NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models

Chankyu Lee*, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, Wei Ping*
contact: {chankyul, wping}@nvidia.com

Overview

Decoder-only LLM based embedding model to outperform existing encoder-based models in general-purpose text embedding tasks such as retrieval, clustering, classification.

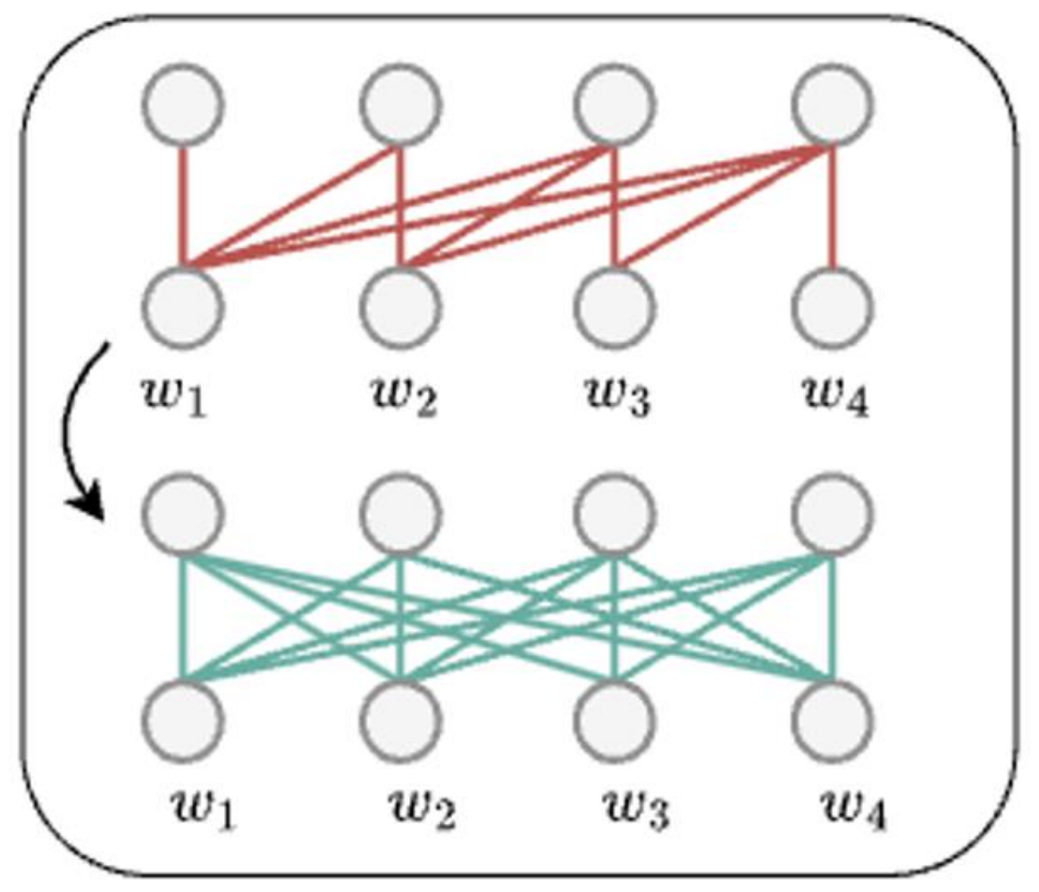
(Content-1) Bi-directional mask
(Content-2) Trainable latent attention layer
(Content-3) Two-stage instruction contrastive training
(Content-4) Curation of training data

Content-1

- **Bi-directional Mask**

- Modifying casual attention to bi-directional mask

Enabling Bidirectional Attention



Content-2

- **Two–Stage training**

- Retrieval task presents greater difficulty compared to the other tasks, so our training strategy focuses on fine-tuning the model for retrieval initially.

