aitheria
partners

Driving Business Value with Generative AI in the Enterprise

A Whitepaper for Business Decision Makers

Patrick Ward

Richard Jones

Version 2.0

Sept 2024

aitheriapartners.com

CONTENTS

EXECUTIVE SUMMARY

Generative AI (GenAI) is a classification of AI that is capable of creating new content including text, software code, sounds, images and video. Large Language Models (LLMs) are the sub-classification of Generative AI that deal with text, and are widely applicable across the different functions of every company.

Relatively recent innovations in AI architecture and hardware have led to the emergence of new services and software offering companies the opportunity to use GenAI to radically enhance and optimise diverse capabilities and functions like knowledge management, software development, customer support, marketing communications and many more.

A large - and growing - range of options are now available to buy or build these capabilities. Software companies are building GenAI features into their packaged software, typically as a premium paid option. Hyperscale cloud vendors and other players offer the opportunity to buy Generative AI as a service. A range of Open Source models has emerged enabling services to be built in-house. An ecosystem of vendors is rapidly developing to offer supporting tools. A broad range of start-ups are training models to bring niche capabilities to the market.

In summary, the range of options is already very broad and is growing fast. While different approaches will be appropriate for different functions and use cases across the organisation, the criteria and framework for assessing the different options should be consistent.

One objective of this White Paper is to outline the different options, weighing the pros and cons of each. We also examine the factors influencing the Build vs Buy decision.

As with the introduction of most new technologies - perhaps especially so with AI - new risks arise in their adoption. These need to be understood and mitigated for. One key risk is the uncontrolled adoption of Generative AI by individuals across the organisation, in the absence of an organisation-wide Strategy, Policy, Guidelines and Plan.

The capabilities of Generative AI toolsets represent a disruptive force which will have profound implications on the way companies operate. The desire to realise their considerable benefits should not obviate standard Enterprise Governance and Architecture principles. In fact, the power and utility of Generative AI – and the risks it gives rise to - make it even more important that they are assessed, procured, integrated, supported and managed using solid governance principles and processes.

Aligned with good governance, the use of GenAI Services and Toolsets should be rooted in a solid understanding of what outcomes we are seeking to achieve: the value to our business, our objectives and the measures we will use to assess progress.

Purpose of this Document

We wrote this document is to achieve the following objectives:

1. Create a living reference point to help business decision makers understand key concepts, common use cases, business value together with critical risks and mitigations for use of GenAI based tools.
2. Outline options to GenAI, and the factors to consider in evaluating them.
3. Ensure that principles of Governance and Change Management are highlighted as key considerations required to maximise the benefits and minimise the risks.

A "Living" Reference Document

The Generative AI domain is developing rapidly. As the capabilities innovate and as we continue to work with clients on the use of Generative AI, we will continue to update this document. Our objective is for the document to become a reference point for clients and others.

> *Our objective is for the document to become a "living" reference point. We strongly encourage your feedback and insights.*

We therefore strongly welcome feedback, insights, links to good content for inclusion in future versions of the document. Updates made, and names of those providing input, will be recorded in "Appendix 2: Document Control, History".

Feedback can be provided by clicking here.

GENERATIVE AI: A BUSINESS BRIEFING

An introduction to Generative AI

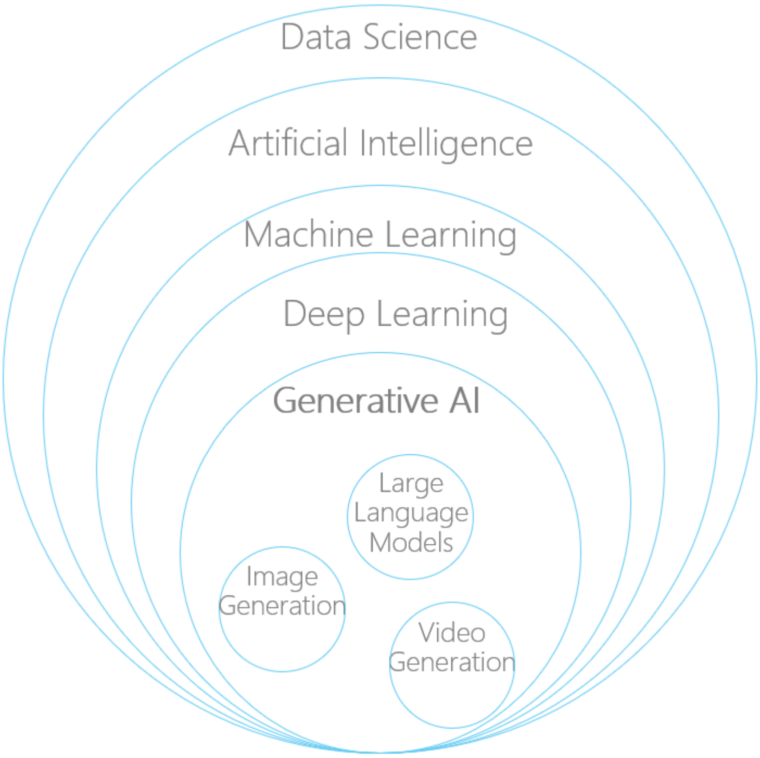


Figure 1 Partial AI Hierarchy shows the classifications of AI into which Generative AI fits

Artificial Intelligence provides the capability for computers to perform many of the tasks of the human brain: to learn, to identify patterns and anomalies, to comprehend and respond in natural language, to understand the content of text, images and sound and so on.

Generative AI is a classification of AI that is capable of creating new content including text, software code, sounds, images and video. It is worth emphasising "new" here: when a prompt is provided to Generative AI, it is not simply retrieving the response from a database or from the internet. It is generating new original content based on the prompt it receives, and the patterns it has learned in its training process.

A Large Language Model (LLM) is a specific type of Generative AI, which generates human-like text based on a prompt. LLMs excel at language-related tasks such as summarising text, translating text, answering questions, explaining concepts and so on.

Before we go further, it’s helpful to understand - at a high level - the key technologies and concepts of AI.

Deep Learning and Neural Networks

Generative AI belongs within the Deep Learning classification, and an AI model built on an underlying structure called an Artificial Neural Network (ANN), or simply a “Neural Network”. A Neural Network is a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain[1]. These artificial neurons are arranged in layers, in such a way that data is passed through them: the output of one layer is the input to the next layer. The ”Deep” in Deep Learning (and ‘Deep Neural Networks’) refers to the fact that there are many layers.

Training GenAI Models

Specifically for a Large Language Model, the Neural Network is trained on a vast amount of text-based data from many sources, typically sourced via the internet. This is a resource-intensive process. The training process creates associations and connections between the neurons in the Neural Network, enabling it to learn how letters make up words, how words make up sentences and how sentences make up paragraphs. Ultimately, the Neural Network is being trained to predict the next word of a sentence. The training process refines the model by comparing what was predicted to the actual next word in the training set, and adjusting the model to minimise incorrect predictions. When training is completed over a vast amount of data, the LLM ultimately becomes capable of interacting in natural language - understanding the context and purpose of the text-based input, and generating a relevant, informed and articulate response.

Models specialising in Image or Video generation are trained to associate words (labels) with recognisable elements within images. Combining the ability to understand these labels with a model trained on an extensive set of existing (and labelled) images provides the ability to generate new images and videos from a descriptive and textual prompt.

[1] Explained: Neural networks – MIT News, April 2017.

Foundation Models

Foundation Models are the starting point for various downstream tasks and applications. As noted above, these models are trained on massive amounts of text and / or image data which enables them to generate coherent and contextually relevant output in response to a prompt. This initial training process is typically referred to as 'pre-training'.

Foundation models provide a baseline of language understanding that can be 'fine-tuned' (explained below) and adapted to specific tasks in areas like natural language processing, chatbots and text / image generation.

Examples of Foundation Models are OpenAI's GPT models, Google's Gemini, Meta's Llama, Stable Diffusion from Stability AI (images) and Runway ML (video).

