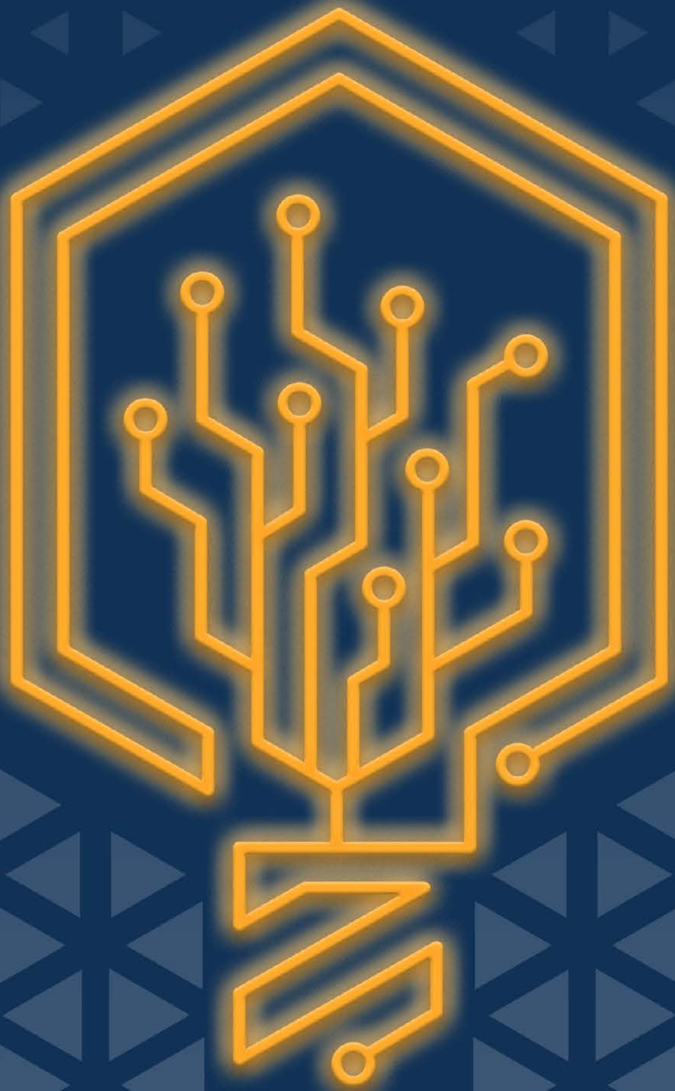
ELDER RESEARCH
DATA SCIENCE · AI · MACHINE LEARNING



Navigating Generative AI:

A Guide for Thoughtful Adoption

Contents

Introduction

Feeling the relentless pressure to jump on the generative AI bandwagon? Whether you're in tech, hospitality, finance, healthcare, or retail, the push to do generative AI is everywhere, and it's not letting up.

From boardrooms to news headlines, it's impossible to ignore the buzz around AI-driven tools like ChatGPT and Midjourney. And executives are pushing for rapid implementation, eager to stay ahead of the competition.

Simultaneously, industry reports tout the transformative power of generative AI, making it seem like an essential component of any forward-thinking organization. Competitors are quick to announce their own AI initiatives, creating a sense of urgency and a fear of falling behind.

Inside your organization, the pressure is twofold. On the one hand, team members are eager for tools that streamline workflows and make their jobs easier. On the other hand, there's a palpable anxiety about both staying ahead of the generative AI learning curve and the potential for AI to automate roles out of existence. Striking the right balance between leveraging AI for efficiency and ensuring job security is a delicate task.

Adding to this complexity are concerns from legal, risk, and compliance teams. The potential for regulatory scrutiny, ethical dilemmas, and data privacy issues means that any AI initiative must be approached with caution and a clear strategy for risk mitigation.

Whether you're at the starting point of using generative AI or a bit further down the path, this is a guide for thoughtful and effective adoption. By maintaining a critical and informed perspective, you can leverage the power of generative AI to drive innovation and efficiency while addressing the concerns and pressures from all sides.

