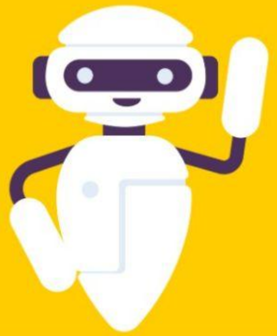
GROUP 32

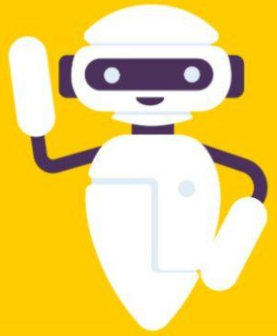
USC Viterbi
School of Engineering

LLM2VEC: LARGE LANGUAGE MODELS ARE SECRETLY POWERFUL TEXT ENCODERS

TEAM MEMBERS:

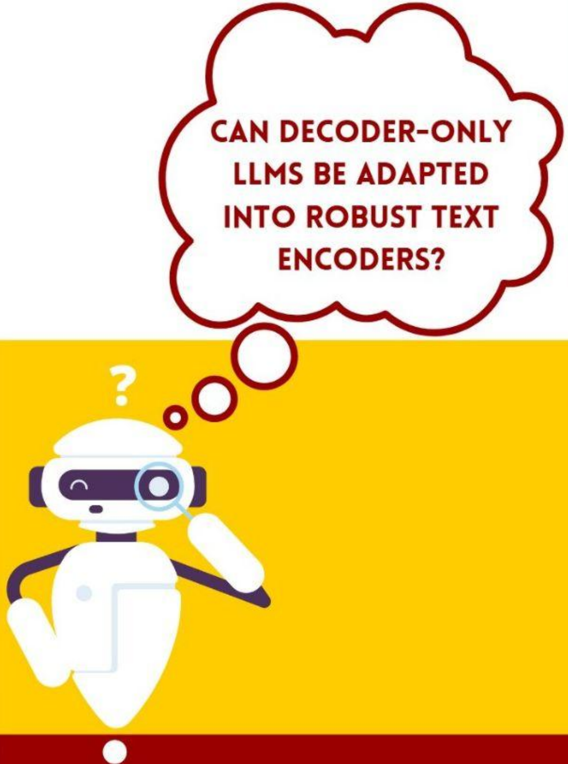


* AANANDHI SONDURI PANTHANGI *
* AKHILAA SONDURI PANTHANGI *
* MOHAN SAI GANESH KANNA *
* MONA TEJA KURAKULA *
* EMMA LEIHE *



MAIN RESEARCH QUESTION IN THE PAPER

USC Viterbi School of Engineering



- **CHALLENGES:**

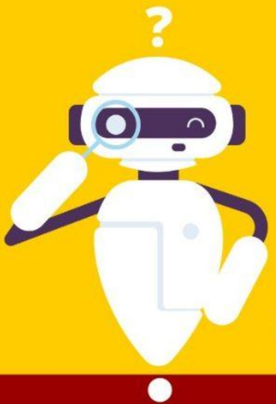
  DECODER-ONLY MODELS ARE TYPICALLY UNIDIRECTIONAL, LIMITING CONTEXT.

- **SOLUTION (LLM2VEC):**

  CONVERTS LLMS INTO TEXT ENCODERS WITH MINIMAL ADAPTATION

STEPS:

1. BI-DIRECTIONAL ATTENTION
2. MASKED NEXT TOKEN PREDICTION (MNTP)
3. UNSUPERVISED CONTRASTIVE LEARNING



MAIN RESULTS SUMMARIZED

USC Viterbi
School of Engineering

Categories → # of datasets →	Retr. 15	Rerank. 4	Clust. 11	PairClass. 3	Class. 12	STS 10	Summ. 1	Avg 56
Encoder-only								
BERT	10.59	43.44	30.12	56.33	61.66	54.36	29.82	38.33
BERT + SimCSE	20.29	46.47	29.04	70.33	62.50	74.33	31.15	45.45
S-LLaMA-1.3B								
Uni + w. Mean	9.47	38.02	28.02	42.19	59.79	49.15	24.98	35.05
LLM2Vec (w/o SimCSE)	15.48	40.99	31.83	50.63	64.54	62.06	26.82	41.43
LLM2Vec	25.93	47.70	37.45	72.21	67.67	71.61	31.23	49.42
Echo	10.36	41.52	30.03	52.08	63.75	59.36	22.79	39.10
LLaMA-2-7B								
Ui + w. Mean	15.16	46.94	36.85	61.41	69.05	63.42	26.64	44.54
LLM2Vec (w/o SimCSE)	19.86	44.74	35.31	61.60	67.94	66.74	26.83	45.70
LLM2Vec	36.75	52.95	40.83	77.89	71.57	76.41	31.38	55.36
Echo	16.16	46.84	34.25	63.54	69.82	67.95	25.57	45.36
Mistral-7B								
Uni + w. Mean	10.43	45.11	35.82	60.28	71.14	58.59	26.57	42.46
Bi + Mean	15.84	47.40	35.55	66.53	72.18	71.04	29.93	46.86
LLM2Vec (w/o SimCSE)	19.74	50.43	40.06	70.95	72.51	71.90	27.84	49.43
LLM2Vec	38.05	**53.99**	40.63	**80.94**	**74.07**	**78.50**	30.19	**56.80**
Echo	22.68	51.07	36.78	75.87	72.69	73.60	29.54	50.26
Meta-LLaMA-3-8B								
Uni + w. Mean	15.17	46.22	36.84	60.94	67.41	62.80	25.51	43.98
Bi + Mean	3.90	34.56	14.27	42.71	57.89	51.15	23.26	30.56
LLM2Vec (w/o SimCSE)	24.75	49.20	39.74	65.91	69.00	67.85	25.59	48.84
LLM2Vec	**39.19**	53.09	**41.99**	78.01	71.88	75.86	**31.45**	56.23
Echo	12.58	49.79	36.32	68.95	70.22	67.43	26.44	45.32

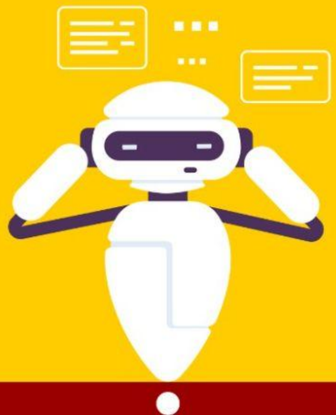
Table 1: Unsupervised results on MTEB. We compare S-LLaMA-1.3B, LLaMA-2-7B, Mistral-7B, and Meta-LLaMA-3-8B with and without LLM2Vec to the unsupervised BERT models of Gao et al. (2021) as well as Echo embeddings (Springer et al., 2024).

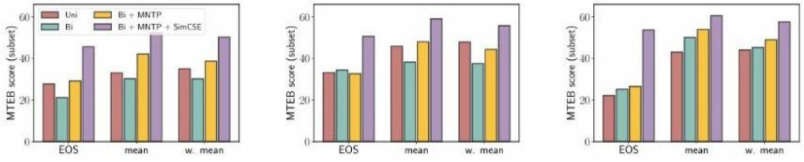
Model	EOS	Mean	W. mean
S-LLaMA-1.3B			
Uni	27.72	33.03	34.99
Bi	21.16	30.26	30.20
Bi + MNTP	29.16	42.10	38.67
Uni + SimCSE	37.44	44.95	47.13
Bi + SimCSE	40.43	44.46	44.83
Bi + MNTP + SimCSE	45.57	**52.40**	50.23
LLaMA-2-7B			
Uni	33.23	45.83	47.85
Bi	34.47	38.22	37.50
Bi + MNTP	32.66	48.00	44.30
Uni + SimCSE	38.47	52.03	53.55
Bi + SimCSE	40.37	44.13	44.08
Bi + MNTP + SimCSE	50.61	**58.97**	55.75
Mistral-7B			
Uni	22.12	43.00	44.01
Bi	25.17	50.07	45.20
Bi + MNTP	26.54	53.89	48.93
Uni + SimCSE	34.60	52.04	53.95
Bi + SimCSE	49.73	60.29	56.56
Bi + MNTP + SimCSE	53.67	**60.50**	57.55

Table 5: Unsupervised results on MTEB subset for different models.

SUMMARY:

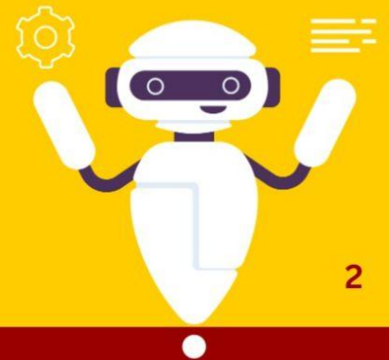
LLM2VEC TRANSFORMS DECODER-ONLY LLMS INTO COMPETITIVE TEXT ENCODERS, OUTPERFORMING SEVERAL ENCODER-ONLY MODELS WITH MINIMAL COMPUTATION, CONFIRMING THE EFFECTIVENESS OF BIDIRECTIONAL AND CONTRASTIVE LEARNING ENHANCEMENTS.





