Safe Harbor and Disclaimers

Other than statements of historical fact, all information contained in these materials and any accompanying oral commentary (collectively, the “Materials”), including statements regarding (i) Snowflake’s business strategy, plans or priorities, (ii) Snowflake’s new or enhanced products, services, and technology offerings, including those that are under development or not generally available, (iii) market growth, trends, and competitive considerations, (iv) our vision for Snowpark, the Data Cloud, and industry-specific Data Clouds, including the expected benefits and network effects of the Data Cloud; and (v) the integration, interoperability, and availability of Snowflake’s products, services, or technology offerings with or on third-party platforms or products, are forward-looking statements. These forward-looking statements are subject to a number of risks, uncertainties and assumptions, including those described under the heading “Risk Factors” and elsewhere in the Annual Reports on Form 10-K and the Quarterly Reports on Form 10-Q that Snowflake files with the Securities and Exchange Commission. In light of these risks, uncertainties, and assumptions, the future events and trends discussed in the Materials may not occur, and actual results could differ materially and adversely from those anticipated or implied in the forward-looking statements. As a result, you should not rely on any forwarding-looking statements as predictions of future events. Forward-looking statements speak only as of the date the statements are first made and are based on information available to us at the time those statements are made and/or management's good faith belief as of that time. Except as required by law, we undertake no obligation, and do not intend, to update the forward-looking statements in these Materials.

Any future product or roadmap information (collectively, the “Roadmap”) is intended to outline general product direction. The Roadmap is not a commitment, promise, or legal obligation for Snowflake to deliver any future products, features, or functionality; and is not intended to be, and shall not be deemed to be, incorporated into any contract. The actual timing of any product, feature, or functionality that is ultimately made available may be different from what is presented in the Roadmap. The Roadmap information should not be used when making a purchasing decision. In case of conflict between the information contained in the Materials and official Snowflake documentation, official Snowflake documentation should take precedence over these Materials. Further, note that Snowflake has made no determination as to whether separate fees will be charged for any future products, features, and/or functionality which may ultimately be made available. Snowflake may, in its own discretion, choose to charge separate fees for the delivery of any future products, features, and/or functionality which are ultimately made available.

The Materials may contain information provided by third-parties. Snowflake has not independently verified this information, and usage of this information does not mean or imply that Snowflake has adopted this information as its own or independently verified its accuracy.

WARNING: Preview Features In Use

Scenario

Audio files in call centers offer rich insights to aid business objectives such as customer satisfaction, customer retention, quality metrics, efficiency scoring, and staffing models. With modern AI tools and frameworks we may easily (relative term) extract call summary details, measure sentiment, and discover patterns to help in enhancing customer experience. By transcribing audio to text and applying analytics, call centers gain actionable insights on agent responses and proactive issue resolution, ultimately driving better customer satisfaction.

Our fictional scenario is for a vehicle insurance company. Call Center Supervisors need agent performance details such as Average Handle Time(AHT), first call resolution count, and Sentiment scoring. We, the IT organization, have been tasked with transcribing text from Call Center audio (call) recordings, calculating call duration, and calculating various metrics. We must further build+deploy an interactive application for the supervisors. BTW, supervisors have also requested the ability to ask questions of the recordings via a chatbot.

This is our prototype…

Play Along/Clone Away!

https://signup.snowflake.com/

https://github.com/sfc-gh-psheehan/kcdc-call-center-analytics-with-snowflake-cortex-and-spcs

Components

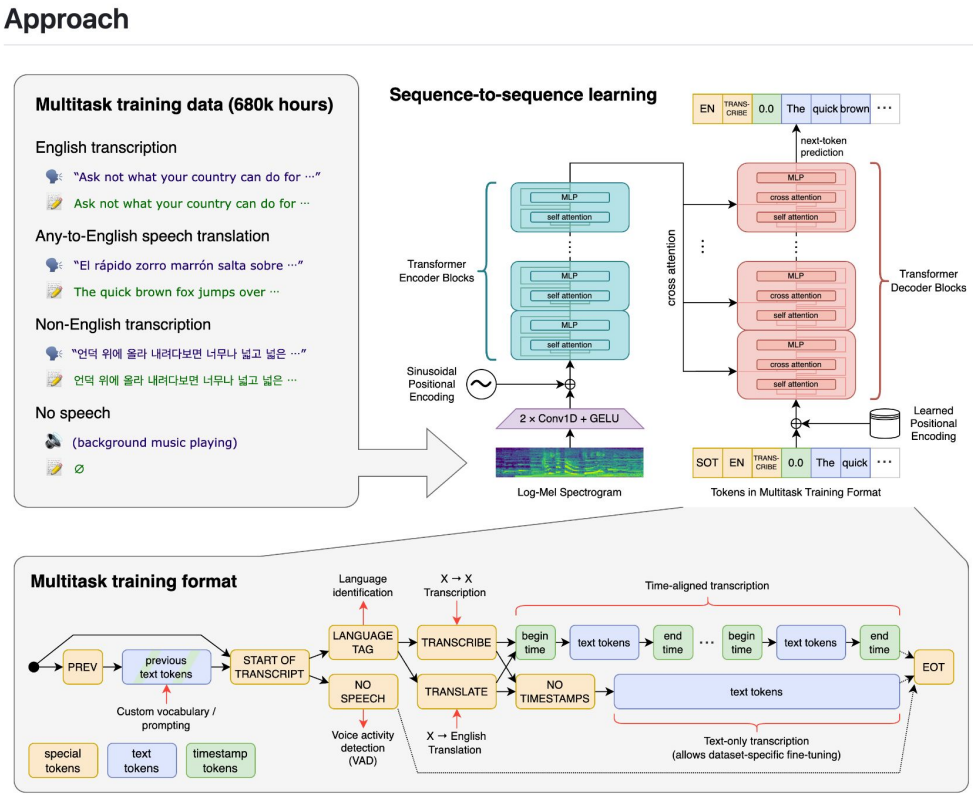
- **Audio Files in Object Storage:** Recordings in file format (.mpeg) stored in an object storage location (Internal Stage) for easy access and retrieval purposes.
- **Audio-to-Text Conversion:** OpenAI Whisper running in k8s (SPCS) to transcribe insurance call center audio files into text, extract call duration facilitating for efficient analysis.
- **Insight Capture:** Extract and compile customer details, agent interactions, sentiment analysis, summary, resolution, next steps, duration, and intent for every call using Cortex LLM functions.
- **Supervisor Dashboard:** Streamlit dashboard to showcase metrics, enabling users to gain a holistic view of various kpis and user experiences.
- **RAG-Based Chatbot:** Contextual responses for enhanced user engagement.
- **Text2SQL Functionality:** Empower users with a personalized copilot, allowing for natural language queries and output tailored to tables in context, to enhance user experience and analytical capabilities. We will leverage a custom-trained instance of HuggingFace's nsql-llama-2-7B running in k8s (SPCS) for this task.