Understanding Generative AI

Before we dive in, here's a quick refresher. Generative AI refers to a class of algorithms that can create new content—like text, images, or code—based on patterns learned from existing data. Key technologies include Large Language Models (LLMs) such as GPT-4, image generation models like DALL·E, and coding assistants like GitHub Copilot. These models are trained on vast datasets, enabling them to generate coherent and contextually relevant outputs.

Historical Context and Evolution

Generative AI isn't a new concept; its roots can be traced back to early AI research in the 20th century. However, recent advancements in computational power (e.g. faster and cheaper GPUs), the availability of large datasets (due to cheaper storage and more internet usage), and better algorithms (e.g. attention-based neural networks) have significantly accelerated its development. The release of models like GPT-3 marked a turning point, demonstrating the ability to generate highly sophisticated text and images. These advancements have opened new possibilities for businesses across various industries, including yours.

"Generative AI refers to a class of algorithms that can create new content–like text, images, or code–based on patterns learned from existing data."

Key Technologies in Generative AI

Understanding the key technologies behind generative AI can help you determine how to apply them effectively.

Large Language Models (LLMs)

These models, such as Gemma and Claude, are capable of understanding and generating human-like text. They are trained on diverse datasets, enabling them to perform tasks ranging from writing and translation to question answering and summarization.

Image Generation Models

Tools like DALL·E and Midjourney can create realistic images from textual descriptions. These models are revolutionizing fields like marketing, design, and entertainment by enabling the rapid creation of visual content.

Coding Assistants

AI tools like GitHub Copilot assist developers by generating code snippets, automating repetitive tasks, and providing intelligent suggestions. These models enhance developer productivity and accelerate software development.

Vector Databases

These databases, such as Pinecone and Weaviate, offer an efficient way to store and search through different materials. For instance, many text documents can be converted to embeddings and stored in a database, allowing for extremely efficient search across the database.

Infrastructure

Navigating the infrastructure needed for generative AI can feel daunting, especially with the pressure to get it right. The infrastructure requirements can vary based on the model's size, complexity, and how you plan to use it. Here are some key considerations to help guide you through this process.

Cloud-Based Solutions

Large models like Claude and GPT-4o typically run on cloud-based platforms due to their substantial computational requirements and high barriers to set up and maintain. These platforms offer the scalability and resources needed to handle extensive datasets and complex computations. However, using cloud services can raise concerns about data privacy, security, and compliance, particularly when dealing with sensitive or proprietary information.

API Solutions

For small-scale use cases or proofs-of-concept, API-based solutions provide a commitment-free way to take advantage of state-of-the-art generative AI models without investing large amounts of time and money upfront.

On-Premises Deployment

For organizations with stringent data security requirements, deploying AI models on-premises can be a viable option. This approach ensures all data remains within the organization's secure environment, mitigating risks associated with data breaches and unauthorized access. However, setting up and maintaining the necessary infrastructure can be costly and resource intensive.

Edge Computing

In some cases, smaller and more efficient models can be deployed on edge devices, such as laptops, mobile phones, or IoT devices. This approach allows for real-time data processing and decision-making without relying on constant cloud connectivity. Edge computing can be particularly useful for applications requiring low latency and high privacy.

Tradeoffs Between Size, Cost, and Convenience

There is often a tradeoff between the size and cost of a model and the convenience it offers. Open-source models that are free to use and can run on a laptop are becoming increasingly sophisticated and are frequently updated. In contrast, interfaces like those offered by OpenAI and Google are large, powerful, and user-friendly but come with higher costs, resource requirements, and data exposure.

Identifying Potential Benefits

As you explore ways to integrate generative AI within your organization, you may have already seen some promising results. **But you're aiming to reap long-lasting benefits that drive your team forward. And that starts with identifying the key ways generative AI can be applied.**

We categorize the primary use cases into three broad areas: knowledge retrieval, content creation for non-technical cases, and technical assistance (coding). By tapping into these areas, you can boost efficiency, enhance the quality of your team's work, and ultimately deliver more value to your organization. Let's explore the benefits of each use case.