(a) S-LLaMA-1.3B (b) Llama-2-7B (c) Mistral-7B

Figure 3: Unsupervised results on our 15 task subset of the MTEB dataset. We ablate three different pooling choices: EOS, mean pooling, and weighted mean pooling. LLM2Vec is compatible with all three approaches and works best with mean pooling.

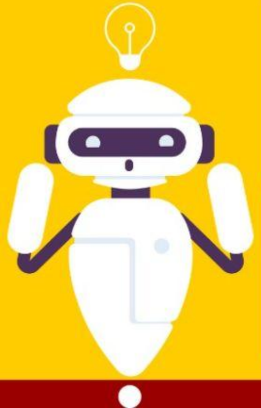


HOW THIS INFLUENCES OUR PROJECT

USC Viterbi
School of Engineering

CONTEXTUAL CANDIDATE MATCHING: AN INTELLIGENT APPROACH TO RESUME RETRIEVAL FOR A GIVEN JOB DESCRIPTION

1. LLM2VEC'S BIDIRECTIONAL ATTENTION AND CONTRASTIVE LEARNING CAN CAPTURE NUANCES LIKE "AWS AND GCP" AS EQUIVALENT TO "CLOUD-BASED SYSTEMS."
2. OUR PROJECT COULD INCORPORATE THESE METHODS WITHOUT THE NEED FOR HIGH COMPUTATIONAL RESOURCES.
3. THE TRANSFORMATION PROCESS ALLOWS THE USE OF EXISTING MODELS WITHOUT MAJOR RETRAINING.
4. DPR CAN BENEFIT FROM CONTRASTIVE LEARNING TO PRODUCE HIGHER COSINE SIMILARITY SCORES, ENSURING BETTER RESUME-TO-JOB DESCRIPTION MATCHING.
5. FURTHER FINE-TUNING WITH LLM2VEC OPENS DOORS TO HYBRID MODELS (E.G., BM-25 + DPR) FOR EVEN HIGHER RETRIEVAL PRECISION.

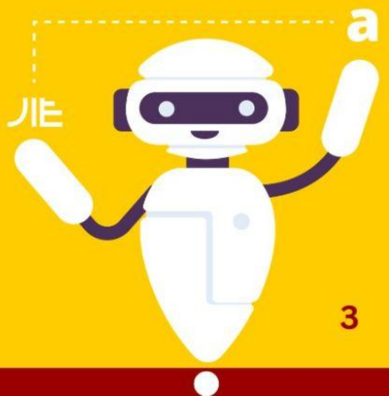


METRIC	BM-25	DPR-BASE	DPR-FT
ACCURACY	63.0%	48.6%	63.0%
MAP	0.9368	0.7101	0.8549
MMAP	0.8052	0.5459	0.7327

Table 1: Results on Test Data of 100 Job Descriptions

```
from transformers import (
    DPRContextEncoder,
    DPRQuestionEncoder,
    DPRContextEncoderTokenizer,
    DPRQuestionEncoderTokenizer,
)
```


```
[
    {
        "label": "Security_Analyst",
        "pos": [ ···
        ],
        "neg": [ ···
        ],
        "description": "The System Support Analyst II position requires a detail-oriented professional with 17+ years of experience supporting enterprise user communities ranging from 1000 to 15,000. The ideal candidate will have extensive knowledge of Microsoft OS Windows 7,8,10 Pro/Enterprise, XP Professional, Windows Server 2003/2008/2012, and Microsoft Office 2003-2016/365 software suites. Responsibilities include troubleshooting hardware and software issues, installing and configuring Windows based applications and peripherals, and creating/applying images on windows machines. Additionally, the candidate will be expected to utilize remote tools such as NetOp, DameWare, and SCCM to connect and troubleshoot software applications. Active Directory Users and Computers experience is required, as well as strong customer service skills.\n\nThe successful candidate will have experience managing telephone and ticket queues, providing first and second-level support, configuring and troubleshooting workstations and network issues, and administering Office 365. Desirable skills include familiarity with VOIP applications, self-starter mentality, and a positive attitude. This position offers opportunities for professional growth and advancement within the organization."
    },
```

THANK YOU!