Leading through innovation: 7 Strategic considerations for CIOs implementing Generative AI

In the era of Generative AI, Chief Information Officers (CIOs) stand at the forefront of a technological renaissance, shaping the future with every strategic decision.

With the rapid pace at which the frontier of technology is moving, leading through innovation has never been more important for CIOs. As the leaders of technological transformation, CIOs are increasingly turning to generative AI to reimagine traditional workflows and spur unprecedented growth. However, the implementation of generative AI is not without its challenges. It requires strategic foresight, meticulous planning, and a deep understanding of the technology's potential and shortfalls. This paper aims to guide leaders on this journey, presenting a comprehensive set of strategic considerations that should be considered when integrating generative AI into their business.

Understand your long-term strategic goals and how Generative AI aligns with them

In the current digital age, generative AI has become a buzzword that many organisations are eager to incorporate into their business strategies. However, it's crucial to understand that while this technology holds immense potential, it's not a silver bullet that can solve all business challenges. Many IT leaders are buying into the hype of generative AI without fully understanding how the technology works and how it can truly transform their business. This lack of understanding often results in disappointing outcomes to generative AI projects. The consequence of this is that it can lead to wasted resources, unmet expectations, and a loss of confidence in the technology. It's essential to comprehend that generative AI is a tool that, when used correctly, can drive innovation, efficiency, and growth. However, without the proper understanding of its capabilities and strategic application, it can lead to a lack of trust and disillusionment, ultimately resulting in missed opportunities for the business.

When considering the implementation of generative AI, it's vital to look at your organisation's long-term strategy. The use of AI should align with the broader goals of the organisation. Figure out what drives your organisation and then identify the use cases where AI can add value and contribute to achieving your strategic objectives.

Look for the low hanging fruit, internal use cases that offer the highest business value yet have the lowest level of complexity should be adopted first. This approach allows for quick adoption and testing, enabling the organisation to gain early wins and learn from the implementation process. It's

important to develop a roadmap capturing the journey from the low complexity initiatives through to the more complex high value solutions. The road map should look to build capability as the organisation delivers each solution while also building confidence and garnering support from the wider organisation along the way.

Ensure your data captures your business information – well!

As with any AI model, the quality of the data that goes into your model will determine the success of your solution. It's important to assess the completeness of your unstructured data, whether it be internal knowledge bases or policy documents. Ensuring your data captures all your business information at a sufficient quality is crucial. Business value from generative AI models is fully realised when the models have access to your organisations data. Confirm your dataset has all the information required for the identified use case and the types of prompts that will be asked of the model.

In the event you find your existing datasets do not capture all your business information, think about how you can systemize the information of your organisation. Engage the key stakeholders of your organisation who are responsible for the management of information, define robust knowledge management systems that capture and organise your knowledge effectively. By ensuring your data accurately captures your business information, you're able to set a solid foundation for getting the most value out of generative AI.

Get a lay of the land

OpenAI's GPTs, Google's Gemini, Anthropic's Claude, Mistral's Large, Metas' Llama so many AI models to choose from! With the extreme competition in the frontier model space, we are seeing families of generative AI models constantly outperforming one another week to week. What this ultimately means for you as a consumer of these models is there is healthy competition. These frontier models have come to be known as LLM's. Start with the family of models supported by your cloud provider. If you're on Azure, the preferred provider is Azure OpenAI GPTs. If you're on AWS you can consume Anthropic's Claude, and if you're on GCP you can start with the Gemini family. Starting with the preferred partner of your cloud provider does not mean you are locked into any given provider. Each cloud provider is constantly evolving to enable hosting of various foundation models as well as open-sourced models (Llama, Mistral, Cohere). These platforms also offer the flexibility to fine-tune custom models - more on this later.

Fig.1 Cloud providers and their Generative AI partners

So, you've determined your cloud provider and preferred foundation partner to start with what's next? While the generative AI foundation models have all been trained on vast amounts of data and have a tremendous amount of knowledge, these models don't have your business' data and context. Introducing retrieval augmented generation (RAG).

The primary role of a RAG system is to retrieve the most relevant business data which will aid in responding to a given input prompt. This retrieved information will be provided to the LLM as reference material in which the model will use to generate a response to the given prompt. Building a strong RAG pipeline is critical to the success of implementing generative AI within your organisation. Look for existing frameworks or accelerators which have developed standard processes for implementing state of the art RAG solutions, this will enable you to fast track the benefits of generative AI.

