DIGITAL
BUSINESS Institute

The Application of Generative AI for Business Solutions

by Andy He

As the world grapples with the recent boom in popularity of Generative Artificial Intelligence (GenAI) following OpenAI's release of ChatGPT, savvy businesses have started to explore the applications of such technologies to help their businesses operate more efficiently and maximize profits. This article provides a brief overview of new, promising variants of GenAI, and makes recommendations to business owners for how and when they should be considered.

Following its launch in November of 2022, OpenAi's ChatGPT became the fastest-growing internet application in history, reaching 100 million users in just 2 months. With its undeniable utility, ChatGPT quickly launched Generative AI into the mainstream, with everyone from every-day citizens, to students, to CEOs leveraging some form of GenAI.

Subsequent projections of GenAi's potential economic impacts showed trillions of dollars being added to the global economy. The speed of the development of the industry along with its obvious capabilities demanded the attention of business owners seeking to enhance their businesses. Pre-2023, AI adoption among businesses had stagnated at around 50% for around five years from 2018-2022. However, after ChatGPT was introduced in late 2022, AI adoption surged. By May of 2024, AI adoption had jumped to 72% among businesses, with Generative AI use not far behind at 65%.

While it remains clear that businesses have and will continue to leverage Generative AI to strengthen their business solutions, it remains less clear how successfully they will be able to do so. A recent Harvard Business Review Article revealed a "sobering reality": as many as 80% of all AI projects will fail.

Many businesses struggle to get AI projects off the ground because there are so many ways a project can have its wings clipped. The hype of GenAI and AI in general becomes a double-edged sword. Companies are prone to throwing themselves at AI projects that should've never existed at all, just to avoid the business equivalent of "FOMO", the fear of missing out. This results in ill-conceived projects that are more-or less doomed to fail. Another obstacle is the GenAI itself. The vast majority of models, including the fabled ChatGPT suffer from an issue that has been dubbed as "hallucinations", whereby the model would generate nonsensical or false responses in response to prompts that regard topics in which the model has gaps in its "knowledge". Another issue is copyright. Models have been observed creating potentially copyrighted materials after imitating the works of an artist or a writer whose works it was trained on. Another issue is privacy, another issue is with biased data pools, and on, and on, and on.

If Generative AI is to deliver on the promises it has made, there are immense costs, obstacles, and challenges to overcome, and each business and each industry will need to tackle these issues in a unique fashion.

Scopes of Generative AI Adoption

Off-The-Shelf Models

Many companies have already chosen to adopt various forms of Generative AI into their business practices, each to different degrees. One of the foremost distinctions between different scopes of GenAI adoption is the choice between so-called “Off-The-Shelf” and Customized Models.

For the purposes of this article, an Off-The-Shelf model refers to a model that is already available and doesn't require any modifications before usage. Examples of this include ChatGPT, Claude, and Meta’s LLaMA. The models are already available and have capabilities that businesses can use as the models are right now, with no modifications.

The main draw of using such a model is the cost to benefit ratio. As the definition implies, there is minimal startup cost for an Off-The-Shelf model. While each Off-The-Shelf model is different, they can be split into two main camps.

First, some closed source models like ChatGPT can be accessed by businesses through the developer itself. OpenAI for instance, offers its own API, from which ChatGPT can be deployed or customized. The main cost in this instance is in tokens. Tokens are like a form of currency, which are used to pay for the use and training of (if customized) models that are inside a third-party API like OpenAI’s. If this is a true Off-The-Shelf model, you won’t need to worry about the token cost of customizing the model. This means that there is very little startup cost, but instead, a scaling cost that increases or decreases based on how much you choose to use that model.

On the other hand, there are “open-source” models. Models like Meta’s LLaMA are open source, meaning that it is open for anyone to access, use, and edit for free. If this is the type of Off-The-Shelf model you’re using, you won't need to worry about tokens, but you will need to be able to provide computational resources, as the model will be running on your servers, not Meta’s, any maintenance expenses associated with the model or your server are your responsibility.

Either way, Off-The-Shelf models have capabilities that are both limited and non-specialized. That is to say, a basic version of ChatGPT, or any other Off-The-Shelf model, lacks the use cases that a more specialized, custom built model will have. A customized model can be constructed with a specific goal or task in mind, making it more suited to fill certain needs, whereas an Off-The-Shelf model has set, unchangeable and un-adjustable capabilities. Therein lies the tradeoff: cost is traded for stronger capabilities and an ability to build a model that plays to the strengths you need.

