DELL Technologies

Artificial Intelligence

# Generative AI for Organizations

**Dr. Jeffrey Lancaster, Chief Innovation and Technology Strategist, Dell Technologies**
**Adam Robyak, Field CTO, Dell Technologies**

# Introduction

It may be a fool's errand to try to codify a technology that's currently undergoing an explosion in both the public's eye and its underlying technology – like trying to make sense of the dot-com era mid-boom – but generative AI is having a moment.[1] The confluence of algorithms, capabilities, riskiness, and society-altering potential means that people in every industry are asking themselves whether this technology can be used in their world; will it change how their work is done, how their customers are engaged, how their decisions are made? With the recognition that, by the time this whitepaper is styled and published it may already be nominally out-of-date, we submit to you a framework to help not only decide whether generative AI is something you might be interested in, but to provide a set of questions to help you think through how you might best implement it, the ecosystem of tools and technologies that exist (today) to help you achieve your objectives, and the issues that may drive you toward (or away from) one technical decision or another. There is no single way to implement generative AI and new ways are appearing daily. And although the technology itself might be neutral, the human decisions which inform and shape any AI are necessarily fraught with complexity.

## Implementation

Before jumping to implementation, consider three important qualities of generative AI:

1. the type of training data used to build a model
2. the source of that data
3. the source of model reinforcement

For text-based AI, there are many large language models (LLMs) which have been pre-trained to produce generative textual outputs; the corpus of text used to train each model largely determines the utility of the subsequent outputs: training a model only with code will be useful if the output will surely be code, but training only on code for an output that is intended to be poetry or narrative wouldn't be very successful. Additionally, for non-text-based models, the best training data is made up of whatever the intended target of the model is: images train a model that would generate images, sounds train a model that would generate sounds, etc. Although intended outputs can be stitched together – a voice model can generate text that could in turn generate an image – single models are best trained on the same data type as the intended output.

Perhaps the most important initial decision you will make after determining the type of training data you will leverage is whether you intend to use pre-existing "open" models as they are, whether you can and should build your own "private" models, or whether you can take advantage of the extensive investment of time and energy already made into existing open models while "blending" in your own institutional content, context, and guard rails.

[1] Wiles, Jackie "What's New in Artificial Intelligence from the 2022 Gartner Hype Cycle" https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2022-gartner-hype-cycle, published September 15, 2022.

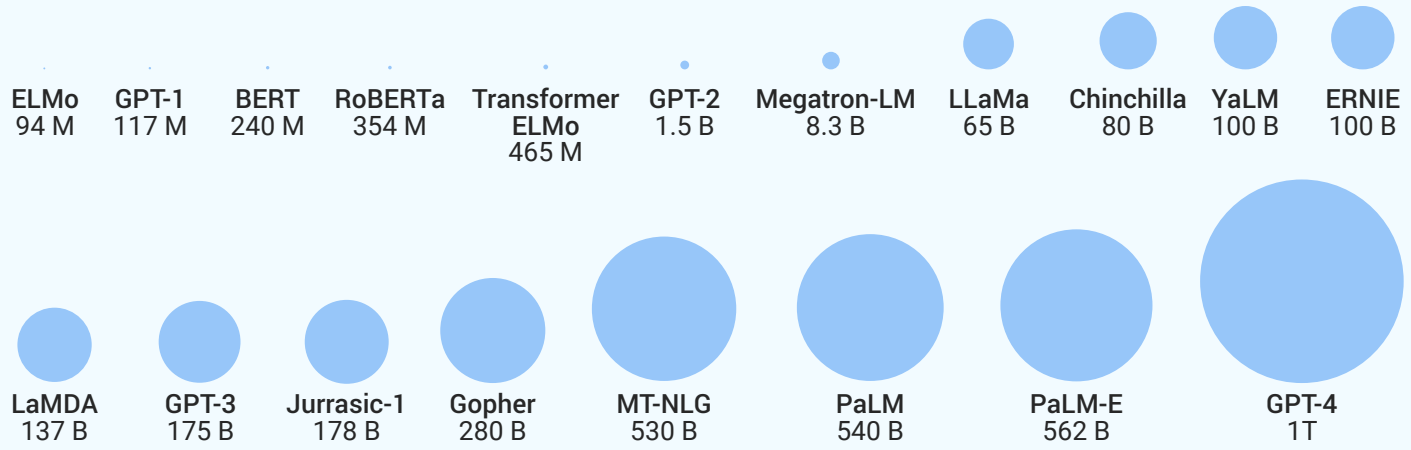


*Figure 1. Numbers of parameters in several small- (≤100B, top row) and large- (>100B, bottom row) language models.*[2]

