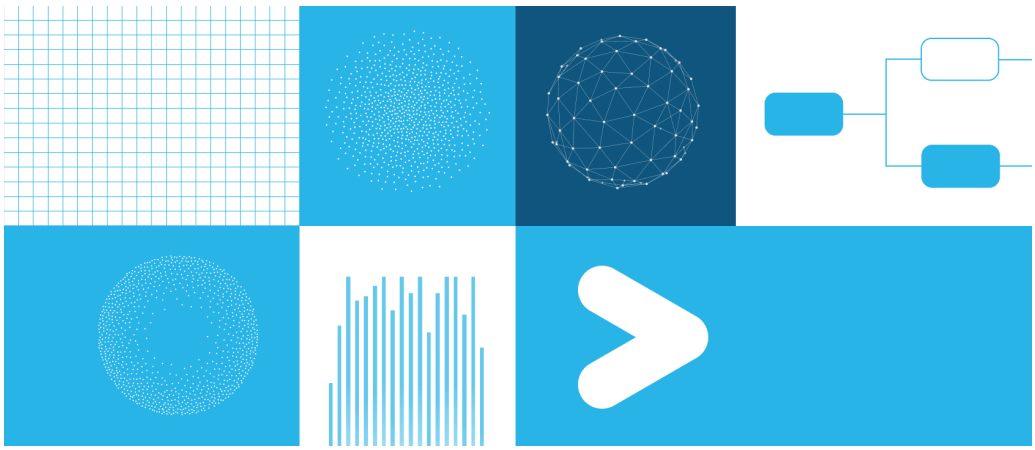
DEC 04, 2024

Snowflake's Arctic Embed 2.0 Goes Multilingual: Empowering Global-Scale Retrieval with Inference Efficiency and High-Quality Retrieval

Gen AI



AUTHOR

Puxuan Yu

Luke Merrick

Gaurav Nuti

Daniel Campos

Snowflake is excited to announce the release of *Arctic Embed L 2.0* and *Arctic Embed M 2.0*, the next iteration of our frontier embedding models, which now empower multilingual search. While our previous releases have been well received by our customers, partners and the open source community, leading to millions of downloads, we have consistently received one request: Can you make this model multilingual? Arctic Embed 2.0 builds on the robust foundation of our previous releases, add[illegible] multilingual support without sacrificing English performance or scalability, to address the needs of an even broader user base that spans a wide range of languages and applications.

AUTHOR

Puxuan Yu

Luke Merrick

Gaurav Nuti

Daniel Campos

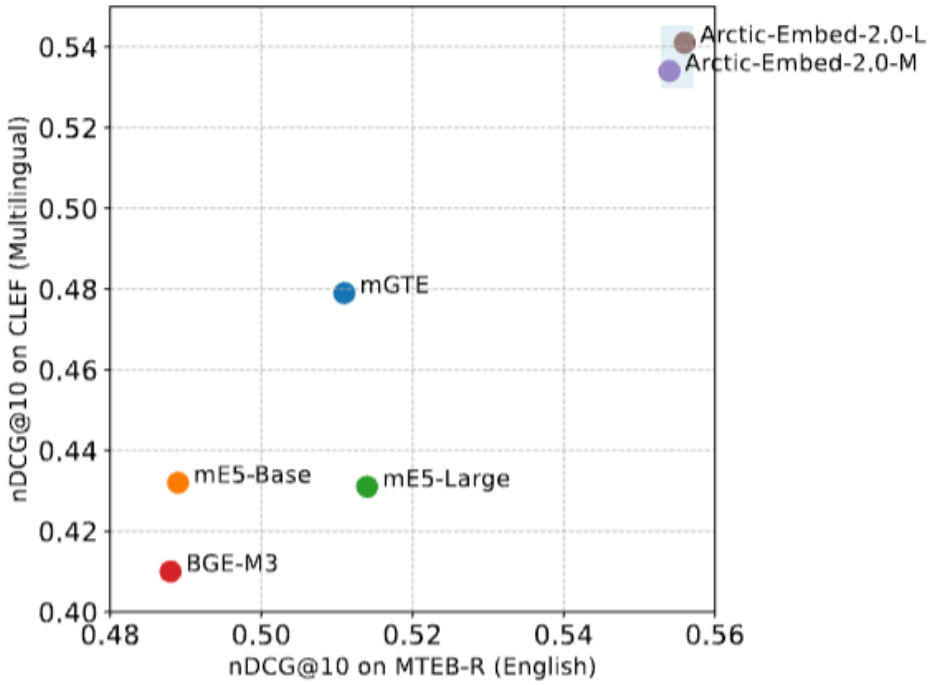


Figure 1. Single-vector dense retrieval performance of open source multilingual embedding models with fewer than 1B parameters. Scores are average nDCG@10 on MTEB Retrieval and the subset of CLEF (ELRA, 2006) covering English, French, Spanish, Italian and German.