If the capabilities of a basic Off-The-Shelf model will fill a need your business has, it is preferable to a customized option thanks to its low cost. The main draw of an Off- The-Shelf model is automating tasks currently performed by human employees. To illustrate this, let's look at a few of the capabilities of the most popular Off-The-Shelf Model, OpenAI's ChatGPT.

Customer Service Tasks

Many Off-The-Shelf models have a skillset that can be leveraged to serve a customer service function. Thanks to a recent development known as In Context Learning, models like ChatGPT are capable of having full, coherent conversations with humans, and are able to use past interactions in the same "conversation" as context to guide future responses. This capability makes it a great tool for customer service. An example of this is the British renewable energy company Octopus Energy. Soon after the company integrated ChatGPT into its customer service channels, the company's CEO Greg Jackson revealed that 44% of the company's customer service inquiries were handled by ChatGPT, replacing the need for 250 human customer service employees. Not only that, but ChatGPT also managed to surpass human customer service representatives in terms of customer satisfaction, achieving an 80% satisfaction rate compared to the human average of 65%.

Coding

ChatGPT is also able to write code. Many models including ChatGPT can write code that incorporates outside libraries like pandas, numpy, and pytorch just to name a few, and are fluent in many different coding languages. It can even translate code between languages while preserving its original intent. These applications are also already being used by real world businesses. The CEO of software company Freshworks stated that workers at his company had begun using ChatGPT to write code. As a result, they cut the time needed to complete software development tasks from up to 9 weeks to just a few days.

Education

Recently, the CEO of Khan Academy Sal Khan announced in March of 2023 that his education company would begin to pilot a new virtual tutor named Khanmigo to teach students (replacing a human tutor) that uses the then new GPT-4. Now, a version of the software has been made publicly available for a cost. Most recently, a video was uploaded on Khan Academy's YouTube channel featuring the all new GPT-4o teaching Khan's son Imran how to do math problems on Khan Academy, with GPT-4o verbally conversing with Imran, walking him through problems. A new feature harnessing GPT-4o by Khan Academy has yet to be released.

But there's a caveat with each of these examples of Off-The-Shelf AI usage. For the first 2 examples, it's the scale of adoption. Neither example has ChatGPT working to answer emails or write code autonomously. Instead, a human worker simply uses ChatGPT to assist in their tasks. In the case of Octopus Energy, the CEO later revealed that agents only used ChatGPT to draft email responses to customers who had complaints. The human agent would then evaluate the response written by ChatGPT and decide whether or not to use it. The story is similar for Freshworks, whose employees used ChatGPT on a small scale to help them write code. In both cases, the use of a base version of ChatGPT was limited to use by individuals to assist them. In the case of Khan Academy's “Khanmigo”, Khanmigo was never ChatGPT at all. It simply runs on “ChatGPT technology”. In reality, it's extremely difficult to truly use an unaltered version of ChatGPT, or any other off the shelf model in a large capacity.

Ultimately, this is what makes customization important. It allows companies to tailor the capabilities, or a model, so that it can be used for a very specific use case with less and less human intervention or supervision, and thus, less cost from human labor

Customization Options

Retrieval-Augmented Generation (RAG)

RAG is a technique that allows a GenAI model, usually a Large Language Model, to draw on separate data outside of its original training dataset to generate responses. Generally, the outside data will be the internal data of a business. This is useful to reduce the occurrence of hallucinations. Hallucinations refer to when an LLM generates a response that is either partially or entirely made up by the model. This usually happens when the AI receives a prompt that requires knowledge of subjects that it does not have. In response, it will make up a reasonable sounding answer that is false. In the context of use on the part of businesses, RAG helps LLM avoid hallucinations, which could prove to be common because of the nature of prompts it may receive when in use by a business.

Take an LLM that is handling a customer service complaint for an Internet Service Provider, a company that sells Wi-Fi. If an unaltered version of ChatGPT receives a complaint asking why a customer's bill for the month is higher than it usually is, ChatGPT will have no way to accurately answer this question, because it doesn’t know what the ISP companies' billing policies are. In many cases, ChatGPT will “hallucinate” and generate a reasonable sounding answer that is entirely false. As a business, this would be unacceptable if a human employee gave a false answer to a customer, so it should be equally unacceptable if a GenAI model did.

This means that RAG can improve the relevance, quality, and accuracy of responses for a business.

However, the extent to which RAG can fix issues like hallucinations is still being hotly debated. At present, most experts believe that hallucinations by GenAI will never be completely eradicated, with RAG or with any other method. However, consensus still exists that RAG can at least limit hallucinations.

