snowflake ENGINEERING BLOG CATEGORIES

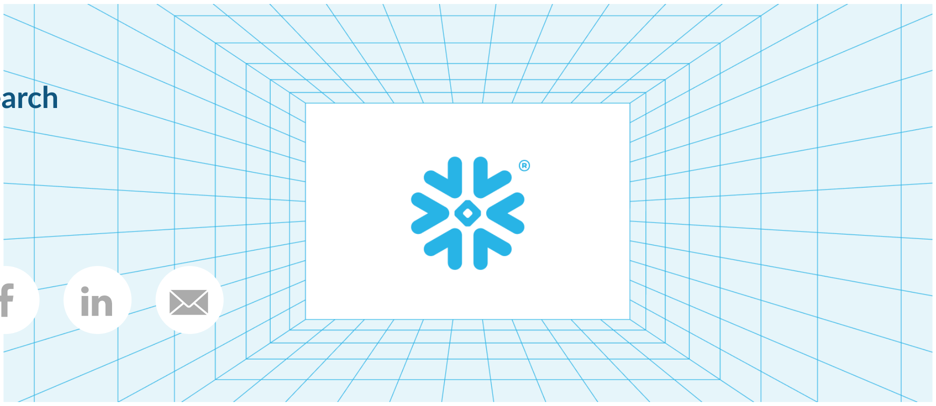
JUL 18, 2024

AUTHOR

Luke Merrick

Snowflake AI Research

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



SHARE

Today Snowflake released the world's most pragmatic text embedding model for English-language search: arctic-embed-m-v1.5. Our new model can deliver tiny embedding vectors — packing nearly 8 million per GB of storage — enabling up to a 24x improvement in retrieval system scalability while still delivering uncompromising retrieval quality.

Building on **our v1.0 open source release**, this model continues to run just as fast as its predecessor and is released with a permissive Apache 2.0 license.

Making Scale Affordable Without Compromise

With Arctic Embed M v1.5, we sought to augment the strong value proposition of v1.0's production-friendly model sizes with a lower total cost of ownership by compressing the embedding vectors, since we believe that it is the combination of efficient embedding model inference and efficient embedding storage that will enable pragmatic engineers to deliver high-value products at affordable technical and financial costs.

Same Great M Size

For many enterprises, being able to comfortably scale to embedding tens or hundreds of millions of documents, using widely available and budget-friendly hardware (e.g., NVIDIA A10 GPUs rather than H100s), is critically important. For this reason, our v1.5 model development focused on the M size of Arctic Embed, which sits in the sweet spot between size and quality.

