VIANAI

FROM HYPE TO REALITY -

Generative AI in the enterprise

The following white paper contains Vianai Confidential and Proprietary Information and references Vianai intellectual property rights subject to legal protection under applicable U.S. and international laws.

TABLE OF CONTENTS

Hype To Reality

We all know the hype. Bloomberg claims there is a $1.3 trillion market opportunity for generative AI by 2032. ChatGPT saw 1.7 billion visits in November of 2023. ChatGPT exceeded 100 million users in two months — a number that took the internet seven years to hit.

Regardless of the figure, generative AI has become the most hyped technology in decades. Yet, the many ways that this will play out inside of an enterprise are still being determined. It's clear:

Enterprises are still working to understand the technology — does it even work?

Next, they want to understand the use cases that provide value, not just the "Co-pilots" that accompany software — how can it meaningfully improve the work of a business user?

Finally, specific to their environments — what do we have to do to make it work? How do we ensure that it meets our security, privacy and other standards?

Taking generative AI from hype to reality requires several core capabilities that Vianai has built around the LLMs — the guardrails, additional feature sets and cross-checks — that enable reliable, consistent responses.

In this white paper, we will detail these core capabilities in our generative AI platform hila and our application Conversational Finance. Specifically, how these components enable enterprises to make generative AI relevant and useful.

VIANAI

A foundation for generative AI in the enterprise — hila Platform

The hila Platform enables reliably building and running domain-specific generative AI applications on structured data in an enterprise.

It offers the following key capabilities to enable enterprise success with generative AI:

- NL2SQL
- Anti-hallucination
- Custom and local LLMs
- Agentic orchestration
- Flexibility and extensibility
- LLMOps and monitoring
- Enterprise-grade security

NL2Code and anti-hallucination for structured data

Performance: It often takes a long time to receive an answer from the LLM.

Reliability: Responses can be inconsistent or incorrect, even the same question can sometimes yield different answers.

The most performant models and papers in Yale's Spider Leaderboard for NL2SQL have a credible 91.2 percent accuracy, but that remains too low for an enterprise. Indeed, using a combination of models and chain-of-thought reasoning, the next closest technique from the Alibaba Group achieved only 86.6 percent accuracy. Further (see details below) this performance drops even further when domain specific (e.g. finance) language or "jargon" is used, or company specific information (e.g. chart of accounts, fiscal years, etc.) is referenced in the natural language question.

For an enterprise to trust a system, it needs to be significantly more accurate and consistent. We have achieved this through a combination of techniques. To start, we process the question to see that it can be answered by the available dataset. We then cross-reference the query against a vast database of knowledge that is both domain-specific (such as finance-specific for an ERP system) and company-specific (such as which months constitute a fiscal year for a particular company). The response then gets post-processing after the LLM returns the SQL to ensure that the SQL answers the initial query. These various guardrails around any model, public, private or fine-tuned, ensure we raise its SQL generation accuracy to a level acceptable for enterprise use cases.

The key to our process is a robust set of knowledge that guides the LLM in making the correct calls. This knowledge is based on both domain-specific understanding and customer-specific data. Both contribute to the synthetic data generation that constitutes the backbone of our system.

ENSURING CORRECT RESPONSES

There has been a lot of emphasis and focus on various LLM models hallucinating, or confidently providing an answer that is not accurate. While this remains true, in our testing, several models, such as GPT4, categorically fail to provide any answer at all. In particular, GPT4, and many other SOTA in natural language to SQL generation models, often do not provide any response for several types of questions, such as those with domain-specific vocabulary, and for complex questions that may result in relatively complex SQL.

In a recent paper from Postech and Google, which assessed the current state of NL2SQL, “Natural Language to SQL: Where are we today?” came to the same assessment, “The existing [NL2SQL] techniques are still at a basic level and need to be significantly enhanced… Real-world databases may have a lot of rare words which cannot be found in the pre- trained word embedding… Those become even worse on complex queries.”

In our own testing, we found that consistently, GPT4 refused to provide an answer. We’ve provided the same data schema and information to both GPT4 and to our own hila Conversational Finance application. The results are clear — the techniques implemented in hila Conversational Finance (on top of the LLM) are necessary, given the current maturity of even the best LLMs. And our techniques not only ensure answers, but those answers are also reliable.