Fine-Tuning

Once a model is pre-trained it can be further enhanced through a process called fine-tuning to improve performance in specific tasks or using information from specific sources.

In a Business context, this fine-tuning can add very significant value where it is done using internal information of a company. This can include documents, slideware, emails, Instant Messages, data from internal systems like CRM's and ERP's, and operational data from systems related to Manufacturing or Logistics, and so on. In this way, the fine-tuning process enables the Model to then take prompts and generate responses *specific to the company*. This can include product specifications, product roadmaps, sales data or operational data, for example. It can include the terminology and acronyms commonly used across the company, and Brand Guideline documentation that guide on the brand's tone of voice.

> *Fine-tuning enables the Model to generate responses specific to your company, including the terminology and acronyms commonly used across the company, and content aligned to the brand's tone of voice.*

Once fine-tuned, the AI model begins to "speak the company's language" and provide responses that includes information buried deep in company documentation.

Similarly, an image generation model may be fine-tuned based on a training set that 'pushes' the model towards a specific artistic style when generating new images from a user prompt.

As part of Fine-Tuning, **Reinforcement Learning from Human Feedback (RLHF)** can be used to prompt the end-user to rate an LLMs responses. This acts a as a feedback mechanism to further refine the accuracy of model. This process was used by OpenAI to create their 'InstructGPT' models[2].

Low Rank Adaptation, LoRA (not to be confused with LoRa, the long-range low-power radio network) is a technique that enables organisations to fine-tune LLMs without a requirement for extensive storage or computing power (i.e. significantly lower than that required to train the foundational model). Quantized LoRA (QLoRA) further extends the ability to reduce required memory usage during the fine-tuning process.

Retrieval-Augmented Generation (RAG)

RAG has become an extremely popular method to enhance a GenAI model without requiring re-training or fine-tuning. It allows an LLM to retrieve relevant information from external sources and build this into the response. This can be a very effective approach for organisations to leverage the power of LLMs against their own content repositories (of documents, customer data, etc.). It is commonly used in conjunction with Enterprise Search engines and / or 'Vector databases' which provide an enhanced level of correlation of results against a specific user prompt.

Transformers

A transformer is a type of Neural Network architecture that uses a technique called "self-attention" to look at different parts of a sentence or text and understand how they relate to each other. A breakthrough technology that dramatically improves a computer's ability to understand and generate human language, Transformers have become the foundation for many state-of-the-art NLP models including OpenAI's GPT (Generative Pre-trained Transformer) and Google's BERT models (Bidirectional Encoder Representations from Transformers).

Plug-Ins

A relatively new concept even by the standards of Generative AI, Plug-Ins offer the ability to enhance the capabilities of LLMs by enabling them to interact with real-time information and

[2] Aligning language models to follow instructions – OpenAI, January 2022

business data. For example, they can use APIs (Application Programming Interfaces) to retrieve real-time information and can perform actions such as checking a stock price or booking a flight.

Multi-modal Models

From a user perspective, a ‘multi-modal’ modal can work with non-text-based media such images, sound and video. These media can be used as input as part of a prompt (e.g. “what is in this image?”), and can also be generated as part of its output (e.g. “create a video of a cat riding a horse”).

Prompt Engineering

The practice of carefully designing and crafting input prompts for language models to obtain desired responses is often referred to as Prompt Engineering. It involves formulating prompts that effectively guide the model's behaviour, specifying the desired output format, adding instructions or constraints, and providing context to elicit desired responses. Prompt engineering plays a crucial role in controlling and shaping the output of language models, improving their usability.

Example – text generation:

- Basic: “Write a user story about uploading photos to their profile.”

- Better: "Act as an experienced Product Owner. Write a user story for a feature that allows users to upload photos to their profile. Use a professional tone. Use less than 250 words."

Example – image generation:

- Basic: “Generate an image of a cat riding a horse”
- Better: “Generate an image of a cat riding a horse, medieval armour, regal posture, detailed fur texture, high quality, historic, detailed eyes, proud stance, noble, majestic, medieval setting, detailed armour, regal colours, luxurious, detailed lighting”

Supporting Tools – A Growing Ecosystem

An LLM should be considered a component in an overall solution. Supporting tools are usually necessary to compliment the LLM capability and apply it to specific scenarios. Some examples:

- LangChain - an open source framework for Developers to create applications powered by language models for a broad range of Use Cases.
- Hugging Face - an open source based platform and community that enables people to collaborate to build applications using Machine Learning.
- MosaicML - provides open source foundation LLM models that are and available for commercial use which can be fine-tuned against a customer's own data in their own secure environment.

Manage AI Tools Under Your Enterprise Architecture Governance process

Tools like those outlined above should be acquired, managed and governed in the same manner that an Enterprise manages all developer tools. Individuals and teams may have their own preferences as to which tools are used, but costs, risks and duplication can escalate in the absence of adequate diligence given to their introduction to the company.

It is recommended that the use of these tools is agreed and documented as part of existing Enterprise Architecture / Software Development governance processes. Tools should be assessed for their Total Cost of Ownership (TCO), the value they deliver and potential risk they introduce to the business. The assessment should seek to avoid duplicating what has already been adopted elsewhere in the organisation.

See also the "Policy, Ethics and Managing Change" Section below.

Closed Source vs. Open Source Models

As with many other technologies, GenAI solutions can be implemented using one or a combination of closed-source and open-source solutions.

Open and Closed Source LLM technologies

Despite the "Open AI" name, Open AI's GPT-x are 'closed source' models. This means that the details of their structure, training mechanisms and information sources are not publicly available and cannot be re-used or modified outside the company. Instead, users must interact with the models via APIs (based on a paid subscription) to include interacting with those models as a component in their solutions.

Businesses need to understand whether their prompts will be kept private, or will be used to further train the Service Provider's model, in which case significant IP risk arises.

It is strongly recommended that the use of data collected by the LLM owner during the usage of a LLM service should be checked by the procuring organisation during an RFI/RFP process. Specifically, it should be made clear whether prompts or API queries sent will be used to further train the Model. If they are, there is a risk that proprietary data will be made available in the public domain, possibly showing up in the responses to prompts or API calls from users outside the organisation. See "Key Risk 5: Protection of Internal IP" in the Risks section below.

Some services such as ChatGPT Enterprise[3] include options to ensure users' prompts sent via API remain private – that is, not be used to further train the model. This typically comes at an increased cost.

Many Open-Source Models are available for developers to download, modify and include in their own solutions (subject to open source licensing agreements).

A level of technical expertise is required to modify, host and include them as components in a solution.

It is recommended that organisations wishing to use Open-Source software as part of their solutions familiarise themselves with open-source licensing restrictions[4].

The recent 'explosion' of GenAI: Why now?

Since January 2023, every week has seen big announcements involving GenAI models. Although major milestones were achieved in previous years, it hasn't really been until relatively recently that the world has become obsessed with them, and how they can be used.

[3] Introducing ChatGPT Enterprise – OpenAI, 28 August 2023

[4] See Open Source Initiative FAQ for more details.

The concept and practice of using Neural Networks has been around for decades, which have enabled various kinds of AI such as Machine Learning and Computer Vision.

The emergence of GenAI capability relies on Neural Network developments, and has been further enabled by 3 key innovations:

1. **Transformers** – A transformer is a type of neural network architecture. When they were first introduced in 2017[5], transformers represented a significant breakthrough for computers to understand and generate human language.

2. Sufficient **compute power** to train the models. Graphics Processing Units (GPUs) excel at performing mathematical operations on massive tables of numbers (known as matrices or tensors). These operations that are core to the process of training an GenAI models (or indeed any neural network). The ability to distribute the training process across thousands of GPUs has meant that larger models than ever before can be trained on enormous amounts of data, including multiple sources across the Internet.