To fine tune or not to fine tune? Earlier we mentioned cloud providers are evolving to enable organisations to fine-tune their own generative AI models. Fine tuning is the process of training a model to improve performance on specialised tasks, for example question answering, text summarisation, or classification. Most large generative AI models, such as GPT-4 or Llama 3 will be fine-tuned for question answering out of the box. This is why early progress can be achieved with off-the-shelf LLM's combined with your business data, a process made possible with RAG. If a RAG solution doesn't provide the performance you need, or if you need to control the language style or formatting of the model's outputs, you can resort to fine-tuning. An example of fine-tuning is through X.ai's LLM, Grok, which embedded humor and wit in its responses. This was done by providing the LLM with training data of question answer pairs that demonstrated the expected response style. However, it is important to note that fine-tuning can be a costly and time-consuming process and should be considered as a last resort if RAG solutions fail to meet performance expectations.

Think about transforming traditional workflows with the introduction of AI agents

As an IT leader, it's essential for you to challenge the traditional understanding of how business workflows are conducted within your organisation. The advent of generative AI provides an opportunity to reimagine how teams operate, moving beyond the conventional one-step approach to a more dynamic, iterative process. With 2023 we saw the rapid growth of LLM's, most notably OpenAI's GPT-3.5 & GPT-4 which exhibit human-level performance on various professional and academic benchmarks. What does this mean? AI can now understand and generate human-like text with unprecedented accuracy. But that's 2023, with how rapidly technology is evolving, that might as well have been 2003.

Looking at emerging trends within the field we can see the rise of AI agents. The current mode of usage of LLM's is through what's known as zero-shot mode. The idea is that the model will generate a response in one go. It's essentially like asking someone to compose a business case in one hit, typing continuously without stopping, and expecting a polished document on the other side. We ask a question, and the AI generates an answer in one 'stream of consciousness'. As humans, we think, reflect, plan, and research to achieve an outcome. We might go to the internet to do some research or grab a tool to access data and generate insights. We might then break down a task into a series of

subtasks, these are all steps in the workflow for how we would traditionally achieve a particular outcome. This is where AI agents come in.

AI agents allow complex problems to be broken down into several simpler problems each of which may be solved by a tool – or agent – that is specific for the purpose. In this model, AI systems don't just respond to prompts or questions, rather they actively engage with a series of tasks, have agency to interact with tools, to retrieve information, offer drafts, and improve upon them iteratively. Earlier we mentioned GPT-3.5 & GPT-4 have demonstrated human-level performance on various benchmarks. To put this in context GPT-3.5 can score 48.1% on the human-eval benchmark, whereas GPT-4 is able to score 67%. Researchers at DeepLearning.AI synthesized results from several research papers, exploring the benefits of agentic workflows, and found that if you broke down the tasks to use an iterative agent workflow, you could get a GPT-3.5 model to achieve scores over 95%[1], clearly demonstrating the power of AI agent workflows.
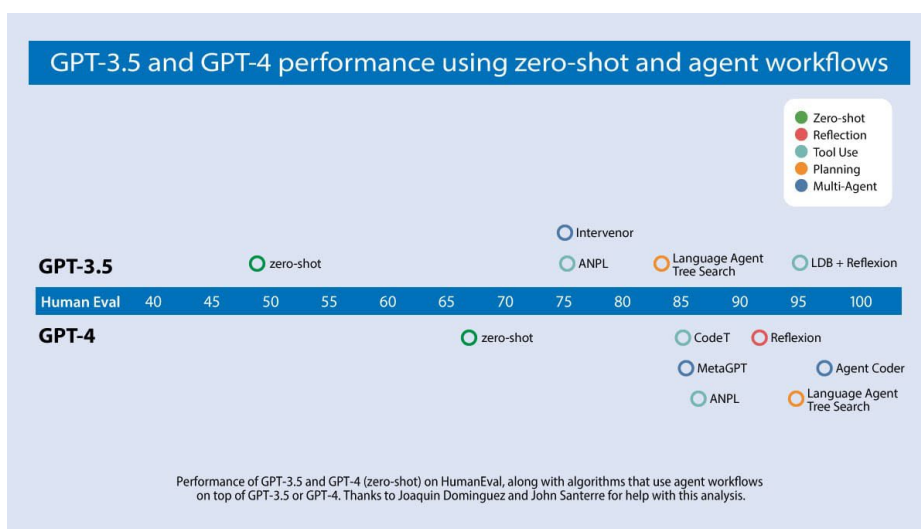

Fig.2 GPT benchmark performance with agent workflows[1]

One of the main areas in which generative AI is transforming organisations is through saving time, allowing your employees on the ground to find the information they need in a faster more efficient manner. While these improvements are necessary and should be realised immediately as low hanging fruit they may only result in incremental improvements. As an innovative leader within your organisation look for the opportunities where generative AI can support decision making as opposed to time saved. Instead of having a knowledge discovery tool, how can you leverage generative AI agents to generate business critical reports within your organisation to support decision making?