## Open

Many of the current batch of news-making algorithms – OpenAI's GPT-3 and GPT-4, Google's BERT, Facebook's RoBERTa – have been trained on very extensive open LLMs[2], e.g. the entire text of the Internet[3]. Similarly, algorithms which generate images – Dall-E, Midjourney, Stable Diffusion – have been trained on millions of images scraped from the web[4]. Despite their gargantuan digital underpinnings, what makes these models seem magical to most lay users is their general-purpose application: a user can type some text or upload an image and get relevant, albeit sometimes non-sensical, results. And these models are already being embedded in other tools: ChatGPT is being added to Microsoft's Azure stack, Bing search, and Edge browser.

Open generative AI algorithms typically reinforce themselves via user feedback loops. When text or an image is generated, the human user can indicate if that output aligns to their expectation (leading to positive reinforcement of the model) or if it is not what they wanted (leading to negative reinforcement of the model). By continuously tuning itself, these open systems can adapt based on the collective usage of very large groups of people. These models can also retrain based on new input/query data, which has already led to private data appearing in public model outputs (see security concerns in blended model below).[5]

Building from such an extensive open dataset may be good for applications such as ChatGPT, Google's Bard, and Baidu's Ernie Bot, but it may not provide the degree of specificity or security that organizations are looking to bake into their generative AI.

## Private

On the opposite end of the spectrum is generative AI that has been trained only on an organization's data, e.g. a knowledgebase, FAQ, internal documents, sales reports, patents, etc. Although the outputs will likely be much more aligned to an organization's standards, a completely private generative AI would likely suffer from both a limited training dataset and a small user base to provide reinforcement. Private generative AI may provide a conversational interface which can interpret a user's intent, but organizations adopting this type of AI are essentially just making a human-friendly layer to access their existing data within the context of a broader conversation.

To guide the conversation, controlled vocabularies funnel a user toward an existing piece of data or an existing solution via a dialogue routing engine: the decisions embedded into the engine are human decisions, often embedded as code, which are meant to control how the AI will respond to pre-selected intents or keywords. By overriding the algorithmic model, the routing engine provides the predictable outcome that the organization wants without needing to prioritize the users' inputs or feedback for reinforcement. The shift from purely scripted or fixed response systems (chatbots) to enterprise class generative AI systems is a response to the time and effort required to support a fixed script (Figure 2).

[2] Based on Momentum Works, "The emergence of Large Language Models (LLMs), https://thelowdown.momentum.asia/the-emergence-of-large-language-models-llms/

[3] See CommonCrawl: https://commoncrawl.org/

[4] Oftentimes without the original creator's consent, e.g. Growcoot, Matt "Midjourney Founder Admits to Using a 'Hundred Million' Images Without Consent" https://petapixel.com/2022/12/21/midjourny-founder-admits-to-using-a-hundred-million-images-without-consent/, published December 21, 2022.

[5] Maddison, Lewis "Samsung workers made a major error by using ChatGPT" https://www.techradar.com/news/samsung-workers-leaked-company-secrets-by-using-chatgpt, published April 24, 2023.