Knowledge Retrieval

Generative AI can help you get the most out of your internal knowledge bases, making it easier to find and use in-house resources. By serving as advanced search engines, these tools can provide your team with contextually relevant information from vast data repositories, improving decision-making and operational efficiency.

For example, an AI-driven knowledge retrieval system can quickly surface relevant documents, emails, and reports, enabling employees to find the information they need without sifting through large volumes of data.

Additionally, Retrieval-Augmented Generation (RAG) models combine the strengths of retrieval-based methods with generative models to provide more accurate and contextually aware responses. RAG models retrieve relevant information from a database and use it to generate more precise and informative outputs, making them particularly useful for knowledge-intensive tasks.

Because RAGs require a database, there are also several important considerations related to infrastructure, including selecting the right vector storage types, configuration, and admins. While they can provide very efficient and specified knowledge retrieval, just having a folder full of PDFs doesn't mean you're ready to implement a RAG database.

KEY TAKEAWAY

Generative AI, particularly Retrieval-Augmented Generation (RAG) models, can help your team quickly access relevant internal resources. RAG models can boost decision-making and operational efficiency, but they need the right infrastructure to be effective.

Content Creation

Generative AI can help with writing, improving workflows, and creating content. These tools can take care of routine tasks so your team has more time to focus on important, strategic work. Additionally, workflow augmentation tools can streamline repetitive and tedious business processes.

For example, marketing teams can use AI to generate engaging content for campaigns, reducing the time and effort required for content creation while maintaining high-quality standards. And workflow augmentation tools can assist administrative staff by automating scheduling, email responses, and report generation, significantly improving productivity.

Technical Assistance (Coding)

Tools like GitHub Copilot act as smart assistants for developers, boosting productivity by automating code generation and offering intelligent suggestions. By handling routine coding tasks, these tools allow developers to focus on more complex and creative work, helping your team bring new software solutions to market faster.

For example, AI can help generate boilerplate code, suggest optimizations, and identify potential bugs, improving the overall efficiency and quality of software development.

In practice, developers can use AI to quickly prototype new features, troubleshoot code issues, and streamline their workflows. By automating repetitive tasks, coding assistants enable developers to focus on innovative solutions and critical problem-solving, driving faster and more efficient software development.

The technical assistance also includes traditional natural language processing tasks, such as document clustering or classification. Data scientists undergo painstaking efforts to analyze text, and large language models offer impressive capabilities with little additional effort. For example, in classification, constrained generation can be a very useful control to place around models. This prunes all possible predictions of the model to only a constrained subset the user has identified as appropriate.

Selecting the Right Use Case

Let's pause for a second. With so many use cases for generative AI, we know how easy it can be to get bogged down in all the details. In reality, the best path forward starts with a close look at your organization's goals and capabilities. It's about finding what fits and what doesn't, ensuring you make decisions that truly benefit your team and projects.

Perhaps the most important consideration to keep in mind when evaluating the use case is that generative AI excels in scenarios where there is a clear tradeoff between speed and accuracy. When evaluating potential projects, think through how this tradeoff impacts your business needs and goals. Here are some key criteria to consider.

Business Value and ROI

Prioritize projects that promise clear business benefits and a high return on investment. Consider the potential impact on revenue, cost savings, and overall operational efficiency.

For instance, implementing AI-driven chatbots for customer support can significantly reduce operational costs by automating responses to common inquiries. While these chatbots may occasionally provide inaccurate responses, the tradeoff is much faster response times, which can lead to a net positive impact on customer satisfaction.

By analyzing the cost savings from reduced human labor and the improved customer experience, you can calculate the ROI of such projects and make informed decisions about their value.

But here's a warning: There are an increasing number of vendors that provide very pretty user interfaces with the same basic modeling capability that is freely available. This isn't to say that user interface isn't important, but be aware that an impressive sales pitch doesn't necessarily equate to impressive technical capability.

