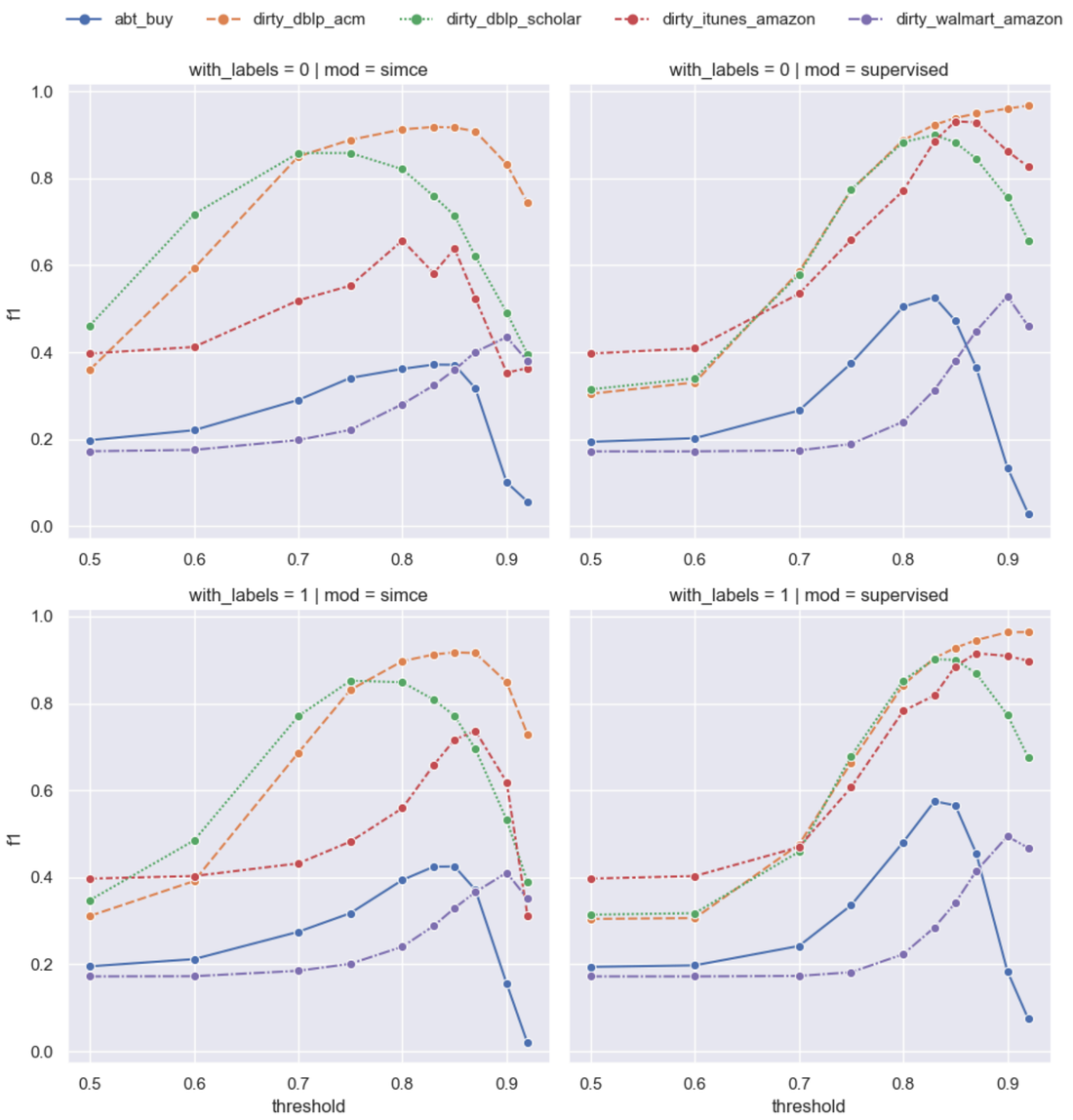
Performance of the LLM2Vec + Cosine Similarity:

- threshold: 0.5 – 0.93;
- modality: simCSE or Supervised model;
- with_labels:
  - 0: no attribute separator in the record
  - 1: otherwise.