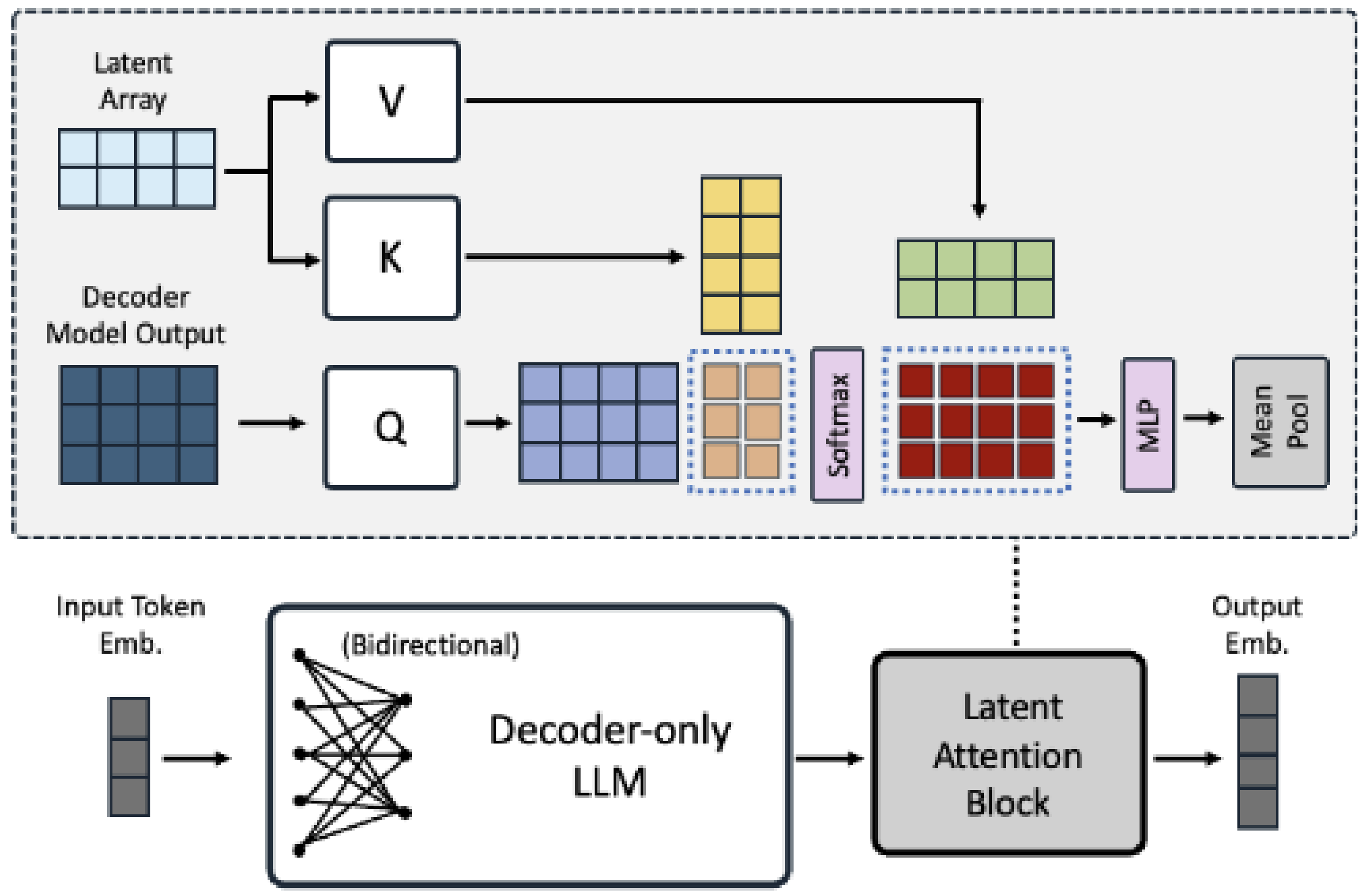
  - **First stage**: conducts contrastive training with instructions on retrieval datasets, **utilizing in-batch negatives**
  - **Second stage**: perform contrastive instruction-tuning on a **combination of retrieval and non-retrieval datasets** without in-batch negatives.

Embedding Task	Retrieval	Rerank	Cluster.	PairClass.	Class.	STS	Summ.	Avg.
Single Stage (Inbatch Enabled)	61.25	60.64	57.67	87.82	86.6	83.7	30.75	70.83
Single Stage (Inbatch Disabled)	61.37	60.81	58.31	88.3	90.2	84.5	30.96	71.94
Two Stage Training	**62.65**	60.65	58.46	88.67	90.37	84.31	30.70	**72.31**
Reversed Two Stage	**61.91**	60.98	58.22	88.59	90.26	83.07	31.28	**71.85**

Content-3

- **Trainable Latent attention layer**

- Mitigate the information dilution caused by averaging the output embeddings.