AUTHOR

Luke Merrick
Snowflake AI Research

SHARE

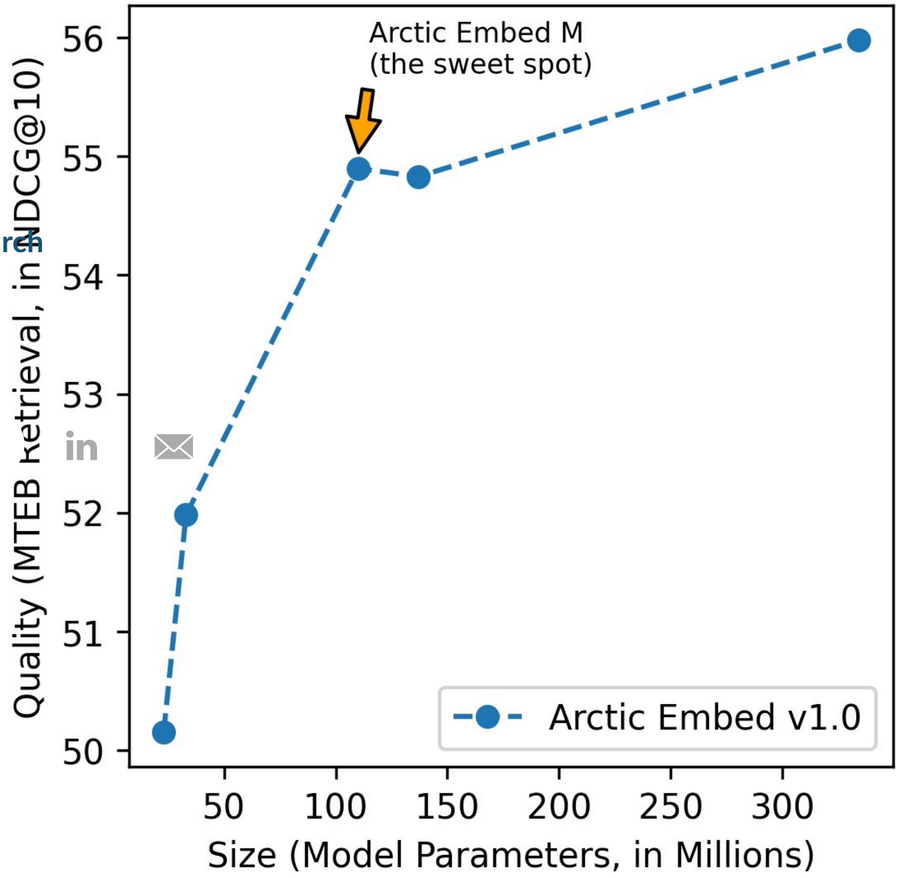


Figure 1. Arctic Embed M achieves both small model size and high retrieval quality, leading us to select this form-factor as the backbone for our compressed-embedding model Arctic Embed M v1.5.

Building on the enterprise workhorse that is the BERT-base-uncased model architecture, this 110m param v1.5 model packs quite a punch while remaining friendly to inference efficiency. Besides its continued accuracy and efficiency, we are also releasing this model with the same permissive Apache 2.0 license that enables organizations to confidently adopt Snowflake Arctic Embed in their production systems.

In fact, from the perspective of running the model, v1.5 is just a drop-in upgrade for v1.0. The differences (and by differences, we mean improvements), only show up during retrieval.

Scaling Retrieval

Practitioners know that scaling text embedding is only half the battle toward scalable retrieval systems — the systems that store and leverage text embedding vectors for search also need to contend with large-scale corpora. While the MTEB retrieval leaderboard abounds with models outputting vectors with 1,024 or even 4,096 dimensions, in practice such exorbitant embedding vector sizes can chew through the memory capacity of a retrieval system quite quickly.

AUTHOR

Luke Merrick
Snowflake AI Research

Consider for a moment a modest data set consisting of the 37 plays authored by William Shakespeare. These weigh in around 800,000 words and gzip down to just under 2 MB on disk, but if we split these plays into page-sized chunks of 400 words each and embed them into 4,096-dimensional vectors, we suddenly need to store more than *8 MB of vectors in RAM* to support retrieval over a corpus that weighs as little as *2 MB of documents on disk.* While this works out fine for many demos and weekend tinkering, scaling this setup from dozens of plays to millions of documents starts to become quite expensive, both in terms of performance and dollar cost of hardware in a production setting!

SHARE

Although Arctic Embed M v1.0 already uses a more modest 768-dimensional embedding space, we realized there was still room to improve the retrieval-time footprint of this model, and that's what we have done with Arctic Embed M v1.5. Our v1.5 model was trained with Matryoshka Representation Learning (MRL) during both pretraining and fine-tuning stages to deliver minimal degradation in performance after truncating down to just 256 dimensions. Additionally, we concurrently designed a global scalar quantization scheme alongside the model to enable even further vector compression while minimizing both quality degradation and operational complexity. Complementing the release of the v1.5 model's weights, we are also publishing the details of this model-centric scalar quantization, which allows practitioners to easily achieve a high degree of compression while maintaining good retrieval quality. As shown in the table below, this global int4 uniform scalar quantization scheme enables Arctic Embed M v1.5 to scale down to just 128 bytes per vector while achieving a full

98% of the quality score associated with uncompressed embedding vectors that weigh in a full 24x larger. Compared to other compression schemes, like product quantization, our model-centric compression approach delivers lower-quality degradation while being operationally simpler, enabling straightforward and fast integer-arithmetic-based similarity computations, and requiring no quantization parameters to be tuned to the embedded corpus.

Model Version	Dimensionality	Scalar Quantization	Bytes Per Vector (fraction of baseline)	MTEB Retrieval Score (fraction of baseline)
v1	768	None (float32)	3072 (100%)	54.9 (100%)
v1	768	int8	768 (25%)	54.9 (100%)
v1.5	768	int8	768 (25%)	55.1 (100%)
v1.5	256	int8	256 (8.3%)	54.2 (99%)
v1.5	256	int4	128 (4.2%)	53.7 (98%)

Table 1. Snowflake Arctic Embed M v1.5 delivers up to 24x smaller embeddings compared to the already-modest-sized embeddings of Arctic Embed M v1.0, with minimal degradation in retrieval quality.