3. Further **advances in 'fine-tuning'** techniques to refine models. Generated responses from foundational models may not provide desired results for organisations. 'Fine-tuning' allows model outputs to align more closely to the information held within an organisation. Other techniques such as using vector databases and 'plug-ins' to reference external data sources have further increased the capability to include well-structured and canonical content in the responses.

[5] Attention Is All You Need - Vaswani et al., July 2017

FOUR KEY USES OF GENERATIVE AI IN A COMMERCIAL SETTING

In this Section, we focus on four popular Use Cases that are not specific to a particular industry but are applicable to most organisations.

1. Knowledge Management

According to McKinsey, knowledge workers spend about a fifth of their time searching for and gathering information[6]. GenAI models fine-tuned on the Enterprise's own documentation and content can make high-value summarised content (e.g. Product Roadmap, support information, operational data) easier and quicker to retrieve.

Using these models, Information workers can:

- Use natural language queries to retrieve information that is embedded in internal documentation.
- Condense long documents into a shorter summary that is easier and faster to read and understand.
- Identify and highlight real-world objects such as Persons, Locations, Organisations, etc. within documents to assist with information classification. This is a Natural Language Processing function known as Entity Extraction, or Named Entity Recognition (NER).

Key considerations

- Existing Knowledge Management programmes such as Enterprise Social Networking (ESN), Enterprise Content Management (ECM) and Enterprise Search (ES) can prove very useful to build upon when establishing a GenAI-based Knowledge Management. See "Deep Dive: GenAI for Knowledge Management" at the end of this section.
- An assessment or audit of the Enterprise's document stores is a critical initial step to consider what documentation should be made available to the model, and the mechanism through which they will be ingested.
- Some sources will be considered more reliable than others. Careful consideration needs to be given to how relevance and credibility should be reflected in model output.

[6] "The economic potential of generative AI: The next productivity frontier" – McKinsey, June 2023

- Duplicate information can be problematic. For example, where two copies of annual leave policy exist, one of which is out of date, the use of GenAI models may make this problem worse by responding with out-of-date information.
- While Enterprise Search Engine capability has been refined to limit visibility of results based on Role-Based Access, GenAI models by themselves do not inherently support this. It is therefore critical to consider how Role-Based Access will be managed as part of a GenAI based Knowledge Management solution. Some enterprise GenAI applications such as Microsoft CoPilot honour existing role-based access and privacy settings.
- Where a model is used to create content, there is a risk of IP issues arising. See Section ”Six Key AI Risks – And How to Mitigate” below.

Benefits

- GenAI models present an opportunity to create an ‘Expert System’ allowing end-users across the Enterprise to access information embedded in documents and other information sources using natural language queries.
- Such solutions can present a summary of information across multiple documents, including differing viewpoints from different sources. This is a significant advantage over existing Enterprise Search capability which typically presents results as a list of links.
- As workers in an organisation leave, either through movement or attrition, years of accumulated knowledge leaves with them, leaving less experienced workers to duplicate old mistakes in order to learn efficient and effective working practices. GenAI models can form part of a Knowledge Management solution that captures this knowledge (in documents and other artefacts) and makes it available to all.

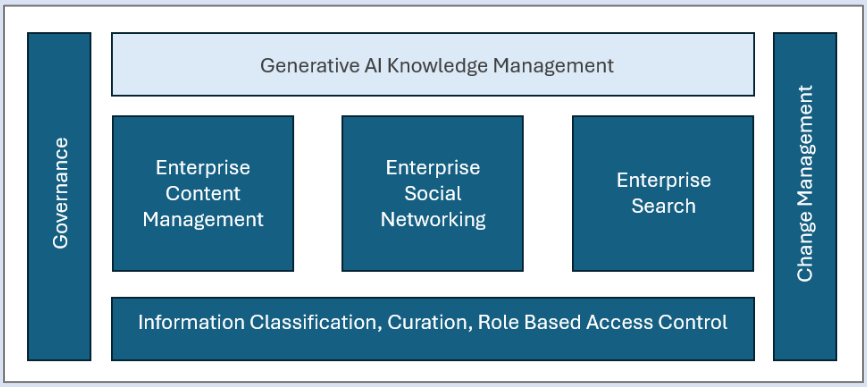
> *Gen-AI models present on opportunity to create an ‘Expert System’ allowing end-users across the Enterprise to access information embedded in documents and other information sources using natural language queries.*

Deep Dive: GenAI for Knowledge Management – learnings from Customer projects

We have found it helpful to apply some of the learnings and insights gained from three common Knowledge Management programmes which have become well established over the last decade or so: Enterprise Social Networking (ESN), Enterprise Content Management (ECM) and Enterprise Search (ES).

Knowledge Management Programme	Insights gained	GenAI implications
Enterprise Social Networking (ESN) Project / technical teams collaborate internally by posting questions & receiving responses from experts across the org. Breaks down departmental silos. Makes expertise and experience more widely available, now & for future use.	Getting users to engage represents a cultural shift that needs to be managed through communication, training, trials, together with ongoing monitoring and adapting. Human feedback on responses helps identifies experts, promoting future responses.	ESN engagement is a rich source of information to enhance the GenAI Model. Manage the change through a strong communication and training plan. Use human feedback to further refine the solution.
Enterprise Content Management (ECM) Content may be tagged with metadata as part of a formal enterprise structure. Users encouraged to manage document lifecycle from "draft" to "published" to "retired".	Active information classification by end-users is critical to implement effective access control. Solid Business Change Management required so users can effectively implement content lifecycle workflows as part of their daily activities.	Enterprise Content Management is a necessary foundation on which to build GenAI-based knowledge management efforts. Information classification and role-based access are critical.
Enterprise Search (ES) Search capability to find documents and information. 'Promoted Search Results' enables organisations to prioritise content based on relevance and importance.	Again, information classification and RBAC are essential. Content curation role(s) critical, as search makes it easier to find documents and information that is out of date, has been superseded (but not archived) or is incorrect.	GenAI can replace Enterprise Search, with easier to access information via natural language queries. Information classification and role-based access are critical. Some enterprise GenAI applications such as Microsoft CoPilot honour existing role-based access & privacy settings.

Efforts to establish a GenAI-based Knowledge Management system should build on the programmes listed above, underpinned by Enterprise Governance and organisational Change Management. See "Policy, Ethics and Managing Change".


Governance
Generative AI Knowledge Management
Enterprise Content Management
Enterprise Social Networking
Enterprise Search
Information Classification, Curation, Role Based Access Control
Change Management

2. Guided Software Development

Models specifically trained on large bodies of existing development code can be used to speed up development activities by writing new software to address requirements that are expressed in natural language. Models can also be used to review existing software modules to detect bugs or identify opportunities for optimisation of the code.

Key considerations

- The GenAI model should be thought of, not as a replacement for a Developer, but rather as a toolset to help accelerate their software development and increase the quality of code.
- Developers using these tools still need to understand programming fundamentals and the overall ALM (Application Lifecycle Management) concept.
- Developers are accustomed to using internet-based resources to provide frameworks, reusable code snippets etc., while remaining accountable for the quality of the code that is entered into the organisation’s code repository. This should remain the case for code generated by GenAI model-based tools: it should be clear that accountability rests with the Developer.

In Software Development, the Gen-AI model should be thought of, not as a replacement for a Developer, but rather as a toolset to help accelerate their software development and increase the quality of code.

Benefits

- Can be used to create code in a desired programming language using a natural language prompt.
- Reduction in development time. A study by McKinsey[7] suggests coding can be completed in about half the time using suitable GenAI tools and training.

[7] Unleashing developer productivity with generative AI – McKinsey, June 2023

- Increase in code quality and security through inspection of code.

Example

GitHub Copilot is a subscription-based service providing AI based code generation to developers. According to the GitHub site: "Research shows developers using GitHub Copilot code up to 55% faster—and report feeling more productive, more fulfilled, and better able to focus on more satisfying work."[8]

3. Customer Support and Chatbots

Virtual Assistant support can be provided internally to Customer Support Agents (CSAs) to help them generate accurate, relevant content for interaction with customers in the customer's own language. Virtual Assistants can also interface directly with customers through a chatbot.

