JUNE 14, 2023

Improvising with AI: The Jazz of Data Democratization

By Mark Grover

CO-FOUNDER, CEO OF STEMMA

AI will help data teams and data users perform like a tight ensemble (image generated using Midjourney)

SHARE THIS ARTICLE

I've been in this industry long enough to see Big Data, cloud, and now AI, and I am an AI optimist. I think AI will help us achieve some of the goals with data that we have been working towards for years but have always seemed just out of reach. Here I will describe three of them, talk about why AI will usher in change, and how AI and related products need to evolve to enable those changes, based on my experience building data products at Stemma and beyond including our recently released Discover Assistant powered by GPT.

Chat will become the dominant interface for business user analytics

Every week, another two solutions pop up that allow a user to ask a question where ChatGPT translates a natural language query to SQL and returns an answer.

It makes sense. The basic chat template has become ubiquitous due to the popularity of messaging apps. Combine that familiarity with ChatGPTs ability to power responses within the interface and there appears to be a clear path for LLMs to reduce friction between end-users and data. No need to learn SQL, or understand the organization-specific context about what data is available and trustworthy, or even care if you have access.

What we still need to work on:

We still need to show that AI can provide a trustworthy answer more often than not. Typically the conversation here turns to the hallucination problem but there are other even more persistent problems in data. Even today, translating the natural language question to SQL is relatively easy but what is not easy is:

- Using the right set of tables and columns to generate the answer
- Using the right logic (aka metric definition) to arrive at the answer

Data catalogs help with answers to such questions for humans, but they will need to evolve to serve LLMs with similar information.

The true test of trust is validation by a human. It takes transparency and time to build trust. At the end of the day, the user has deep intuition about the area being analyzed. It will take consistent answers and transparency about the work underneath the hood, to earn that users trust.

In the meantime, we have to build fault-tolerant interfaces[1] to account for lack of trust in the answers given by LLMs.

- Affordances: Unlike a graphical user interface, a chatbot interface doesn't provide clues about how it should be used, what data it has access to, or what it can/can not do. Users are left to figure all of that out, adding a new type of learning curve. In the short term, design can help with guided instructions and examples that lead end-users to work more effectively with chat.
- Context aware: Different questions need to be answered with varying degrees of accuracy and performance. The chatbot needs to answer differently based on the role of the person (new vs. experienced) and the intent of the question (back of the envelope calculation vs. company-wide metric). The product, in the short term, and LLM itself should leverage what it knows about the user and intent to contextualize the responses.

- Enable flow state: The cycle of asking a question and waiting for a response can disrupt the user's flow state. While AI may become fast enough in the long term, in the meantime, we are left to build experiences (notifications and such) that tolerate the delay.

For a deeper-dive into this topic, R&D designer at GitHub, Amelia Wattenberger, wrote an excellent summary of the downsides of chat UIs.

Data will be truly democratized

Data democratization has been an important goal for companies fostering a data-informed culture. The effort involves making collected data readily available to all members within the organization, along with the skills and training needed to use data for decision-making, process development, and metric creation. In practice, it hasn't worked out. There are really only two self-serve options - the business user either leans on a data analyst for their analytics needs, or they become a pseudo data analyst themselves. Both of these options are challenging.

Leaning on data analysts means staffing enough analysts to meet demand. As any data team knows, ad hoc requests from business users are time sensitive and unsuited to waiting in weeklong queues for the attention of a professional analyst. Hiring and onboarding analysts has its own limits to scale as these analysts must navigate a complex landscape, answering a diverse range of analytical questions relevant to their departments before they can contribute to the organization's business strategy.

Enabling business users to act as pseudo data analysts themselves necessitates that they have access to the data, know how to write performant SQL, and most importantly, know the organization-specific context and gotchas of the data. Simply providing them access to tables without these support skills can lead to bad decisions due to misinterpreting, or misunderstanding of what data to use and how to use it.

What we still need to work on:

The big challenge here is hallucination. Until that fundamental issue is alleviated, which may be a while, the path of progress is to approach LLMs as a *translator* of existing content instead of a *generator* of new content. This practically means providing the LLM specific context and asking it to mutate in a very specific way.

For example, Stemma's Discover Assistant uses generative AI by passing a prompt to a LLM to translate the SQL query used to generate the table to plain English, as opposed to providing a lot of metadata about the table and queries and asking the LLM to simply describe the table. LLMs can be very creative so the first guardrail for useful content is to give them tasks without much room for invention. In line with fault-tolerant interfaces, our AI-generated content is separate from but alongside curated content and metadata so the user can weigh the value of generated content against standard signals. We also allow users to easily provide feedback on AI content that is not useful.

Demonstrations of LLMs for analysis are typically limited to exploratory analysis - asking a question and getting a rough answer to make a product or business decision. Another common future use-case for AI, could be metric triage (why were rides down 10% week or week?). However, LLMs are particularly bad at statistics, and therefore would need to get better over time before they can be used for metrics like metric triage and reporting.

The helpdesk will be less of a burden on data teams

The responsibilities of the modern data team have evolved beyond managing storage and compute resources to delivering more data for the business, and addressing the more "socially" complex issues such as communication barriers, aiding discovery in overwhelming amounts of data, and managing shifting data ownership. Day-to-day requests about data in the warehouse can add to this complex mix of work.

There's no doubt that LLMs will make it easier for a data team member to answer common helpdesk questions in some cases, and completely eliminate categories of questions in others.

What we still need to work on:

At Stemma, we explored using GPT to answer questions commonly asked in data helpdesk channels on Slack. We found that GPT was particularly good at two things:

1. Simple classification: Classify whether a question was for access request, request for new data, or asking for existence of some data.
2. Summarization: Describe assets that didn't have a lot of existing context or documentation, using proxy signals (like query that was used to generate the data)

We found GPT to be particularly bad at convergence. GPT would diverge by adding options when asked a helpdesk question, instead of reducing options and pushing towards an answer.

I think it is unrealistic to see AI replacing the helpdesk in the near future but it can automate elements of work to address problems where an answer exists in the system, or to make more complex problems easier for the engineers to address by providing more context. Many of the capabilities being developed to give business users access to data and analytics also enable the streamlining of workflows for data experts. For example, data engineers can read SQL queries directly to understand how a table was created, they do not need an LLM to translate it for them. But, the translation from an LLM can add that context to other parts like the user's team or start date, saving them significant time as they look for parts of code where their focused time will directly contribute to solving a data problem.

Looking forward

It’s clear AI will change data, but insofar as the purpose of using data is to make better informed decisions, there’s at least one thing that’s really important today and will become even more important with AI - asking the right questions - asking the eigenquestion. An eigenquestion is the question, if answered, likely answers the subsequent questions as well. For a deeper dive, see this post by Shishir on Eigenquestions. Asking the right questions comes from experience and judgment and that becomes even more valuable in the world of AI.

[1] For a deeper discussion, listen starting 19:03 of Lenny’s podcast with Gustav Söderström