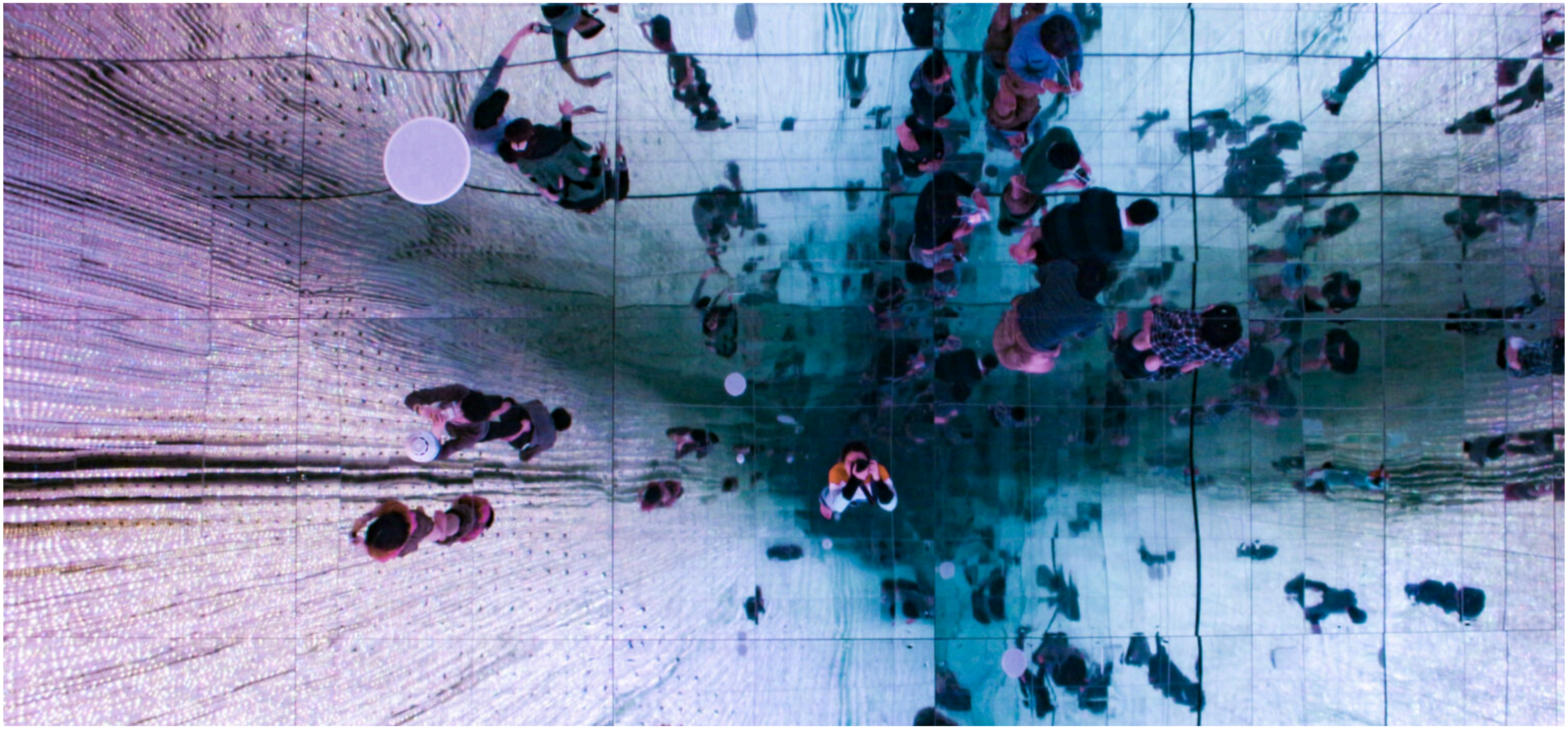
Ines Chami

Research

Three Lessons to Bring Generative AI into Enterprise Production Systems

Despite the hype around LLMs and their potential, enterprises struggle to get these models in production. In this blog, we outline the challenges presented in today's environment and share key pillars to enable enterprises to successfully bring generative AI into production to supercharge analytics.



