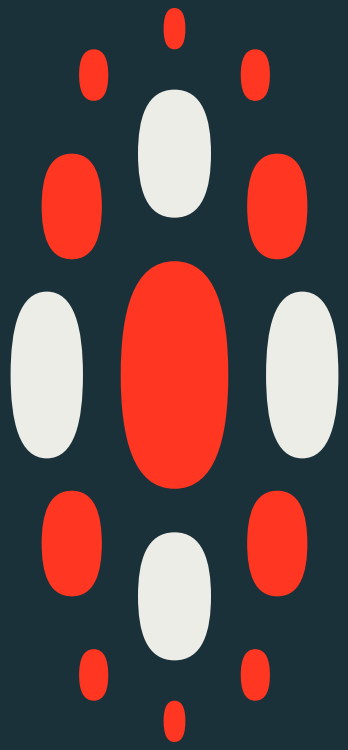

Stage 1: Prompt Engineering

Many companies still remain in the foundational stages of adopting generative AI technology. They have no overarching AI strategy in place, no clear use cases to pursue and no access to a team of data scientists and other professionals who can help guide the company's AI adoption journey.

If this is like your business, a good starting point is an off-the-shelf LLM. While these LLMs lack the domain-specific expertise of custom AI models, experimentation can help you plot your next steps. Your employees can craft specialized prompts and workflows to guide their usage. Your leaders can get a better understanding of the strengths and weaknesses of these tools as well as a clearer vision of what early success in AI might look like. Your organization can use things like the Databricks AI Playground to figure out where to invest in more powerful AI tools and systems that drive more significant operational gain and even use LLMs as a judge to help evaluate responses.

PRACTICAL APPLICATIONS OF GENAI TECHNOLOGY

Let's delve into a compelling use case that illustrates the power of prompt engineering with off-the-shelf LLMs. Consider the challenge many businesses face: sifting through vast amounts of product reviews to glean actionable insights. Without a dedicated team of data scientists or a clear AI strategy, this task might seem daunting. However, leveraging the flexibility of LLMs through prompt engineering offers a straightforward solution.

Prompt Engineering Use Case

Automated Analysis of Product Reviews Using Large Language Models

Keep track of customer feedback at scale

Check out our LLM Solution Accelerators for Retail for more details and to download the notebooks.

While conversational AI has garnered a lot of media attention in recent months, the capabilities of large language models (LLMs) extend well beyond conversational interactions. It's in these less prominent capabilities such as query response, summarization, classification and search that many organizations are finding immediate opportunities to supercharge their workforce and up-level customer experiences.

The potential of these applications is staggering. By one estimate, LLMs (and other generative AI technologies) could, in the near future, address tasks that today occupy 60%–70% of employees' time. Through augmentation, numerous studies have shown that the time to complete various tasks performed by knowledge workers such as background research, data analysis and document writing can be cut in half. And still other studies have shown that the use of these technologies can dramatically reduce the time for new workers to achieve full productivity.

But before these benefits can be fully realized, organizations must first rethink the management of the unstructured information assets on which these models depend and find ways to mitigate the issues of bias and accuracy that affect their output. This is why so many organizations are currently focusing their efforts on focused, internal applications where a limited scope provides opportunities for better information access and human oversight can serve as a check to errant results. These applications, aligned with core capabilities already residing within the organization, have the potential to deliver real and immediate value, while LLMs and their supporting technologies continue to evolve and mature.

PRODUCT REVIEW SUMMARIZATION COULD USE A BOOST

To illustrate the potential of a more focused approach to LLM adoption, we consider a fairly simple and common task performed within many online retail organizations: product review summarization. Today, most organizations employ a modestly-sized team of workers to read and digest user feedback for insights that may help improve a product's performance or otherwise identify issues related to customer satisfaction.

The work is important but anything but sexy. A worker reads a review, takes notes, and moves on to the next. Individual reviews that require a response are flagged and a summary of the feedback from across multiple reviews are compiled for review by product or category managers.

This is a type of work that's ripe for automation. The volume of reviews that pour into a site mean the more detailed portions of this work are often performed on a limited subset of products across variable windows depending on a products importance. In more sophisticated organizations, rules detecting course or inappropriate language and models estimating user sentiment or otherwise classifying reviews for positive, negative or neutral experiences may be applied to help identify problematic content and draw a reviewer's attention to it. But either way, a lot is missed simply because we can't throw enough bodies at the problem to keep up and those bodies tend to become bored or fatigued with the monotony of the work.

LARGE LANGUAGE MODELS CAN AUTOMATE PRODUCT REVIEW ANALYSIS

By using an LLM, issues of scale and consistency can be easily addressed. All we need to do is bring the product reviews to the model and ask:

- What are the top three points of negative feedback found across these reviews?
- What features do our customers like best about this product?
- Do customers feel they are receiving sufficient value from the product relative to what they are being asked to pay?
- Are there any reviews that are especially negative or are using inappropriate language?

Within seconds we can have a tidy response, allowing our product managers to focus on responding to issues instead of simply detecting them.

But what about the problem of accuracy and bias? Standards for identifying inaccuracies and bias in LLM output are evolving as are techniques for better ensuring that outputs align with an organization's expectations, and the fine-tuning of models using approved content can go a long way to ensure models have a preference to generate content that's at least aligned with how an organization prefers to communicate.

This is a long-winded way of saying there is no ideal solution to the problem as of yet. But when compared to where we are with human-driven processes and more simplistic models or rules-based approaches, the results are expected to be better or at a minimum no worse than what we currently experience. And given that these review summaries are for internal consumption, the impact of an errant model can be easily managed.