Fine Tuning

Fine Tuning is a process of adapting a pre-trained model for a more specific use case. Think of this process like a worker going to a specialized skills workshop, or in the case of heavy fine tuning, like a high school student going to college. They will concentrate on a new set of skills so that they will be able to work on more specialized tasks. In the same vein, Fine Tuning allows a model to “fine tune” its capabilities to make it more effective for a specific task.

Take the realm of code generation as an example. One could Fine Tune a version of an already existing Off-The-Shelf model to specifically be able to code at a higher level. Think of this like a student going to school and studying computer programming. This way, a model can be trained to excel beyond its current capacity in almost any skill. This also lets us leverage the past knowledge of the pre-trained model for the new task without having to start from scratch.

As a result, this is a much more cost and resource effective option as compared to training a model from scratch, as the foundations of the Fine-Tuned model have already been laid out in the form of the original pretrained model.

Continual Learning

Continual Learning refers to the process of an AI model being able to constantly adapt, improve, and learn new things over the course of its lifetime with minimal human interference, all while remembering the past skills it had learned before.

If implemented correctly, a continual learning model will be able to constantly take in new streams of data, primarily from user interactions, and use this data to continually improve itself and train itself from this newly acquired data. This also in turn helps the model adapt to user needs and preferences over time, keeping the model more or less modern.

The main application of this is more or less to keep your model current. In a 2020 survey of bankers in the UK, it was revealed that a substantial portion of respondents, 35%, reported that the COVID-19 pandemic has negatively impacted the performance of their Machine Learning models. This was because their models, which had only been trained on data from before the pandemic, were unable to adapt their predictions to the economic conditions brought on by the pandemic. This is a good example of the potential importance of Continual Learning. In this case, a Machine Learning model equipped with continual learning could have potentially collected new data during the pandemic, adapting its predictions to the new economic conditions rather than making outdated predictions based on pre-pandemic datasets.

This same idea can be applied to Generative AI, with new data informing a model, which could then adjust its responses. This would give a model a “timeless” quality to a certain extent, because a model could adapt to changes over time, relieving the need for a brand-new model to be constructed every time a major (or minor) change to conditions arises.

Third-Party API vs Private Deployment

Aside from choosing customizations to the model itself, there also remains a question of which API a business should embed the model in. The main distinction is in whether or not the business should use a third-party API or privately deploy the model on their own. Both options have their own costs and rewards, meaning it's not clear cut which option your business should choose. Below are some of the main selling and sticking points of both options.

Benefits of Third-Party APIs

Cost
Private Deployment for a business is a costly ordeal, which can require a high upfront and maintenance cost. Additionally, building and maintaining functionalities can be more expensive. Alternatively, hosting on a third-party API can eliminate this concern. Third Party APIs often offer relatively more affordable pricing structures to businesses looking to host a model.

Convenience (Time)
Establishing an API in house can be very time-consuming depending on the size and scope of the project, and the capabilities the business requires. Opting for the third-party option mitigates this, as the API will already be set up for a business to utilize.

Scale
Because the third-party API has already been cultivated, the scale of a project hosted on a third-party API can be much higher than private deployment, simply because the capabilities of the third-party API or (generally) more varied and fleshed out.

Benefits of Private Deployment

Customization
On the other hand, the adjustability of a project can be limited by a third-party API. In a similar fashion to how customizing a model allows for the model to have more specific capabilities compared to an Off-The-Shelf option, using an in-house API allows for greater flexibility than a third-party option.

Security and Privacy
Many businesses will be training models with sensitive information and data. A business like a bank or a hospital, for instance, will likely be using sensitive, private financial or medical records that are highly regulated. As a result, it would be preferable to keep all this data private rather than have it exported to a third party to mitigate security and privacy concerns.

Independence and Dependence
Using a third-party API can also leave your business and your project at the whims of the third party. If the third party's API crashes, goes offline, falters, or in any way loses function, your project will be at the mercy of the third party, leaving you without control in such a situation. As a result, it could be preferable to build an API in-house to mitigate such concerns of potential unreliability.

Cost of GenAI Variants

Of course, one of, if not the biggest concern for a business owner would be the cost of developing and implementing a new feature, in this case, GenAI. Below is a general breakdown of the cost of implementing each different type of GenAI discussed above.

Third Party API and Private Deployment

Before anything else, there must be a distinction between the cost of each type of GenAI on either a Third-Party API or Private Deployment. The actual cost of deploying a model can vary based on which you choose. Everything below will be a **generalization**. Every cost mentioned will be on a general level, not taking into consideration the context of the deployment itself.

