Prolifics

Using Gen AI for a 10-Fold Increase in Output - Going Beyond Prompt Limits

At issue – working around token limits for Gen AI for productivity gains

Most generative AI (Gen AI) platforms, like ChatGPT, work on a "prompt and response" approach – you ask it a question (prompt) and it responds. A prompt, however, has what are called "token limits." Generative AI token limits are like a character or word count maximum. Imagine you're chatting with a super-smart friend, but they can only handle a certain number of words at a time. "Tokens" are basically chunks of text, and Gen AI has a limit on how much it can process in one go. If you throw too many words at it, things might get cut off or jumbled. So, when you're asking your question (creating your prompt), it's similar to respecting a text message or tweet character limit, but on a much higher level. This becomes a problem when trying to apply Gen AI to processes or a series of sophisticated steps that would a require a complex, multiple part prompt or series of prompts.

Let's apply this to a real scenario

Our client is a startup consulting firm that creates custom immigration visa plans for their customers. Their customers submit source documents (such as resumes, industry statistics, and employment reports), and the client extracts the necessary information and aligns it with a templated plan from its catalog of examples to create a tailored plan. This process enables the client to generate one plan per day.

There are two main challenges that the customer seeks to overcome:

- *Creating one plan per day is not enough. The client is working to accelerate the sales process and would like to use Generative AI (Gen AI) to improve how immigration plans get done.*
- *However, in current Gen AI basic models, there's a token limit on how much the language model can process in a single interaction. The client's desired process would exceed token parameters.*

The Prolifics team designed a solution that allows the client to upload their customers' source documents to an IBM Cloud database. From this database, relevant facts can be automatically extracted based on the client's input. The solution then sends the information to IBM watsonx, which pulls the data, arranges it in the desired output as defined by the client, and correlates it with the catalog template, generating the custom plan one section at a time.

Let's apply this to a real scenario

During the engagement, our team navigated several client pain points to ensure an optimal outcome. Recognizing that financial constraints are often common with startups, we tapped into the expertise available at the Prolifics Innovation Center, a virtual sandbox environment where experts work together. We also collaborated with our engineering specialists and the IBM ecosystem team to develop a prototype and a minimum viable product that met all the requested standards and achieved it at a cost the client could commit to.

To execute our unique solution, we assembled a cooperative team that included our Innovation Center, onshore and offshore engineering experts, and IBM consultants. In tandem, we continuously turned to the client for their invaluable insights to design a compelling prototype that aligned with their needs.

Following a comprehensive briefing, the clients were so enthusiastic that we moved directly to the proposal phase, streamlining our timeline and bypassing the investment needed for an interim pilot step.

The benefits

The Gen AI-powered solution went beyond prompt and response; it exactly met the needs of the client in such a way that the client wanted a proposal immediately. The vision is for the solution to allow them to reinvent the way business gets done and accelerate their sales cycle.

With their original system, the client could generate only one custom plan a day. They now estimate the Gen AI solution will allow them to create a minimum of ten unique plans every day. We delivered a lower-than-expected price point by leveraging our Innovation Center and our offshore expert engineering teams and avoiding unnecessary tuning steps for the model.

We designed a custom solution combining IBM Database and IBM watsonx to overcome the Gen AI token limits. Our solution uses Gen AI to orchestrate steps and pull items together to create new content. This produces value way beyond how the client would have been able to use basic prompt-and-response Gen AI previously. Information is passed into the Gen AI foundational model when needed, meaning there's no fine-tuning or customization of the model itself, which saves time and money.

For this client, greater immigration plan through-put will mean that individuals will get placed with the host businesses more quickly, meaning greater productivity for those businesses and, in turn, better service for their customers.

Beyond this client, this advanced solution will create cost and time efficiencies for any organization that wants to use Generative AI for more complex processes.

Generative AI is already changing the world by reinventing customer experience, productivity, and products. In this case, our solution is reinventing the product. The logic behind this solution is transferable and sustainable and could easily be applied to other customers' productivity needs – helping to make the world work better.

Learn more about **Gen AI** at Prolifics

Why Prolifics for your Gen AI solutions?

We've made considerable investments in Gen AI use cases at the virtual Prolifics Innovation Center, so we are well positioned to help you understand what you would need to first invest in and implement.

We'll help you understand what the best technology is for your Gen AI use cases – because we've already done it with a range of different technologies. We're not tied into one specific technology solution.

And, you'll have the full benefit of the Prolifics' "Power of 1." We'll generate impactful ideas in one day; we'll give you proof of value in your context in one week; we'll deliver an MVP with tangible business value in one month; and scale to deliver 10x ROI in one year.