> "Prioritize projects that promise clear business benefits and a high return on investment by considering their impact on revenue, cost savings, and overall operational efficiency."

Feasibility and Alignment

Assess the technical feasibility of the project and ensure alignment with organizational capabilities and strategic goals. Evaluate whether the necessary data, infrastructure, and expertise are available to support the implementation of generative AI. For example, a company with a robust IT infrastructure and a wealth of historical data might be well-suited to implement AI-driven predictive analytics.

Pilot Testing and Evaluation

Conduct rapid pilot tests to validate the potential impact and refine the approach before full-scale implementation. The pace of technological change in AI is swift, and taking too long to test can mean missing out on the speed advantages generative AI offers.

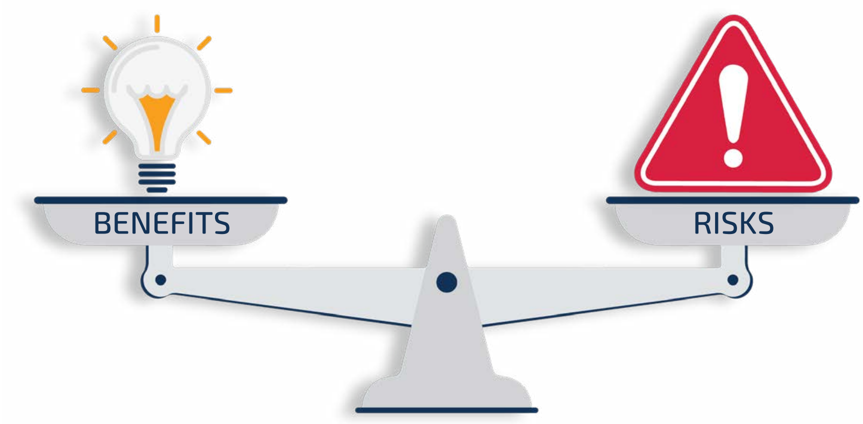
Pilots allow you to experiment with different models, gather feedback, and make necessary adjustments without committing significant resources. For instance, a pilot project for an AI-driven content generation tool can help identify potential challenges and areas for improvement before rolling it out company-wide.

Realizing the benefits of AI requires a willingness to move quickly, learn fast, and iterate.

Speed vs. Accuracy Tradeoff

The tradeoff between speed and accuracy is crucial in determining how suitable generative AI is for given use cases. Here are two scenarios where this tradeoff is particularly evident.

1. **Low-Stakes Outputs:** Use generative AI when the accuracy of the output is not critical. For example, brainstorming sessions are renowned for having plenty of content that is fringe, infeasible, or just plain wrong. And it's often a slog for humans to come up with enough ideas worth pursuing. Generative AI can produce a wide variety of content so quickly that using it to brainstorm can be very productive as long as you accept many of the outputs will be incorrect.
2. **Known Outputs with Validation:** Use generative AI when you can validate the outputs against known results. For instance, when a developer is using it as a coding assistant it's important for them to write and run tests to ensure the code is doing what it should. An example is turning transactional data into summary metrics. Here, the developer knows what the output should look like and can rigorously test the AI-generated code to ensure its accuracy. This approach leverages the speed of AI while mitigating the risk of errors.



Choosing the Right Model Size

When experimenting and testing, it's best to choose the smallest model that can accomplish the job. Small models are faster, cheaper, and easier to deploy locally, making them ideal for rapid iteration and experimentation. Once you have validated the concept and refined the approach, you can consider scaling up to more powerful models if necessary. This strategy helps manage costs and risks while allowing for flexible and agile development.

Any model that uses stochastic processes can make mistakes.

Generative AI is stochastic because it can produce different outputs each time—even when given the same inputs. Therefore, it's vital to approach generative AI outputs with a healthy dose of skepticism. Always take the time to critically evaluate and validate these outputs to make sure they meet your standards.