First, Third Party APIs are generally seen to be cheaper than Private Deployment. Both options share costs in the form of data and coding talent, but each also has distinct costs in other respects. Third Party APIs generally offer pricing structures in tokens. This structure limits upfront costs but necessitates a constant token cost to repay the third party's inference cost. On the other hand, Private Deployments' main unique costs are the cost of hardware and maintenance. In order to have an in-house private deployment, a company must develop and maintain its own API, and that API needs hardware to be stored and run on. This requires high upfront costs, which still force the company to pay its own inference cost.

Off-The-Shelf Cost Structure

As discussed before, Off-The-Shelf models generally follow two different pricing structures. First, if you choose a Third-Party API, you will pay the token cost that a Third- Party will charge you for using their API. The price of tokens can generally be found online on the Third Parties website if you wish to use a Third-Party API. Alternatively, select models that are Open Source can be directly used by any entity for commercial use. In this case, it's possible to Privately Deploy an unaltered Open-Source model on your own API, but this comes with hardware and maintenance costs.

Customized Options

All three of the customization options that were presented earlier have more-or-less the same types of costs. The overall cost will vary depending on many factors like the scale of the model, who you hire to build the model, the maintenance costs, and more. But generally, here are the main costs associated with customized models.
First, programming talent. Obviously, you need to find a team of programmers to help physically write the code and train the model. The cost of "labor" will depend on which company or team you choose.

Second, data. Most companies will not have immediate access to the immense amounts of data that can be necessary to train certain types of models. Collecting and storing this data will come at a cost, both in terms of money and time.

Third, security and privacy. This is something that will be touched on in further sections, but the cost of mitigating legal liabilities, and making sure that private information stays private, will be a challenge in and of itself, requiring a potentially large investment in legal services.

Finally, maintenance. This applies more so to Fine Tuning and RAG, but making sure that the model is constantly being trained on up-to-date data, and making prediction with the current set of facts will be exceedingly important for many types of businesses, especially in industries where rapid changes are expected, like finance, health, etc.

Liabilities

The power of Generative AI for businesses is obvious, yes, but equally as complex and important are the risks that GenAI comes with. If GenAI is to be implemented, business owners must first understand, and then mitigate or manage the risks. Below are a few of the most important.

Harmful or False Outputs

One of the biggest red flags of GenAI from the start has been the risk of Hallucinations. As discussed previously, a hallucination is when an LLM generates an output that contains misleading or incorrect information. Naturally, a company cannot have an LLM providing false information to customers or employees, depending on the use case. Such mishaps have already caused major losses for businesses, even if indirectly.

In February of 2023, Google posted a demo of their brand new LLM, Bard to twitter. Bard then hallucinated, producing a factually inaccurate response to the prompt. Following this error, Google parent company, Aplhabet’s shares fell 7.7%, translating to $100 Billion of lost value.

Although this isn't an instance of an LLM hallucinating when responding to a customer, it does show the risk that businesses take on when using an LLM, and the devastating costs when mistakes happen.

As previously mentioned, experts disagree on how well hallucinations can be limited. Many experts claim that hallucinations will never be completely eradicated. However, even if they can't be cured, they can be handled. Employees in charge of “quality control” of a GenAI model's responses would help limit the impact of hallucinations by ensuring false outputs are never used.

Data Leaks

In April of 2023, Samsung shared with the world that it would be banning the use of ChatGPT by its employees. This news followed the discovery that one of the company's engineers had shared sensitive information, part of the company's internal source code, with ChatGPT. As a result, Samsung's proprietary information was at risk of being used to train ChatGPT, or even be revealed to other users in multiple different ways.

In November of 2023, a team of Google researchers revealed that through the use of specific keywords, users could force ChatGPT to reveal some of the LLMs training data to users, including the PII (Personally Identifiable Information) of real people, data which the model had been trained on.

Alternatively, such information could be accessed another way, through a data leak on the part of the company. In September of 2023, a Microsoft AI research team inadvertently exposed 38 terabytes of private data. Hacks or accidents like this one have been a concern in every industry that uses sensitive data, and the story is no different for businesses that are using GenAI. It is therefore important to understand how to mitigate the risk of a business's internal data being leaked.

A solution would involve the education of employees that are utilizing such tools, to make them aware of the risk of proprietary and private data being exposed to unwanted third parties.

Regulatory Risk

Recently, companies like OpenAI have been under fire from regulatory agencies, claiming that their practices do not align with regulations. In 2023, Italian privacy watchdog Garante accused OpenAI of violating the European Union's data privacy laws, resulting in a brief ban of ChatGPT in Italy. While only a brief shutdown, this incident speaks to the possibility of both governmental and non-governmental authorities regulating or even barring use of such services. In such an event, a business that is reliant on an LLM would be effectively crippled in the short term until a workaround could be found.