Key considerations

When looking at the application of Generative AI to Customer Support functions, the following key aspects should be given consideration:

- Start by considering your Customer Support objectives and KPI's, then define the options for GenAI-based technology to support them.
- There may be options around integrating the model with key systems used to provide Customer Support such as Customer Relationship Management (CRM), Interactive Voice Response (IVR), Automating Call Distribution (ACD), Enterprise Resource Management (ERP), Ticketing Systems and Knowledge Bases. Such integration can provide value by making customer-specific data or other operational data available to improve the response. Integration can also be used to improve workflow.
- The range of integration options can be overwhelming. We recommend defining the end-state, then identifying and prioritising the planned steps to get there over time. This is covered in more detail in the "

[8] GitHub Copilot website

- Policy, Ethics and Managing Change" section below.
- The balance of human- and machine-based interaction with the end-customer is likely to change over time. Machine-based interaction may grow as the model becomes more fine-tuned with human feedback, and as confidence grows in its capabilities.
- Consider and define the impact on workflows for each system and role in the Customer Support function, and the training that will be required by each role.
- Seek a feedback loop to further fine-tune the model, based on Customer Support provided, customer feedback, new Product developments, operational data, etc.
- Consider the extent of near real-time information that will be required, and how this information will be managed and provided.
- There is a Risk that a model divulges sensitive internal company information, information about other customers or other information protected by privacy regulation. Human review of content provided externally will be required and may change as time progresses. In initial stages, 100% of all content could be reviewed until confidence increases. Human oversight may reduce somewhat over time but is unlikely to diminish anywhere close to zero in the foreseeable future.
- This risk should be further mitigated by including a customer feedback loop on content provided, with a clear escalation process in place where sensitive or protection information is divulged.

Human oversight of generated content may reduce somewhat over time, but is unlikely to diminish anywhere close to zero in the foreseeable future.

- Where a GenAI-based solution is not capable of understanding or responding to a complex or multi-faceted issue, the workflow should include a hand-off to a human CSA.

Benefits

- Models fine-tuned on the Enterprise's own documentation and content can make high-value summarised content (e.g. Product information) easier and quicker to retrieve and utilise for Customer Support purposes.
- Improved Customer Support Agent (CSA) productivity by providing CSAs accurate relevant information provided by the model increasing the throughput of customer interactions per CSA.
- Linking to additional internal data sources provides a channel – via the CSA, or directly - to the customer, of highly relevant timely content (e.g. a known Service issue), improving customer contact resolution times.

- Language translation provides an opportunity for natural language interaction in the customer's native language and an opportunity to recruit Customer Service Agents speaking different languages to that of the Customer.
- Automated chatbot capability can reduce cost by reducing the number of CSAs required to serve a given support workload, and extend hours of support service. Over time, only more complex service requests might need to be handed off to a CSA.
- Improved customer satisfaction and loyalty by providing personalised service, reduced waiting times and accurate and timely information.

4. Marketing and Internal Communications

Marketing Departments have the opportunity to embrace GenAI models as a key tool to assist Marketers with their craft. These models are already impacting speed, efficiency, creativity, personalisation and quality in Marketing functions of many companies.

According to Gartner, "By 2025, 30% of outbound marketing messages from large organizations will be synthetically generated, up from less than 2% in 2022."[9]

A significant characteristic of a company's Brand Identity is its Tone of Voice (ToV). This guides the creation of external communications to potential or existing customers and other stakeholders such as employees, partners and investors. External communications can include adverts, Press Releases, website copy, Product Documentation, and so on. Larger companies often specify the characteristics of the Brand ToV in Brand Guidelines using words like "playful, warm, informal, confident, trusted advisor", and so on.

> *A GenAI solution can be used to create the copy for external communications, guided by the Tone of Voice specified in the Enterprise's Brand Guidelines.*

A GenAI solution can be used to create the copy for external communications, guided by these qualifying words. The fine-tuning process can be used to "tune" the underlying model to the Brand ToV. The solution can generate website and email content personalised to the end-user. Content can be tailored to the segment, demographic and language of the potential customer, and can also take into account their search, browsing and purchasing history and other end-user data such as preferences, support history and previous customer feedback.

A GenAI model can also be used as a creative source, for example to create Ad Campaign concepts, given a set of Objectives for the campaign. Social Media, Blog, email and website copy can be quickly drafted with a well-defined prompt which guides the model on the objective, style and audience of the message.

Increasingly, GenAI models that focus on images and video are also being used to create supporting photography, illustrations, sound and video to support a given Campaign.

[9] Beyond ChatGPT: The Future of Generative AI for Enterprises – Gartner, 26 Jan 2023

Key considerations

- Brands are based on trust: Brands have a trusted relationship with their customers. It's essential that trust is central to the AI Strategy, including data security, data privacy and transparency about how customer data is handled and processed. The customer must also have clarity as to when the customer is dealing with an AI, rather than a human.
- Defining a Brand Tone of Voice is key to guide the solution in a consistent way for the creation of copy that is consistently aligned to the overall Brand identity.
- There is a Risk of Copy or images produced infringe Copyright protections. For example, text created by an model is the same as or similar to text from a Copyright-protected source. See "Key Risk 5: Protection" in the "Six Key AI Risks – And How to Mitigate Them" section below.
- With increasing availability of GenAI tools to create realistic video, audio and images, companies increasingly need to prepare for the risk that their brand and key members of management could be mis-represented through the creation of "deepfake" content. See "Key Risk 6: Use of Video Generation AI tools to create "Deepfake" material purporting to represent your brand" section below.

Benefits

- GenAI models can be a good source of "brainstorming" for the generation of new, innovative, creative ideas.
- They can be a rapid, low cost approach to the generation of copy aligned to a Brand's Tone of Voice. This should be reviewed and refined by Brand and Marketing experts.
- Experiment and refine the prompt to include objectives, target audience details, brand tone of voice descriptors and the media that will be used to deliver the copy. See "Prompt Engineering" section above.
- Image- and Video-based services such as DALL.E, Midjourney and Runway can be used as a rapid, low cost source of photography, illustration, video and graphics which for use in marketing communications. In February 2024, OpenAI first previewed their new Sora offering by releasing various clips of high-definition videos that it created. At the time of writing, a Release Date has not been announced.

SIX KEY AI RISKS – AND HOW TO MITIGATE THEM

As with the introduction of any new technology, new risks arise. It is strongly recommended to develop an understanding some of the key risks of GenAI based solutions, and how best to mitigate them. In this section, we consider the most pressing, which apply to most Use Cases.

Key Risk 1: Unclear AI Policy Leading To Inconsistent, Uncontrolled Adoption of AI-Based Tools

The Risks

There is a risk that the use of GenAI model-based solutions infiltrate the organisation in an uncontrolled way, exposing the Organisation to risks in the following areas:

- Brand Trust – for example where an end-user is provided incorrect information by a such a solution or is not aware that they are interacting with one.
- Duplication and inefficiencies – resulting from siloed approaches in different units across the organisation.
- IP Protection – where internal information gets divulged externally, or content is generated that closely resembles content (images, text etc. that is protected by Copyright).
- Regulatory Risk – where a regulatory requirement is not taken into account or mis-understood.
- Business Continuity – where the execution of a key process becomes reliant on a GenAI-based service, and that service becomes unavailable for some reason.
- Retention of Employees – where employees are not clear on the boundaries of GenAI capability, or feel that their role will ultimately be replaced by that capability.

Mitigations for these Risks

- Establish a clear top-down-sponsored Policy on the use of AI across the organisation, rather than leaving its use to infiltrate in a patchy way. This Policy should cover the use of all classifications of AI, including GenAI. Be specific on which Functions in the organisation will use GenAI capabilities, and

An AI Policy is essential, and should include the measures the organisation will take to ensure its ethical use.

for what purposes. Ensure regular communication of the Policy, and key changes as it inevitably evolves over time.

