

AUG 14, 2024

AUTHOR

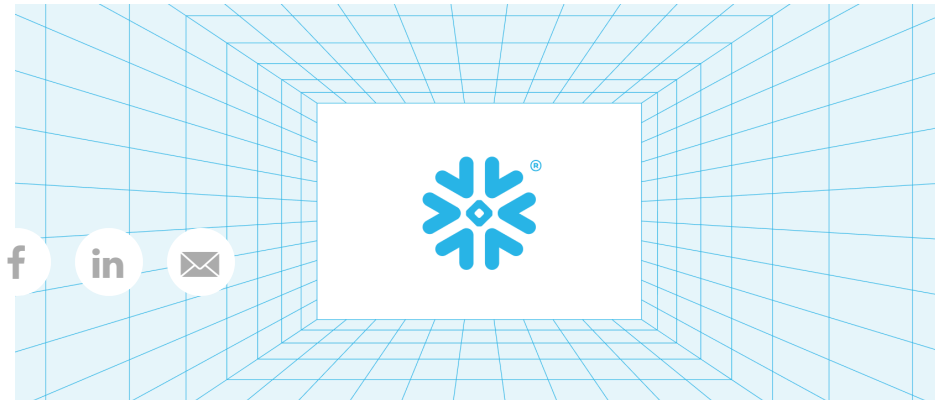


Nipun Sehrawat

Snowflake Cortex Analyst: Behind the Scenes

Gen AI

SHARE



Building a conversational self-service analytics product for business users is a complex and challenging endeavor. Trust and accuracy have to be at the core of such a product, as business users won't rely on a tool that might provide inaccurate insights. Incorrect data insights can lead to severe business ramifications, thus making accuracy paramount. In this post, we present the technical architecture of Snowflake Cortex Analyst™, Snowflake's AI feature that enables business users to ask data questions in natural language and receive reliable answers. We will also walk through some of the technical details that enable Cortex Analyst to achieve 90%+ accuracy on real-world use cases.

Overview

Cortex Analyst is an agentic AI system that uses a collection of state-of-the-art LLMs, including Meta's Llama and Mistral AI models, to reliably answer users' data questions. Answering a question involves a complex workflow where multiple agents interact with each other, with guardrails at every step to prevent hallucinations and to provide highly accurate and trustworthy

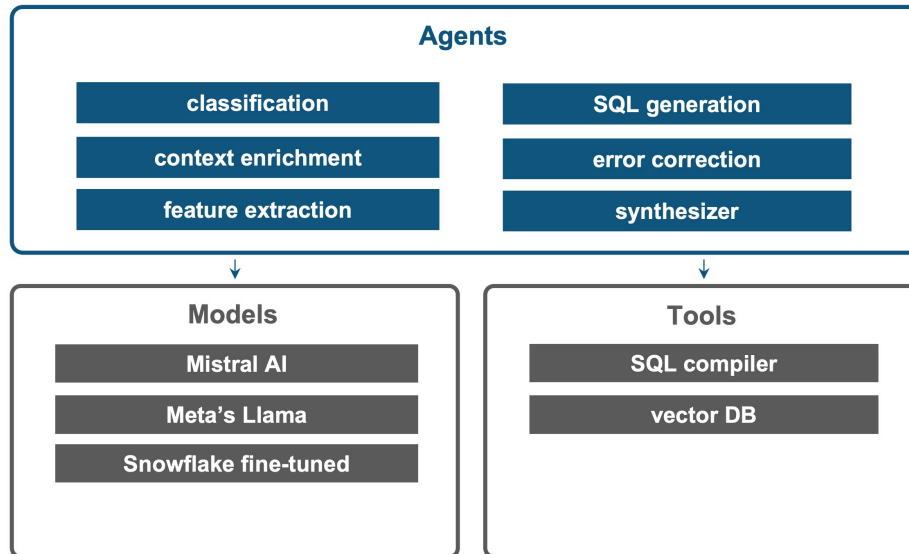
answers. While LLMs are an important component, building a data product that business users can trust requires developing agents with a deep understanding of data analytics and business intelligence.

AUTHOR



Nipun Sehrawat

SHARE



f g+ in K ️ Components of Cortex Analyst

Need for a semantic model

An LLM alone is like a super-smart analyst who is new to your data organization — they lack the semantic context about the nuances of your data. Given only the raw schema, it's challenging for any analyst to write SQL accurately to answer data questions. This is because raw schemas are often messy and lack the semantic information needed for accurate data analysis.