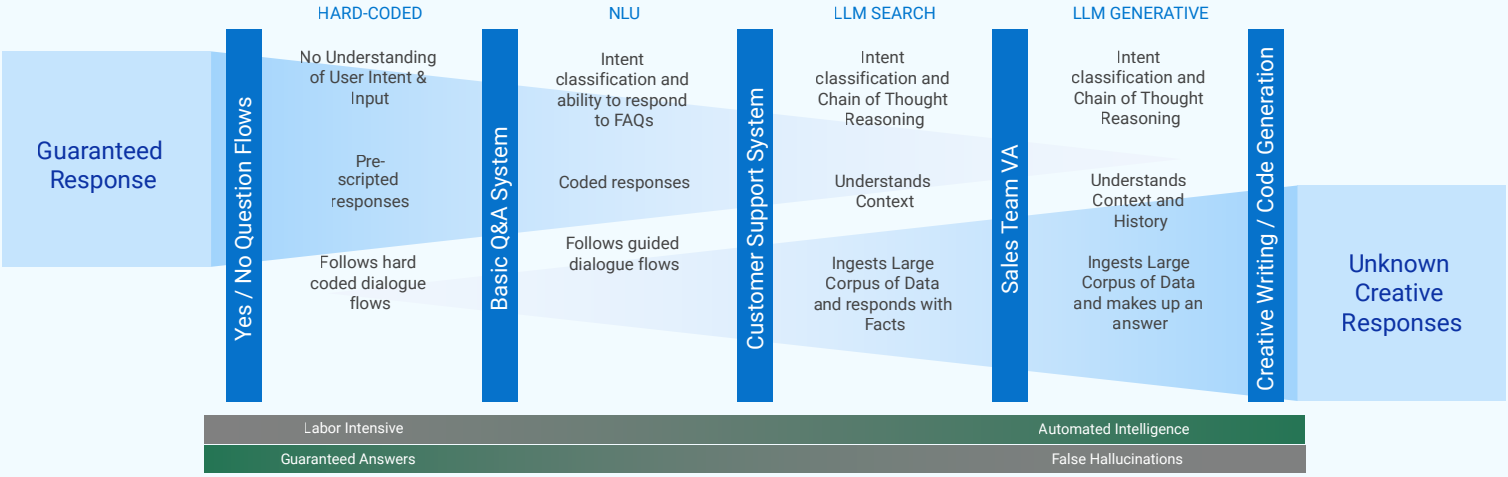


*Figure 2. A generative approach improves deploy/run/maintain capabilities.*

### Blended

To counteract the limitations of a completely private generative AI while also avoiding some of the potential pitfalls of a completely open generative AI, some organizations have layered their own knowledgebases and internal documents on top of a pre-trained open LLM. This technique gives the generative AI access to the various associations of the open LLM while still providing some guard rails and organizational context via the proprietary data layer.

In cases where this might be of interest, security concerns are quickly introduced into the conversation. If queries and prompts are fed back as reinforcement, subsequent responses could contain breaches of proprietary data fed into an algorithm as a prompt. Additionally, the hallucinations of an open model may provide fabricated information that is out of line with an organization's intended objectives.

## Trends

The evolution of the generative AI ecosystem is happening impossibly fast with new developments releasing daily. An accurate snapshot of the current landscape is best rendered instead as a constellation of trends with a call-to-action to keep an eye on this space.

**Beware the monoliths.** Solutions are not built within a single platform, but are instead compiled from various platforms, often leveraging vendor partners to stitch tools together; different tools can be connected to improve the quality and diversity of subsequent outputs.

Triaging at scale. Especially for customer service and troubleshooting uses where chatbots had found widespread utility, generative AI is capable of diagnosing problems in a human-like manner; anything an algorithm cannot resolve is then passed to a human.

**Improved internationalization.** LLMs make it easier to build a solution in a single language and then rely on various algorithms to convert to other languages.

Removing hallucinations. New algorithms are working to improve upon unwanted or inappropriate hallucinations while still maintaining creative serendipity.

**Smarter data management**. LLMs and generative AI are fast-tracking data management by algorithmically cataloging and tagging data for the knowledge economy.