Don’t Compromise On Quality

Although vector size can sometimes feel like an afterthought when scrolling through the MTEB leaderboard, a handful of both proprietary and open models also offer 256-dimensional truncated embeddings via MRL. Unfortunately, however, these models tend to suffer a sharp quality degradation after truncation — possibly due to incorporating MRL only during their fine-tuning step, rather than during both pretraining and fine-tuning (the creators of Google Gecko and Nomic Embed both mention using MRL specifically during fine-tuning).

AUTHOR

Luke Merrick

Snowflake AI Research

SHARE

As shown in the table below, Arctic Embed M v1.5 can deliver uncompromised retrieval accuracy while decreasing the total cost of ownership. When compared to comparable text embedding offerings, Arctic Embed M v1.5 delivers better retrieval with less quality degradation, all in a model which is more inference efficient.

AUTHOR

Luke Merrick
Snowflake AI Research

Model Name	Truncated Embedding Dimensionality	Model Parameters	MTEB Retrieval Score(fraction of Snowflake Arctic Embed M v1.5 score)	Fr M A D
Snowflake arctic-embed-m-v1.5	256	109M	54.2 (100%)	9
Googlegecko	256	1,200M	52.4 (97%)	94
OpenAI text-embedding-3-large	256	Not Published	51.7 (95%)	93
Nomicnomic-embed-text-v1.5	256	138M	50.8 (94%)	9

SHARE

Table 2. Compared to other embedding models, Snowflake Arctic Embed M v1.5 offers higher quality and lower degradation when targeting a reduced embedding dimensionality.

Making the Secret Sauce an Open Secret

Although this v1.5 is an incremental release that will not come with a complete technical report like v1.0, Snowflake's commitment to AI openness extends to this model release as well. Below we cover the relatively straightforward adjustments we used to deliver v1.5.

MRL Only to 256 + Int4 Quantization Offers Optimal Quality at the Same Level of Compression

After reading the MRL paper, it's tempting to follow the recipe exactly and include several nested dimensions of truncations in your model training. However, we found that truncation below 256 dimensions led to a sharp decline in hard-earned retrieval performance, making them of limited value. On the other hand, quantizing the scalars that make up each embedding vector offers a separate and often superior opportunity for compression with minimal quality degradation.

Combining these observations, we designed a holistic compression scheme consisting of both vector truncation and quantization. While MRL experiments in a vacuum suggest that including many nested MRL dimensions in training leads to little degradation in 256-dimension-truncated embedding performance, our combined MRL + quantization evaluations identified a major pitfall: MRL with many layers of nesting can substantially harm quality after scalar quantization. In the figure below we illustrate the crux of the issue by plotting the ranges of scalar values across the embedding vector after we have included lower-dimensionality truncations in MRL training. We see that MRL pushes the range of values wider and wider in the smallest subvectors, a trend which unfortunately increases quantization error and decreases retrieval quality after quantization.