Ensuring Quality & Mitigating Risks

As you work to determine the generative AI use cases that will drive business value, **there are significant risks that you need to consider**.

Legal and Ethical Considerations

Stay informed about the evolving legal landscape and ensure compliance with relevant regulations. Ethical considerations should guide the use of AI to prevent misuse and bias. For example, ensure the data used to train AI models is ethically sourced and that the models do not perpetuate harmful biases. Additionally, be aware of intellectual property issues related to AI-generated content.

For some large models, the size and variety of the training data make it infeasible to identify all potential ethical issues. Instead, conduct post-hoc tests to identify and minimize unethical or biased outputs. Well-placed guardrails can help mitigate risk of those undesirable outputs.

> "Ethical considerations should guide the use of AI to prevent misuse and bias, ensuring that data used to train models is ethically sourced and that models do not perpetuate harmful biases."

Data Privacy and Security

Protect sensitive information by following best practices for data handling and storage. Avoid sharing proprietary data with external AI tools without proper safeguards. Implement robust security measures to prevent data breaches and unauthorized access. For instance, consider encrypting data and using secure APIs for data transfer.

When considering data privacy and security, it is important to note that smaller, open-source models are increasingly competitive with larger models and can often run locally. This allows for greater control over data privacy, as sensitive information does not need to be transmitted to external servers. For example, many of the models from Hugging Face can be deployed on-premises, providing similar capabilities to larger, cloud-based models without compromising data security.

Another option is to use managed services like Amazon Bedrock. Amazon Bedrock is a fully managed service that offers a choice of high-performing foundation models (FMs) from leading AI companies. Using Bedrock, you can experiment with and evaluate top FMs for your use case, privately customize them with your data using techniques such as fine-tuning and RAG, and build agents that execute tasks using your enterprise systems and data sources.

Since Amazon Bedrock is serverless, you don't have to manage any infrastructure, and you can securely integrate and deploy generative AI capabilities into your applications using familiar AWS services. However, Bedrock is just one example of the many evolving tools and platforms available, so staying updated on the latest options is crucial.

Accuracy and Reliability

Remember that models using stochastic processes, like all generative AI, *will* yield incorrect results. Establish robust validation and monitoring processes to maintain the accuracy and reliability of AI outputs. For example, implement automated testing frameworks to regularly evaluate the performance of AI models and detect any deviations from expected behavior.

Generative AI models like CriticGPT can help assess the accuracy and reliability of AI outputs, but they can't be counted on solely. Be aware you'll also need people dedicated to designing and monitoring the frameworks that evaluate your AI outputs.

KEY TAKEAWAY

Protect sensitive information by using best practices for data handling and storage, opting for secure, local AI models or managed services to maintain data privacy and security.

Ownership and Control

Consider the implications of relying on third-party providers for AI technologies. Evaluate options for developing and maintaining in-house models to retain control and flexibility. While third-party AI tools can offer quick deployment, in-house models can be tailored to specific business needs and provide greater control over data privacy and security.

For example, although third-party solutions like OpenAI's GPT-4o offer powerful capabilities, developing in-house models by tuning smaller, open-source models can provide similar benefits with added control over customization and data management. The state-of-the-art small models are constantly changing, and checking the `transformers` repository can give you a great starting point on potential small models for your use case. This approach allows you to align AI capabilities with your organization's unique requirements and ensure compliance with internal security policies.

Achieving and Measuring Results

For you to maximize the impact of generative AI initiatives, **it's crucial to set realistic expectations and define clear success metrics**. Here are some key steps to achieving and measuring results.

Defining Success Criteria

You're feeling pressure from around the organization to dive into generative AI, but before you start making API calls to ChatGPT, make sure you have defined and aligned on metrics for success. This is the most important point to consider when embarking on a generative AI project.

"Before you start making API calls to ChatGPT, make sure you have defined and aligned on metrics for success. This is the most important point to consider when embarking on a generative AI project."

