

JUN 17, 2024

AUTHOR

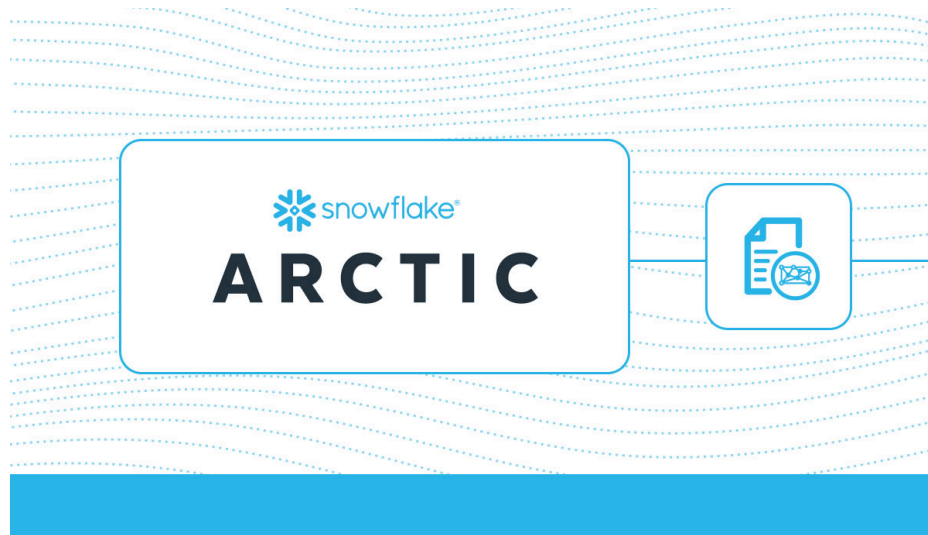
Snowflake AI Research

Snowflake's Arctic-TILT: A State-of-the-Art Document Intelligence LLM in a Single A10

SHARE



Machine Learning



The volume of unstructured data — such as PDFs, images, video and audio files — is surging across enterprises today. Yet documents, which represent a substantial portion of this data and hold significant value, continue to be processed through inefficient and manual methods.

To help organizations derive more value from unstructured data, Snowflake has introduced Document AI. Currently in private preview, this new feature easily extracts content, like invoice amounts or contract terms, from documents via a proprietary, built-in, multimodal large language model (LLM) we call Snowflake Arctic-TILT (Text Image Layout Transformer). Arctic-TILT joins [Snowflake Arctic](#), an efficient and open source LLM we [announced April 24](#), as another testament to the Snowflake AI

Research Team's commitment to continuously improving LLM accuracy and making purpose-built products for enterprises.





Now, we are proud to share a breakthrough benchmark achievement for Arctic-TILT. The model, containing 0.8B parameters, has secured a top score in the [DocVQA benchmark test](#), the standard for visual document question answering. It even beat GPT-4, which, by comparison, is believed to be a 8x220B MoE model.

AUTHOR

Snowflake AI Research

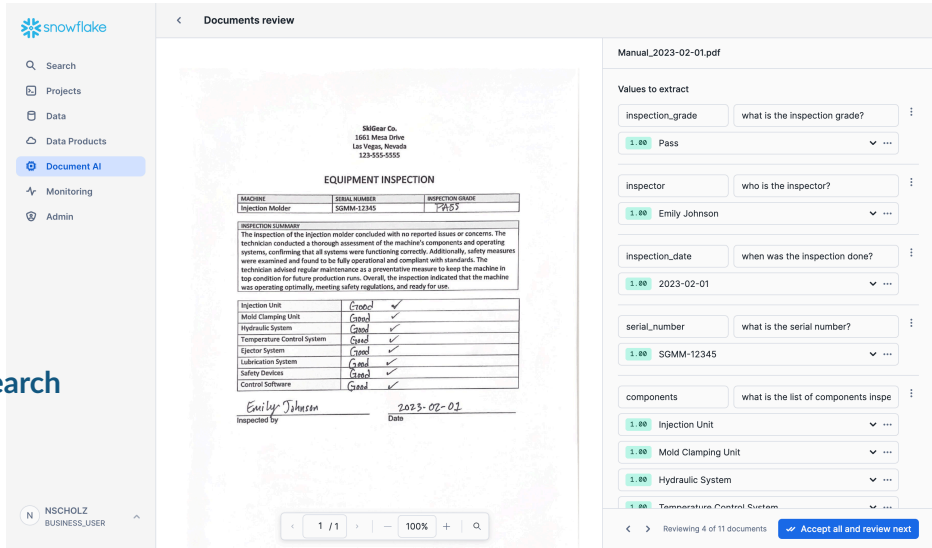
This demonstrates that smaller and more efficient models that are specifically tailored for targeted applications can outperform

SHARE

    general purpose models. It also means that the Arctic-TILT model can fit into a single A10 GPU instance, making it cheaper and more accessible.

What is Arctic-TILT?

Arctic-TILT is a Snowflake-grown LLM that leverages a proprietary and unique transformer architecture, tailored to understand and extract data from documents. By combining multiple data modalities, Arctic-TILT offers unparalleled versatility and performance in document-understanding tasks. The model powers Snowflake's [Document AI feature](#), an intelligent document-processing solution that allows users to interact with the model via a natural language interface; users can ask questions of their documents, evaluate and annotate the responses and optionally fine-tune the model with a simple click of a button. Part of Snowflake's platform, this solution allows users to structure their unstructured data, level it side-by-side with existing data in tables, and produce automated workloads and analytics previously out of reach.



AUTHOR
Snowflake AI Research

SHARE

Document AI users can interact with the Arctic-TILT model via a natural language interface

Key Features and Capabilities

- Multimodal Understanding:** Arctic-TILT can understand, analyze and extract information from text, images and spatial layouts simultaneously, providing a holistic understanding of content and its context.
- State-of-the-Art Performance:** On benchmarks such as [DocVQA](#), Arctic-TILT demonstrates Visual Question Answering capabilities on par if not better than models like GPT-4 with orders of magnitude more parameters.
- Extended Context Window:** Arctic-TILT features an exceptionally large context window of 375,000 tokens. This capability is crucial for grasping the full context of multimodal content.
- Efficient Inference:** Arctic-TILT is designed to handle both small- and enterprise-scale document volumes, while maintaining performance and, more importantly, accuracy — both of which are critical when it comes to business document processing.
- Adaptability:** Designed for a wide range of applications and industries, Arctic-TILT requires no previous knowledge of a

given document or format and is easily fine-tuned if needed.

Punching above its weight: Arctic-TILT on DocVQA

AUTHOR

Snowflake AI Research

SHARE

Average Normalized Levenshtein Similarity (ANLS) score is a metric used to provide a comprehensive assessment of a model’s performance in handling various textual inputs. In our most recent evaluations of the DocVQA data set, Arctic-TILT achieved an impressive 90% ANLS score despite being substantially smaller with far fewer parameters — and therefore cheaper — than other LLMs. Unlike other models, Arctic-TILT efficiently leverages far fewer parameters to match state-of-the-art results, often [further from](#) more resource-intensive models. This efficiency is a product of its purpose-built and sophisticated design, which optimally balances performance with resource utilization and enables cost-effective training at enterprise scale. In short, Snowflake customers see the benefits of that efficiency in lower costs.

Like Snowflake Arctic, Arctic-TILT is built to deliver top-tier outcomes with a lean parameter set, reflecting not only Snowflake’s dedication to pushing the envelope in AI capabilities but also our commitment to developing scalable enterprise AI solutions. This paradigm — combining robust results with low resource usage — makes advanced AI more accessible, nimble and effective.

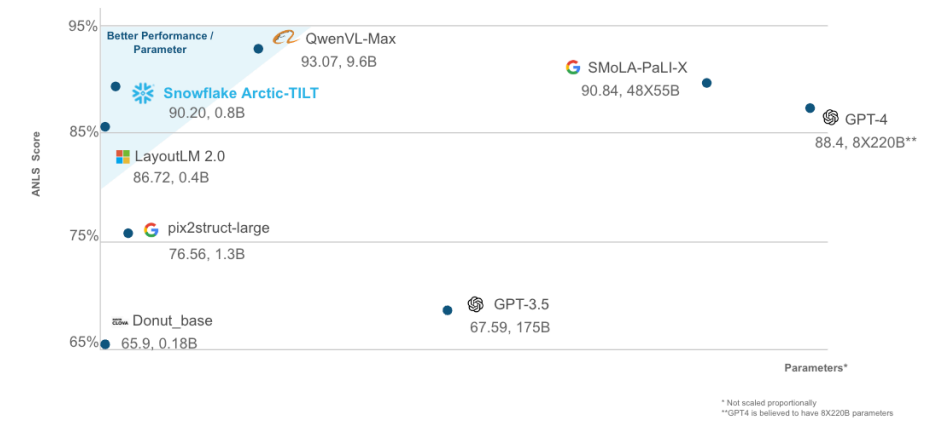


Figure 2: DocVQA result from notable companies, source: [Robust Reading Competition](#)

Why DocVQA?

Document Visual Question Answering (DocVQA) is a recognized benchmark for evaluating the capabilities of models to handle document-centric questions and answers. It provides comprehensive challenges across a data set of 50,000 questions defined on 12,000+ document images that test a model's ability to understand different types of information conveyed by a document. This includes, but is not limited to:

AUTHOR

Snowflake AI Research

Textual content (handwritten or typewritten)

Non-textual elements (marks, tick boxes, separators, diagrams)

SHARE

f L: in t (✉ structure, forms, tables)

By excelling in DocVQA, Arctic-TILT proves its effectiveness in real-world scenarios where complex document understanding is critical.

Document AI use cases and applications

With Snowflake Document AI, powered by Arctic-TILT, users have a new way of interacting with unstructured data. Document AI simplifies the setup and deployment of the Arctic-TILT model, enabling users without any machine learning background to effortlessly package model builds for high-value, enterprise-scale extraction tasks. Through an intuitive natural language interface, document owners can use their domain expertise to prepare models for their specific use case and, if necessary, train the model with a single click of a button.

Once the model is prepared, they pass the reins to pipeline or data engineers, who embed these models into operational workflows and frameworks. This seamless integration not only enhances efficiency, but it also scales up the potential applications of AI within enterprise environments.

Private preview customers across industries are already using Document AI to gain more value from their documents — including patient records and insurance claims in healthcare, tax filings and loan applications in financial services, licensing agreements in technology, and talent and copyright contracts in media, just to name a few.

Looking ahead and getting started with Document AI

Join us at the [Data Cloud Summit](#) from June 3-6 to learn more about Document AI and how you can put the power of enterprise AI into the hands of your entire business.

AUTHOR

[Snowflake AI Research](#)

To try Document AI firsthand, Snowflake customers can reach out to their account team for more information and enablement. You can also start today by using Snowflake Arctic models, such as [Arctic LLM](#), an efficient and truly open model optimized for enterprise intelligence, and [Arctic embed](#), the world's best embedding model for retrieval available in [Hugging Face](#) or [Snowflake Cortex](#).

SHARE

[f](#) [in](#) [te](#) [✉](#)

SHARE

[f](#) [in](#) [✉](#)

Start your 30-day free trial

START NOW!



PLATFORM	SOLUTIONS	RESOURCES	EXPLORE	ABOUT
Cloud Data Platform	Snowflake for Financial Services	Resource Library	News	About Snowflake
Pricing	Snowflake for Advertising, Media, & Entertainment	Webinars	Blog	Investor Relations
Marketplace	Snowflake for Retail & CPG	Documentation	Trending	Leadership & Board
Security & Trust	Healthcare & Life Sciences	Community	Guides	Snowflake Ventures
		Procurement	Developers	Careers
		Legal		Contact

Data Cloud
Snowflake for
Marketing
Analytics

AUTHOR
Snowflake AI Research

Sign up for
Snowflake
Communications

diana.shaw@sno United States

By submitting this form, I understand Snowflake will process my personal information in accordance with their **Privacy Notice**. Additionally, I consent to my information being shared with Event Partners in accordance with Snowflake’s **Event Privacy Notice**. I understand I may withdraw my consent or update my preferences **here** at any time.

SHARE



SUBSCRIBE NOW

[Privacy Notice](#) | [Site Terms](#) | [Cookie Settings](#) | [Do Not Share My Personal Information](#)

© 2024 Snowflake Inc. All Rights Reserved | If you’d rather not receive future emails from Snowflake, [unsubscribe here](#) or [customize your communication preferences](#)