Moreover, there's typically a gap between the vocabulary used in business users' questions and the database schema. Business users' vocabulary aligns more with business terms, while the database schema vocabulary is closer to the ETL pipelines. This gap makes it difficult to build a product that answers data questions with high accuracy. In addition, data teams tend to be comprehensive — a raw table may include multiple versions of "sales" number, while you only want to present one version to your end business users.

Cortex Analyst introduces the concept of a semantic model as a way to capture and provide the missing semantic information that LLMs need to correctly answer user questions — similar to what a

human analyst would require. With semantic models, data teams can:

1. Capture and provide semantic information about the schema through more descriptive names, synonyms, freeform descriptions about tables and columns, and instructions on how best to utilize them.
2. Perform data modeling by exposing only the relevant columns, defining common metrics, filtering conditions, etc.

AUTHOR



Nipun Sehrawat

SHARE

By leveraging semantic models, Cortex Analyst achieves more than 90%+ SQL accuracy on real-world use cases, while allowing enterprises to customize their Cortex Analyst experience according to their unique requirements. However, from a technical perspective, working with a semantic model makes the task significantly more challenging due to its complexity and the demands for precise instruction-following, as compared to a typical text-to-SQL task on raw schemas. This is because all the information and instructions from the semantic model must be used correctly to answer users' questions accurately.

In the following section, we will delve into the details of how Cortex Analyst is architected to excel at answering questions based on semantic models.

Agentic workflow for answering a question

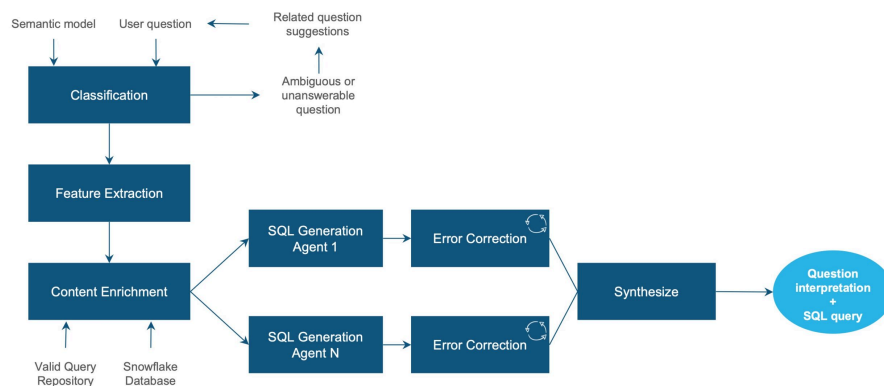


Figure 2. Workflow for answering a question.

We will now describe the key components of the above architecture in detail.

Classification agent

User questions can sometimes be vague and difficult to answer reliably. For example, a question like, “What was the best product last year?” can be ambiguous if the semantic model contains multiple metrics that can be used to compute the “best” product. To avoid hallucinations and ensure a more trustworthy experience for business users, we believe that it’s crucial to reject such questions upfront, rather than responding with potentially misleading answers.

AUTHOR



Nipun Sehrawat

The classification agent categorizes incoming questions into classes, such as *ambiguous*, *non-data question*, *non-SQL data question*, etc. It only answers questions that are unambiguous and can be answered using SQL. Other classes of questions are rejected, and the user is presented with a list of similar questions that can be answered confidently. This prevents users from f tti in :u d enables them to continue getting reliable answers to their data questions.

SHARE

Feature extraction agent

After classification, a feature extraction agent analyzes the question to understand its specific features. For instance, is it a time-series question? Does it ask for a period-over-period calculation? Does it involve a rank calculation? Leveraging Snowflake’s extensive experience in running data analytics workloads, Cortex Analyst has a vast collection of features to choose from.

Answering different types of questions requires different skills, and thus the extracted features influence the behavior of downstream agents that generate SQL. For example, the downstream agents craft tailor-made prompts with instructions most suitable for answering questions with the extracted features.

Context enrichment agent

Next, a context enrichment agent processes the semantic model with additional context relevant for answering the question. Such enrichment is crucial to answering business users’ questions, which sometimes lack necessary context. The following are two key types of context retrieved:

AUTHOR



Nipun Sehrawat

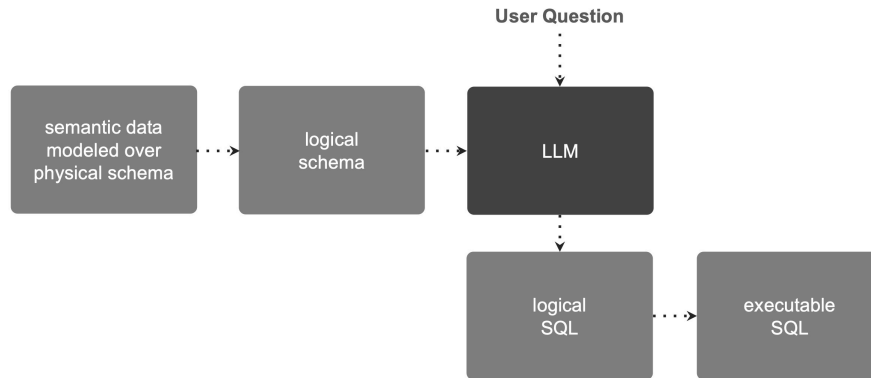
SHARE

1. **Relevant verified queries:** To further enhance Cortex Analyst's trustworthiness, data admins can specify *verified queries* as part of the semantic model setup. These are queries that have been verified as correct by human analysts. When answering a new question, Cortex Analyst retrieves semantically similar verified queries and uses them to generate trustworthy answers. Verified queries bring human analysts into the loop, which is an important pillar for building a reliable and trustworthy product for business users.
2. **Relevant literals:** There's often a vocabulary gap between business users and the data values stored in a database. For example, a user might ask for "month over month profit growth in USA," but the actual value stored in the database might be "United States of America". In fact, incorrect literal generation is a common failure mode for text-to-SQL products. Cortex Analyst performs a semantic search to retrieve relevant literals for answering the user's question, greatly improving the accuracy in real world use cases.

SQL generation agents

The enriched context is then passed to a collection of SQL generation agents, each utilizing a different LLM. In our experience, different LLMs excel at answering different types of questions. For instance, some handle time-related concepts better, while others are more effective at multi-level aggregations. Hence, using multiple LLMs makes the overall query-generation process more robust and accurate.

As mentioned above, the SQL generation agents use the features extracted from the question to generate prompts tailored to answering the specific question at hand. These prompts contain a smaller and more specific set of instructions, thus reducing the chances of the LLM forgetting some of the instructions¹.



AUTHOR



Nipun Sehrawat

Figure 3. Two-step SQL generation approach.

Another key challenge with SQL generation is that LLMs struggle with complex schemas. To address this, the SQL generation agents use a two-step process:

- 1. Logical Schema Construction:** The agents first construct a logical schema over the underlying physical schema and ask the LLM to generate SQL against this logical schema. This simplification makes the task easier for the model.
- 2. Post-Processing:** The agents then post-process the generated SQL to make it executable on the underlying physical schema.

This two-step approach boosts SQL generation accuracy by hiding schema complexity from the LLMs, thereby enhancing their SQL generation performance.

Error correction agent

No matter how powerful, LLMs do make mistakes. However, they are also good at correcting their errors when pointed out. The error correction agent takes the generated SQL and checks for both syntactic and semantic errors by utilizing core Snowflake services, such as the SQL compiler. If any errors are found, the agent runs an error correction loop to have the LLM fix them. This error correction module also addresses hallucinations, where the model might invent nonexistent entities or SQL functions.

Synthesizer agent

Once all the generated SQLs have been checked and corrected, they are forwarded to a synthesizer agent. The synthesizer agent,

SHARE



akin to an expert data analyst, receives multiple candidate SQL queries and additional context, such as relevant literals and verified queries. Leveraging the work done by the previous agents, the synthesizer agent generates a final SQL query that accurately answers the question at hand.

Benchmarks

AUTHOR



Nipun Sehrawat

Benchmarking has been a key focus of our team since the early days. We have a comprehensive internal benchmark suite that is representative of real-world, business intelligence-style questions that we expect our users to ask.