While it is early days, and the AI community is only just getting started, if the current rate of technological advancement is anything to go by, AI agents will be transforming businesses before we know it. The key here lies in leading from the front, as businesses embrace the benefits of AI, organisations will be looking to the future to innovate and stay ahead of the curve.

Don't forget to measure success

One of the early benefits of generative AI is its ability to boost productivity, reduce the time taken to achieve certain tasks, and ultimately reduce costs. However, applying standard performance metrics can be challenging due to the inherent non-deterministic nature of the outputs produced by generative AI models. Venture capital firm a16z, interviewed 70 top enterprise leaders understanding how they were using and buying generative AI, and found that 60% of leaders were not measuring ROI at all.2. They found that many leaders were taking a leap of faith trusting their employees when they say they are more productive with achieving their work.
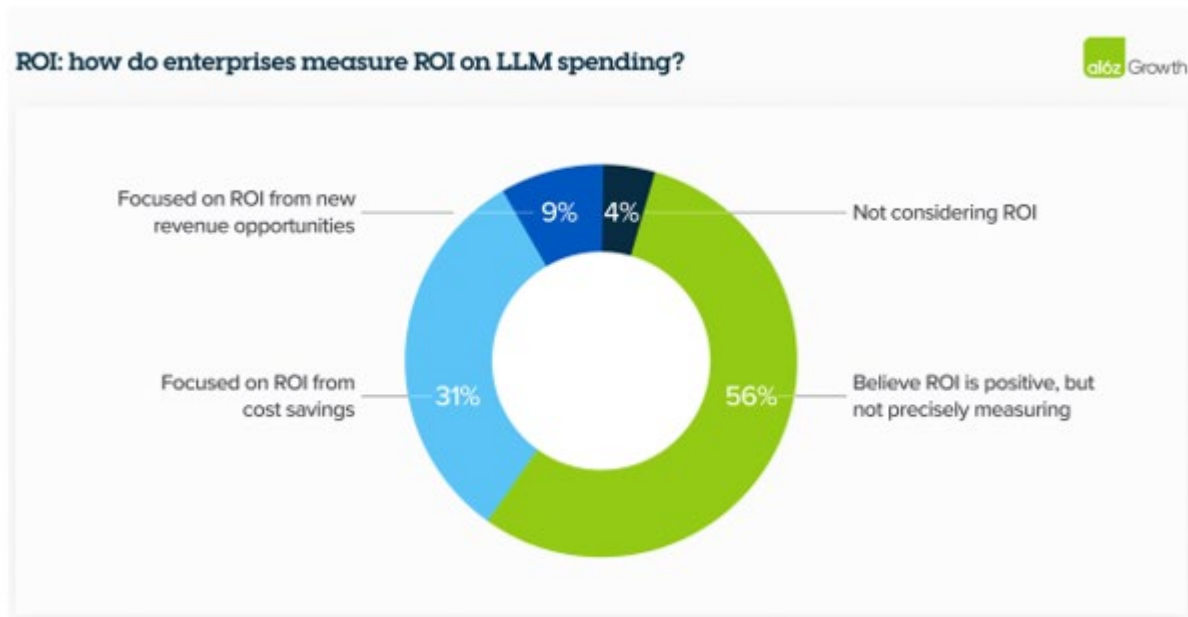

Fig.3 Distribution of Enterprises measuring ROI[2]

When defining metrics of success for your generative AI use cases, it's important to be creative and flexible. Consider metrics that capture the diversity, novelty, and usefulness of the generated outputs, rather than just accuracy or efficiency. For instance, user engagement or satisfaction could be a valuable proxy metric. A crucial component to this process is understanding the key pain points of your business, having a firm understanding of this will aid in the definition of key success measures to evaluate your generative AI use case.