Introduction

Numbers Station is building the next generation of self-service analytics by providing a natural language interface to data. This enables non-technical business users to chat with their data and get answers to their ad-hoc questions, without having to wait weeks or months for someone

You May Also Like

Product

applications and to avoid being left behind. However, after all the ChatGPT hype earlier this year, people quickly started realizing the challenges of going beyond demos to bring this amazing technology into production. In this blogpost, we'll share our learnings building an enterprise vertically integrated AI solution for self-service analytics.

LLMs are the future of self-service analytics

The goal of self-service analytics is to start from a business question (e.g. what was our sales funnel conversion in Q3 2023 by state?) and produce an answer. The answer can take various forms such as a chart or a table (e.g. csv), but we will abstract it for now and only consider the SQL that produces the final results. This problem has been long standing and various actors have tried solutions such as templated query languages, visual interfaces etc. However, none of these existing solutions successfully made it into production, they were too brittle and not powerful enough from a technical standpoint.

With the LLM revolution and their ability to generate SQL code, it became clearer that the technology to solve this problem was there. With the release of ChatGPT, it became extremely easy to build prototypes and try out ideas, and many builders decided to leverage LLMs like to build products for self-service analytics. We've seen a lot of

Research

DuckDB-NSQL: How to Quack in SQL

Research

Text-to-SQL That Isn't

some of our key learnings.

Limitations of API-powered demos

Before we dive into the Numbers Station solution, let's talk about why API powered demos have failed on enterprise use cases.

1. Lack of ownership

Enterprises that want to use generative AI in their pipeline face some major challenges when it comes to using black-box third party APIs because the LLMs are not owned by the enterprise using them. These models are typically hosted on third party cloud services which means that for any request, data has to travel outside of the enterprises' firewalls. While this might be acceptable for some enterprises that deal with non-sensitive data, it causes privacy and security concerns for enterprises that manage sensitive customer data (e.g. SSN numbers). In analytics in particular, the majority of the data that analysts deal with is internal data that is usually confidential. Additionally, third party models are subject to change which removes some control over the output of the model and introduces risks for models already deployed in

2. Lack of accuracy

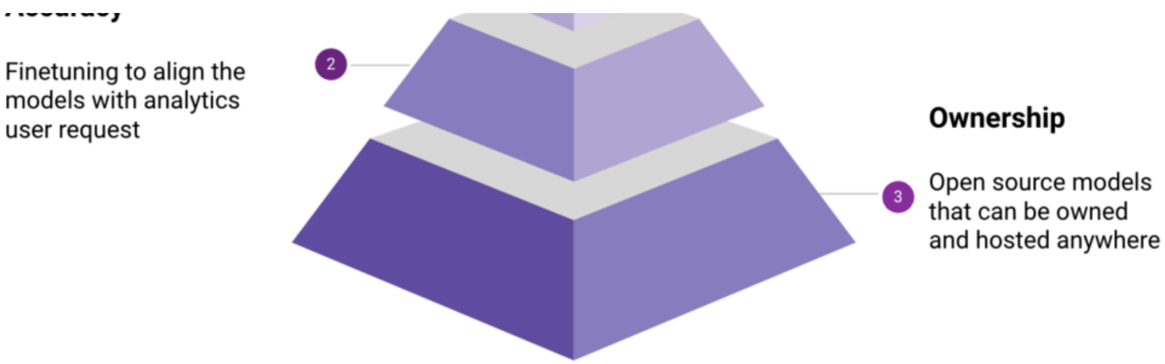
One important limitation of using off-the-shelf (closed and open) LLMs is that their quality is far from perfect. These models are usually general purpose models trained on publicly available data and while some models yield incredible performance on public benchmarks, performance can drastically drop on enterprise specific tasks because the models have not been trained to align with enough instructions from these tasks. For instance, while many models have some knowledge of SQL (SQL is available online), they can perform poorly at the task of transforming natural language questions to SQL. They can hallucinate columns, generate invalid code, or even worse, generate valid code that does not align with the input questions. Low accuracy and brittleness in the results prevent enterprises from deploying these models into production.

