

embedding模型下载

1. 下载方法

embedding的下载方法和大语言模型下载方法和途径是一样的

- ModelScope
- huggingface

具体的使用详见【[大语言模型是如何下载.md](#)】

- https://git.imoooc.com/coding-920/RAG_full_stack_course_notebooks/src/master/install/大语言模型如何下载.md

可以在ModelScope和huggingface中进行模型搜索

2. 课程中使用embedding模型下载路径

- gte-large-zh
 - <https://huggingface.co/thenlper/gte-large-zh>
 - <https://hf-mirror.com/thenlper/gte-large-zh>
 - <https://www.modelscope.cn/models/AI-ModelScope/gte-large-zh>

```
huggingface-cli download --resume-download thenlper/gte-large-zh --local-dir gte-large-zh
```

- bge-m3
 - <https://huggingface.co/BAAI/bge-m3>
 - <https://hf-mirror.com/BAAI/bge-m3>
 - <https://www.modelscope.cn/models/BAAI/bge-m3>

```
huggingface-cli download --resume-download BAAI/bge-m3 --local-dir bge-m3
```

```
git clone https://www.modelscope.cn/BAAI/bge-m3.git
```

3. MTEB 评测榜单地址

- https://huggingface.co/spaces/mteb/leaderboard_legacy

最新地址:

- <https://huggingface.co/spaces/mteb/leaderboard>

We only display models that have been run on all tasks in the benchmark

Summary Performance per task Task information															
Rank (Overall)	Model	Zero-shot	Number of Parameters	Embedding Dimensions	Max Tokens	Mean (Task)	Mean (TaskType)	Bitext Mining	Classification	Clustering	Instruction Retrieval	Multilabel Classification	Pair Classification	Re	Re
1	gemin-embedding-exp-02-07	100%	Unknown	3072	8192	68.32	59.64	79.28	71.02	54.99	5.58	29.16	83.83	65.58	
2	Llm-Embed-Mistral	99%	7B	4096	32768	61.47	64.21	70.34	62.24	51.27	6.04	24.77	80.43	64.37	
3	gpt-0002-v7b-instruct	NA	7B	3584	32768	62.01	66.00	72.92	61.05	53.30	4.94	25.48	85.33	65.50	
4	multilingual-e5-large-instruct	100%	820M	3824	8192	63.16	63.17	68.13	61.46	51.54	-0.18	22.41	85.08	67.61	
5	SFR-Embedder-Mistral	96%	7B	4096	32768	60.93	64.08	70.88	60.82	52.57	6.26	24.55	80.29	64.10	
6	Grillm-7B	100%	7B	6944	6944	60.43	63.03	70.83	63.13	58.91	3.58	22.77	79.46	63.78	
7	text-multilingual-embedder-002	99%	Unknown	768	2048	62.13	64.32	70.73	64.04	48.47	4.68	22.88	81.14	61.22	
8	Grillm-8x7B	99%	57B	4096	4096	60.58	63.39	68.17	63.05	58.88	2.44	24.43	79.73	62.61	
9	e5-mistral-7b-instruct	100%	7B	6944	32768	60.28	63.18	70.88	60.31	51.39	-0.82	22.28	81.12	63.82	

10	cohere-embed-multilingual-v3.0	 NA	Unknown	512	Unknown	61.18	53.31	76.58	62.95	47.61	-1.89	22.74	79.88	64.87
11	gte-Qwen2-1.5B-instruct	 NA	18	8960	32768	56.47	52.75	62.51	58.32	52.59	6.74	26.62	81.58	62.58
4														
Download Table														

下一节：实战：embedding模型加载和使用对比(22:25)

下一节