Multilingual without compromise

In this Arctic Embed 2.0 release, we make two variants available for publish usage, a medium variant focused on inference efficiency built on top of Alibaba's GTE-multilingual with 305 million parameters (of which 113 million are non-embedding parameters), and a large variant focused on retrieval quality built on top of a long-context variation of Facebook's XMLR-Large, which has 568 million parameters (of which 303 million are non-embedding parameters). Both sizes support a context length of up to 8,192 tokens. In building Arctic Embed 2.0, we recognized a challenge that many existing multilingual models have faced: Optimizing for multiple languages often ends up sacrificing English retrieval quality. This has led many models in the space to have two variants for each release: English and multilingual. The Arctic Embed 2.0 models are different. They deliver top-tier performance in non-English languages, such as German, Spanish, French, while also outscoring their English-only predecessor Arctic Embed M 1.5 at English-language retrieval.

By carefully balancing multilingual needs with Snowflake's commitment to excellence in English retrieval, we built Arctic Embed 2.0 to be a universal workhorse, useful for a wide range of global use cases. Throughout this article, all qualitative evaluations refer to the average NDCG@10 score across tasks, unless otherwise indicated.

AUTHOR

Puxuan Yu

Luke Merrick

Gaurav Nuti

Daniel Campos

SHARE

Table 1. Our Arctic Embed L v2.0 model achieves high scores on the popular English-language MTEB Retrieval benchmark, while simultaneously achieving high retrieval quality across multiple multilingual benchmarks. Prior iterations of Arctic Embed perform well for English but suffer multilingually, while popular open source multilingual models see degradation in English performance. Scores across all models and data sets reflect an average NDCG@10. CLEF and MIRACL scores reflect average across German (DE), English (EN), Spanish (ES), French (FR) and Italian (IT).

The diverse and powerful feature set of Arctic Embed 2.0

1. **Enterprise-ready throughput and efficiency:** The Arctic Embed 2.0 models are built for large-scale enterprise demands. Even our "large" model weighs in well under 1B parameters and delivers fast, high-throughput embedding capabilities. Based on internal testing, it easily handles more than 100 documents per second (on average) on NVIDIA A10 GPUs and achieves sub-10ms query embedding latency, enabling practical deployment on budget-friendly hardware.

2. **Uncompromising quality for English and non-English retrieval:** Despite their compact sizes, both Arctic Embed 2.0 models achieve impressive NDCG@10 scores across a variety of English and non-English benchmark data sets, demonstrating a capability to generalize well even to languages not included in the training recipe. These impressive benchmark scores position Arctic Embed L 2.0 as a leader among frontier retrieval models.

3. **Enabling scalable retrieval through Matryoshka Representation Learning (MRL):** The Arctic Embed 2.0 release includes the same quantization-friendly MRL functionality introduced in Arctic Embed 1.5, allowing users to reduce cost and optimize scale when performing searches over large data sets. With both model sizes, users can achieve high-quality retrieval with as few as 128 bytes per vector (96x smaller than uncompressed embeddings from OpenAI's

popular text-embedding-3-large model[1]). Just like Arctic Embed 1.5, the Arctic Embed 2.0 models also outshine several MRL-supporting peers with substantially lower quality degradation and higher benchmark scores in the compressed regime.

4. **Truly open source:** The Arctic Embed 2.0 models are released under the permissive Apache 2.0 license.

AUTHOR

Puxuan Yu

Luke Merrick

Gaurav Nuti

Daniel Campos

The flexibility of open source meets enterprise-grade reliability

As with their predecessors, the Arctic Embed 2.0 models are released under the permissive Apache 2.0 license, empowering organizations to modify, deploy and scale under a familiar license. Out of the box, these models support applications across verticals with reliable, multilingual embeddings that generalize well.

"Multilingual embedding models are crucial for enabling people worldwide — not just English speakers — to become AI builders," Hugging Face CEO Clément Delangue said. "By releasing these state-of-the-art models as open source on Hugging Face, Snowflake is making a tremendous contribution to AI and the world."

Indeed, especially among open source options, the Arctic Embed 2.0 family deserves special attention due to its observed generalization across multilingual retrieval benchmarks. By licensing the 2000-2003 test suite for the Cross-Lingual Evaluation Forum (CLEF), our team was able to measure out-of-domain retrieval quality for a variety of open source models and discover an unfortunate trend of underperformance, as compared to scores achieved on the in-domain MIRACL evaluation set. We hypothesize that some earlier open source model developers may have inadvertently tuned their training recipes too aggressively toward improved MIRACL performance at the cost of generality, possibly by overfitting the MIRACL training data. For more details about how the Arctic Embed 2.0 models were trained and what we learned in the process, look out for the forthcoming technical report.

AUTHOR

Puxuan Yu

Luke Merrick

Gaurav Nuti

Daniel Campos

Table 2. A comparison of multilingual retrieval models on several data sets from the out-of-domain CLEF benchmark.