AUTHOR

Luke Merrick
Snowflake AI Research

SHARE

AUTHOR

Luke Merrick

Snowflake AI Research

SHARE

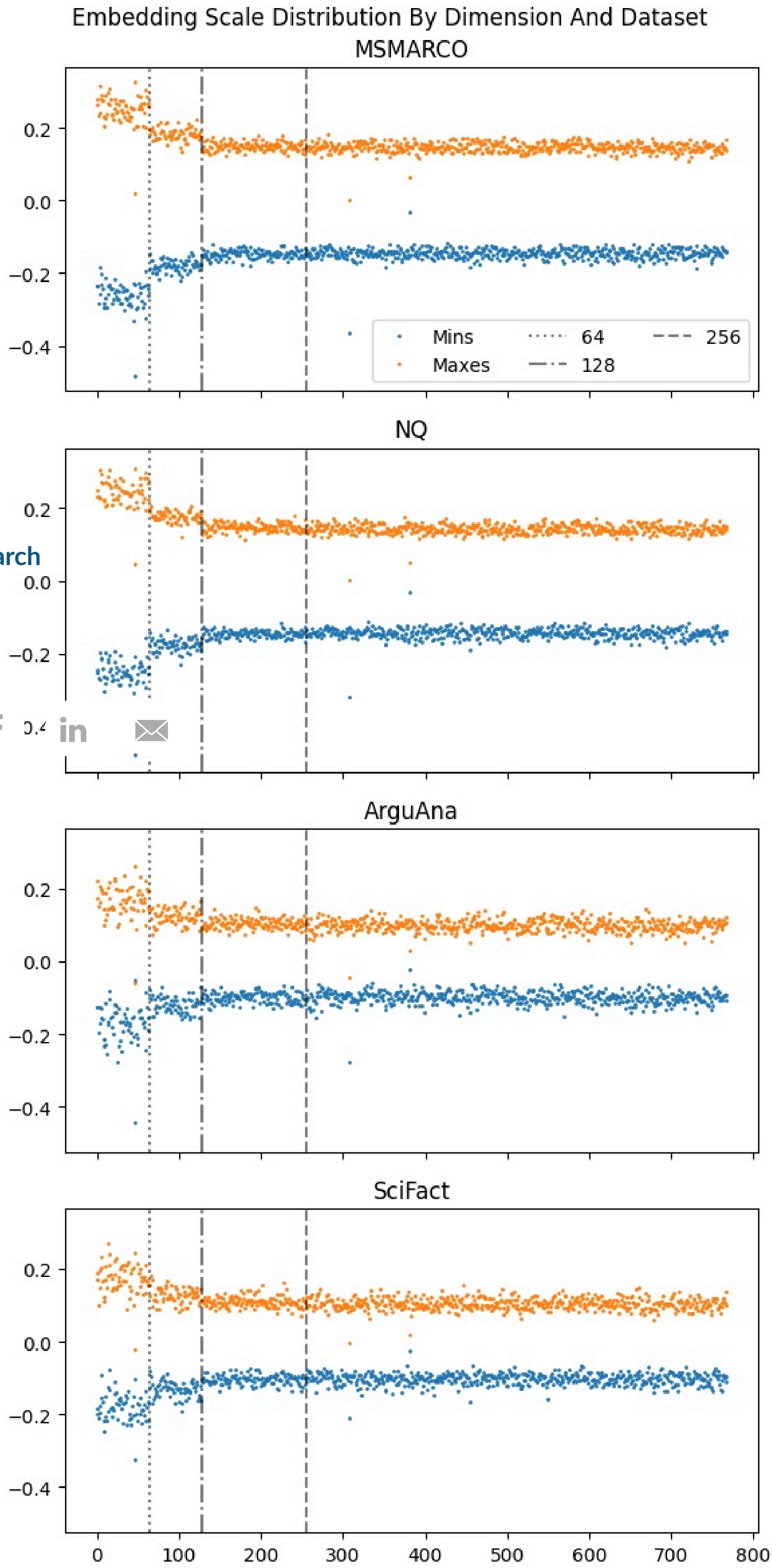


Figure 2. With true nested MRL, the scalar ranges of vectors become uneven and trickier to quantize well.

On the other hand, when we pare down the MRL to only one 256-dimensional subvector, we achieve a much more uniform set of ranges for the individual values within the truncated subvector, as shown in the figure below. These more consistent ranges make it much easier to implement an effective global int4 quantization scheme without sharply degrading retrieval quality.