Requirements

- Familiarity with Python, Pandas, DataFrame API
- Familiarity with Snowflake
- Familiarity with Docker & Kubernetes (k8s)
- Familiarity with YAML
- File storage location (we'll use a Snowflake internal stage for this session)
- Access to GitHub/GitLab (source code repo)
- Integrated Development Environment (IDE; we'll use VSCode for this session)
- Ability to follow directions

WTH is Whisper? https://openai.com/index/whisper/

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification.



WTH is NumberStation llama-2-7b?

NSQL is a family of autoregressive open-source large foundation models (FMs) designed specifically for SQL generation tasks.

In this repository we are introducing a new member of NSQL, NSQL-Llama-2-7B. It's based on Meta's original Llama-2 7B model and further pre-trained on a dataset of general SQL queries and then fine-tuned on a dataset composed of text-to-SQL pairs.

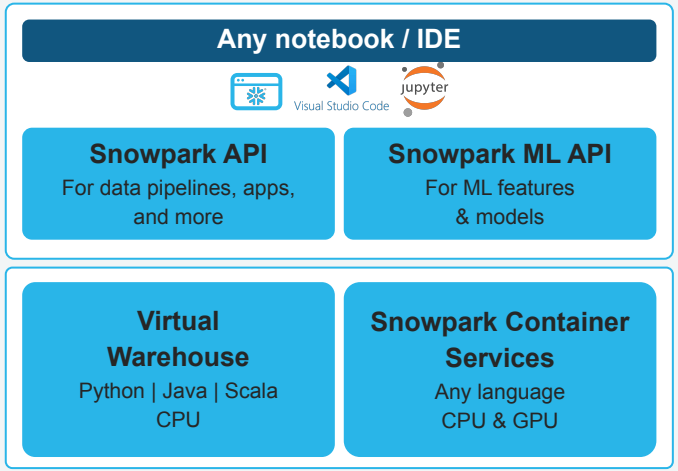
https://huggingface.co/NumbersStation/nsql-llama-2-7B

WTH is Snowpark?

What is Snowpark?
Set of libraries and runtimes that securely deploy and process Python and other programming languages in Snowflake to develop data pipelines, machine learning models, apps, and more.