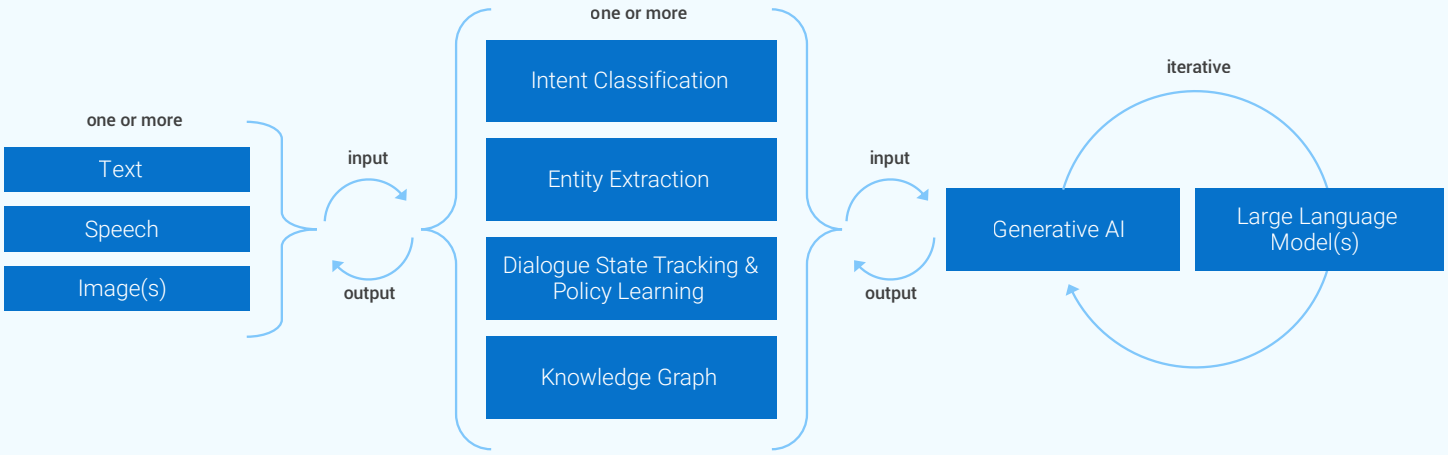


*Figure 3. A generalized architecture for a generative AI system.*

## Common architectures for success

While it's not imperative that every leader fully understand the technical underpinnings of a generative AI system, a decision-maker must not consider generative AI to be a magical technology. By thinking through how the pieces fit together, you will be better able to avert potential roadblocks and missteps, you will be able to ask better questions of your implementation teams and potential vendors, and you will be more likely to foster magical experiences for your end users.

### Inputs

The critical first step when implementing generative AI is to ensure users can interact with the system via one input or some combination of inputs. Today it is common for a user's query to be input using text, speech (sound), and/or a set of images, and in cases where the generative AI is meant to mimic a human interaction, chatbots, virtual assistants, or digital humans might mediate the delivery of the input into the system.

Once the initial prompt is submitted, however, it may need to be transformed into a different format using an AI model that is separate from the generative model itself: e.g. speech-to-text (STT) models are used in applications such as dictation software, automated voice assistants, and speech recognition for video subtitles[6]. As transformation algorithms continue to develop, more complex inputs and more diverse inputs will be candidates for generative AI algorithms.

Visual inputs, importantly, extend beyond sets of static images. Eye movement tracking and facial recognition are common inputs that convert a video feed into sets of image frames and determine where a user is looking or how a face is oriented. Additionally, facial expression extraction, lip movement tracking, gesture tracking, and motion capture can be extracted from images in order to act as inputs to generative AI. By blending visual input with STT and TTS, generative AI platforms are enhancing their usability and "natural feel" for users.

I'm a foodie and I want to take a vacation to somewhere warm with a beach, with less than a 4 hour plane trip, not during hurricane season, and for less than $1,000 total. Please provide a table of options and calculate the cost of each to fly and stay for 3 nights. For each destination, list in the table the major tourist attractions.

*Figure 4. An example prompt to a generative AI like ChatGPT.*

## Intent Classification

Once an input is provided by a user, that data is typically sent to an intent classification algorithm; intent classification is a natural language understanding (NLU) technique (NLU is a branch of Natural Language Processing, NLP) used in chatbots, voice assistants, and customer service systems that works to identify the purpose or intention behind an input with the goal of assigning a predefined category or label that represents that user's intent. The accuracy of intent classification depends on the quality and quantity of the labeled data used to train the algorithm; a well-trained intent classification system can significantly improve the user experience by enabling faster and more accurate responses to user queries. Many LLMs contain a fully trained intent classification system, but it may be necessary to implement and train your own when deploying a customized LLM.