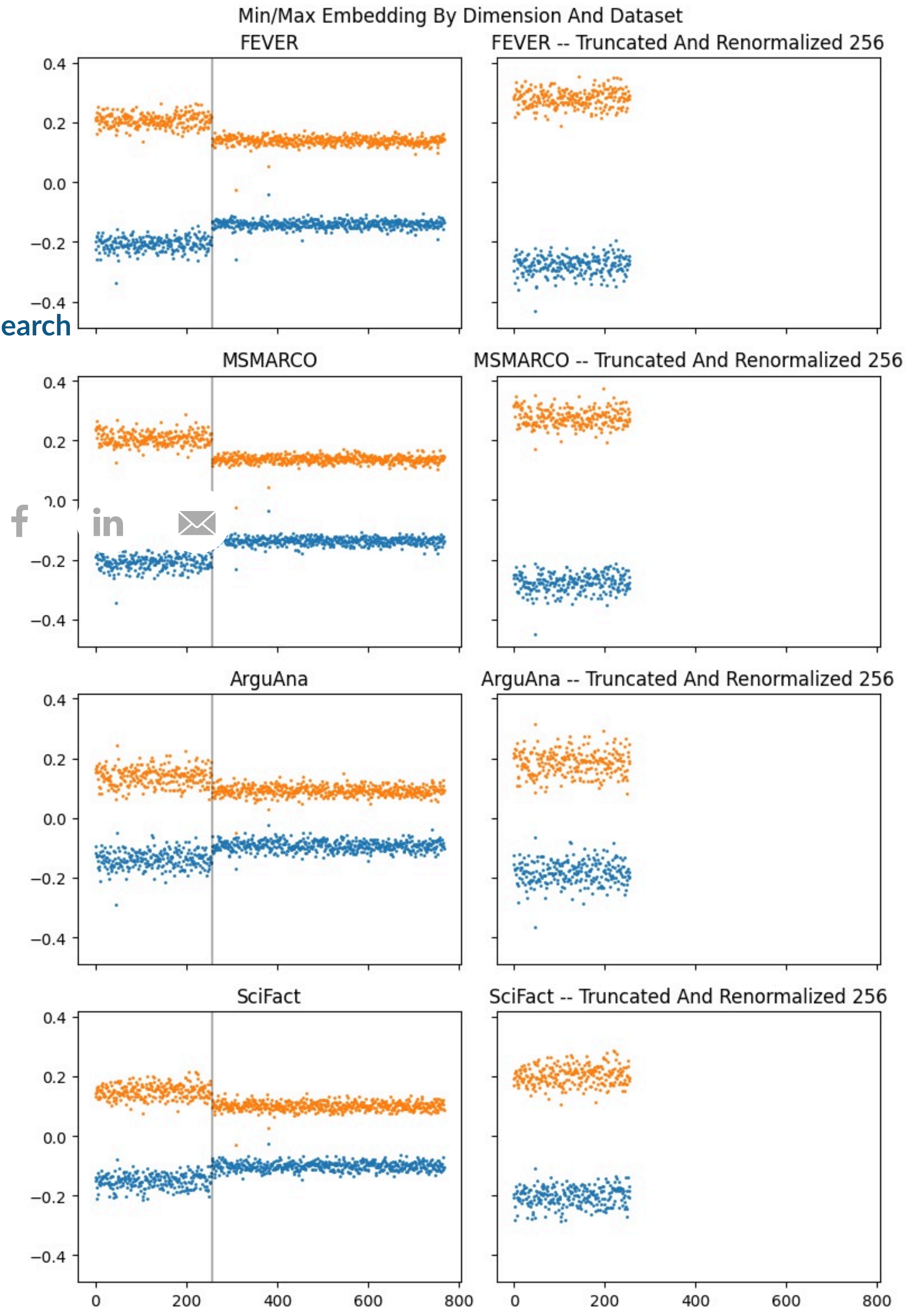


Figure 3. The single-dimension MRL strategy used in our v1.5 model achieves a more homogenous set of scalar ranges, which are easier to compress effectively with a global scalar quantization scheme.

AUTHOR

Luke Merrick

Snowflake AI Research

SHARE

This was our first trick to making v1.5 work well — by concurrently designing truncation and quantization to work well together, we were able to identify and sidestep the MRL + quantization pitfall, achieving higher quality at 128-byte compression than a more naive compression scheme (e.g., 128-dimension MRL with int8 quantization) could accomplish.

AUTHOR

Luke Merrick
Snowflake AI Research

Train Bigger and Longer

By focusing on just the M model size, it became easy to increase our pretraining duration from two to three epochs, as well as to increase our batch size by borrowing from the implementation of our v1.0 L-size model training. We found that the model did better overall with the larger batch size and that it continued to improve slightly with the extra epoch, as shown in the figure below. We also saw that these improvements carried over through fine-tuning and helped to offset the slight performance degradation we encountered when using MRL loss.