Code development & deployment
client-side libraries

Code execution
elastic compute runtimes



Language of Choice on a Single Platform

Scalability Without Operational Complexity

No Governance and Security Trade-offs

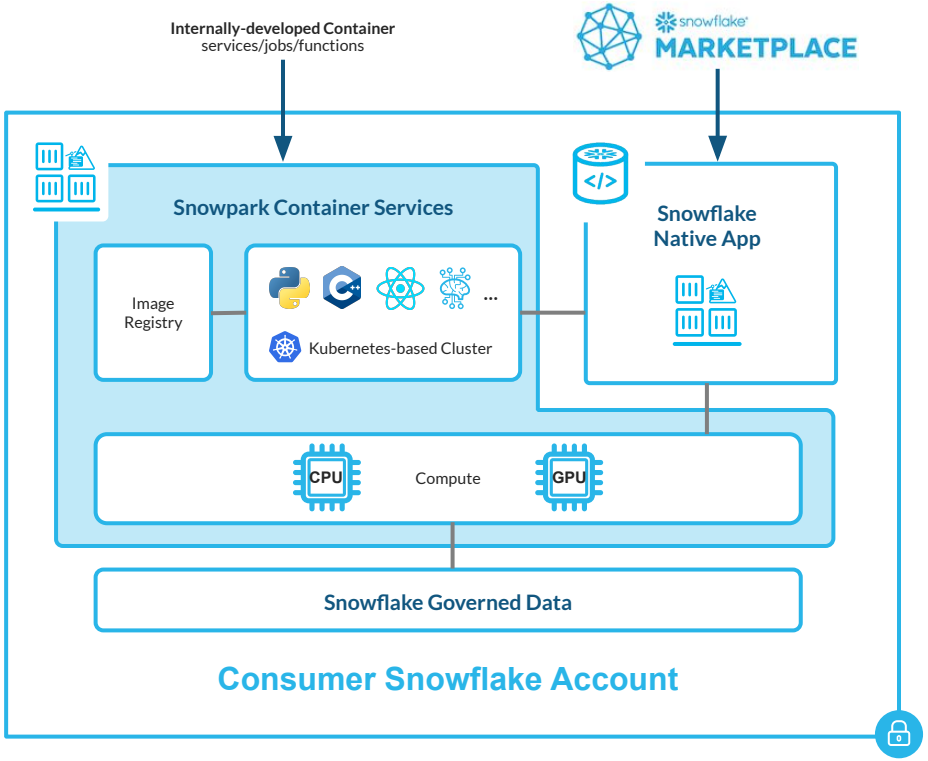
WTH is Snowpark Container Services (SPCS)?

What is it

Additional Snowpark runtime that helps developers register and deploy container images in Snowflake

Why use it

- **Language & hardware flexibility:** Build in any programming language, package as a container image and deploy in configurable CPUs & GPUs
- **Unified services experience:** Effortlessly deploy with integrated image registry, elastic compute infrastructure and managed Kubernetes-based cluster
- **Bring sophisticated apps to the data:** Run entire containerized applications from third-party developers in your account as Snowflake Native Apps via Snowflake Marketplace

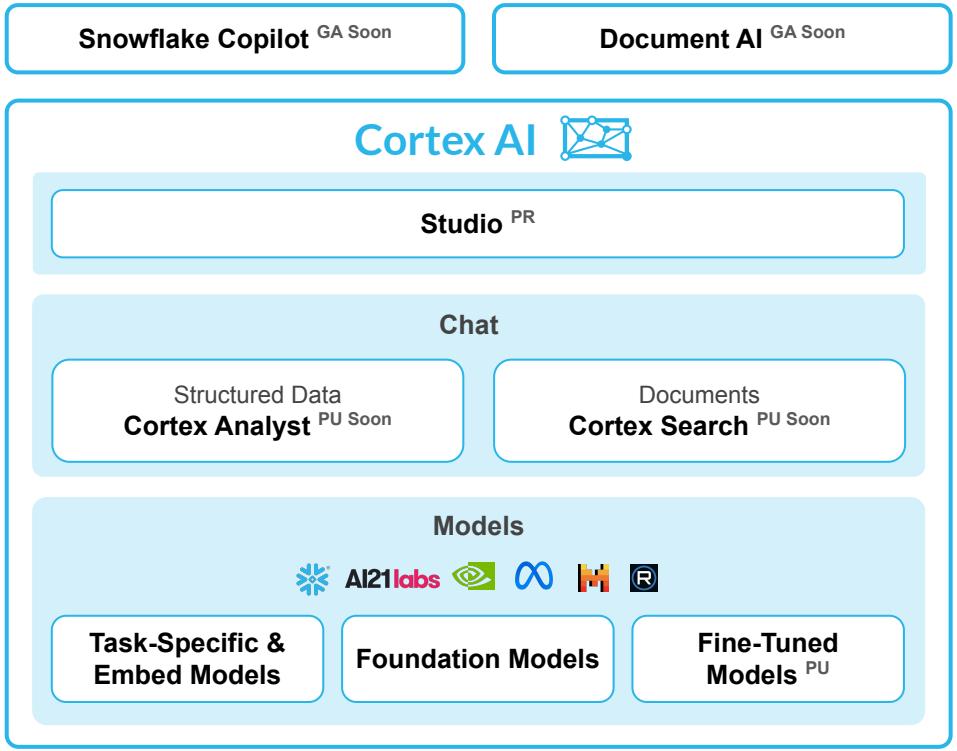


https://docs.snowflake.com/en/developer-guide/snowpark-container-services/working-with-services

https://docs.snowflake.com/en/developer-guide/snowpark-container-services/working-with-compute-pool