Establish specific, measurable objectives to track the performance and impact of AI implementation. Consider metrics such as accuracy, efficiency, user satisfaction, and financial impact. For instance, measure the reduction in customer support response times after implementing an AI-driven chatbot. Additionally, consider metrics related to speed, such as the time saved in creating marketing copy or developing code.

While speed often isn't tracked rigorously, estimating baseline speeds for various tasks can help measure the impact of AI. For example, compare the time taken to manually create marketing copy versus using an AI tool.

It's important to focus on organizational speed rather than individual performance metrics. Improved individual performance doesn't always translate to improved organizational performance. Metrics should capture the overall efficiency and effectiveness of the organization.

Continuous Improvement

Use feedback and results from your initial projects to refine and enhance your AI applications, ensuring continuous improvement. Implement a feedback loop to gather insights from users and stakeholders and make necessary adjustments to improve the effectiveness of AI solutions. For example, regularly update AI models with new data to maintain their relevance and accuracy.

Continuous improvement is essential for adapting to evolving business needs and technological advancements. Regularly asking for feedback from users and stakeholders helps identify areas for enhancement and ensures AI solutions stay aligned with goals.

Scaling and Integration

Once validated, scale successful projects and integrate AI solutions with existing systems and processes. Develop a phased implementation plan to gradually roll out AI solutions across your organization, ensuring minimal disruption to business operations. For instance, start by integrating AI-driven analytics into a specific department before expanding to other areas.

Speed is a critical factor in scaling AI initiatives. Rapidly pilot testing and validating AI projects allows organizations to quickly realize the benefits of AI. However, it's important to understand tradeoffs between speed and quality assurance. Scaling should be done in a controlled and systematic manner, with a focus on maintaining the required level of accuracy and reliability for AI outputs.

KEY TAKEAWAY

Scale successful AI projects gradually with a phased implementation plan, balancing speed with quality assurance, and ensure that the necessary infrastructure is in place to support large-scale integration.

These costs and efforts are easy to overlook at the beginning of a project—especially in large organizations. Take care to understand where the tool will be used. An internal tool built from an open-source model that helps a small number of developers might be easy to integrate. A RAG for thousands of employees to review documentation is going to need dedicated infrastructure behind it.

Implementing Initiatives

As you balance existing data initiatives with ideas for new ones, we realize implementing generative AI may feel like no small task. **This structured approach will help you determine how generative AI can enhance your team's capabilities**—ultimately propelling you toward your goals.

Initial Assessment and Readiness

Evaluate your organization's readiness and identify key areas where generative AI can provide value. Conduct a comprehensive assessment of the current state of AI adoption, including data infrastructure, technical expertise, and strategic goals.

While top-level executives are often ready and eager to implement AI solutions, it's crucial to focus on the tactical implications for end users. Consider how end users will integrate AI-generated outputs into their existing workflows and processes. Assess the availability of clean and relevant data for training or tuning AI models, and make sure the necessary infrastructure is in place to support AI initiatives.

> "While top-level executives are often ready and eager to implement AI solutions, it's crucial to focus on the tactical implications for end users."

Proof of Concept

Start with small-scale pilot projects to test the feasibility and impact of AI applications. Prioritize use cases where the data is less sensitive and stakeholders or users are more excited and engaged.

For example, piloting an AI-driven marketing campaign can be effective because the data involved is typically less sensitive, and marketing teams are often enthusiastic about new tools that can enhance creativity and efficiency. Be prepared for some iteration and potentially humorous or unexpected results, which can be valuable learning experiences.

Scaling and Integration

Once validated, scale successful projects and integrate AI solutions with existing systems and processes. Develop an implementation plan that includes timelines, resource allocation, and risk management strategies. A critical aspect of this phase is deciding where the model will run, as this can inform the type of model you select.

For instance, small, local models may be suitable for on-premises deployment, while more powerful models might require API calls to cloud-based services. This decision will affect data privacy, security, and performance considerations.