3. Lack of personalization

Third party LLMs are amazing at capturing public knowledge found on the internet. However, when it comes to business intelligence, these models do not have the necessary contextual knowledge to generate “number of active users” might produce a technically accurate answer (e.g. “SELECT COUNT(*) FROM USER WHERE LOGGED_IN_THE_LAST_30_DAYS=TRUE”) but that answer is wrong in the context of a specific organization where active users are defined as users that sent a message in the last 10 days. These business-specific definitions are quite complex and nuanced, but without capturing them, there is no hope to bring this technology in the hands of non-technical users.

The Numbers Station solution: Your Models, Your Tasks, Your Knowledge

At Numbers Station, our core technology is based on three pillars that drive our innovation for the next generation of analytics: Ownership, Accuracy and Personalization. We believe that bringing AI in enterprise production workflows can be done leveraging the following solutions.


2
Finetuning to align the models with analytics user request
Ownership
3
Open source models that can be owned and hosted anywhere

1. The solution to the lack of ownership: Open source models

At Numbers Station, we are firm believers that enterprises should own their AI models. This all starts with the foundation of building with open source models. We build our core product by leveraging some of the amazing open LLMs (e.g. Llama from Meta) which serve as the starting point to the Numbers Station technology. Leveraging open source models (as opposed to closed source models) not only solves the aforementioned challenges around data privacy (open models can be self-hosted on VPC or on prem), this also enables enterprises to better control their models output and avoid surprises that may arise with third-party owned models when the underlying weights change. By owning the models and their weights, enterprises can now securely deploy these models in their organizations without incurring risks for sensitive data breaches.

in the wild. Additional finetuning to better align these models with instructions from these tasks is key to get enterprise-grade accuracy. At Numbers Station, we developed a proprietary finetuning platform that takes in any open model and finetunes them to better follow instructions for various enterprise tasks such as converting business questions to SQL queries.

In particular, it has been shown that on complex tasks, it is possible to achieve higher quality by helping the model think step-by-step. We proposed to break down the highly complex and open-ended task of generating SQL queries into multiple, simpler subtasks and finetuned models to align with instructions from each of these simpler tasks. Our research team found that this yields significantly higher performance than simply finetuning for the text to SQL tasks.

3. The solution to the lack of personalization: Highly curated semantic catalogs integrated with RAG

The last pillar that powers the Numbers Station technology is a semantic catalog that acts as an interface between the LLM and the private knowledge of an organization. As discussed earlier, models trained on public knowledge lack some enterprise-specific knowledge and context that is crucial to generating accurate and trustworthy answers for

However, a big challenge to get this architecture to work is creating that knowledge source. For applications of LLMs on top of unstructured data corpurses, then traditional retrievers (e.g. text encoders) can suffice to retrieve that knowledge without modifying the text corpus. However, for analytics applications, or any application in the modern data stack, that knowledge needs to be highly structured and relational. For instance, taking the “active user” example from above, the textual definition of an active user is not enough for a model to get a query right. The model needs to know which table to use, which columns to look at etc.

Similar to semantic layers (e.g. dbt), our knowledge layer stores information about common metrics, dimensions, filters that business users need to answer their questions. This layer can be manually created, or enriched with weak signals extracted from a variety of information stored in enterprise query logs, dashboards, data catalogs etc. In particular, our research team developed technology that can automatically parse Tableau dashboards and extract contextual enterprise knowledge from these dashboards. This can help enterprises accelerate their journey to create a semantic layer and democratization access to self service analytics.

While the barriers to getting LLMs into production are present in today’s environment, Numbers Station is here to overcome these challenges and empower your analytics journey with generative AI. To learn more about

About Us Blog See a Demo

TAKE THE NEXT STEP

Conversational analytics is just a question away.

What's your work email? Get Started

YOUR AI FOUNDATION
FOR DATA ANALYTICS

About Us

Ready to learn more?

Get started today with Numbers Station Cloud or Enterprise.

See a Demo

Join our newsletter.

Keep updated, keep learning, keep in touch.

What’s your work email? Join

Terms of Use Privacy Policy