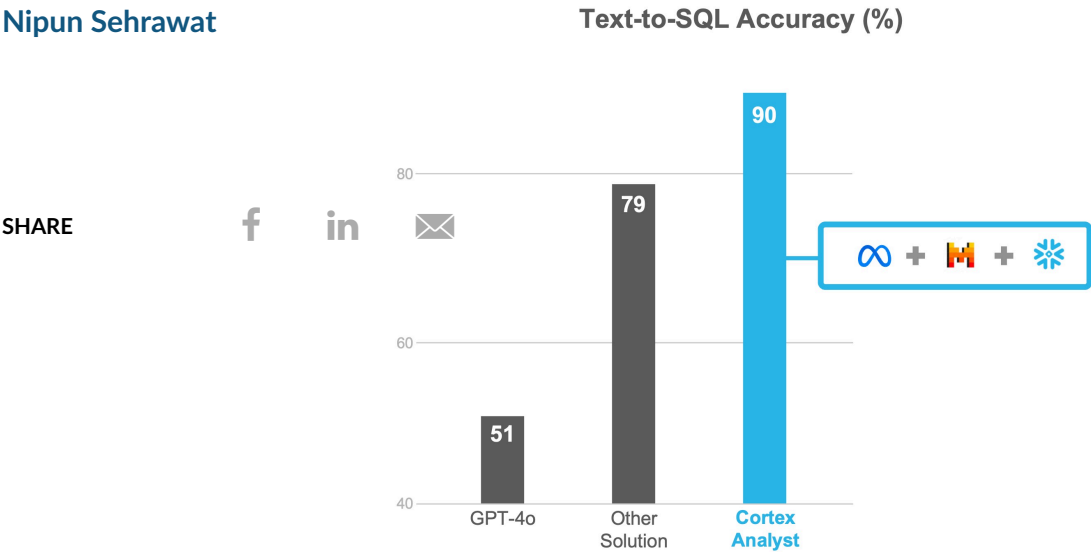


Figure 4. SQL generation accuracy for Cortex Analyst vs. alternatives

Our benchmarking results show that Cortex Analyst is close to 2x more accurate than single-shot SQL generation using a state of the art LLM, like GPT-4o, and delivers roughly 14% higher accuracy than another text-to-SQL solution in the market. Stay tuned for an upcoming engineering blog post, where we will further delve into the benchmark details and results.

Conclusion

Given the vast productivity gains it could provide for business users, a conversational self-service analytics product holds immense potential in an enterprise setting — but only if it is accurate and trustworthy. Faulty data insights can have

debilitating effects, so it has been our top priority to deliver a product that lives up to that potential.

Cortex Analyst leverages a collection of AI agents, built with deep understanding of data analytics and business intelligence, to mimic a human analyst and deliver trustworthy responses with an extraordinary SQL accuracy of over 90%. We are excited to see you put Cortex Analyst in the hands of business users at your company.

AUTHOR



Nipun Sehrawat

SHARE

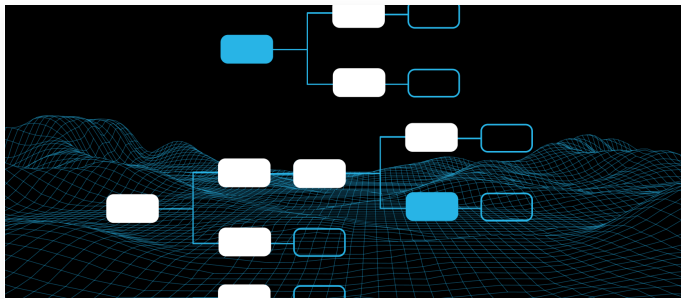


¹ <https://arxiv.org/abs/2307.03172>

SHARE



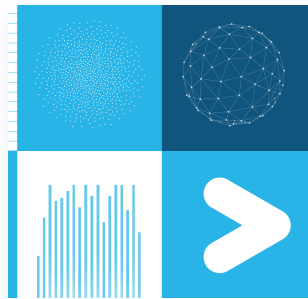
RELATED CONTENT



AUG 08, 2024

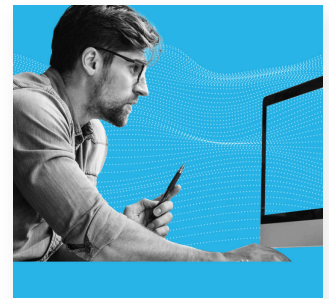
Snowflake Cortex Search: High-Quality, Performant Search and Retrieval for Enterprise AI

Search and retrieval systems have always been a critical backbone for knowledge management in enterprises. These systems cater to use cases ranging from site search, product catalog or feed search,...



JUL 23, 2024


Achieve Low-Latency and High-Throughput Inference with Meta's Llama 3.1 405B using Snowflake's



JUN 17, 2024

Snowflake Launches the World's Best Practical Text-Embedding Model for Retrieval Use Cases

[Delve into the details](#)



Nipun Sehrawat

Optimized AI Stack

Meta's Llama 3.1 405B represents a groundbreaking milestone for open-weight large language models (LLMs), pushing...

[Have a look](#)

Today Snowflake is launching and open-sourcing with an Apache 2.0 license the Snowflake Arctic embed...

[Full Details](#)

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		Webinars	Blog	Investor Relations
Marketplace	Snowflake for Advertising, Media, & Entertainment	Documentation	Trending	Leadership & Board
Security & Trust	Snowflake for Retail & CPG	Community	Guides	
	Healthcare & Life Sciences Data Cloud	Procurement	Developers	Snowflake Ventures
	Snowflake for Marketing Analytics	Legal		Careers
				Contact

Sign up for
Snowflake
Communications

diana.shaw@sn 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.

AUTHOR



SUBSCRIBE NOW

Nipun Sehrawat [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](#)



SHARE