Table 3. A comparison of several open source multilingual retrieval models on the in-domain MIRACL benchmark.

As seen in Tables 2 and 3, several popular open source models score on par with **Arctic Embed L 2.0** on the in-domain MIRACL evaluation but fall short on the out-of-domain CLEF evaluation. We also benchmarked popular closed-source models like OpenAI's text-embedding-3-large model and found that Arctic L 2.0's performance is in line with leading proprietary models.

As seen in Table 4, existing open source multilingual models also score worse than **Arctic Embed L 2.0** on the popular English-language MTEB Retrieval benchmark, forcing users who seek to support multiple languages to choose between lower English-retrieval quality or more operational complexity from using a second model just for English retrieval. With the Arctic Embed 2.0 release, practitioners are now able to search single open source model without sacrificing English-language retrieval quality.

Table 4. A comparison of several top open and closed source multilingual retrieval models on the in-domain MTEB Retrieval Benchmark.

Compression and efficiency without trade-offs

Snowflake continues to prioritize efficiency and scale in its embedding model design. With **Arctic Embed L 2.0**, users can pack the quality characteristic of larger models into compact embeddings requiring as little as 128 bytes per vector for storage. This makes it possible to serve retrieval over millions of documents at a lower cost on low-end hardware. We achieve efficiency in embedding throughput as well by squeezing Arctic Embed 2.0's impressive retrieval quality into just over 100M and 300M non-embedding parameters in its two sizes (medium and large) respectively — just a slight increase from our earlier English-only versions.

Indeed, the scale-focused regime is where **Arctic Embed L 2.0** truly shines, achieving better quality under compression than other MRL-trained models, such as OpenAI's text-embedding-3-large.

AUTHOR

Puxuan Yu

Luke Merrick

Gaurav Nuti

Daniel Campos

Table 5. A comparison of OpenAI's text-embedding-3-large performance with truncated embeddings compared to Arctic Embed L 2.0 on English only (MTEB Retrieval) and multilingual (CLEF).

Conclusion: A new standard for multilingual, efficient retrieval

With Arctic Embed 2.0, Snowflake sets a new standard for multilingual, efficient embedding models. Additionally, we make the frontier of text-embedding quality not only efficient but permissively open sourced as well. Whether your goal is to expand reach to multilingual users, reduce storage costs or embed documents on accessible hardware, Arctic Embed 2.0 offers capabilities and flexibility to meet your needs.

Our soon-to-be-released technical report will dive deeper into the innovations behind Arctic Embed 2.0. In the meantime, we invite you to start embedding with Snowflake today.

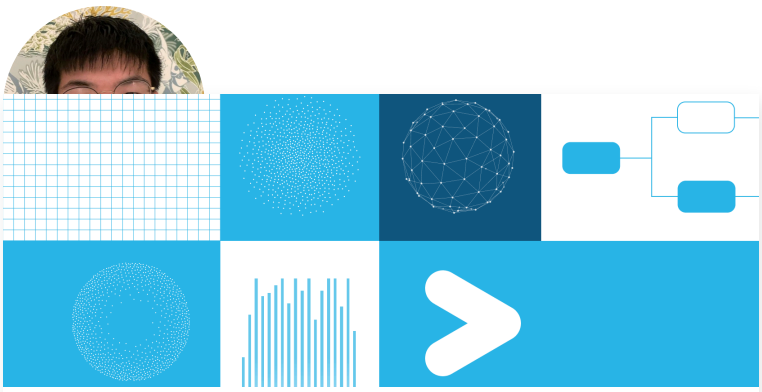
---

[1] This calculation uses float32 format for the uncompressed baseline, i.e. 3,072 numbers of 32 bits each for a total of 98,304 bits per vector, exactly 96x larger than the 1,024 bits per vector (equivalent to 128 bytes per vector) used when storing MRL-truncated 256-dimensional vectors from the Arctic Embed 2.0 models in int4 format.

SHARE

RELATED CONTENT

AUTHOR



NOV 19, 2024

Benchmarking LLMs on Writing Feature Engineering Code

Today, the limitations of LLMs are predominantly assessed using benchmarks focused on language understanding, world knowledge, code generation or mathematical reasoning in separation. This approach, however, overlooks some critical capabilities…

Delve into the details

SEP 12, 2024

LLM Interactive Workloads: Optimizing GPU Capacity for Interactive and Batch Workloads

At Snowflake, we offer a wide variety of LLM-powered features in Cortex AI, including Cortex…

Learn More

JUL 18, 2024

Snowflake Arctic Embed M v1.5: Hitting the ROI Sweet Spot for Enterprise Retrieval

Today Snowflake released the world's most pragmatic text embedding model for English-language search: arctic-embed-m-v1.5. Our…

Expand your knowledge

AUTHOR

Inc.

Puxuan Yu

Luke Merrick

Gaurav Nuti

Daniel Campos

AUTHOR

Puxuan Yu

Luke Merrick

Gaurav Nuti

Daniel Campos

SHARE