Fixed:

- Instruction: “Represent the text for finding another product description for the same product”

	dataset_name	threshold	mod	with_labels	recall	precision	f1
DSM1	abt_buy	0.83	supervised	1	0.71	0.48	0.57
DSM2	dirty_itunes_amazon	0.85	supervised	0	1.00	0.87	0.93
DSM3	dirty_dblp_acm	0.92	supervised	0	0.96	0.97	0.97
DSM4	dirty_dblp_scholar	0.83	supervised	1	0.94	0.87	0.90
DSM5	dirty_walmart_amazon	0.90	supervised	0	0.56	0.50	0.53



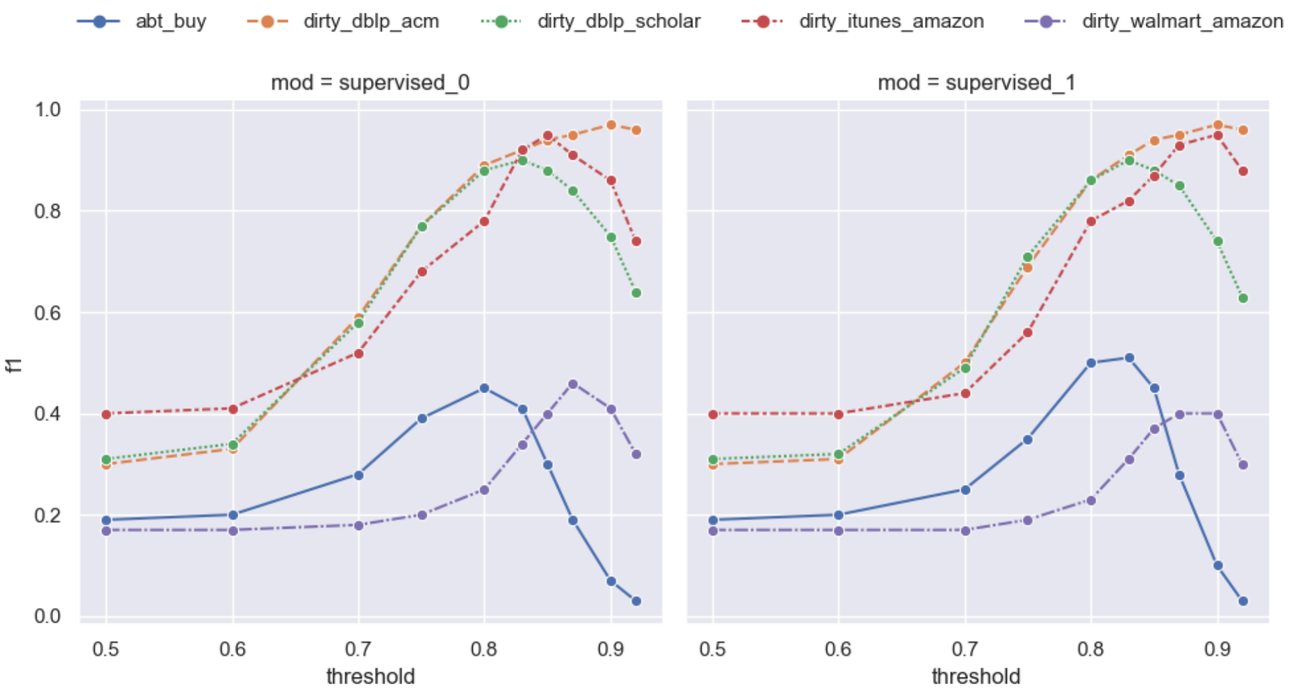
Results

Performance of the LLM2Vec + Cosine Similarity:

- threshold: 0.5 – 0.93;
- modality: simCSE or Supervised model;
- with_labels:
  - 0: no attribute separator in the record
  - 1: otherwise.