- AI Policy should address the ethical considerations of AI use, and the measures an organisation will take to ensure ethical use. This should include:
    - Criteria for assessment of AI-based services and capabilities
    - Compliance with laws, regulations, and industry standards (see "Emerging Standards and Regulations" Section below)
    - Transparency on the use of AI
    - Data privacy and security
    - Human oversight and review
    - Accountability, responsibility and escalation
    - Capability development and training
- Establish an "AI Steering Council" (or similar) to monitor developments in the field and identify best practices, and guide the different Functions on how they might best be used across the Organisation. Include representation from all key Functions (Product, Marketing, Sales, HR, Legal & Risk, Finance, etc.).
- Establish a Capability Development plan specific to each Function (Marketing, IT, Product, HR, etc.), since the needs of each function, the approach to AI adoption and the level of AI knowledge will differ significantly across the Functional teams.

For more, see the "Policy, Ethics and Managing Change" Section below.

Key Risk 2: Hallucination Risk: Incorrect information confidently expressed

The Risk

GenAI models are not perfect! They can produce responses that are simply incorrect - often referred to as "hallucinating". This can be made more problematic by the fact that the model gets most information right, giving the impression of a highly authoritative source. Furthermore, the impressive natural language capabilities can also be a factor: erroneous information which is expressed well can lead to a misappropriated level of trust in the content of the response.

Mitigations for this Risk

Human oversight should be maintained to review GenAI-generated content. Companies should over-index on human oversight at the start, checking most/all content – especially if it is externally facing (website, social media, email, etc.). Human oversight may reduce somewhat over time but is unlikely to diminish anywhere close to zero in the foreseeable future. GenAI-based solutions should be considered "a toolset used by humans", as opposed to a replacement for human engagement.

Adjust prompts to request a response only if the solution has a high degree of certainty of the response.

Establish a process to review GenAI content and take appropriate action where incorrect information has been used (e.g. customer engagement, further fine-tuning, increased sampling, etc.).

Ensure customers and employees (e.g. Customer Support Agents) have a clear mechanism to flag erroneous information. Include rapid escalation in the workflow to review content and act fast where this flag is raised.

Example of this Risk

Recent press coverage[10] of a legal case in New York has highlighted the risks associated with model hallucination. A lawyer working for a plaintiff in a legal case used ChatGPT to research case history. The suggested cases returned were found to be 'bogus' following research performed by the opposition's legal team.

This is a great example of 'hallucination': content that *looks* perfectly reasonable in terms of language and structure is actually an erroneous invention. It highlights the significant difference between the results that GenAI model and a search engine might return for a given request.

It also emphasises how essential human review and fact-checking is, especially when model-generated content is used for public-facing purposes (in this case, to create a case history for a trial). We believe a more fitting headline would be "Here's What Happens When Your Lawyer Uses ChatGPT *Blindly, And Doesn't Check Key Facts*".

[10] Here's What Happens When Your Lawyer Uses ChatGPT – New York Times, 27 May 2023

Key Risk 3: No One Left Behind

The Risk

There is a Risk of individuals are left feeling left behind as their - perhaps more tech-savvy – colleagues embrace new Generative AI toolsets to improve their effectiveness and efficiency. This risk is more likely in organisations in which GenAI spreads organically without a stated Policy and in the absence of managing the adoption of GenAI toolsets in line with good governance.

Mitigations for this Risk

Establish a Capability Development plan specific to each Function (Product, Marketing, Sales, HR, Legal & Risk, Finance, etc.) to bring all roles up to speed on:

- Generative AI fundamentals
- Organisational Policies
- Toolkits approved for use; how they will be used within the Function, and limitations to their use
- Risks of using Generative AI in an uncontrolled manner
- Hands-on training and experimentation

For more, see the "Policy, Ethics and Managing Change" Section below.

Key Risk 4: Copyright Infringement

The Risk

There is a Risk that Generative AI creates new content (text, images, video) that is similar to (or based heavily on) content that is protected under copyright.

The content used to train models is not always made public, therefore it is extremely difficult (if not impossible) for an organisation to judge if use of generated content could result in future litigation.

Mitigations for this Risk

- Monitor emerging litigation in this area. A number of cases including the reference here[11] are emerging where content creators are suing AI companies where they believe their Intellectual Property has been used without permission and/or payment.
- To be absolutely sure a company will not be potentially liable for including what turns out to be IP protected content, it may be necessary to mandate that such content must not be used in external facing offers that are monetised by the organisation, at least in the short term as the legal landscape develops.
- Providers of GenAI services may be willing to underwrite the risk. For example, Microsoft announced in September 2023 a new 'Copilot Copyright Commitment for customers'[12] providing an undertaking to 'assume responsibility for the potential legal risks involved'.

> *Microsoft announced in September 2023 a new undertaking to 'assume responsibility for the potential legal risks involved' of the creation of content that is similar to content protected under copyright.*

Key Risk 5: Protection of Internal IP

The Risk

When using a publicly hosted GenAI-based solution in an Enterprise, the information given to it in the form of prompts may be used by the service provider to enhance or re-train the model. At the very least, the prompt information entered will most likely be stored, allowing people outside of the Enterprise to access it.

If the information added as part of the prompt is commercially sensitive, leakage could result in competitors gaining commercial advantage.

[11] AI Art Generators Hit With Copyright Suit Over Artists' Images – Bloomberg Law, January 2023

[12] Microsoft announces new Copilot Copyright Commitment for customers – Microsoft Blog, Sept 2023

A recent example where employees at Samsung entered company proprietary information into ChatGPT[13] got significant attention. Samsung is said to have subsequently banned use of public GenAI tools to prevent further leaks.

Mitigations for this Risk

- Establish clear Policy and Training on the use of Generative AI. For more, see the "Policy, Ethics and Managing Change" Section below.

- Consider using privately hosted instances of LLM solutions that are created and managed exclusively for the consuming enterprise.
- Some LLM Services come with data protection guarantees from the provider. For example, Enterprise ChatGPT includes the following commitment: "We do not train on your business data or conversations, and our models don't learn from your usage"[14].
- In some cases, it may be necessary to ban or block access to public Generative AI services where users may be tempted to add commercially sensitive information as part of a prompt. This measure may be used in the short term while other mitigations such as those above are put in place. However, this mitigation may only apply to PCs; be aware that end-users may circumnavigate this control by using their mobile phone. Policy, training and awareness are often a more effective control for this reason.

Key Risk 6: Use of Video Generation AI tools to create "Deepfake" material purporting to represent your brand

The Risk

With text-to-video AI tools like Meta's Make-A-Video and OpenAI's Sora, it is becoming increasingly straightforward to create a video that appears to have been created by a company or featuring an individual that has not sanctioned the video.

Video material of this nature could be used for various purposes, including seeking to damage a brand's reputation. As synthetic voice and video creation becomes more sophisticated, it may also be used to execute fraudulent transactions. This appears to be happening already: in February 2024, according to Hong Kong police, a member of the Finance team at a

[13] Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak – Bloomberg, May 2023
[14] Introducing ChatGPT Enterprise – OpenAI, 28 August 2023

multinational firm was tricked into paying out $25 million to fraudsters using deepfake technology to pose as the company's chief financial officer in a video conference call[15].

Unlike the risks outlined earlier in this Section, this is an external risk. It will be difficult – perhaps impossible – for a company to prevent videos of this nature from being created. Instead, companies need to prepare a plan of action to execute in the event that such a video is circulated.