Roles and Responsibilities

Define roles and responsibilities for successful implementation. Ideally, establish a cross-functional team with representation from IT, data science, business units, and senior management. Clearly outline the responsibilities of each team member and ensure effective collaboration and communication. For instance, assign a project manager to oversee the implementation and coordinate with various stakeholders.

Importantly, instill a culture of curiosity and proactive thinking across the team. While it's easy to get enamored with positive results, it's crucial to critically evaluate all outputs and maintain realistic expectations. Encouraging team members to approach AI-generated outputs with a critical eye will help ensure the technology is used effectively and responsibly.

Part of the responsibility of each team member is to rigorously track results, not just the outputs of the AI models but also the results of process changes. By closely monitoring the impact of AI on workflows and outcomes, you can identify areas for improvement and ensure AI implementations are delivering the intended benefits. This comprehensive tracking helps in making data-driven decisions and adjusting strategies as needed to optimize performance and value.

KEY TAKEAWAY

To effectively implement generative AI, start with a comprehensive assessment of readiness and infrastructure, run small-scale pilot projects to test feasibility, and gradually scale successful initiatives while defining clear roles and responsibilities for team members.

Case Studies and Examples

Your wheels are probably turning at this point. Let's explore how generative AI has been applied across different industries, focusing on practical benefits and lessons learned.

While no single example can show the full picture, we hope these case studies provide useful insights.

Case Study: Knowledge Retrieval

Overview:

A large government agency was struggling to train customer service staff to reference the correct documents from a vast and complex knowledge base. They wanted to improve the speed and accuracy of information retrieval for training purposes.

A Good Generative AI Use Case?

This has great potential. There is a large amount of data available; there is a way to measure ROI (time involved training staff); the effort aligns with the organizational goals; and there is minimal risk in data security, since all of the data is publicly available government data.

Challenges:

It's easy to measure speed, but it's difficult to measure accuracy of retrieved documents. Several experts needed to manually label whether certain documents were relevant for specific customer questions in order to validate whether the generative AI tooling had any impact on accuracy.

The infrastructure was also a challenge. All of the data is publicly available, but the compute resources are on-site in air-gapped environments. This meant that any model had to ensure certain compliance to be used, and it was difficult to iterate on different model sizes and architectures.

Results:

The speed was significantly increased, and the change in accuracy was negligible. This provided a positive ROI, as agents were able to retrieve relevant documents in less time without sacrificing accuracy. This is a mixed result, as ideally the accuracy would have improved as well. With a process in place, it will be easier to continue to label data, and thus test and tune models and measure the accuracy. If the accuracy reaches a certain threshold, this can be made publicly available.

Lessons Learned:

This emphasized that speed is often the advantage offered by generative AI (Page 10) and highlighted the challenges of on-premise deployments (Page 6).

Case Study: Content Creation

Overview:

A design company needed to help their customers in the design process. How can they capture customers' preferred styles and offer appropriate design suggestions? The current process either shows only popular designs in the same category or requires substantial time and expertise.

A Good Generative AI Use Case?

There is certainly potential. There is a moderate amount of data available; there is a way to measure ROI (e.g., changes in customer satisfaction, upselling or cross-selling based on suggestions, reduced labor cost from design experts); the effort is aligned with organizational goals; and there is a large customer focus on speed: "I want my design suggestions now!"

Challenges:

There is risk in the inability to control the outputs that are delivered to customers. The design company reputation is at risk if they deliver inappropriate design suggestions.

Results:

The opportunity was divided into two phases:

1. Using generative AI to classify customer images into pre-defined "styles."
2. Generating new images based on the customer preferences.

The first case was completed with generative AI, and the natural guardrails were ensuring designs could only be classified into specific styles. Anything else would be manually classified.

The second case was implemented internally and delivered to design experts but was not delivered directly to customers. This mitigated the risk of customer-facing generative AI while reaping some of the creative and time-saving benefits.

Lessons Learned:

While the overall tool didn't do everything that was desired by delivering directly to customers, having clear success criteria (Page 14) in terms of time-saving and cross-selling meant the project was an overall success. This helped in scaling across multiple design areas and keeping stakeholders engaged.