We've also broken out "simple" queries, such as "Total amount by fiscal year between 2019 and 2024 for category Sales Revenue," with complex questions, such as, "What was the total amount for the account category Gains Price Difference on each of the underlying general ledger accounts in fiscal year 2024?"

	Simple	Complex	Total
Answered	92%	26%	59%
No response	8%	74%	41%

It's important to note that this does not include times when the model returned the wrong answer, which does happen, especially for complex questions.

Our data set used for the above evaluation is specific to financial systems. This also is the target for hila Conversational Finance, and the key to the domain knowledge that we've placed in the system. This domain knowledge aids in overcoming the barrier the Postech and Google paper points out — the "rare words that cannot be found in the pre-trained word embedding."

This domain knowledge also enables us to handle many questions in our dataset that GPT4 fails on, such as ratios, percentages, CAGR formulas, year-end and quarter contexts, and more.

GPT4, as all advanced LLMs, continues to face challenges at overcoming the last-mile aspects of SQL generation. With the hila Platform, and hila Conversational Finance, we've done this work for our customers, so they can get started with useful generative AI right away.

FLEXIBILITY AND EXTENSIBILITY

We provide a system that can flexibly work across structured data, enabling users to ask questions in natural language on their diverse enterprise data via a single, unified user interface. Whether users are asking questions of data that resides in systems of record, spreadsheets, PDFs, or simple text, hila can respond.

Subsequently, applications like Conversational Finance can be built on top of this platform by adding GenAI and other content – not code.

This is further enhanced by customization on user-preferences and customer-specific data. This customization extends across all dimensions specific to an enterprise, such as the style of questions asked, the jargon used, and the knowledge specific to the company, such as the fiscal year. For example, one dataset may consider "profit centers" a core part of the business, while others may call them business units or otherwise. In that case, the system can interpret the following question to maintain reliable responses based on the company-specific knowledge.

What was the total amount for the account category Interest Expense in fiscal year 2024 by Profit Center

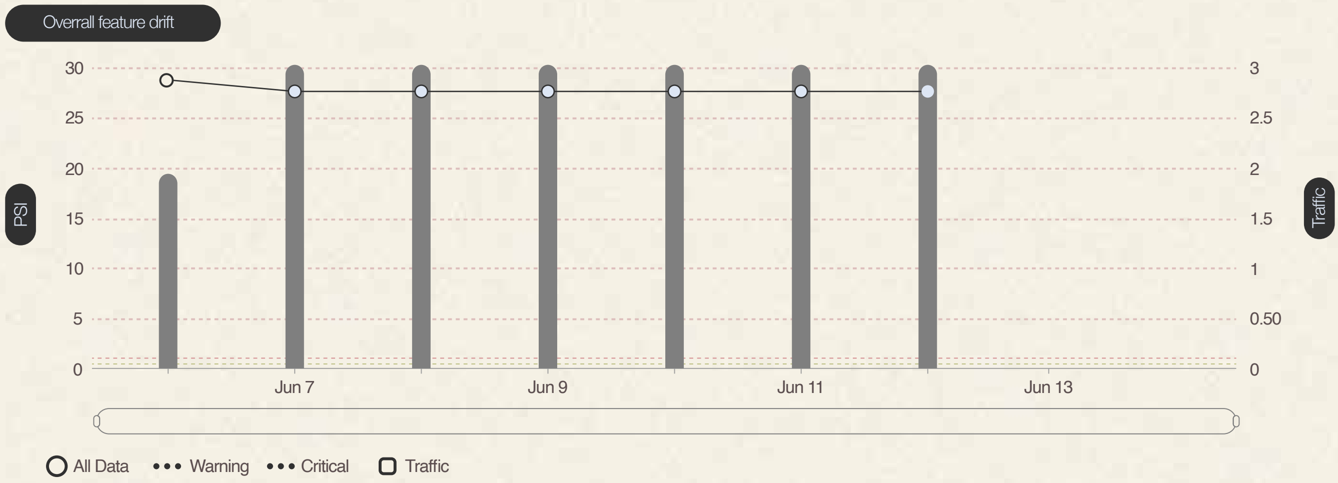
The SQL this question generates changes based on the knowledge of the company. For example, if the company's fiscal year runs from April 2023 to March 2024, the hila Platform will automatically generate code to get results for that timeframe.