Mitigations for this Risk

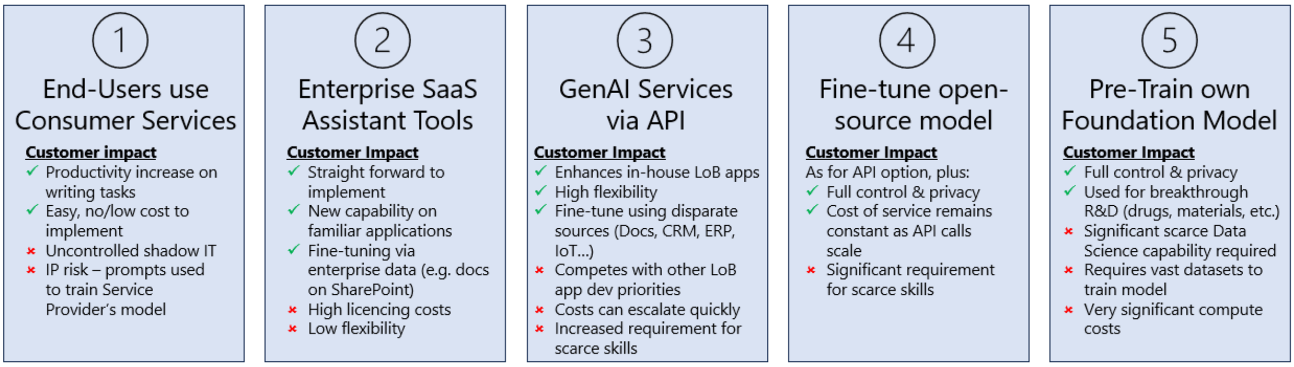
- Ensure AI policy, governance & training are put in place, and that such risks are considered and mitigated against
- Tighten identity verification protocols: voice- or video-based instruction from a "familiar person" should not suffice for the execution of material financial transactions
- Incorporate deepfakes into incident response training, scenario planning and crises comms plans
- Update existing cyber-awareness training to include deepfakes, mitigations & policies
- Consider technology solutions (e.g. Pindrop) where appropriate to detect synthesized voices and prevent fraud

[15] "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'" – CNN, Feb '24

ADOPTING GENERATIVE AI IN BUSINESSES: FIVE APPROACHES

> *These 5 options are not mutually exclusive. The blend of approaches should be considered in its entirety under a single organisation-wide AI Policy, managed though the organisation's Enterprise Architecture governance.*

In this section, we guide on how to adopt GenAI to leverage its capabilities within an organisation. We outline five approaches, starting with the most straightforward (individuals in the organisation using public consumer-based services for the purposes of their role) through to an organisation training its own Foundation Model from scratch.



1 End-Users use Consumer Services

Customer impact
- ✓ Productivity increase on writing tasks
- ✓ Easy, no/low cost to implement
- × Uncontrolled shadow IT
- × IP risk – prompts used to train Service Provider's model

2 Enterprise SaaS Assistant Tools

Customer Impact
- ✓ Straight forward to implement
- ✓ New capability on familiar applications
- ✓ Fine-tuning via enterprise data (e.g. docs on SharePoint)
- × High licencing costs
- × Low flexibility

3 GenAI Services via API

Customer Impact
- ✓ Enhances in-house LoB apps
- ✓ High flexibility
- ✓ Fine-tune using disparate sources (Docs, CRM, ERP, IoT…)
- × Competes with other LoB app dev priorities
- × Costs can escalate quickly
- × Increased requirement for scarce skills

4 Fine-tune open-source model

Customer Impact
As for API option, plus:
- ✓ Full control & privacy
- ✓ Cost of service remains constant as API calls scale
- × Significant requirement for scarce skills

5 Pre-Train own Foundation Model

Customer Impact
- ✓ Full control & privacy
- ✓ Used for breakthrough R&D (drugs, materials, etc.)
- × Significant scarce Data Science capability required
- × Requires vast datasets to train model
- × Very significant compute costs

Figure 2 Adopting Gen-AI in Businesses: 5 Approaches

These options are not mutually exclusive – many organisations are adopting a number of approaches. However, the blend of approaches should be considered in its entirety under a single organisation-wide AI Policy, managed though the organisation's Enterprise Architecture governance.

1. Individuals using public consumer-based services

If a staff-member in an organisation has a browser on their desktop connected to the public internet, they are able – and increasingly likely - to take advantage of consumer-based tools that use Generative AI including LLMs. An example of this is Microsoft's Bing platform, which uses OpenAI's GPT-4 Model.

Employees commonly use these tools day-to-day to draft documents, emails and other communications, summarise text-based content and create images for use in presentations.

Because the end-user is using a public service unconstrained by controls like a Commercial Contract or Privacy Agreement, there is a risk that individuals may inadvertently expose internal IP in their prompts which could be used by the provider to further train their Model, exposing this IP to the wider world. See "Six Key AI Risks – And How to Mitigate" Section for more information on this risk.

It is important for an organisation to have a clearly defined Policy on the use of these tools to make staff aware of the potential risks of using Generative AI. In some cases, companies may choose to monitor or block URL level access to specific sites that host these services while they determine an organisational Policy.

Pros:

- Can enhance productivity and quality for writing tasks. One study shows a reduction in time of 40% and an increase in output quality of 18%[16].
- Easy to implement.
- Intuitive to use with little or no training required for most employees.
- Low or zero cost.

Cons:

- IP Risk, as outlined above.
- Where some employees adopt GenAI Services themselves, other employees with little understanding of GenAI tools may feel left behind in the absence of clear Policy or a perceived need for training.
- Services are used "as is", without the protection of Commercial Contract or Service Level Agreement and are subject to hallucination risk. See "Key Risk 2: Hallucination Risk: Incorrect information confidently expressed" for more details.

[16] Experimental evidence on the productivity effects of generative AI – Science, July 2023

2. GenAI-Powered Enterprise SaaS Assistant Tools

Many commercial software products now come with GenAI capability embedded, often as a paid premium option. Examples are Microsoft 365 Co-Pilot, Salesforce Einstein and Oracle Digital Assistant - conversational bots natively integrated into these products.

This is a low friction approach to adopting GenAI: effectively a new capability included with software that employees already use in their daily work. These tools typically have access to the relevant subset of the organisation's data. For example, with Microsoft 365, this data includes the documents, spreadsheets, presentations, emails etc. that are already managed by the application suite. For Salesforce Einstein, the data includes customer and transactional data in the Salesforce CRM system.

SaaS Assistant tools represent a low friction approach to adopting GenAI: effectively a new capability included with software that employees already use in their daily work. These add-ons are relatively expensive however, sometimes almost doubling the per-user licence cost.

These add-ons are relatively expensive. At the time of writing, Microsoft 365 Copilot costs $30 per user per month on top of the Microsoft 365 Licence[17]. This will almost double the cost of an E3 licence which is currently $33.75[18].

Like the adoption of any software, but particularly with this level of price uplift, it is critically important to gain a deep understanding of benefits, evaluate the RoI and plan the change:

- Evaluate the benefits and compare with Total Cost of Ownership (TCO), including costs associated with licencing, deployment, training and support.
- Agree clear success criteria across organisational leadership, and determine how success will be measured.
- Determine the roll-out approach; resource and plan accordingly.
- Raise awareness amongst end-users of new capabilities available, providing training and support where necessary to maximise utilisation and benefits realisation.
- Measure the value and adjust plans to maximise.

[17] Copilot for Microsoft 365 pricing – Microsoft CoPilot website
[18] Find the right Microsoft 365 enterprise plan for your organization – Microsoft 365 website

Pros:

- Straightforward to implement, since the capabilities are built into an existing software suite.
- Training and support options are available from the vendors of the software and their Partner ecosystem.
- Can add significant value by utilising the organisation's own data within the application suite.

Cons:

- Can be relatively expensive, in some cases almost doubling the existing cost of licencing.
- Relatively low in flexibility – the underlying GenAI capability is limited to the specific application suite.

3. Enable In-House Applications to Access GenAI Based-Services using APIs

Organisations generally use a portfolio of in-house applications created for specific parts of the business. These are generally built or bought over time using in-house development teams or through third party Partners. There is now an opportunity to build in GenAI-based components to unlock the benefits of Generative AI as part of an application's capability.