SHARE

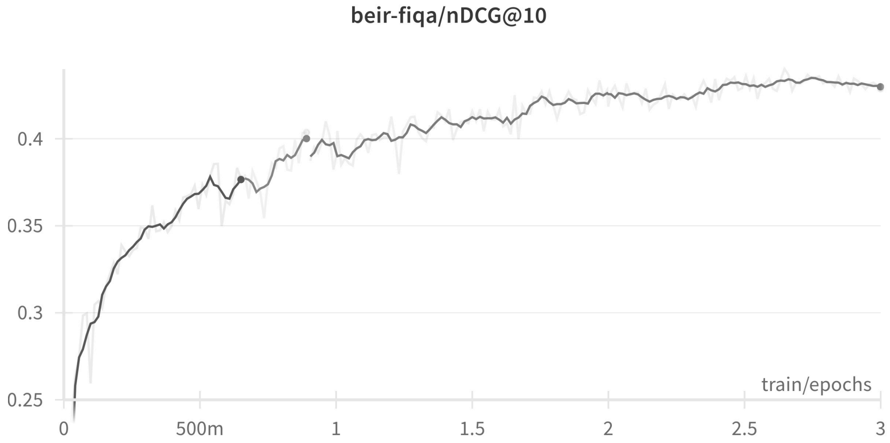


Figure 4. Model quality continues to improve slightly with three epochs of pretraining instead of two. The rolling average is shown over faded original values, with some discontinuities appearing as a result of training with automatic preemption and resumption.

Train with Smarter Negatives

In addition to training longer, we also trained smarter. During pretraining we leveraged a novel data-clustering approach to improve the quality of in-batch negatives. (Those interested in this method, please keep an eye out for a forthcoming technical report discussing this aspect.)

AUTHOR

Luke Merrick

Snowflake AI Research

SHARE

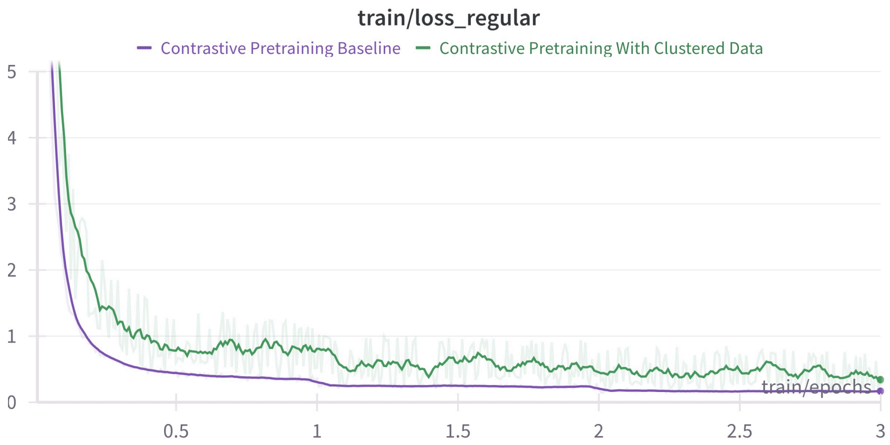


Figure 5. Clustered pretraining data leads to harder in-batch negatives. We can measure this by looking at the loss during training. Image from the coming technical report.

Additionally, during fine-tuning, we leveraged the curriculum-learning trick that we studied in our Arctic Embed v1.0 technical report (see: Figure 5) but did not use it when training Arctic Embed v1.0. The image below (reproduced from the report) demonstrates the gains from the curriculum-learning trick.