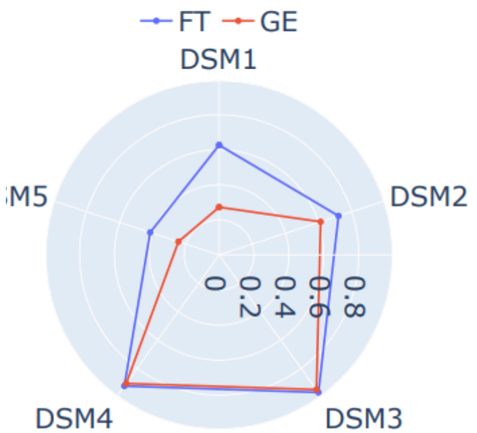
Fixed:

- Instruction: “Retrieve semantically similar text”

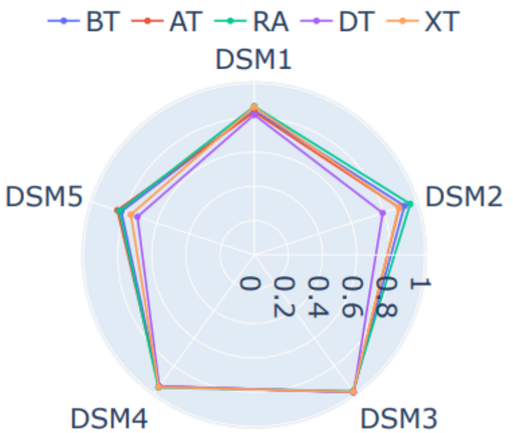


	dataset_name	threshold	mod	with_labels	recall	f1	precision
DSM1	abt_buy	0.83	supervised	1	0.49	0.51	0.55
DSM2	dirty_itunes_amazon	0.90	supervised	1	0.96	0.95	0.93
DSM3	dirty_dblp_acm	0.90	supervised	0	0.97	0.97	0.96
DSM4	dirty_dblp_scholar	0.83	supervised	0	0.90	0.90	0.91
DSM5	dirty_walmart_amazon	0.87	supervised	0	0.56	0.46	0.39

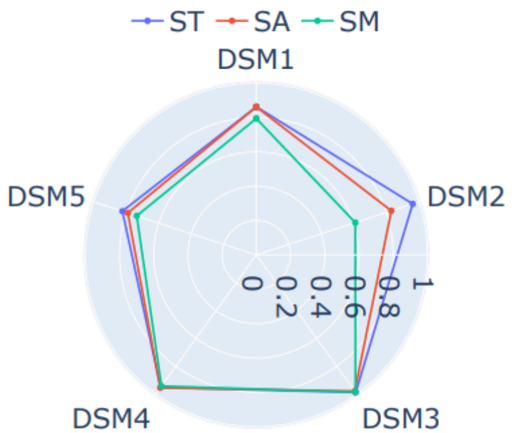
Results SoTA



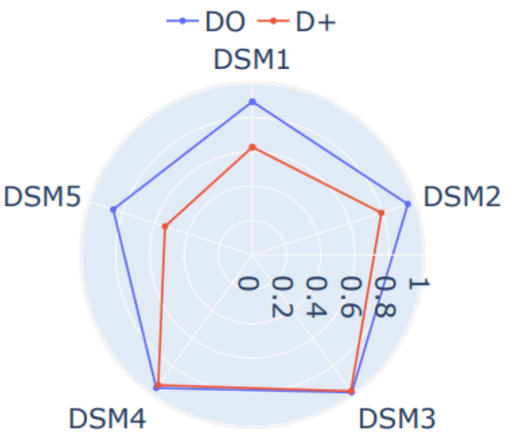
(a) static



(b) BERT



(c) SBERT



(d) SotA

https://dl.acm.org/doi/abs/10.14778/3598581.3598594

Results Final

In the table we have the merged result of the experiments.
We can see that the prompt, threshold and record structure impact on the F1 score (hence the LLM2Vec embeddings) with no a priori dominance of those but the Supervision finetuning of the LLM.

	dataset_name	threshold	mod	with_labels	recall	precision	f1	instr
DSM1	abt_buy	0.83	supervised	1	0.71	0.48	0.57	0
DSM2	dirty_itunes_amazon	0.85	supervised	0	1.00	0.90	0.95	1
DSM3	dirty_dblp_acm	0.90	supervised	1	0.97	0.96	0.97	1
DSM4	dirty_dblp_scholar	0.83	supervised	1	0.94	0.87	0.90	0
DSM5	dirty_walmart_amazon	0.90	supervised	0	0.56	0.50	0.53	0

We can see that the LLM are able to generalize more with dirty dataset, while are more strictly with the clean records.