Such regulations, especially in the United States, are still relatively underdeveloped, but developing. Businesses will need to evaluate how their models, and how their usage of models, interacts with new upcoming, and constantly changing regulations.

Recommendations

It can be overwhelming as a business owner to choose between different options of customization and implementation. This article only covers a few of the most recent major developments, and there are many more ways to use GenAI for your businesses. However, a list of general tips can be helpful to get you in the right ballpark. Below is a compiled list of the most important.

Is This a Problem for AI?

The most fundamental question a business must ask itself before using AI is if it's even a problem that AI can, or should, be solving in the first place.

Part of the reason so many AI projects fail is because businesses use AI to solve problems that AI isn’t suited to solve. The ultimate issue is whether AI can or cannot provide a solution to the problem you have. Past that, if AI can provide a solution, is that solution better than the other possible solutions you have?

Further, you must consider not only the efficacy the model could have, but the “side effects” that AI comes with. Does this problem require moral or ethical considerations? Does it require creativity or innovation? Emotional intelligence? Cultural or Emotional Sensitivity? Does AI introduce new liabilities that threaten your business that cannot be effectively managed? If the answer to any of these questions is yes, AI is almost certainly not the answer.

For the time being, AI should also be seen as a tool instead of as a chief. AI can rapidly speed up the speed of certain tasks and offer far superior cost efficiency, but it's important that humans still have some level of oversight. Almost any real-world application of AI has humans overseeing the AI in some way, because without humans, AI is prone to breaking down, making mistakes, and therefore, introducing liability.

All these issues need to be considered, and many cannot be built around. Before making a commitment, you must ask if AI is the best answer, or just an answer.

Is Your Business Ready?

Apart from considering the capabilities and limits of the AI itself, you also must consider if or when your business will be ready to use AI. This goes back to many of the requirements and liabilities of AI, and how your business can adapt to them.

First and foremost, does your business require the requisite amount of quality data to train an AI model, and is that data organized, unbiased, and accessible? Without this, your business is not ready for AI if you are looking at customized options.

Another step is assessing all your risks and liabilities. Are you in an industry that requires data to remain private? How much human oversight can you provide? Are you located in an area that has or could have strict regulations that would limit your model’s capabilities? If you can't resolve or address one or more of these issues, then your business is not ready to adopt AI.

For any of these concerns, it's important to weigh the tradeoff of the potential power of AI and the time and costs it would require in order to transform your business or its operations to meet these requirements. In this sense, it isn't just the cost of creating, storing, and maintaining the model that counts. Let’s say your business lacks the proper data it needs before being able to develop a GenAI model. You will have to consider the cost of collecting that data as a cost of adopting AI.

Take a Trial Run

Like any other change to your business, it would be wise to start with a smaller scale prototype, and then implement it in a way that is as unobtrusive as possible. First, starting small helps you gather insights without serious harm to your business if it doesn't work the way you intended. Implementing a model on a “personal” level, allowing employees to experiment in using a GenAI model to assist them instead of having it be a major piece of your business ecosystem, could help mitigate costs and major risks. This could, however, limit innovation and prevent major breakthroughs. As with everything, this is a tradeoff you will need to consider.

However, experimentation should also have its limits. Experimenting with a model that disrupts the human employees in your business should be avoided, if possible, in order to preserve the functioning of your business while the AI is being tested. Of course, this is a fine line to walk, in that the more disruptive (in both a good and bad way) the AI is in testing, the more valuable observations you can make.

Consult With Industry Experts

At the end of the day, no one paper can contain the answers for every specific business in every possible scenario. Therefore, it's imperative that before any major decisions are made or resources are used, you consult an expert in the AI industry.

As the AI industry has boomed, hundreds of AI consulting groups have entered the market. This ranges from small groups of programmers and visionaries to massive consulting corporations that have expanded into the AI space as it gained steam. Companies like McKinsey (QuantumBlack), the Boston Consulting Group, and IBM, offer consultations about how your business can most effectively leverage AI to your advantage. This way, all the specifics of your business can be accounted for effectively.

Conclusions

Every major technological advancement in history has been both tantalizing in its possibilities, and equally terrifying because of its possibilities. The rapidly accelerating pace of the development of AI has and will provide businesses with a vast source of potential applications. However, AI also comes with its fair share of risks, fears, and liabilities that must shape the way business leaders utilize AI.

Ultimately, it is the job of business owners to make informed and educated decisions about the use of AI in the business space, as the norms that are set today could come to govern the world for decades to come.

Andy He
Summer Research Intern
Digital Business Institute
Boston University Questrom School of Businesses