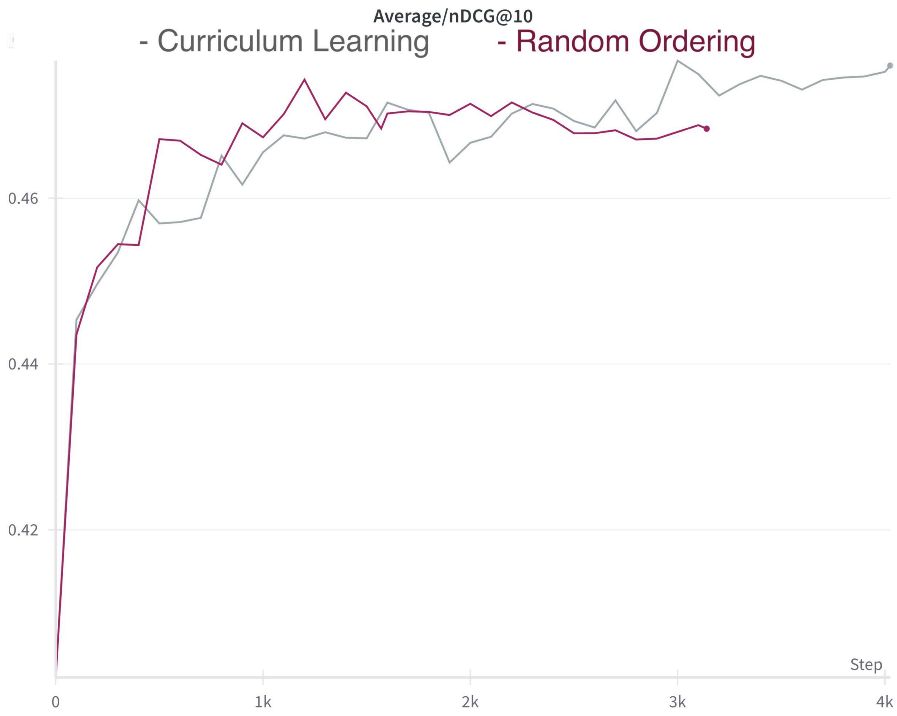


Figure 6. Curriculum learning using harder-and-harder negatives during fine-tuning mitigates quality plateau during fine-tuning. Image from our original v1.0 technical report – Arctic-Embed: Scalable, Efficient and Accurate Text Embedding Models.

Both of these “smart negatives” tricks, like the longer pretraining schedule, also helped mitigate the slight performance degradation that came from using MRL loss, pushing v1.5's uncompressed quality score above that of the v1.0 model despite the slight setback.

Summary and Getting Started

Snowflake Arctic Embed M v1.5 represents a significant leap forward in scalable, high-quality text embeddings, empowering enterprises to achieve efficient and cost-effective retrieval systems. By combining advanced compression techniques and extensive training optimizations, this model ensures unparalleled performance without compromising on quality when evaluating enterprise-grade text embedding models. Snowflake's commitment to AI openness is underscored by releasing this model under the permissive Apache 2.0 license, inviting developers and organizations to adopt and leverage Arctic Embed M v1.5 to unlock new possibilities in their large-scale search and retrieval applications. We encourage you to give it a try yourself on Hugging Face or to look at our usage example on GitHub.

AUTHOR

Luke Merrick
Snowflake AI Research

SHARE

SHARE

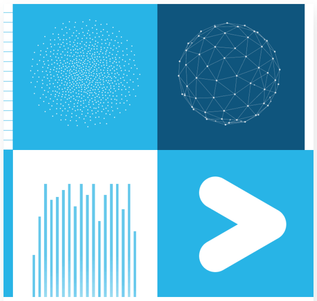
RELATED CONTENT

JUN 17, 2024

Snowflake Launches the World's Best Practical Text-

JUN 17, 2024

Snowflake Arctic: The Best LLM for



JUL 11, 2024

Snowflake Arctic Cookbook

Embedding Model for Retrieval Use Cases

Today Snowflake is launching and open-sourcing with an Apache 2.0 license the Snowflake Arctic embed family of models. Based on the Massive Text Embedding Benchmark (MTEB) Retrieval Leaderboard, the largest...

Learn More

Enterprise AI — Efficiently Intelligent, Truly Open

Building top-tier enterprise-grade intelligence using LLMs has traditionally been prohibitively expensive and resource-hungry, and often...

Explore

Series: A Deep Dive into LLM Evaluation Standards

What level of astronomy knowledge should an accountant have? Today's evaluations of LLMs assess their...

Learn More

SHARE

Snowflake Inc.

PLATFORM
- Cloud Data Platform
- Pricing
- Marketplace
- Security & Trust

SOLUTIONS
- Snowflake for Financial Services
- Snowflake for Advertising, Media, & Entertainment
- Snowflake for Retail & CPG

RESOURCES
- Resource Library
- Webinars
- Documentation
- Community
- Procurement
- Legal

EXPLORE
- News
- Blog
- Trending
- Guides
- Developers

ABOUT
- About Snowflake
- Investor Relations
- Leadership & Board
- Snowflake Ventures
- Careers

Healthcare &
Life Sciences
Data Cloud

Snowflake for
Marketing
Analytics

Contact

AUTHOR

Luke Merrick

Snowflake AI Research

SHARE

Privacy Notice | Site Terms | Cookie Settings | Do Not Share My Personal Information