*Intents: reservation, airline reservation, hotel reservation, travel, menus, weather, tourism*

*Figure 5. The derived intents for the prompt in Figure 4.*

## Entity Extraction

As data flows into an LLM, important pieces of information must be identified from unstructured text data: this process is called entity extraction. Entities might include named entities, such as people, organizations, and locations, or other types of entities, such as dates, times, and numerical values. In a generative AI system, entity extraction is incredibly valuable because it helps the system understand the context of the text it is generating. By identifying the important entities in the text, the generative AI can generate more coherent and relevant responses by leveraging relationships between different entities identified in the prompt. Entity extraction reduces the amount of data that the AI system needs to process; by only extracting relevant entities from the text, the system focuses on the most important information and avoids data processing of irrelevant text.

a) *I'm a (foodie)[NOUN] and I want to take a (vacation)[NOUN] to (somewhere)[PLACE, VARIABLE] (warm)[TEMPERATURE] with a (beach)[NOUN], with less than a (4 hour)[TIME] (plane)[NOUN] trip, not during (hurricane)[WEATHER] season, and for less than ($1,000)[CURRENCY] total. Please provide a (table)[NOUN] of options and calculate the cost of each to fly and stay for (3 nights)[TIME]. For each destination, list in the (table)[NOUN] the major (tourist attractions)[PLACE].*

b) *Entities: foodie (NOUN), vacation (NOUN), somewhere (PLACE, VARIABLE), warm (TEMPERATURE), beach (NOUN), 4 hour (TIME), plane (NOUN), hurricane (WEATHER), $1000 (CURRENCY), table (NOUN), 3 nights (TIME), tourist attractions (PLACE)*

*Figure 6. a) The entities derived from the prompt in Figure 4. b) The full list of extracted entities from Figure 6a.*

Entity extraction ensures that when a user asks a question about a specific person, place, or thing, the AI system identifies that noun and provides an answer that is relevant to it while not generating responses for other persons, places, or things. Similarly, if a user provides some information about themselves, such as their name or occupation, the AI system can extract that information and use it to personalize responses for that user alone.

## Dialogue State Tracking and Policy Learning

For conversational AI systems, dialogue state tracking and policy learning help guide the back-and-forth between the user and the generative AI. Dialogue state tracking, as the name suggests, keeps track of the current state of a conversation so that the AI system can respond appropriately; the state of the conversation may include the user's preferences, any previous requests, and the user's current intent.

To guide the generative AI in furthering the conversation, policy learning develops a strategy for how to respond based on the current dialogue state. Policy learning algorithms can train the AI on how to respond in different scenarios based on the desired outcome(s) of the conversation.

In combination, dialogue state tracking and policy learning enable a generative AI system to engage in natural and useful conversations with users.

I'm a foodie and I want to take a vacation to somewhere warm with a beach, with less than a 4 hour plane trip, not during hurricane season, and for less than $1,000 total. […]

Before I can offer suggestions, I'll need a bit more information.

Can you tell me where you'll be departing from on your vacation?

Oh, right. I'll be leaving from Atlanta.

Ok. Here are a few destinations you might consider.

| Destination | Overall Cost | Attractions |
|---|---|---|
| San Juan, Puerto Rico | $945 USD | Old San Juan tours<br>Visit El Morro<br>Visit La Fortaleza<br>Paseo de la Princesa promenade<br>… |
| … | … | … |

For each destination, please include travel advisories and popular food items.

| Destination | Overall Cost | Attractions | Travel Advisories | Popular Foods |
|---|---|---|---|---|
| San Juan, Puerto Rico | $945 USD | Old San Juan tours<br>Visit El Morro<br>Visit La Fortaleza<br>Paseo de la Princesa promenade<br>… | None | Mofongo, tostones, pasteles, arroz con gandules, tembleque, and coquito |
| … | … | … | … | … |

*Figure 7. Dialogic responses to the prompt in Figure 4. Note: the AI did not need the prompt repeated after the user indicates their starting location; the prompt is kept in memory in order to formulate the response.*

## Knowledge Graphs

Knowledge graphs provide a structured representation of data where connections (edges) between entities (nodes) describe the type of relationship between the connected entities. For example, in a knowledge graph of geographic locations, a node might represent a particular city, state, or country, and edges might represent relationships such as "located in," "has a population of," or "famous for." For generative AI, knowledge graphs guide the generation of new content based on known entities and their relationships. Additionally, knowledge graphs help generative AI more easily understand context; understanding relationships between entities helps generate more accurate and more relevant content.

## APIs

Application Programming Interfaces (APIs) to external services enable LLMs to integrate data, functionality, and services with other systems and applications. APIs can help an LLM keep up with the latest trends on social media, check the weather, translate text into another language, check flight options, book hotels, or even see restaurant menus. They also can be leveraged to pull in or reference external data for continued training of the LLM.