Case Study: Technical Assistance

Overview:

A large organization had several developer teams that were working on the same projects. This caused delays and frustrations with code compatibility and documentation.

A Good Generative AI Use Case?

The overall project is high priority to the organization, and technical assistance is something that generative AI can handle well.

Challenges:

The organization wanted to ensure the code being developed to process and analyze sensitive data remained in their secured cloud environment.

While inconsistent code formatting and documentation is frustrating for development and testing teams, it isn't always straightforward to measure the ROI of improvements. The most appropriate way was by measuring the relevant ticket items related to code integration.

Results:

The time needed to integrate code was markedly reduced using small, local generative models (Page 13). Developers tested several models that were able to take code as input, rewrite in a consistent style, and include robust documentation. There was no quantitative measurement of style or documentation, but the qualitative results were that developers had a much easier time working across teams.

Lessons Learned:

This emphasized that speed is often the advantage offered by generative AI (Page 10) and that small models can often be very effective (Page 13). Since code integration was already subject to testing, there was little risk of incorrect results in style or documentation.

Conclusion

As you're well aware, generative AI holds significant potential for transforming business operations and driving innovation. **We hope this guide has given you the impetus to let the AI bandwagon roll by while you take a moment to think strategically.**

We know you're aiming to make thoughtful choices that work with your team and your overall goals. As you keep planning, here are three takeaways for leveraging generative AI effectively.

Expect and Manage Mistakes:

Generative AI, regardless of the use case, will make mistakes. This is inherent in models that use stochastic processes. It is crucial for everyone involved to understand this and approach AI-generated outputs with a healthy dose of skepticism. Implement robust validation processes and maintain realistic expectations to manage and mitigate the impact of these inevitable errors.

Focus on Organizational Performance:

While generative AI can reliably speed up and improve individual tasks, it is essential to track and measure its impact on overall organizational speed and performance. Identify bottlenecks and areas where your organization is slow, and consider how generative AI can help streamline these processes. By focusing on organizational metrics rather than individual performance, you can ensure AI initiatives contribute to broader strategic goals.

Experiment with Small Models First:

Big models may grab headlines, but small models offer significant advantages: they are faster, cheaper, more secure, and easier to experiment with locally. If you have the necessary personnel, start by running experiments with small models to achieve quicker and more cost-effective iterations. This approach allows you to refine your AI applications before deciding which solutions to scale and integrate into larger systems.

By remembering these points, generative AI can be the catalyst for meaningful improvements in your organization. The path may include some stops and starts along the way, but these tips can help you navigate effectively and responsibly.

As you plan for what's next, strategic thinking will help ensure generative AI produces the most value and momentum for your team. We wish you success in every step ahead.

Appendix

Glossary

Generative AI: A class of AI algorithms that create new content based on patterns learned from existing data.

Large Language Models (LLMs): AI models capable of understanding and generating human-like text that are trained on large datasets, typically with many billions of parameters.

Small Language Models (SLMs): Similar to LLMs, but significantly fewer parameters, typically ranging from a few million to a few billion. Less memory and computational power required.

Stochastic Processes: Random processes that can yield different outcomes given the same initial conditions.

Vector Database: A collection of data stored as mathematical representations that correspond to objects. Very fast and efficient search and retrieval.

Retrieval-Augmented Generation (RAG): A process used to improve the output of generative AI models by specifically referencing a known knowledge base outside of its training data sources.

Resources and Further Reading

"Approaches to Regulating Artificial Intelligence: A Primer" | National Conference of State Legislatures

"Attention is All You Need" | Cornell University

"Evaluating Large Language Models" | Medium

"Insights on Generative AI" | Elder Research

"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" | Cornell University

"Training Compute-Optimal Large Language Models" | DeepMind

Contact Information

Want to take the next step with generative AI? Let's see how we can come alongside you.

Reach out to us at elderresearch.com/contact

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING