JUN 17, 2024

AUTHOR
Snowflake AI Research

Snowflake Arctic: The Best LLM for Enterprise AI — Efficiently Intelligent, Truly Open

SHARE

Building top-tier enterprise-grade intelligence using LLMs has traditionally been prohibitively expensive and resource-hungry, and often costs tens to hundreds of millions of dollars. As researchers, we have grappled with the constraints of efficiently training and inferencing LLMs for years. Members of the Snowflake AI Research team pioneered systems such as ZeRO and DeepSpeed, PagedAttention / vLLM, and LLM360 which significantly reduced the cost of LLM training and inference, and open sourced them to make LLMs more accessible and cost-effective for the community.

Today, the Snowflake AI Research Team is thrilled to introduce Snowflake Arctic, a top-tier enterprise-focused LLM that pushes the frontiers of cost-effective training and openness. Arctic is *efficiently intelligent and truly open.*

Efficiently Intelligent: Arctic excels at enterprise tasks such as SQL generation, coding and instruction following benchmarks even when compared to open source models trained with significantly higher compute budgets. In fact, it sets a new baseline for cost-effective training to enable Snowflake customers to create high-quality custom models for their enterprise needs at a low cost.

AUTHOR
Snowflake AI Research

Truly Open: Apache 2.0 license provides ungated access to weights and code. In addition, we are also open sourcing all of our data recipes and research insights.

Snowflake Arctic is available from Hugging Face, NVIDIA API catalog and Replicate today or via your model garden or catalog of choice, including Snowflake Cortex, Amazon Web Services (AWS), Microsoft Azure, Lamini, Perplexity and Together over the coming days.

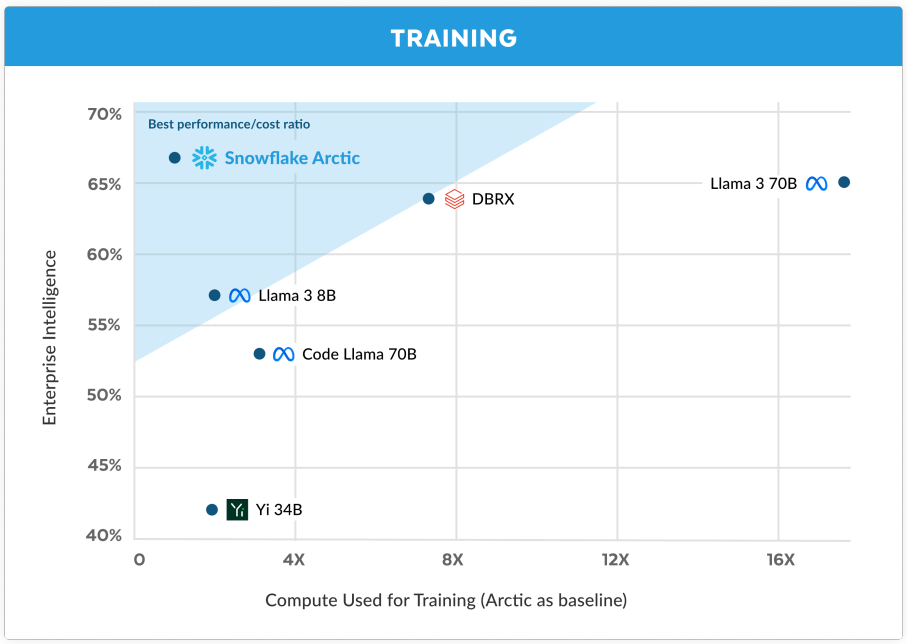


Fig 1. Enterprise intelligence – average of Coding (HumanEval+ and MBPP+), SQL Generation (Spider), and Instruction following (IFEval) – vs. Training cost

Top-tier enterprise intelligence at incredibly low training cost

At Snowflake, we see a consistent pattern in AI needs and use cases from our enterprise customers. Enterprises want to use

LLMs to build conversational SQL data copilots, code copilots and RAG chatbots. From a metrics perspective, this translates to LLMs that excel at SQL, code, complex instruction following and the ability to produce grounded answers. We capture these abilities into a single metric we call **enterprise intelligence** by taking an average of Coding (HumanEval+ and MBPP+), SQL Generation (Spider) and Instruction following (IFEval).