LLMOps and Monitoring

Enterprises need tools to monitor the cost, quality, variability and reliability of their LLMs, so that they can have measurable, trackable returns on their investment.

The monitoring capabilities inside of hila track critical risks that jeopardize a model's reliability and performance including monitoring feature and prediction drift, data quality, and more. This monitoring can occur over tens of thousands of predictions per second, hundreds of features and segments, and sub-segments, with millions or billions of transactions, across multiple time windows, to find and solve problems degrading model performance.

These monitoring capabilities utilize prediction drift and feature drift to indicate model performance when the ground truth is not available, it will also utilize flexible feature drift to determine segment-level data quality issues.

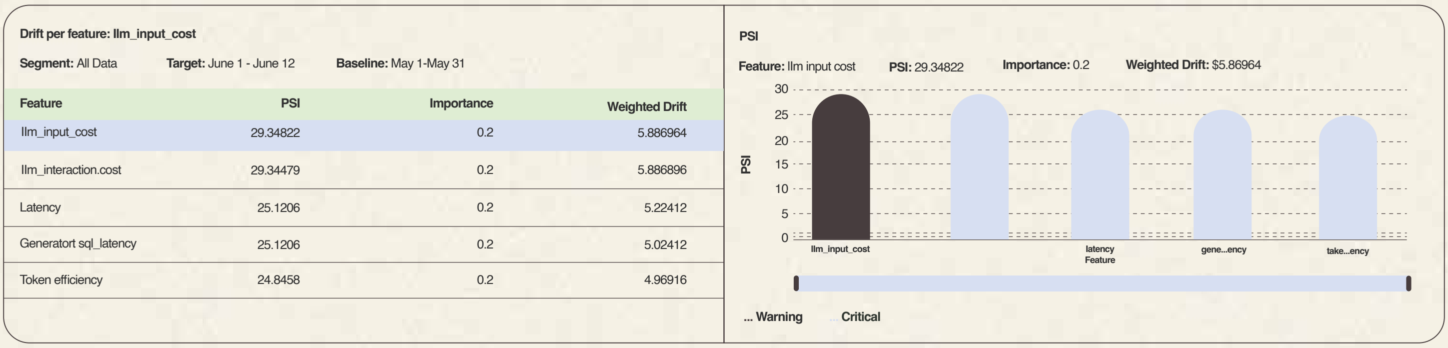


This shows the overall change for the various metrics we're monitoring, including cost, and latency

For example, should the number of users in a system remain the same, but the cost increases, the reason could be an increased number of tokens sent to the model. Put differently, users could be asking far longer or more complex questions.

There are other items that impact cost, of course, such as the type of LLM used (GPT3.5 costs approximately 10X less than GPT4, for example), number of queries, the number of users and the model's efficiency (the length of the answer compared with the length of the question).

In addition to these various aspects of an LLM, hila Monitoring tracks the time that it takes for an LLM to respond, the relevancy of an answer to the original question, and the quality of the SQL generated.



Drift per feature: llm_input_cost

Segment: All Data Target: June 1 - June 12 Baseline: May 1-May 31

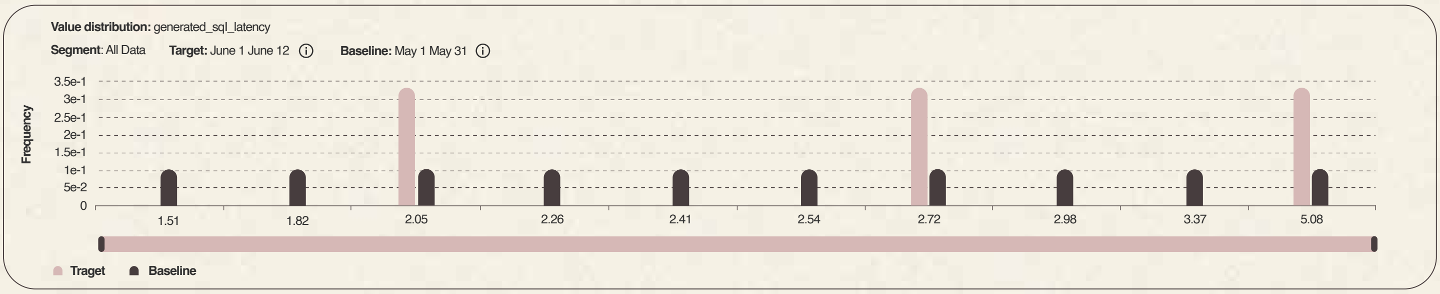
Feature	PSI	Importance	Weighted Drift
llm_input_cost	29.34822	0.2	5.886964
llm_interaction.cost	29.34479	0.2	5.886896
Latency	25.1206	0.2	5.22412
Generatort sql_latency	25.1206	0.2	5.02412
Token efficiency	24.8458	0.2	4.96916

PSI

Feature: llm input cost PSI: 29.34822 Importance: 0.2 Weighted Drift: $5.86964

This is a breakdown of the metrics we are monitoring. The metrics include many dimensions of cost (such as input and interaction cost) and latency (both for the overall response and for the generated SQL).

Should there appear a major change in any of the metrics tracked, hila Monitoring provides compelling visualizations that show context around the shift.



Value distribution: generated_sql_latency

Segment: All Data Target: June 1 June 12 Baseline: May 1 May 31

A closer look at one of our monitored metrics, generated SQL latency, we can see the discrepancy between the target date (month of June) and the baseline date (month of May).

Administrators also receive alerts if costs exceed preset boundaries, the quality drops below a set threshold, or the latency rises above average.

ENTERPRISE-GRADE SECURITY

Data security and privacy is a primary component of hila and our application Conversational Finance. There are several steps that we take at the application level, the platform level and the company level to safeguard our customers' data. And, should the customer want the ultimate level of security, we can deploy entirely on-premises and air-gap our solution from the internet.

DATA SECURITY IN A SYSTEM OF RECORD

Conversational Finance, our application built on the hila Platform, enables users to ask complex, natural-language queries against their systems of record. This data, be it ERP data or otherwise, is often some of the most crucial and protected data inside of the company.

We take significant steps to protect the data from any LLM. Our processes protect the data from public models without inhibiting the performance of Conversational Finance.

Therefore, we have the same processes when the application uses an open-source model for code generation.

When a customer uses a public LLM, only the user question is sent, so that public LLM can generate the code. None of the data from the system of record that accompanies the question goes to the LLM. If we're using a local LLM, then all data, including the question, remains inside of the customer's firewall.

In addition, we do not extract data from the deployment. This includes user data, such as their questions and responses. We do train models and add to our customer's specific knowledge base, but this training is specific to our customer and not shared with other customers.

VIANAI

hila Platform Security

The hila Platform also provides a simple way for administrators to assign roles and data access controls. This enables the limitation of data at the row level, which ensures that the information is not exposed to anyone inside of the company who should not see it.

These monitoring capabilities utilize prediction drift and feature drift to indicate model performance when the ground truth is not available, it will also utilize flexible feature drift to determine segment-level data quality issues.

Vianai Security

At Vianai, we have ISO 27001 and SOC2 certifications. These are important, independent verifications that validate that our company holds itself to the highest standards of cyber security. It also shows that our claims on data security are accurate.

VIANAI

hila Platform capabilities in action with Conversational Finance

Weaving together the various core components of the hila Platform into a simple user interface provides rich value for business users. Conversational Finance has these various components available in an application that can provide real value in weeks.

For example, Conversational Finance provides a method to compare different attributes of revenues and expenses, such as by product, by customer, by profit center, by time, and by GL account, and finance users can quickly drill into the change for each, comparing revenue performance or expense actuals over time to understand the health of a business.

There are a series of features to further enhance the broader use and simplicity of outputs in Conversational Finance, including for management meetings, board meetings or other stakeholder reviews that drive decision-making. These include greater explanations around the table displayed in the result, additional explanations about the query, and various features, such as showing the state of the answer and error identification. That means it’s not just an output that a finance user can understand but business and other finance stakeholders as well.

Important links:

Visit website:

https://www.vian.ai/hila-platform

See Conversational Finance in action

https://youtu.be/eqVD_REIFnw

Get in touch

https://www.vian.ai/contact-us

See hila against GPT-4 wrappers:

https://youtu.be/s8ChqjIr1m8

https://youtu.be/G3F_FNQaeZY