Some common API integrations include:

**Data integration.** Social media platforms, weather data providers, or news sources to enrich data available to the LLM.

**Service integration.** Natural language processing tools, chatbots, or voice assistants to provide more sophisticated language-related services.

**Language translation**. To translate text from one language to another for multilingual support.

Sentiment analysis. To analyze sentiment in text for social media monitoring or market research.

**Text summarization**. To summarize large amounts of text into a more concise form for news aggregation or content curation.

## Generative AI and Large Language Models (LLMs)

Fed by inputs that have been transformed into sets of intents and lists of entities along with the state of the dialogue and any externally sourced data, the generative AI model can finally get to work. The exact inner workings of these models are outside the scope of this whitepaper[7], so simply describing the interaction of the generative algorithm with a LLM as a “next word predictor” or “next pixel predictor” must suffice.

What matters most, however, is that the generative AI builds a response which cannot be precisely predicted given the set of inputs. Said another way: the algorithm might produce different results under the exact same starting conditions. This semi-stochastic process leads to the algorithm’s perceived creativity – a.k.a. hallucination – sometimes responding with little or nothing to do with its provided prompts.

The output of the generative AI model is then fed back through one or more of the preceding transformation algorithms until it is returned to the user.

[6] Just as speech can be converted to text, text can be converted to speech: text-to-speech (TTS) models are used prior to delivering an output in applications such as virtual assistants, audiobooks, and navigation systems.

[7] Dugas, Daniel “The GPT-3 Architecture, on a Napkin” https://dugas.ch/artificial_curiosity/GPT_architecture.html, last accessed April 28, 2023.

## Outputs

Although LLMs like ChatGPT and Bard return text to the user, a final transformation process can also convert textual output to speech using TTS.[6] The goal of a final TTS transformation is to deliver the text in a way that would create the impression of a natural human-like conversation. Extending this approach, responses can also be delivered using Digital Human avatars or other methods like 3-dimensional displays, hologram projection, or 2-dimensional kiosks. The computational processes to render facial expressions, mouth movements, and lip movements can be extremely complex and time consuming, but technologies exist to expedite the creation and implementation of digital human avatars.

## Key Players

The gold rush for generative AI is underway with companies ranging from technology stalwarts to boutique players – and too-many companies in between – all staking a claim to various platforms and component technologies. With too many to list comprehensively, a small subset of key players, arranged by component, include:

| | Example | Company |
|---|---|---|
| **Inputs & Outputs** | *Digital Humans* | UneeQ<br>Soul Machines<br>Veritone Voice |
| **Transformation** | *NLP/NLU* | RASA<br>Kore.ai |
| | *Cognitive Services* | Microsoft<br>IBM Watson |
| **Generative AI** | | OpenAI<br>Bloom<br>Bard |
| **Large Language Models** | | Pryon<br>Llama<br>ASAPP |

# Best practices for implementation

Ensuring success when adopting generative AI is significantly less complicated than understanding the generative AI algorithms themselves. Below is a 5-step framework we have seen lead to successful implementations of emerging technologies:

DEFINE USE CASES

ESTABLISH YOUR MODEL

DEFINE INPUTS AND OUTPUTS

CONSIDER SECURITY IMPLICATIONS

IDENTIFY SKILLS AND GAPS

## Define Your Use Cases

What led you to consider generative AI for your organization's application? Do you even need to use AI, or is it possible to achieve your goals without generative AI? What impact could generative AI have on your users, your customers, or your employees? Where might it improve upon the ways you currently do things? Think about triaging customer service, customer support, or technical support. For healthcare, your use cases might be related to patient intake or engagement. For retail, your use cases might relate to concierge services or sales. For education, your use cases might touch on student engagement, teacher and professor assistance, student enrollment, or campus tours.

Whatever your use case(s), document the use case irrespective of an underlying technological solution, its potential impact, the various stakeholders who might need to be involved, and how you will measure the success of any initiatives.

## Establish Your Model

If you plan to employ generative AI, consider the three basic models described above: open, private, or blended. Which best fits your organization? How much risk are you willing to take when it comes to your data, your intellectual property, and the way you operate? For organizations, we suggest a private or responsibly blended approach, but you know your organization best.

