New and improved embedding model

We are excited to announce a new embedding model which is significantly more capable, cost effective, and simpler to use.

Read documentation↗

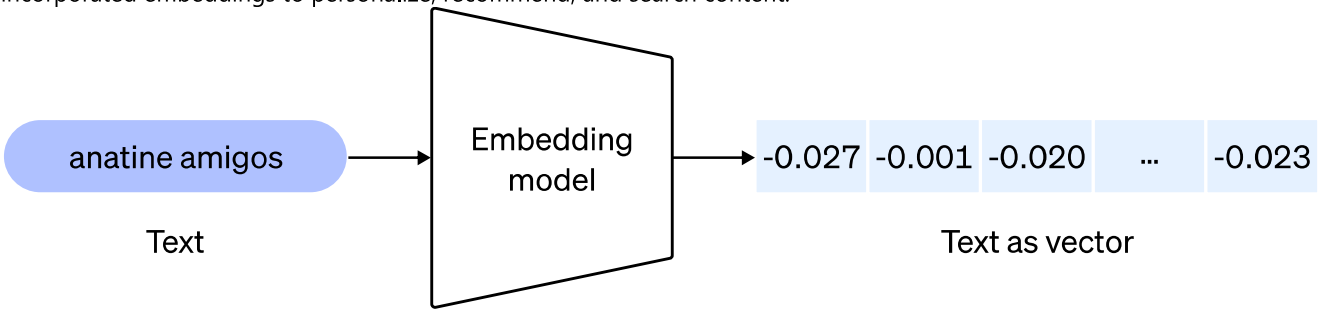
Illustration: Ruby Chen

December 15, 2022**Authors** Product, Announcements

Ryan Greene↓
Ted Sanders↓
Lilian Weng↓
Arvind Neelakantan↓

The new model, `text-embedding-ada-002`, replaces five separate models for text search, text similarity, and code search, and outperforms our previous most capable model, Davinci, at most tasks, while being priced 99.8% lower.

Embeddings are numerical representations of concepts converted to number sequences, which make it easy for computers to understand the relationships between those concepts. Since the initial launch of the OpenAI /embeddings endpoint, many applications have incorporated embeddings to personalize, recommend, and search content.



You can query the /embeddings endpoint for the new model with two lines of code using our OpenAI Python Library, just like you could with previous models:

```
import openai
response = openai.Embedding.create(
  input="porcine pals say",
  model="text-embedding-ada-002"
)
```

Print response→

Model improvements

Stronger performance. `text-embedding-ada-002` outperforms all the old embedding models on text search, code search, and sentence similarity tasks and gets comparable performance on text classification. For each task category, we evaluate the models on the datasets used in old embeddings.

Text search Code search Sentence similarity Text classification

Model	Performance
text-embedding-ada-002	53.3
text-search-davinci-*-001	52.8
text-search-curie-*-001	50.9
text-search-babbage-*-001	50.4
text-search-ada-*-001	49.0

Dataset: BEIR (ArguAna, ClimateFEVER, DBPedia, FEVER, FiQA2018, HotpotQA, NFCorpus, QuoraRetrieval, SciFact, TRECCOVID, Touche2020)

Unification of capabilities. We have significantly simplified the interface of the /embeddings endpoint by merging the five separate models shown above (text-similarity , text-search-query , text-search-doc , code-search-text and code-search-code) into a single new model. This single representation performs better than our previous embedding models across a diverse set of text search, sentence similarity, and code search benchmarks.

Longer context. The context length of the new model is increased by a factor of four, from 2048 to 8192, making it more convenient to work with long documents.

Smaller embedding size. The new embeddings have only 1536 dimensions, one-eighth the size of davinci-001 embeddings, making the new embeddings more cost effective in working with vector databases.

Reduced price. We have reduced the price of new embedding models by 90% compared to old models of the same size. The new model achieves better or similar performance as the old Davinci models at a 99.8% lower price.

Overall, the new embedding model is a much more powerful tool for natural language processing and code tasks. We are excited to see how our customers will use it to create even more capable applications in their respective fields.

Limitations

The new text-embedding-ada-002 model is not outperforming text-similarity-davinci-001 on the SentEval linear probing classification benchmark. For tasks that require training a light-weighted linear layer on top of embedding vectors for classification prediction, we suggest comparing the new model to text-similarity-davinci-001 and choosing whichever model gives optimal performance.

Check the Limitations & Risks section in the embeddings documentation for general limitations of our embedding models.

Examples of the embeddings API in action

Kalendar AI is a sales outreach product that uses embeddings to match the right sales pitch to the right customers out of a dataset containing 340M profiles. This automation relies on similarity between embeddings of customer profiles and sale pitches to rank up most suitable matches, eliminating 40–56% of unwanted targeting compared to their old approach.

Notion, the online workspace company, will use OpenAI's new embeddings to improve Notion search beyond today's keyword matching systems.