Content-4

- **Curation of training data**
  - Multi-class Classification and Clustering Labels: Example-based labels outperforms label-based
  - Adding the positive-aware hard-negative mining technique
  - Synthetic dataset generations

Embedding Task	Retrieval	Rerank	Cluster.	PairClass.	Class.	STS	Summ.	Avg.
[S0] Without HN, Without AD, Without SD	59.22	59.85	57.95	85.79	90.71	81.98	29.87	70.73
[S1] With HN, Without AD, Without SD	61.52	59.80	58.01	88.56	90.31	84.26	30.36	71.83
[S2] With HN, With AD, Without SD	62.28	60.45	58.16	88.38	90.34	84.11	29.95	72.07
[S3] **With HN, With AD, With SD**	62.65	60.65	58.46	88.67	90.37	84.31	30.70	**72.31**

Links

- **QR code**
  - Paper
  - Model

- **More in the paper**
  - LoRA finetuning
  - Compression study (pruning, distillation and quantization)

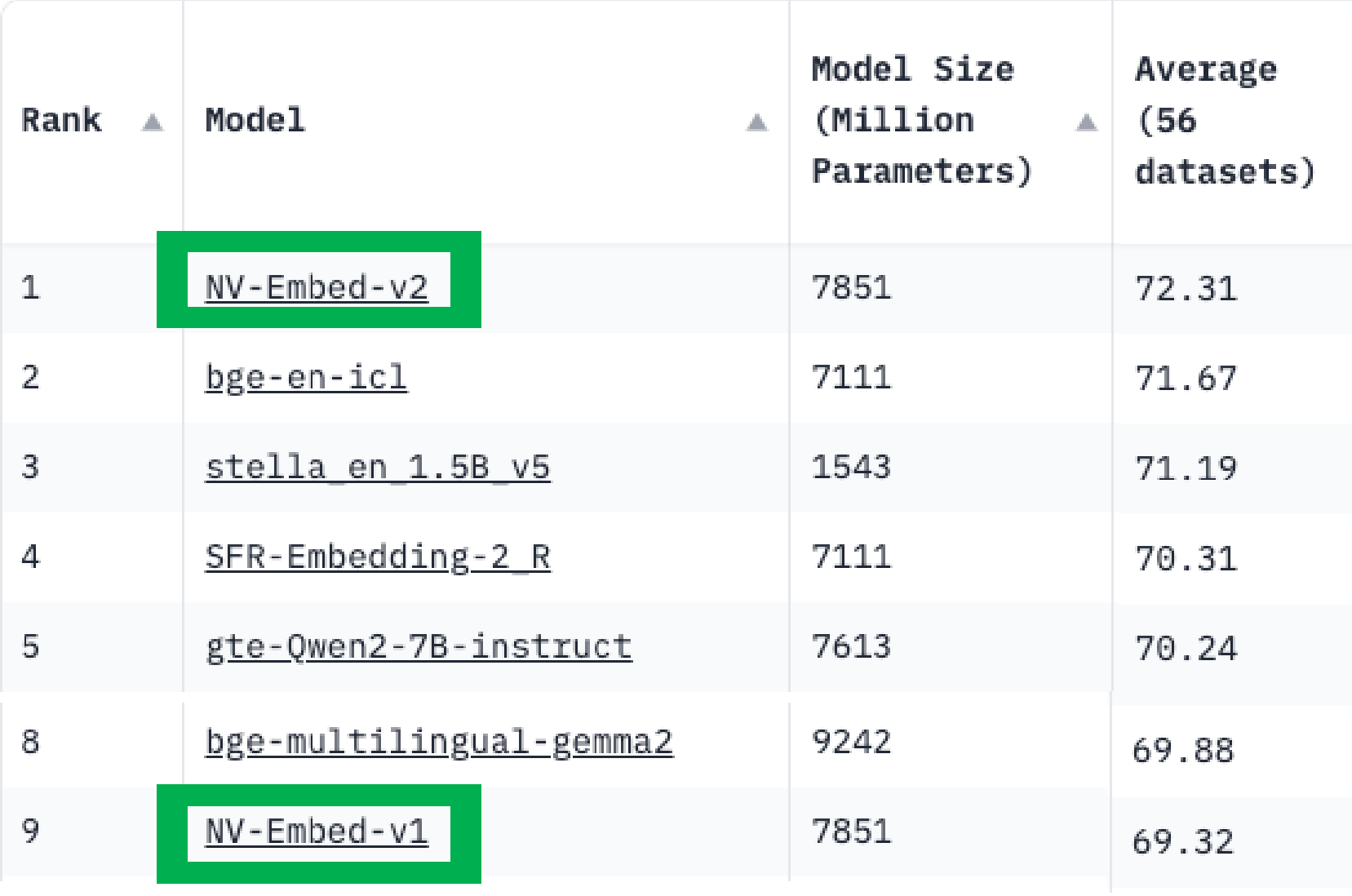
Table- MTEB benchmark

Pool Type	EOS		Mean		Latent-attention		Self-attention	
Mask Type	bidirect	causal	bidirect	causal	bidirect	causal	bidirect	causal
Retrieval (15)	62.13	60.30	61.81	61.01	**62.65**	61.15	61.17	60.53
Rerank (4)	60.02	59.13	60.65	59.10	60.65	59.36	60.67	59.67
Clustering (11)	58.24	57.11	57.44	57.34	**58.46**	57.80	58.24	57.11
PairClass. (3)	87.69	85.05	87.35	87.35	88.67	87.22	87.69	85.05
Classification (12)	90.10	90.01	89.49	89.85	**90.37**	90.49	90.10	90.01
STS (10)	82.27	81.65	84.35	84.35	84.31	84.13	84.22	83.81
Summar. (1)	30.25	32.75	30.75	30.88	30.70	30.90	30.93	31.36
Average (56)	71.63	70.85	71.71	71.38	**72.31**	71.61	71.61	70.6

Result-MTEB

- Twice topping the MTEB benchmark, a competitive leaderboard with 56 retrieval and embedding tasks

  - **NV-Embed-v1** : For the first time, NVIDIA ranked No. 1 MTEB benchmark from May 24-June16 (about 24 days)
  - **NV-Embed-v2** : Reclaimed No. 1 on Aug 30, 2024



Rank	Model	Model Size (Million Parameters)	Average (56 datasets)
1	NV-Embed-v2	7851	72.31
2	bge-en-icl	7111	71.67
3	stella_en_1.5B_v5	1543	71.19
4	SFR-Embedding-2_R	7111	70.31