Now consider the use case(s) that you defined above. Before feeling like you should build everything yourself, consider finding a partner or vendor who brings the level of functionality you require in an easy-to-consume fashion. With so many companies focused on delivering high quality pre-trained and pre-established solutions, you may not need to reinvent the wheel.

## Define Inputs and Outputs

Now define your input and output requirements. How will a generative AI interact with your users? Will the users type or do you plan to support speech and voice? Where and how will users encounter your AI? Is it at a kiosk or display? Via web or mobile? And how will your output be delivered?

We've described several options above, but answers to this question are trending toward more-human-like delivery. Digital Humans are displayed in 2D, 3D, and as holograms. If you're not ready to build that experience, find a partner who brings the level of functionality you require.

[8] Bradley, Tony "Defending Against Generative AI Cyber Threats" https://www.forbes.com/sites/tonybradley/2023/02/27/defending-against-generative-ai-cyber-threats/, published February 27, 2023.

## Consider Security Implications

Data privacy, the protection of intellectual property, and brand reputation must be considered when deploying generative AI systems that will interact with sources external to your organization. Additionally, the threat landscape for generative AI is rapidly evolving – prompt leakage, scraped data rights and usage, and more – and the ability of generative AI to produce human-like content at scale increases the risk of cyber threats like phishing, social engineering, and spear-phishing.[8] Furthermore, generative AI can be coaxed into coding new cyberattacks, can find existing vulnerabilities in digital tools, and may even be able to find and exploit new and unknown vulnerabilities.

Consider finding a partner who can assume some of the risk for these security concerns and who will help mitigate the follow-on effects of any unexpected behavior or breaches; Terms of Service are likely to change as the risk profile of these technologies evolves.

## Identify Skills and Gaps

Lastly, identify the skills and skill gaps of your users, administrators, partners, and other key stakeholders for your generative AI project. Only by first understanding the intended use case(s), establishing a model, defining inputs and outputs, and considering security will you be able to account for the diversity of skills needed to achieve success for your desired outcomes.

# Future Considerations

As generative AI evolves, the way you address the 5 best practices above may change. In fact, the way you adopt and address the best practices should change as the technology evolves. But how will you know when you need to re-evaluate your plan, to iterate your desired outcomes, or to adopt new technologies? Be on the lookout for the things below.

**New inputs. New outputs.** Text, image, and sound inputs and outputs will evolve to include new data types such as virtual environments, 3D models, and more. New types of information will be encoded as text, image, and sound to produce more complex outputs: e.g. musical notation rendered as text or system schematics rendered as knowledge graphs.

**Transparency.** The importance of citing the source of outputs has already led to a drive to include data citations in outputs. Soon, a generative output will include the rationale of the algorithm and its data provenance within its associated model: e.g. opening the black box.

**Improving contextual awareness.** Algorithms will continue to access and consume knowledge graphs to better understand the relationships between ideas; entity maps and taxonomies will contribute to an ever-expansive "index" of the known world.

**Where models live.** Pre-trained models will be embedded in edge devices so the experience of generative AI will become all-pervasive. New inputs will be centrally processed to update the models which will then get re-distributed to the edge.

**Sustainability, green HPC, and an awareness of energy consumption.** A goal of new models will be to require fewer computationally intense resources for training and deployment; already new LLMs like Hyena are striving for energy efficiency.[9]

**Human-AI collaboration.** As generative AI becomes better at some human-like activities, new roles for humans will emerge to work with AI; the need for human companions who provide an ethical foundation to AI may jump start a renaissance in the Humanities.

[9] Ray, Tiernan "This new technology could blow away GPT-4 and everything like it" https://www.zdnet.com/article/this-new-technology-could-blow-away-gpt-4-and-everything-like-it/, published April 20, 2023.

## Summary

Whether or not generative AI ultimately transforms your organization's ways of working remains to be seen, but this set of technologies has tremendous potential to alter your customers' expectations and the ways through which you engage them; generative AI may open avenues for new products, new business models, and new ways of solving problems. As a leader, be curious about that potential and how it might drive your organization's objectives forward. You now have a framework to organize the ever-evolving complexities and someplace to hang your curiosity.

Learn more about Project Helix and Validated Designs for AI

Explore AI Research and Insights

Join the conversation with #genAI



DELL Technologies