AUTHOR

Snowflake AI Research

Arctic offers top-tier enterprise intelligence among open source LLMs, and it does so using a training compute budget of roughly under $2 million (less than 3K GPU weeks). This means Arctic is more capable than other open source models trained with a

budget. More importantly, it excels at enterprise intelligence, even when compared to those trained with a significantly higher compute budget. The high training efficiency of Arctic also means that Snowflake customers and the AI community at large can train custom models in a much more affordable way.

As seen in Figure 1, Arctic is on par or better than both LLAMA 3 8B and LLAMA 2 70B on enterprise metrics, while using less than ½ of the training compute budget. Similarly, despite using 17x less compute budget, Arctic is on par with Llama3 70B in enterprise metrics like Coding (HumanEval+ & MBPP+), SQL (Spider) and Instruction Following (IFEval). It does so while remaining competitive on overall performance. For example, despite using 7x less compute than DBRX it remains competitive on Language Understanding and Reasoning (a collection of 11 metrics) while being better in Math (GSM8K). For a detailed breakdown of results by individual benchmark, see the Metrics section.

	Snowflake Arctic	DBRX	Llama 3 8B	Llama 2 70B	Llama 3 70B	Mixtral 8x7B	Mixtral 8x22B
Active Parameters	17B	36B	8B	70B	70B	13B	44B
ENTERPRISE							
SQL Generation (*Spider*)	79.0	76.3	69.9	62.8	80.2	71.3	79.2
Coding (*HumanEval+, MBPP+*)	64.3	61.0	59.2	33.7	71.9	48.1	69.9
Instruction Following (*IFEval*)	57.4	54.8	42.7	-	43.6	52.2	61.5
ACADEMIC							
Math (*GSM8K*)	74.2	73.5	75.4	52.6	91.4	63.2	84.2
Common Sense (*Avg of 11 metrics*)	73.1	74.8	68.5	72.1	72.6	74.1	75.6
World Knowledge (*MMLU*)	67.3	73.3	65.7	68.6	79.8	70.4	77.5

Table 1 Model architecture and training compute for Arctic, Llama-2 70B, DBRX and Mixtral 8x22B. Training compute is proportional to the product of active parameters and training tokens.

AUTHOR

Snowflake AI Research

SHARE

Training efficiency

To achieve this level of training efficiency, Arctic uses a unique Dense-MoE Hybrid transformer architecture. It combines a 10B dense transformer model with a residual 128×3.66B MoE MLP resulting in 480B total and 17B active parameters chosen using a top-2 gating. It was designed and trained using the following three key insights and innovations:

1) *Many-but-condensed experts with more expert choices:* In late 2021, the DeepSpeed team demonstrated that MoE can be applied to auto-regressive LLMs to significantly improve model quality without increasing compute cost.

In designing Arctic, we noticed, based on the above, that the improvement of the model quality depended primarily on the number of experts and the total number of parameters in the MoE model, and the number of ways in which these experts can be combined together.

Based on this insight, Arctic is designed to have 480B parameters spread across 128 fine-grained experts and uses top-2 gating to choose 17B active parameters. In contrast, recent MoE models are built with significantly fewer experts as shown in Table 2. Intuitively, Arctic leverages a large number of total parameters and many experts to enlarge the model capacity for top-tier intelligence, while it judiciously chooses among many-but-condensed experts and engages a moderate number of active parameters for resource-efficient training and inference.

AUTHOR

Snowflake AI Research

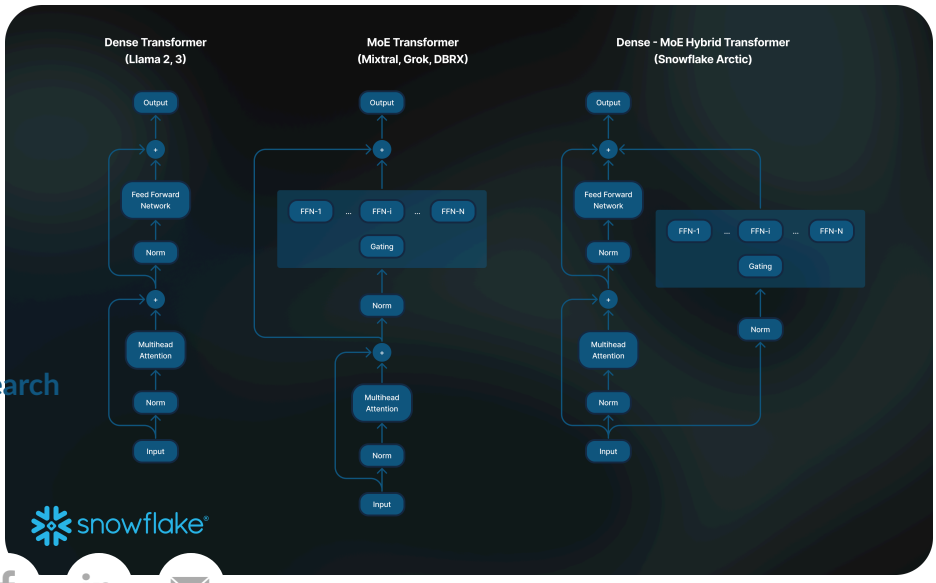


Figure 2. Standard MoE Architecture vs. Arctic

2) *Architecture and System Co-design:* Training vanilla MoE architecture with a large number of experts is very inefficient even on the most powerful AI training hardware due to high all-to-all communication overhead among experts. However, it is possible to hide this overhead if the communication can be overlapped with computation.

Our second insight is that combining a dense transformer with a residual MoE component (Fig 2) in the Arctic architecture enables our training system to achieve good training efficiency via communication computation overlap, hiding a big portion of the communication overhead.

3) *Enterprise-Focused Data Curriculum:* Excelling at enterprise metrics like Code Generation and SQL requires a vastly different data curriculum than training models for generic metrics. Over hundreds of small-scale ablations, we learned that generic skills like common sense reasoning can be learned in the beginning, while more complex metrics like coding, math and SQL can be learned effectively towards the latter part of the training. One can draw analogies to human life and education, where we acquire capabilities from simpler to harder. As such, Arctic was trained with a three-stage curriculum each with a different data composition focusing on generic skills in the first phase (1T Tokens), and enterprise-focused skills in the latter two phases (1.5T and 1T tokens). A high-level summary of our dynamic curriculum is shown here.

Data Category	Phase 1 (1 T)	Phase 2 (1.5 T)	Phase 3 (1T)
Web	75.03%	62.71%	52.46%
Code & SQL	4.50%	19.20%	26.71%
STEM	4.62%	5.80%	8.29%
Other	15.84%	12.29%	12.54%

Table 2. Dynamic data composition for three-phase training of Arctic with emphasis on enterprise intelligence.

AUTHOR
Snowflake AI Research

Inference efficiency

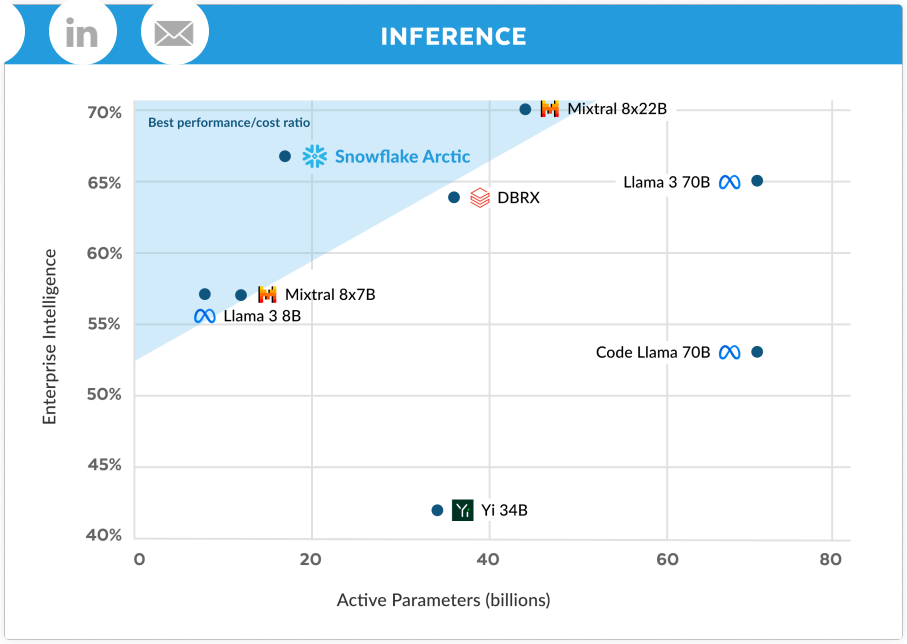


Figure 3. Enterprise intelligence – average of Coding (HumanEval+ and MBPP+), SQL Generation (Spider), and Instruction following (IFEval) vs. Active Parameters during Inference

Training efficiency represents only one side of the efficient intelligence of Arctic. Inference efficiency is equally critical to allow for the practical deployment of the model at a low cost. Arctic represents a leap in MoE model scale, using more experts and total parameters than any other open sourced auto-regressive MoE model. As such, several system insights and innovations are necessary to run inference on Arctic efficiently:

a) At interactive inference of a small batch size, e.g., batch size of 1, an MoE model's inference latency is bottlenecked by the time it takes to read all the active parameters, where the inference is memory bandwidth bounded. At this batch size, Arctic (17B active parameters) can have up to 4x less memory reads than Code-

Llama 70B, and up to 2.5x less than Mixtral 8x22B (44B active parameters), leading to faster inference performance.

We have collaborated with NVIDIA and worked with **NVIDIA TensorRT-LLM** and the **vLLM** teams to provide a preliminary implementation of Arctic for interactive inference. With FP8 quantization, we can fit Arctic within a single GPU node. While far from fully optimized, at a batch size of 1, Arctic has a throughput of over 70+ tokens/second for effective interactive serving.

AUTHOR
Snowflake AI Research

b) As the batch size increases significantly e.g., thousands of tokens per forward pass, Arctic switches from being memory [illegible] to compute bound, where the inference is bottlenecked by the active parameters per token. At this point, Arctic incurs 4x less compute than CodeLlama 70B and Llama 3 70B.

To enable compute bound inference and high relative throughput that corresponds to the small number of active parameters in Arctic (as shown in Fig 3), a large batch size is needed. Achieving this requires having enough KV cache memory to support the large batch size while also having enough memory to store nearly 500B parameters for the model. While challenging, this can be achieved with two-node inference using a combination of system optimizations such as FP8 weights, split-fuse and continuous batching, tensor parallelism within a node and pipeline parallelism across nodes.

We have worked closely with NVIDIA to optimize inference for NVIDIA NIM microservices powered by TensorRT-LLM. In parallel, we are working with the vLLM community, and our in-house development team is also enabling efficient inference of Arctic for enterprise use cases in the coming weeks.

Truly open

Arctic was built upon the collective experiences of our diverse team, as well as major insights and learnings from the community. Open collaboration is key to innovation, and Arctic would not have been possible without open source code and open research insights from the community. We are thankful to the community

and eager to give back our own learnings to enrich the collective knowledge and empower others to succeed.

Our commitment to a truly open ecosystem goes beyond open weights and code but also having open research insights and open source recipes.

AUTHOR

Snowflake AI Research

SHARE

Open research insights

The construction of Arctic has unfolded along two distinct trajectories: the open path, which we navigated swiftly thanks to the wealth of community insights, and the hard path, which is characterized by the segments of research that lacked prior community insights, necessitating intensive debugging and numerous ablations.

With this release, we're not just unveiling the model; we're also sharing our research insights through a comprehensive 'cookbook' that opens up our findings from the hard path. The cookbook is designed to expedite the learning process for anyone looking to build world-class MoE models. It offers a blend of high-level insights and granular technical details in crafting an LLM akin to Arctic so you can build your desired intelligence efficiently and economically — guided by the open path instead of the hard one.

The cookbook spans a breadth of topics, including pre-training, fine-tuning, inference and evaluation, and also delves into modeling, data, systems and infrastructure. You can preview **the table of contents**, which outlines over 20 subjects. We will be releasing corresponding Medium.com blog posts daily over the next month. For instance, we'll disclose our strategies for sourcing and refining web data in "What data to use?" We'll discuss our data composition and curriculum in "How to compose data." Our exploration of MoE architecture variations will be detailed in "Advanced MoE architecture," discussing the co-design of model architecture and system performance. And for those curious about LLM evaluation, our "How to evaluate and compare model quality — less straightforward than you think" will shed light on the unexpected complexities we encountered.

Through this initiative, we aspire to contribute to an open community where collective learning and advancement are the

norms to push the boundaries of this field further.

Open source serving code

We are releasing model checkpoints for both the base and instruct-tuned versions of Arctic under an Apache 2.0 license. This means you can use them freely in your own research, prototypes and products.

AUTHOR
Snowflake AI Research

Our LoRA-based fine-tuning pipeline, complete with a recipe, allows for efficient model tuning on a single node.

In ab ion with NVIDIA TensorRT-LLM and vLLM, we are developing initial inference implementations for Arctic, optimized for interactive use with a batch size of one. We are excited to work with the community to tackle the complexities of high-batch size inference of really large MoE models.

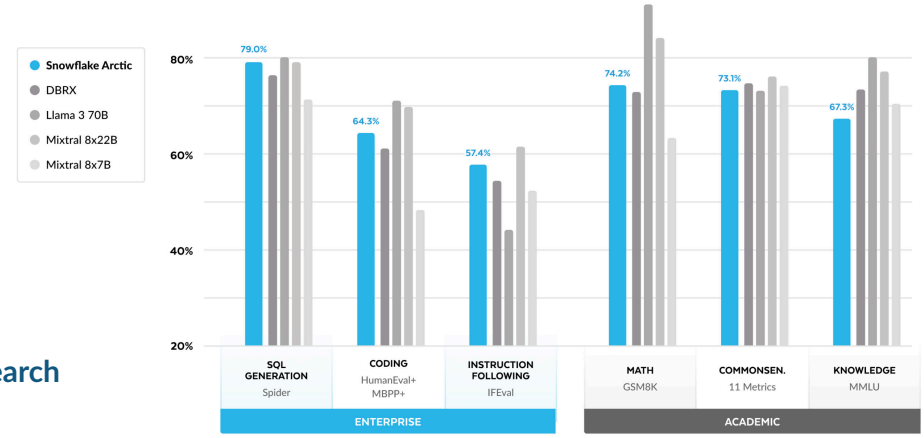
Arctic is trained using a 4K attention context window. We are developing an attention-sinks-based sliding window implementation to support unlimited sequence generation capability in the coming weeks. We look forward to working with the community to extend to a 32K attention window in the near future.

Metrics

Our focus from a metrics perspective is primarily on what we call **enterprise intelligence** metrics, a collection of skills that are critical for enterprise customers that includes, Coding (HumanEval+ and MBPP+), SQL Generation (Spider) and Instruction following (IFEval).

At the same time, it is equally important to evaluate LLMs on the metrics the research community evaluates them on. This includes world knowledge, common sense reasoning and math capabilities. We refer to these metrics as **academic benchmarks**.

Here is a comparison of Arctic with multiple open source models across enterprise and academic metrics:



AUTHOR

Snowflake AI Research

For enterprise metrics, Arctic demonstrates top-tier performance compared to all other open source models regardless of the m cl For other metrics, it achieves top-tier performance at its compute class and even remains competitive with models trained with higher compute budgets. Snowflake Arctic is the best open source model for off-the-shelf enterprise use cases. And if you are looking to train your own model from scratch at the lowest total cost of ownership (TCO), the training infrastructure and systems optimization descriptions in our cookbook should be of great interest.

For academic benchmarks, there has been a focus on world knowledge metrics such as MMLU to represent model performance. With high-quality web and STEM data, MMLU monotonically moves up as a function of training FLOPS. Since one objective for Arctic was to optimize for training efficiency while keeping the training budget small, a natural consequence is lower MMLU performance compared to recent top-tier models. In line with this insight, we expect our ongoing training run at a higher training compute budget than Arctic to exceed Arctic's MMLU performance. We note that performance on MMLU world knowledge doesn't necessarily correlate with our focus on enterprise intelligence.

	Snowflake Arctic	DBRX	Llama 3 8B	Llama 2 70B	Llama 3 70B	Mixtral 8x7B	Mixtral 8x22B
Active Parameters	17B	36B	8B	70B	70B	13B	44B
ENTERPRISE							
SQL Generation (Spider)	79.0	76.3	69.9	62.8	80.2	71.3	79.2
Coding (HumanEval+, MBPP+)	64.3	61.0	59.2	33.7	71.9	48.1	69.9
Instruction Following (IFEval)	57.4	54.8	42.7	-	43.6	52.2	61.5
ACADEMIC							
Math (GSM8K)	74.2	73.5	75.4	52.6	91.4	63.2	84.2
Common Sense (Avg of 11 metrics)	73.1	74.8	68.5	72.1	72.6	74.1	75.6
World Knowledge (MMLU)	67.3	73.3	65.7	68.6	79.8	70.4	77.5

Table 3. Full Metrics Table. Comparing Snowflake Arctic with DBRX, LLAMA-3 8B, LLAMA-3 70B, Mixtral 8x7B, Mixtral 8x22B (instruction-tuned or chat variants if available).[1] [2] [3]

Getting started with Arctic

AUTHOR
Snowflake AI Research

Snowflake AI Research also recently announced and open sourced the Arctic Embed family of models that achieves SoTA in MTEB retrieval. We are eager to work with the community as we develop the next generation in the Arctic family of models. Join us at our Data Cloud Summit on June 3-6 to learn more.

Here's how you can collaborate on Arctic starting today:

- Go to **Hugging Face** to directly download Arctic and use our **Github repo** for inference and fine-tuning recipes.
- For a serverless experience in Snowflake Cortex, Snowflake customers with a payment method on file will be able to access Snowflake Arctic for free until June 3. **Daily limits apply**.
- Access Arctic via your model garden or catalog of choice including Amazon Web Services (AWS), Lamini, Microsoft Azure, **NVIDIA API catalog**, Perplexity, **Replicate** and Together AI over the coming days.
- Chat with Arctic! Try a live demo now on **Streamlit Community Cloud** or on **Hugging Face Streamlit Spaces**, with an API powered by our friends at Replicate.
- Get mentorship and credits to help you build your own Arctic-powered applications during our **Arctic-themed Community Hackathon**.

And finally, don't forget to read the first edition of our **cookbook** recipes to learn more about how to build your own custom MoE models in the most cost-effective way possible.

Acknowledgments

We would like to thank AWS for their collaboration and partnership in building Arctic's training cluster and infrastructure, and NVIDIA for their collaboration in enabling Arctic support on NVIDIA NIM with TensorRT-LLM. We also thank the open source community for producing the models, datasets and dataset recipe insights we could build on top of to make this release possible. We would also like to thank our partners in AWS, Microsoft Azure, NVIDIA API catalog, Lamini, Perplexity, Replicate and Together AI for their collaboration in making Arctic available

AUTHOR

Snowflake AI Research

1. The 11 metrics for Language Understanding and Reasoning include ARC-Easy, ARC-Challenge, BoolQ, CommonsenseQA,)F ell ag, LAMBADA, OpenBookQA, PIQA, RACE and WinoGrande.

SHARE

2. Evaluation scores for HumanEval+/MBPP+ v0.1.0 were obtained assuming (1) bigcode-evaluation-harness using model-specific chat templates and aligned post-processing, (2) greedy decoding. We evaluated all models with our pipeline to ensure consistency. We validated that our evaluations results are consistent with **EvalPlus leaderboard**. In fact, our pipeline produces numbers that are a few points higher than the numbers in EvalPlus for all models giving us confidence that we are evaluating each model in the best way possible.

3. IFEval scores reported are the average of prompt_level_strict_acc and inst_level_strict_acc

SHARE

AUTHOR

Snowflake AI Research

SHARE