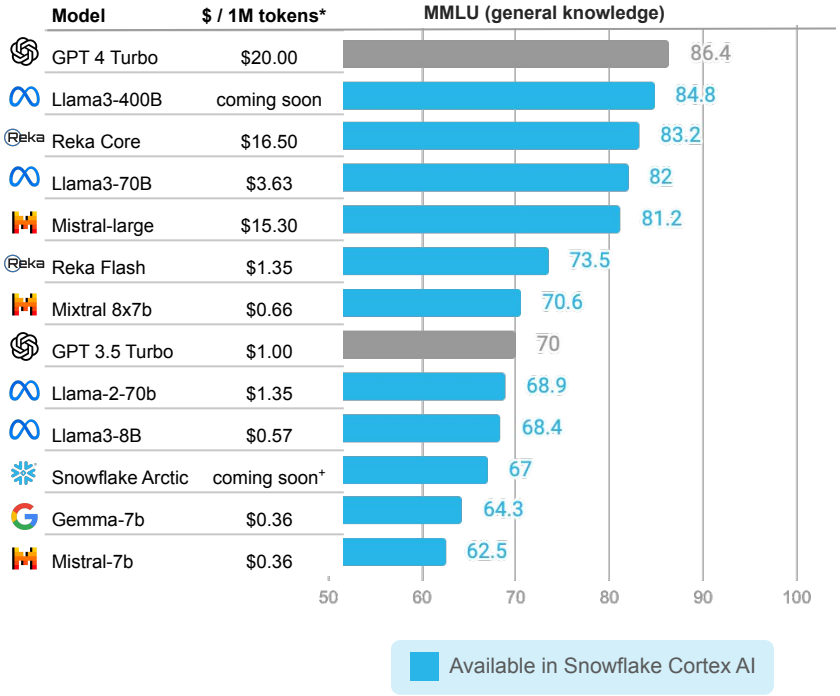
* for select AWS regions

WTH is Cortex?



GA: Generally Available | PR: Private Preview | PU: Public Preview

Foundation Models in Cortex



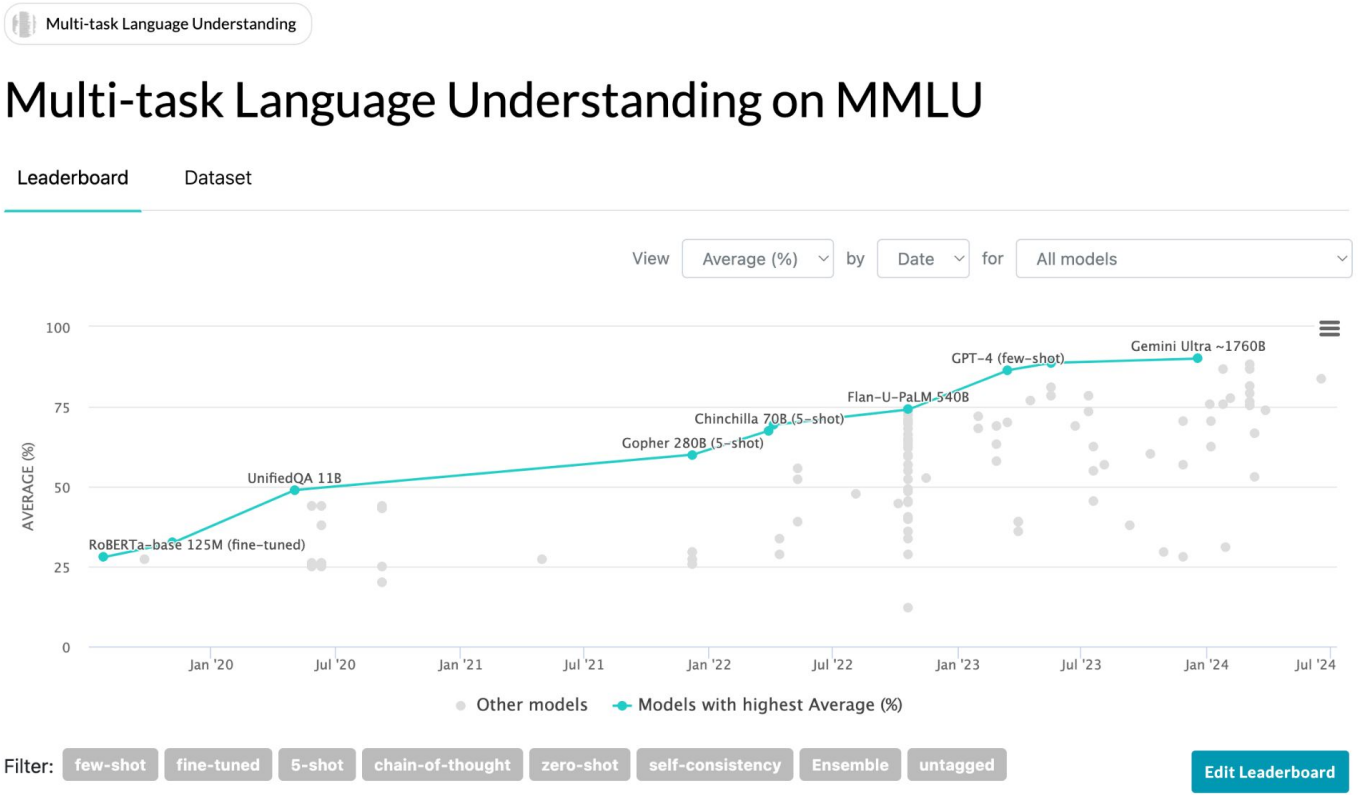
- Flexibility to choose the state-of-the-art model that best fits your task and cost goals
- Run batch operations using SQL or Python serverless functions inside Snowflake
- Use from any application via REST API
- No GPU infrastructure management
- Data stays in Snowflake and is never shared with third-party LLM provider
- RBAC policies limit data access to appropriate roles

*Assumes $3 per Snowflake credit for LLMs in Cortex and blended 50/50 input-output tokens for OpenAI
+ Free for a limited time