This can be particularly valuable where the underlying model is fine-tuned or enhanced with supporting components to include organisational specific data (typically, data considered private to the organisation). This can include documentation and product data, emails and Instant Messages, Customer data from CRMs, and operational data from ERP systems or other operational systems.

The Foundation Model and Services to which applications are integrated may be provided by a third-party (e.g. an API-based integration to OpenAI's GPT-4) or they may be models created and/or hosted internally (see Options 4 and 5 below).

Pay-per-use costs based on API calls can escalate quickly. Ensure costs are well understood and "worst case scenarios" are modelled based on numbers of users, numbers of API calls volume of data exchanged.

Pros:

- Provides significant flexibility as the organisation can decide exactly how to use GenAI capabilities to enhance the value provided in their applications.
- Value is significantly increased where the model is fine-tuned or enhanced with supporting components to include organisational specific data, as outlined above.

Cons:

- Pay-per-use costs on API calls can escalate quickly for large volumes of data or large numbers of users.
- Requires technical capability to design and implement - either through an in-house development team or via a Partner. These capabilities are currently very scarce, even when sourced through a Partner.
- Requires the effort and cost of development and support, competing with all of the other priorities and changes that a business has identified for its applications.

Generative AI technical capabilities are currently very scarce, even when sourced through a Partner.

4. Select a pre-trained open-source model and fine-tune / extend it

An organisation may choose to host a GenAI Model and fine-tune it based on their own specific data sets. They can also chose to extend it by using techniques like vector databases, the detail of which is beyond the scope of this document.

The compute power and infrastructure required for this option is diminishing with recent innovation, and is increasingly within the reach of most organisations.

These options require technical capability which may be recruited, retained in-house or through a partner. The compute power and infrastructure required for this option is diminishing with recent innovation, and is increasingly within the reach of most organisations.

This option provides the organisation with exclusive and private use of this GenAI-based solution. Where a partner is used to implement or host the solution, it is important that the terms of the contract must reflect this.

The organisation will then have control over how the model is fine-tuned / enhanced, although it will not be possible to change the foundational model's behaviour in its entirety without re-training it from scratch (see Option 5 below).

Pros:

- Private organisational data is used to optimise the Model's responses.
- Allows full control and exclusive use of the underlying model while avoiding the cost of fully training a Foundation Model from scratch.

Cons:

- Large datasets and regular fine-tuning will increase the cost of compute power required.
- Requires technical capability to design and implement - either through an in-house development team or via a Partner. These capabilities can be very scarce, even when sourced through a Partner.
- Requires the effort and cost of development, competing with all of the other priorities and changes that a business has identified for its applications.

5. Pre-Train your own Foundation Model

This option is most likely to be used only be larger organisations, or those to whom data provenance, privacy and absolute control is paramount. By way of example, a multi-national Pharma company may select this option to implement their own Foundation Model for use with new drug discovery, a process at the very centre of their R&D efforts, necessitating full control and very high privacy over new IP identified.

As with option 4 above, the use of a specialist partner may be preferrable to acquiring the required skills and capabilities internally. Again, we stress: skills and experience are very scarce, even when sourced through a Partner.

Organisations selecting this option will need to evaluate which model to use. Just as important – perhaps more so - will be the choosing and acquiring the text-based datasets upon which the model will be trained and the quality assurance processes to be applied to those datasets.

Pros:

- Full control over the datasets used to train the Foundation Model from scratch.

- Private organisational data is used to optimise the Model’s responses.
- Allows full control and exclusive use of the trained model.
- Allows the organisation to quickly adopt ground-breaking model innovations.

Cons:

- Very significant compute and infrastructure required, either in-house or through a Cloud provider.
- Significant internal capability required, using skills that can be very difficult to obtain in the market, even via a Partner.
- The vast corpus of data required to train a model from scratch can be difficult to identify, acquire, quality-check and manage.

THE BUILD VS. BUY DECISION

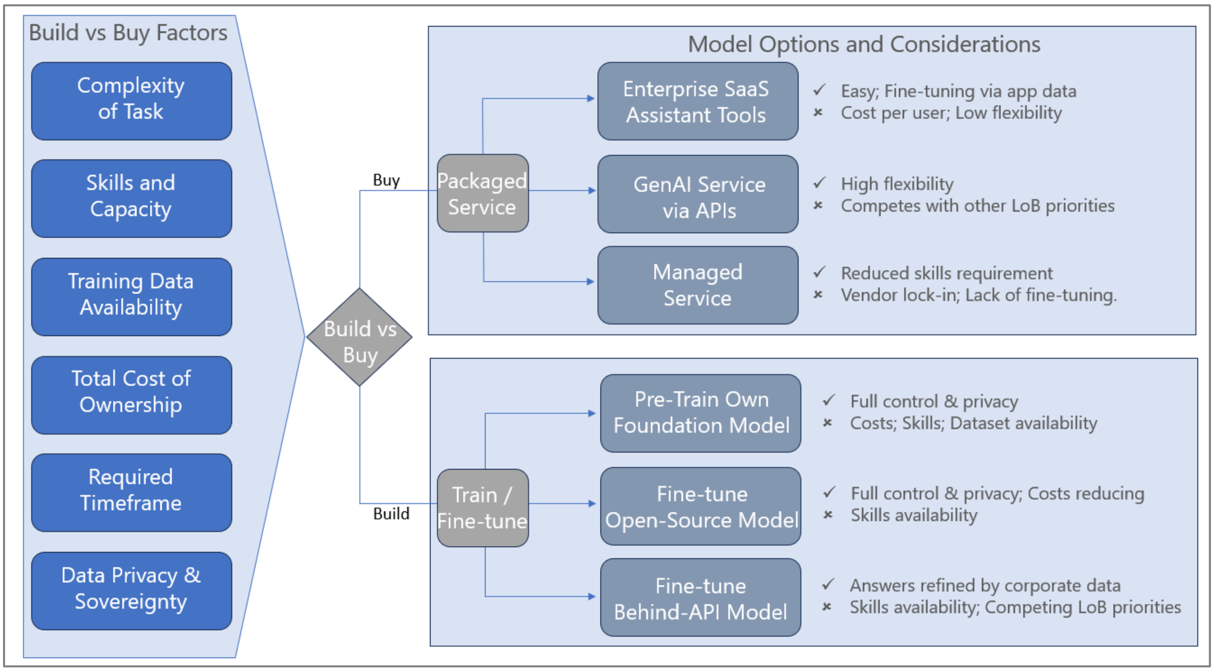

Build vs Buy Factors
Complexity of Task
Skills and Capacity
Training Data Availability
Total Cost of Ownership
Required Timeframe
Data Privacy & Sovereignty
Build vs Buy
Buy
Build
Model Options and Considerations
Packaged Service
Enterprise SaaS Assistant Tools
✓ Easy; Fine-tuning via app data
✗ Cost per user; Low flexibility
GenAI Service via APIs
✓ High flexibility
✗ Competes with other LoB priorities
Managed Service
✓ Reduced skills requirement
✗ Vendor lock-in; Lack of fine-tuning.
Train / Fine-tune
Pre-Train Own Foundation Model
✓ Full control & privacy
✗ Costs; Skills; Dataset availability
Fine-tune Open-Source Model
✓ Full control & privacy; Costs reducing
✗ Skills availability
Fine-tune Behind-API Model
✓ Answers refined by corporate data
✗ Skills availability; Competing LoB priorities

Figure 3 Build vs Buy and Model Choice Considerations

As with many technology implementations, the decision arises as to whether to build the solution (in-house, or using a Partner), or to buy a packaged solution from a vendor or Managed Service Provider. Companies will need to evaluate these options for each Use Case (e.g. Knowledge Management vs. Software Development).

As summarised the diagram above, a number of factors should be considered when choosing the combination of build vs buy as part of your organisation's GenAI strategy. Below we consider each factor, and their impact on the Build vs Buy decision.