YOU CAN BUILD A SOLUTION FOR THIS TODAY

To demonstrate exactly how this work could be performed, we have built a Solution Accelerator for summarizing product reviews. This is based heavily on a previously published blog from Sean Owen that addressed some of the core technical challenges of tuning an LLM on the Databricks platform. For the accelerator, we are using the Amazon Product Reviews Dataset, which contains 51 million user-generated reviews across 2 million distinct books as this provides access to a wide range of reviewer content and presents a scaling challenge many organizations will recognize.

We imagine a scenario in which a team of product managers receives customer feedback through online reviews. These reviews are important for identifying issues that may need to be addressed regarding a particular item and for steering future books to be offered by the site. Without the use of technology, this team struggles to read all the feedback and summarize into a workable set notes. As a result, they limit their attention to just the most critical items and are able to only process the feedback on a sporadic basis.

But using Databricks, they are able to set up a pipeline to collect feedback from a wider range of products and summarize these on a regular basis. Recognizing that positively rated products are likely to highlight the strengths of these books while lower rated products are likely to focus on their weaknesses, they separate these reviews based on user-provided ratings and task an LLM to extract different sets of information from each high-level category of reviews.

Summary metrics are provided to allow product managers an overview of the feedback received and are backed by more detailed summaries generated by the LLM (Figure 1).

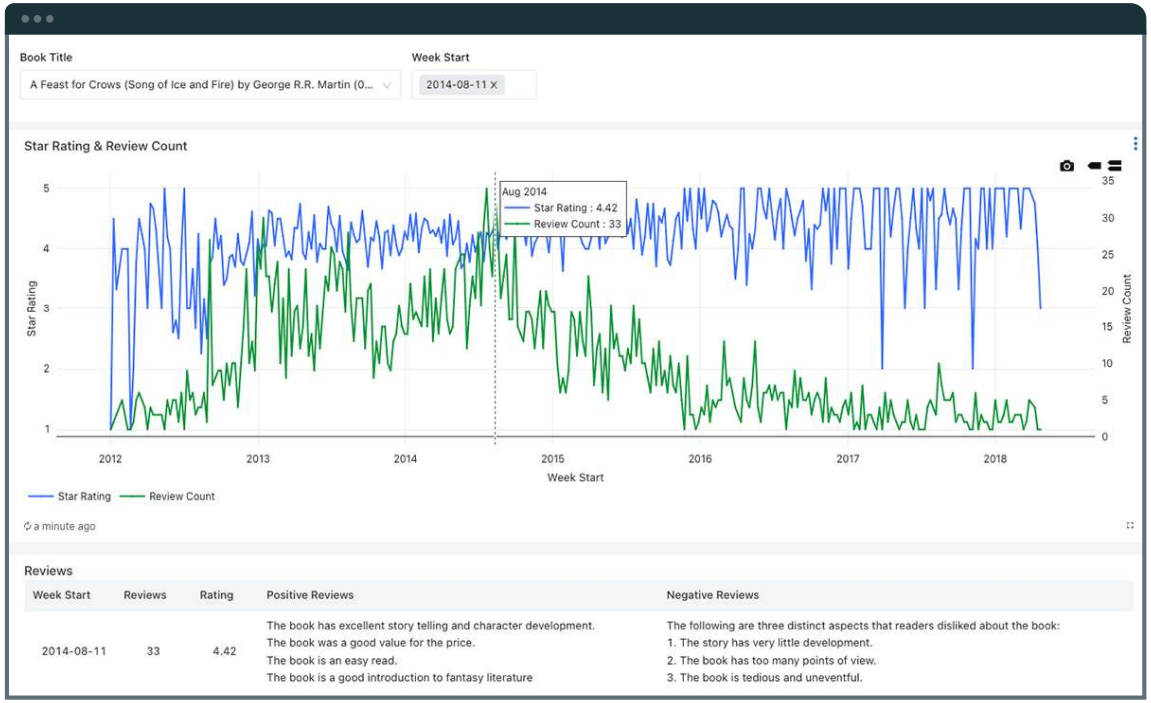

Figure 1: Summary metrics and bullet-point details extracted from user reviews extracted using an LLM

DATABRICKS BRINGS TOGETHER ALL THE COMPONENTS OF A SOLUTION

The scenario demonstrated above depends on the use of an LLM. In months prior, the use of such an LLM required access to specialized computational infrastructures, but with advances in the open source community and investments in the Databricks platform, we are now able to run the LLM in our local Databricks environment.

In this particular scenario, the sensitivity of the data was not a motivating factor for this choice. Instead, we found that the volume of reviews to be processed tipped the cost scales toward the use of Databricks, allowing us to trim about one-third of the cost of implementing a similar solution using a third-party service.

In addition, we found that by implementing our own infrastructure, we were able to scale the environment up for faster processing, tackling as many as 760,000 reviews per hour in one test without having to be concerned with constraints imposed by an external service. While most organizations will not have the need to scale quite to that level, it's nice to know it is there should it be.

But this solution is more than just an LLM. To bring together the whole solution we needed to develop a data processing workflow to receive incoming reviews, prepare them for submission to the model and to capture model output for further analysis. As a unified data platform, Databricks provides us the means to address both data engineering and data science requirements without data replication. And when we are done processing the reviews, our analysts can use their tools of choice to query the output and make business decisions. Through Databricks, we have access to the full array of capabilities for us to build a solution aligned with our business' needs.