Build trust and understand the limitations

While it is important to get the foundational technological pieces right it is equally important to be able to get buy in from your organisation. Building trust and understanding the limitations of your solution is key to getting that buy in. Think about how you can validate the responses of your generative AI solution. Is the correct information being retrieved from your knowledge base? Is the model reading the source information correctly? Is the model answering the question in the first place? Does the response flow and make sense?

The first step in building trust is ensuring that your generative AI solution is providing accurate and relevant responses. This involves setting up a robust validation process. Start with documenting a comprehensive set of common questions that could be asked of the knowledge base.

Comparing the model's responses with the expected answers will help build that initial level of confidence that the solution will perform to a sufficient standard. This will not only help in maintaining the accuracy of the solution but also in identifying areas where the model might need improvement.

During the validation process it's important to understand the key stages of the pipeline in which you can tune to improve the models' responses. These key stages include knowledge retrieval, information interpretation, question understanding, and finally the structure of the response. Existing evaluation metrics such as groundedness, relevancy, and coherence, can be used to measure the performance at each of these stages.

To aid in understanding these evaluation metrics we will use an example where a generative AI model is to answer the question 'How tall is Mount Fuji?' where the ground truth information is 'Mt Fuji's height is 3,776m'.

> **"How tall is Mt Fuji?"**
> **Ground Truth: Mt Fuji is 3,776m**

Question	Response 1	Response 2	Response 3	Response 4
How tall is Mt Fuji?	The height of Mt Fuji is 6,544m tall. ❌	The height of Mt Kosciuszko is 3,776m tall. ❌	Fuji is 3,776m ❌	The height of Mt Fuji is 3,776m tall. ✅

Groundedness

The groundedness of a generative AI's response is defined as its ability to provide a response which is true to the source information retrieved. This can be measured as a score between 1 to 5, where a score of 5 provides a response which is true to the context provided, and a score of 1 does not include any of the ground truth information. A response which is not grounded in the source information would be considered a 'hallucination' of the model, in which it is able to confidently provide a response which is not based in fact. Looking at our example, response 1 would be considered an ungrounded response where the model has provided a confident response which is not based in fact.

Relevancy

The relevancy of a model's response can be defined as its ability to successfully understand the question that is being asked and provide a response which directly answers the question. A response with a high relevancy score of 5 could provide a relevant response but not be grounded in fact, demonstrating the need for assessing a responses' relevancy in relation to its groundedness. In the example, response 2 would be considered a response which is not relevant as it is talking about Mt Kosciuszko, however it is grounded in the fact as it has correctly surfaced the source information of 3,776m.

Coherence

The coherence of a model's response refers to the flow and readability of the response. Coherence can be understood as the overall quality of the response in which it is well structured and has a logical flow. A response with a high coherence of 5 would be a well-structured sentence which makes sense and remains clear and consistent. Whereas a response with a low coherence of 1 will have a poor sentence structure and overall, a lower quality structure. In the example, response 3 can be considered a response which lacks coherency, even though it is both relevant in its answer and grounded in fact.

Understanding these key measures can aid in comprehending the inherent limitations of generative AI models. This understanding is a vital step towards building reliable solutions that can earn the trust of a broader range of stakeholders within your organisation.

Understand how Generative AI will impact your workforce

While there is a significant amount of excitement around the tremendous value that can be realised by generative AI there is equally just as much fear and reservation around how generative AI will impact the workforce and employees' jobs. When introducing technology which fundamentally transforms how people do their work it is important to consider all the impacts you will have. What happens if things go wrong? What are the consequences and how can you mitigate the risks? What does hallucination mean to the business? It's important that the information that is surfaced by generative AI models is based in trust and safety, leveraging the various evaluation techniques discussed in this paper is a good starting point. However, in addition to ensuring accurate and safe information it is also essential to think about how you can prepare your workforce for change. As with any transformation, the way in which that change is introduced can make or break its adoption and ultimately the value realised to the organisation. In the case of generative AI transformation this criticality is amplified due to the magnitude of the shift that will be introduced. With generative AI transforming the world, leaders have a responsibility to ensure that the transformations they introduce enable their workforces, improve their skills, and leverage greater amounts of human creativity and critical thinking from their people.

SOURCES
1 DeepLearning.AI - Four AI agent strategies that improve GPT-4 and GPT-3.5 performance
2 a16z - 16 Changes to the Way Enterprises Are Building and Buying Generative AI