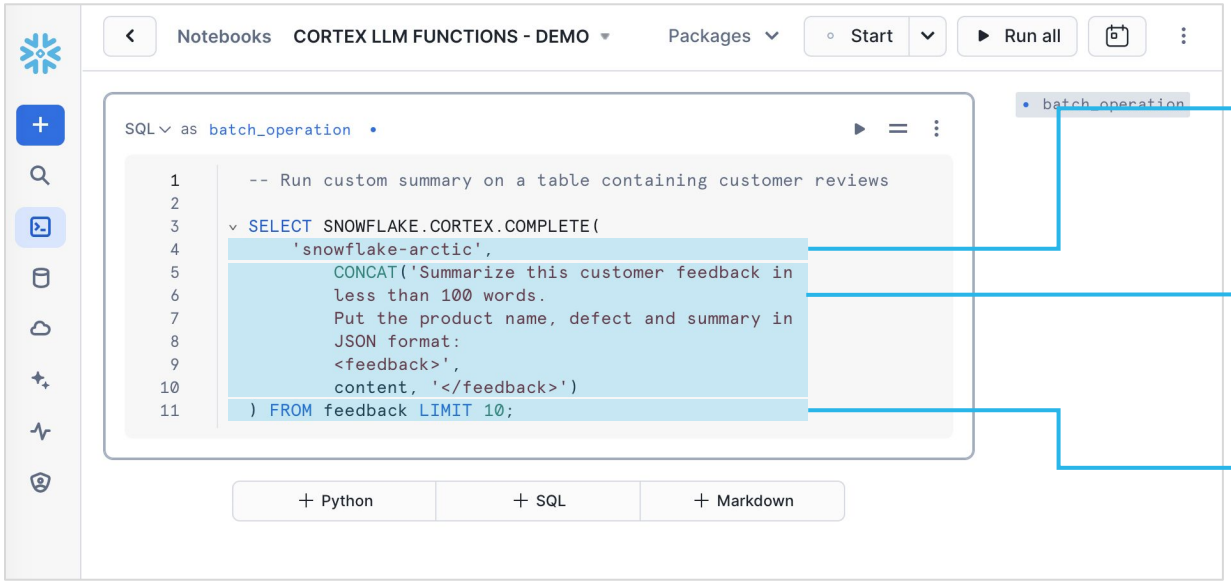
WTH is MMLU?

Measuring Massive Multitask Language Understanding is a benchmark for evaluating the capabilities of language models. It consists of about 16,000 multiple-choice questions spanning 57 academic subjects including mathematics, philosophy, law, and medicine.

https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu



Cortex - How-to



Choose Foundation Model

Add Prompt

Point to Table

Text processing use cases:

- Custom summaries (text to JSON)
- Advanced categorization
- Sentiment analysis, translation & other NLP

Available interfaces:

- SQL
- Python
- REST API (PrPr)

WTH are Tokens?

Tokens: Tokens are the basic units of input and output in a language model. In natural language processing tasks, tokens typically represent words, subwords, or characters. During training and inference, the LLM processes input text as a sequence of tokens, each representing a specific word or symbol in the input text. The model generates output by predicting the most likely token to follow a given sequence of input tokens.

snowflake

Table 6(a): Snowflake AI Features Credit Table

Feature	Snowflake-managed compute[14]
Cortex Complete – reka-core	5.50 Credits / 1M Tokens
Cortex Complete – mistral-large	5.10 Credits / 1M Tokens
Cortex Complete – llama3-70b	1.21 Credits / 1M Tokens
Cortex Complete – llama2-chat-70b	0.45 Credits / 1M Tokens
Cortex Complete – reka-flash	0.45 Credits / 1M Tokens
Cortex Complete – mixtral-8x7b	0.22 Credits / 1M Tokens
Cortex Complete – llama3-8b	0.19 Credits / 1M Tokens
Cortex Complete – mistral-7b	0.12 Credits / 1M Tokens
Cortex Complete – gemma-7b	0.12 Credits / 1M Tokens
Cortex Complete – snowflake-arctic[15]	0 Credits / 1M Tokens
Cortex Translate	0.33 Credits / 1M Tokens
Cortex Summarize	0.10 Credits / 1M Tokens
Cortex Extract Answer	0.08 Credits / 1M Tokens
Cortex Sentiment	0.08 Credits / 1M Tokens
Cortex Embed Text 1024 – nv-embed-qa-4	0.05 Credits / 1M Tokens
Cortex Embed Text 768 – snowflake-arctic-embed-m	0.03 Credits / 1M Tokens
Cortex Embed Text 768 – e5-base-v2	0.03 Credits / 1M Tokens
Document AI[16]	8 Credits per hour of compute

Table 6(b): Snowflake AI Features Credit Table (Fine-Tuning)[16]

Feature	Snowflake-managed compute[14]	
	Fine-Tuning (Training)	Cortex Complete (Fine-Tuned Inference)
Cortex Fine-Tuning – llama3-70b	3.40 Credits / 1M Tokens	2.42 Credits / 1M Tokens
Cortex Fine-Tuning – mixtral-8x7b	3.40 Credits / 1M Tokens	0.44 Credits / 1M Tokens
Cortex Fine-Tuning – llama3-8b	0.64 Credits / 1M Tokens	0.38 Credits / 1M Tokens
Cortex Fine-Tuning – mistral-7b	0.64 Credits / 1M Tokens	0.24 Credits / 1M Tokens

In Dev Private Public GA

Cortex Serverless Fine-Tuning

WHAT IS IT

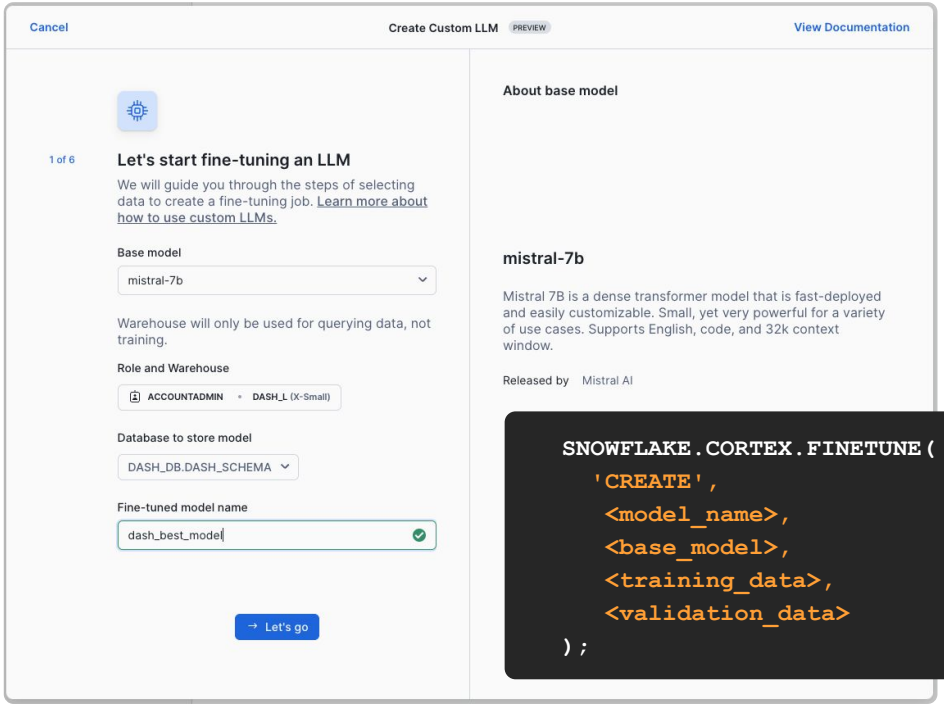
Serverless fine-tuning for subset of Mistral and Llama 3 LLMs available in Cortex AI.

WHY USE IT

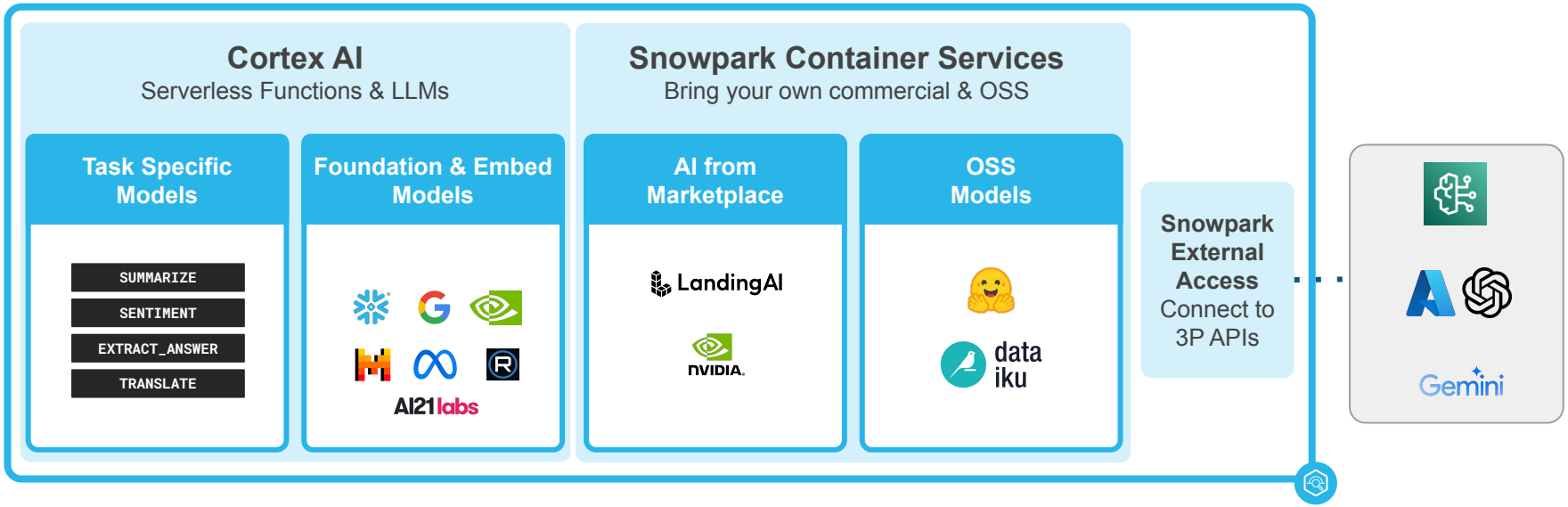
- Customize LLMs securely and effortlessly to increase model accuracy and performance for use-case specific tasks.
- Manage access and governance of custom LLMs with Snowflake Model Registry.

HOW TO USE IT

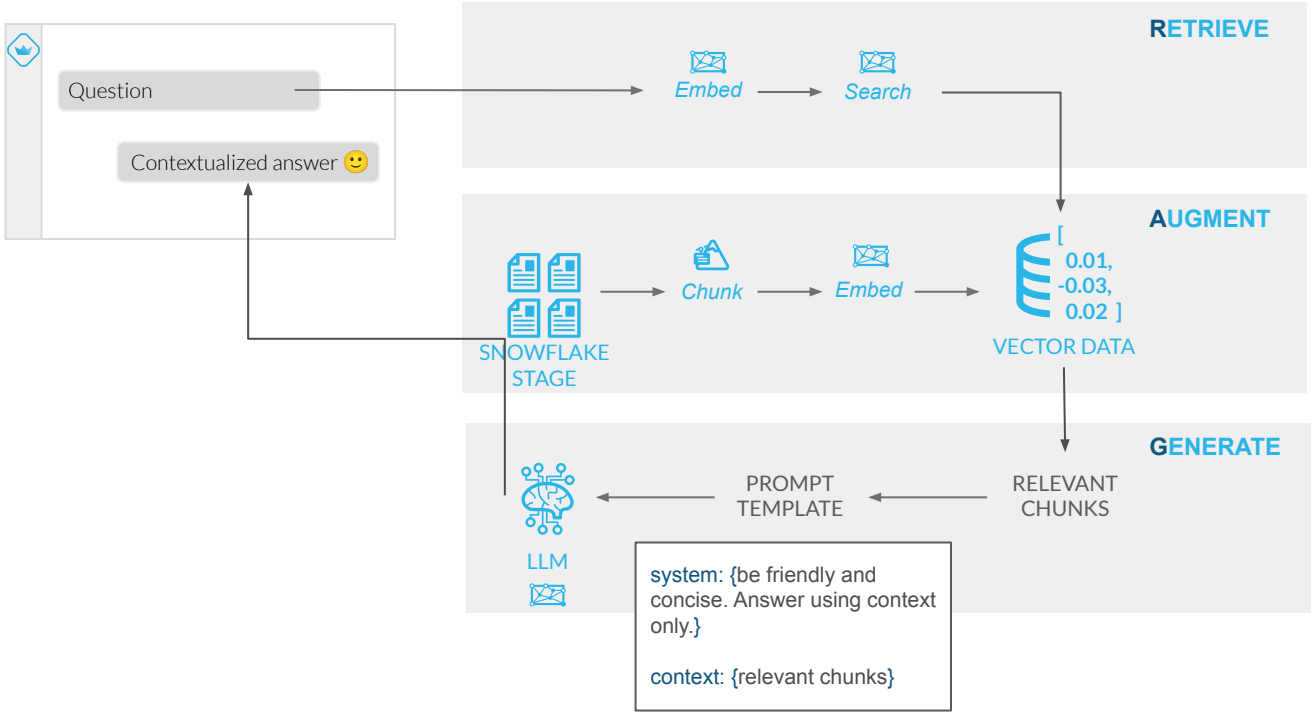
Fine-tune models via a function or directly in the AI/ML Studio. Easily access the fine-tuned models through the COMPLETE function.



Flexibility to adopt wide range of industry-leading AI



WTH is RAG?



STREAMLIT

SNOWFLAKE CORTEX

SNOWPARK

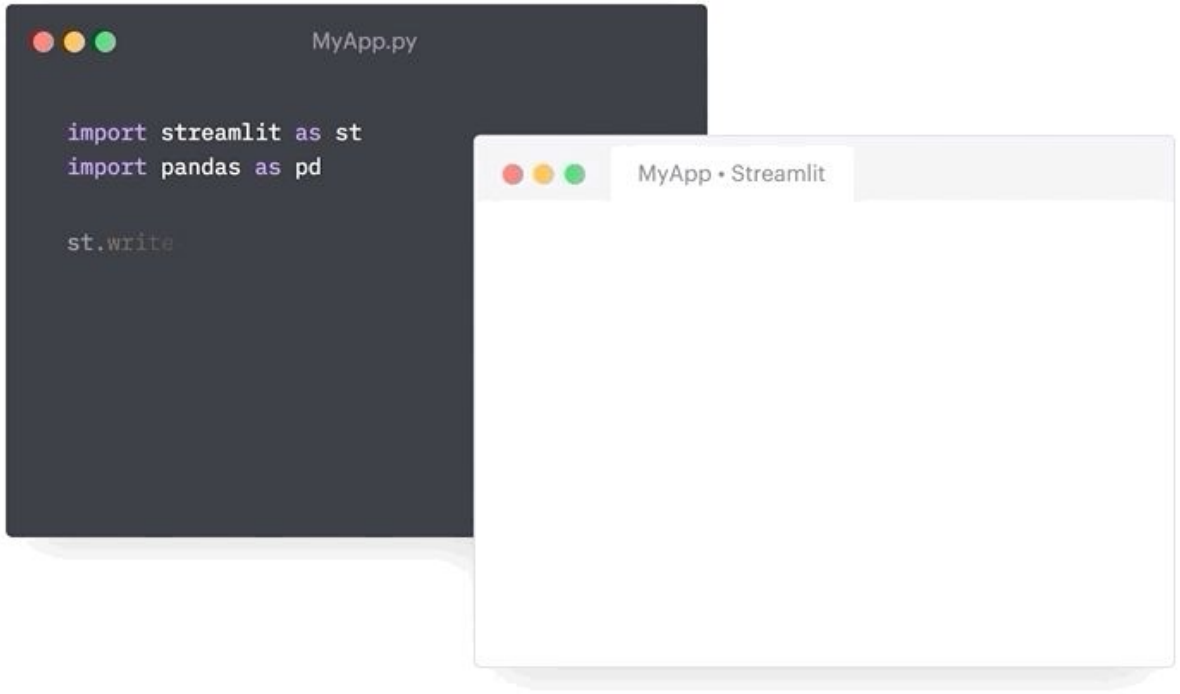
VECTOR DATA TYPE

Knowledgebase chatbot

Find answers in documents or other text data (e.g. wikis) via conversational interface

Build you own document chatbot with this interactive quickstart

WTH is Streamlit?



Streamlit in Snowflake

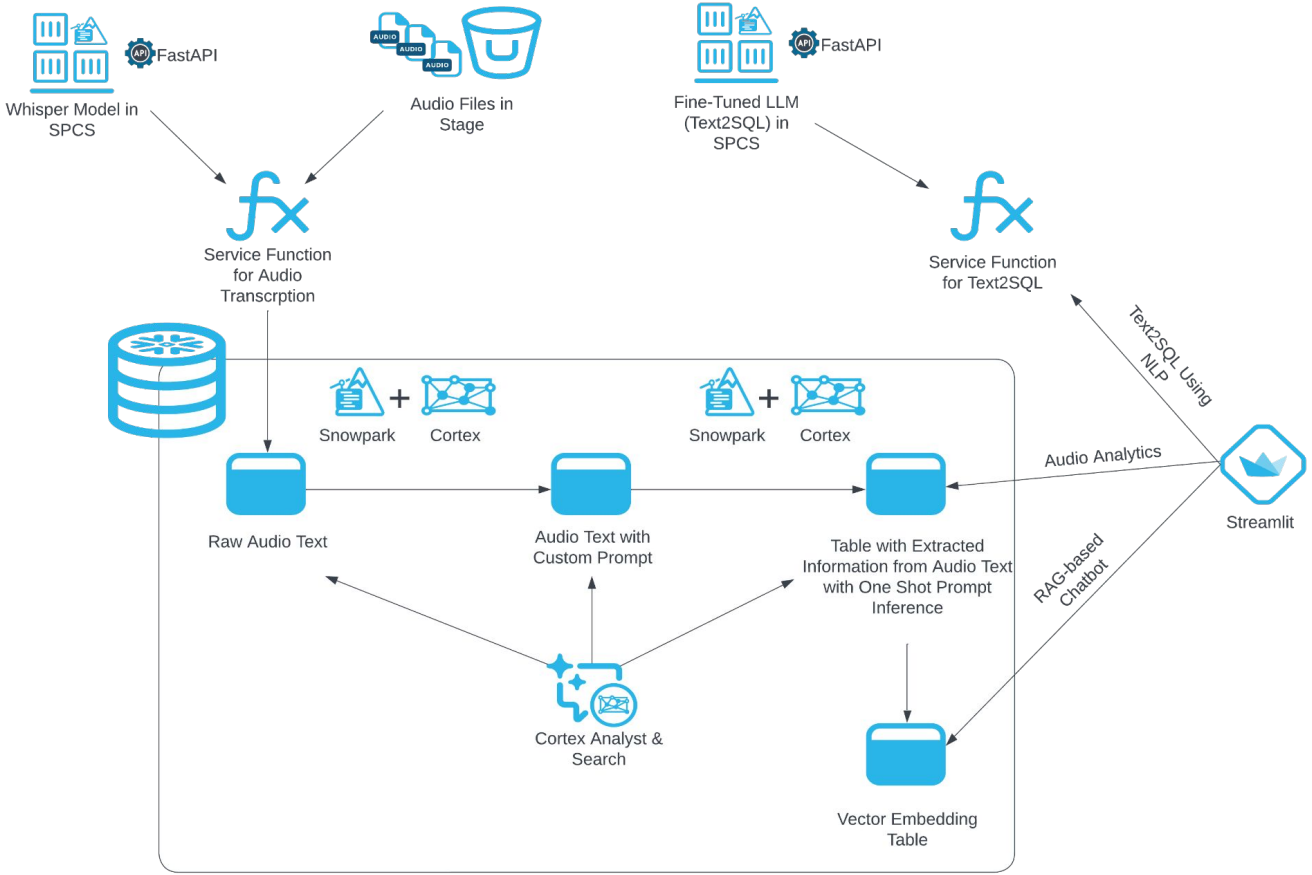
Streamlit

TURN PYTHON SCRIPTS INTO APPS

snowflake

DEPLOY SECURELY AT SCALE WITH ONE CLICK

Target Solution



Feedback & Enter to Win