5	gte-Qwen2-7B-instruct	7613	70.24
8	bge-multilingual-gemma2	9242	69.88
9	NV-Embed-v1	7851	69.32

Result-AIR Bench

- Newly released information retrieval benchmark, helping us to understand the generalization capability
- Majority of different domain samples do not appear in MTEB benchmarks

- **AIR-Benchmark 24.04 scores (QA and Long-Doc).**

  - **Table. QA (nDCG@10):** NV-Embed-v2 achieves the second highest scores in QA section.

Domain	Arxiv (4)	Book (2)	Healthcare (5)	Law (4)	Avg. (15)
NV-Embed-v2	79.27	77.46	73.01	71.18	**74.78**
Bge-en-icl (zero-shot)	78.30	78.21	73.65	67.09	73.75
NV-Embed-v1	77.65	75.49	72.38	69.55	**73.45**
Bge-multilingual-gemma2	71.77	76.46	73.96	70.86	72.88
Linq-Embed-Mistral	75.46	73.81	71.58	68.58	72.11
Stella-1.5B-v5	73.17	74.38	70.02	69.32	71.25
SFR-Embedding-Mistral	72.79	72.41	67.94	64.83	69.0
Text-embed-3-large (OpenAI)	74.53	73.16	65.83	64.47	68.77
E5-mistral-7b-instruct	72.14	72.44	68.44	62.92	68.49
SFR-Embedding-2R	70.51	70.22	67.60	62.82	67.45

  - **Table. Long-document (Recall@10):** NV-Embed-v2 attained the highest scores of 74.78 on the Long-Doc section, surpassing the Bge-en-icl model that requires overheads adding in-context examples to query during training

Domain	Wiki	Web	News	Healthcare	Law	Finance	Arxiv	Msmarco	Avg (8)
Bge-en-icl (zero-shot)	64.61	54.40	55.11	57.25	25.10	54.81	48.46	63.71	52.93
NV-Embed-v2	65.19	52.58	53.13	59.56	25.00	53.04	48.94	60.8	**52.28**
SFR-Embedding-Mistral	63.46	51.27	52.21	58.76	23.27	56.94	47.75	58.99	51.58
Stella-1.5B-v5	61.99	50.88	53.87	58.81	23.22	57.26	44.81	61.38	51.53
Gte-Qwen2-7B-instruct	63.46	51.20	54.07	54.20	22.31	58.20	40.27	58.39	50.26
NV-Embed-v1	62.84	50.42	51.46	58.53	20.65	49.89	46.10	60.27	**50.02**
Linq-Embed-Mistral	61.04	48.41	49.44	60.18	20.34	50.04	47.56	60.50	49.69
SFR-Embedding-2R	63.72	48.77	51.14	55.86	20.98	54.78	42.84	57.66	49.47
E5-mistral-7b-instruct	61.67	44.41	48.18	56.32	19.32	54.79	44.78	59.03	48.56