

Benchmarking embedding models

The background is a dark navy blue space-themed illustration. It features several stylized galaxies: a large, swirling, light blue and white galaxy in the upper left; a smaller, dark blue and white spiral galaxy in the lower right; and a small, dark blue and white spiral galaxy in the lower left. Scattered throughout the background are numerous small, white, four-pointed star-like shapes. There are also a few larger, multi-colored star-like shapes in shades of yellow, orange, and blue.

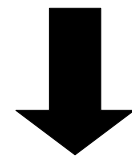
Introduction

The background is a dark navy blue space filled with various celestial elements. In the top left, a large, swirling, light blue nebula or galaxy arm curves across the frame. Scattered throughout are numerous small, white, four-pointed star icons. In the top right, a small cluster of blue and white stars is visible. In the bottom left, there is a group of larger, multi-colored star icons in shades of yellow, blue, and red. Two small, dark brown spiral galaxies are also present, one in the upper right and one in the lower left. The overall composition is a stylized representation of outer space.

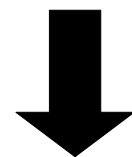
Benchmarking embedding models

- Embedding models **convert text** (or other modalities) to a **dense vector**.
- With embedding models, you can **build RAG and recommendation systems**.
- But **how do you choose an embedding model?**

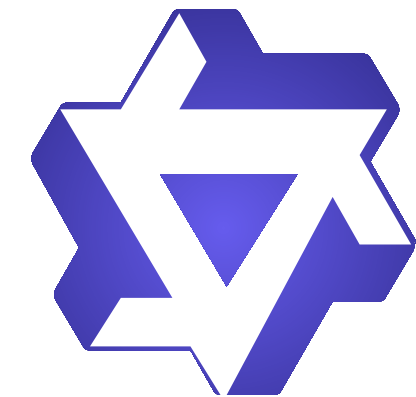
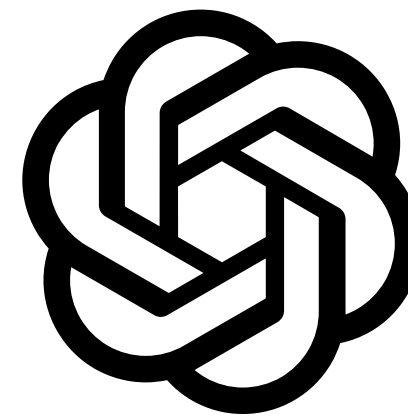
Hi Cassiopeia! Have you
talked with Andromeda?



**Embedding
model**



[0.69, 0.42, ...]



Benchmarking embedding models

- How do you choose an embedding model?
- You look at **benchmarks!**
- **MTEB** ranks models based on their performance on **different benchmarks**.

Rank (Bor...	Model	Zero-shot	Memory U...	Number of P...	Embedding D...	Max Tokens	Mean (T...	Mean (TaskT...	Bitext ...	Classification	Clustering
1	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82	54.59
2	Qwen3-Embedding-8B	99%	28866	7B	4096	32768	70.58	61.69	80.89	74.00	57.65
3	Qwen3-Embedding-4B	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33	57.15
4	Qwen3-Embedding-0.6B	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83	52.33
5	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55	52.77
6	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24	50.60
7	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.22	55.08	80.13	64.94	50.75
8	embeddinggemma-300m	99%	578	307M	768	2048	61.15	54.31	64.40	60.90	51.17
9	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.90	53.92	70.00	60.02	51.84
10	GritLM-7B	99%	13813	7B	4096	32768	60.92	53.74	70.53	61.83	49.75
11	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73	64.64	47.84

The MTEB leaderboard on 20/09/2025.

Benchmarking embedding models

- Should we **trust** these public benchmarks?
- You can also create **your own benchmark** with your private data.
- The dataset must be **clean, diverse, and in your language** (or multilingual).

Rank (Bor...	Model	Zero-shot	Memory U...	Number of P...	Embedding D...	Max Tokens	Mean (T...	Mean (TaskT...	Bitext ...	Classification	Clustering
1	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82	54.59
2	Qwen3-Embedding-8B	99%	28866	7B	4096	32768	70.58	61.69	80.89	74.00	57.65
3	Qwen3-Embedding-4B	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33	57.15
4	Qwen3-Embedding-0.6B	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83	52.33
5	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55	52.77
6	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24	50.60
7	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.22	55.08	80.13	64.94	50.75
8	embeddinggemma-300m	99%	578	307M	768	2048	61.15	54.31	64.40	60.90	51.17
9	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.90	53.92	70.00	60.02	51.84
10	GritLM-7B	99%	13813	7B	4096	32768	60.92	53.74	70.53	61.83	49.75
11	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73	64.64	47.84

The MTEB leaderboard on 20/09/2025.

Benchmarking embedding models

In this course, you will:

- Create a **golden dataset**.
- Run **open source** and **proprietary models**.
- Compute **metrics** to grade the models.
- Perform **statistical tests** to prove if a model is better than another.
- **Automate** some steps in the pipeline.
- **Generate tables** to compare the models.

#	Model	mrr	recall@1	recall@5	ndcg@5
a	OpenAI 3-Small	0.780	0.683	0.901	0.804
b	OpenAI 3-Large	0.778	0.681	0.901	0.802
c	Google Gemini-001	0.810 ab	0.725 b	0.927 ab	0.834 ab

Comparing Gemini embedding to OpenAI's embedding models.