One size won't fit all. Companies will need to evaluate Build vs. Buy options for each Use Case.

Complexity of Task

Is the task you want to address with Generative AI relatively straight forward or highly complex?

- If the task is relatively simple and well-understood, weight towards "buying"

- Where existing Generative AI solutions exist in the market that can meet the requirement, weight towards "buying."
- For unique or highly specialized requirements that off-the-shelf solutions cannot easily accommodate, weight towards "building."

Skills and Capacity

Do you have the Skills and Capacity to train and maintain a custom GenAI solution? If not, do you have the opportunity to acquire them, either through recruitment or through Partnering? Re-training of existing resources may also be possible, but is unlikely to address the needs of more specialist roles.

Required skills include:

- Data Science, Data Engineering, Software Development, Machine Learning Operations (MLOps) and Quality Assurance.
- Legal, Policy, IP Management and Compliance.
- Project Management to manage resources across multi-disciplinary roles and organisational functions.

Build vs. Buy considerations:

- If you have the necessary resources, weight towards "building."
- If not, weight towards “buying”, bearing in mind that this does not remove the requirement for these skills, but it does reduce the requirement.

Data Availability

Do you have access to enough high-quality training data for your Generative AI task? Note that general text data in your required language(s) will most likely be required. Domain specific data will be required for fine-tuning.

- If you have access to the necessary data and expertise to manage it, weight towards "building."
- If not, weight towards “buying”

Total Cost of Ownership (TCO)

Do you have a thorough understanding of the budget required to implement your preferred Generative AI implementation?

- If you are already leaning towards "building", be prepared to conduct a thorough TCO exercise based on time of skilled resources, underlying infrastructure etc.
- If your organisation lacks the skills or capacity – in-house or via Partner - to conduct this exercise, weight towards "buying". A TCO exercise will still be required but calculating the TCO for "buying" is typically easier than that for "building".

Capital Funding Considerations

The underlying infrastructure required to host GenAI solutions may be based on private infrastructure or IaaS (Infrastructure as a Service) provided by either hyperscale cloud providers or specialist AI infrastructure providers. Consideration should be given to whether infrastructure falls under CapEx or OpEx accounting rules, and the availability of the necessary capital over time.

Timeframe

Is there a pressing need for a Generative AI solution in your enterprise?

- If you need a solution quickly, weight towards "buying"
- Where time is less pressing, "building" may become more viable.

Data Privacy & Sovereignty

Most cloud providers (for AI service and infrastructure offerings) state clearly in which sovereign country a customer's data will reside and will often enable their Commercial customer to select the territory themselves.

- If you are satisfied with a cloud provider's availability and guarantees for where your data will reside, then weight towards "buying"
- If not, it may be a factor in weighing towards "building", using infrastructure available in the desired location.

Hardware Availability and IaaS

Most companies seeking to pre-train their own GenAI model will chose to do so on an IaaS (Infrastructure as a Service) basis from either a hyperscale cloud provider or a specialist AI infrastructure provider.

Companies seeking to build their own model training infrastructure should be aware not only of costs, but also of lead times. At the time of writing, the required hardware (specifically powerful GPUs such as NVIDIA’s H100) is extremely difficult to acquire as demand significantly out-strips supply[19]. Hardware lead times may therefore cause a very significant impact to the timeline of the project, pushing companies towards the acquisition of compute power from a cloud provider using an IaaS approach, at least in the short to medium term.

> *Hardware lead times may cause a very significant impact to the timeline of the project, pushing companies towards the acquisition of compute power from a cloud provider using an IaaS approach, at least in the short to medium term.*

When choosing IaaS, it will be important to ensure your provider has an appropriate specification and scale available of powerful GPUs required for model pre-training.

[19] Supply chain shortages delay tech sector’s AI bonanza – Financial Times, 23 Aug 2023

POLICY, ETHICS AND MANAGING CHANGE

As with the rollout of any new technology into an organization, staff will need support and time to adopt and make best use of GenAI-based solutions. It is strongly recommended that a dedicated Change Management programme is created and executed for this purpose.

While a comprehensive guide to change management is beyond the scope of this document, this section includes the elements that we believe to be critical.

Governance, Policy Creation and Ethics

Establishing Governance and Policy around the adoption of GenAI in general is critical to establishing a consistent approach across the organisation, mitigating the risks and maximizing the significant value that AI offers. We recommend the following steps as a minimum:

- Create a clear AI Policy, including a Vision statement outlining what the organization is seeking to achieve through the adoption of GenAI models and tools.
- Include an 'Ethical AI' positioning statement that clearly states what is and what is not considered acceptable use. This should be an external-facing statement and should include:
  - Criteria for assessment of AI services and capabilities
  - Compliance with laws, regulations, and industry standards
  - Transparency on the use of AI
  - Data privacy and security
  - Human oversight and review
  - Accountability, responsibility and escalation
  - Capability development and training
- Think about the governance of AI within your organisation as similar to – and likely a sub-set of - existing data governance processes. This should include:

    1. The overall classification of information.
    2. How Personal Identifiable Information (PII) is used, stored and managed.
    3. How sensitive information is stored and managed, and which roles and 3rd parties may have access.

- Identify the role(s) in the organization that will be accountable for:

1. Setting policies and procedures, including how GenAI is used and where it must not be used.
2. The training users need to use them effectively and safely.

- Ensure it is clear to individuals using GenAI-based tools that they are accountable for the content & responses. For example, an email generated by such a tool must be proof-read before sending it, and all facts asserted should be checked; Code generated by GenAI tools must be subject to code reviews.
- Ensure Vendor & Tool Selection is subject to the organization's Vendor selection processes and policies and is aligned with the organization's Enterprise Architecture (with appropriate modifications made to accommodate, as necessary).

Governance and Feedback: AI Steering Council

We recommend establishing an "AI Steering Council" (or similar) to monitor developments in AI and identify best practices, and guide the different Functions on how they might best be used across the Organisation. Include representation from all key Functions (Product, Marketing, Sales, HR, Legal & Risk, Finance, etc.).

This key governance structure should guide the creation of – and potentially own - the organisation's AI Policy.

> *Establish an "AI Steering Council" (or similar) to monitor developments in AI and identify best practices, and guide the different Functions on how they might best be used across the Organisation.*

Organisations should implement new (or enhance existing) feedback mechanisms from their internal users and customers.

- Use the feedback as an input to AI Strategy and Policy
- This feedback may act as a guide as to where additional fine-tuning and RLHF is required.
- Cultural and Regional Variation in conversational style and sensitivity may be critical for inclusive implementations for companies wishing to maximise the reach of their products and services.

Deployment, Integration and Business Continuity

Ensure the following considerations are clearly understood by the IT Team:

1. Where the model(s) reside (which service providers, geographical location).
2. How 'available' the solution is. Ensure a clear Service Level Agreement is provided by 3rd party Foundation Model provider. Ensure resilience of 'private' infrastructure used for training / fine-tuning / hosting the solution is specified.
3. How they are updated with the latest information by relevant vendors when they make changes to the training/fine-tuning and its APIs, and how those changes are reviewed and implemented in their organization.
4. The cost per interaction
5. The Business Continuity approach in the event the Service becomes unavailable.

Change Management Methodology

Change Management related to GenAI should be aligned with the existing Change Management policies and procedures in the organization. At a minimum, this should include the following considerations:

- Create and execute an overall 'Change Management' programme using a framework such as the 'ADKAR' model from Prosci[20].
- Ensure the programme is user-centric and gives individuals the information they need to safely and effectively use GenAI-based tools to bring maximum benefit to the organisation while mitigating the kinds of risks outlined in the "Six Key AI Risks – And How to Mitigate" Section.

[20] The Prosci ADKAR Model – Prosci website

EMERGING STANDARDS AND REGULATIONS

EU AI Act

The EU AI Act was passed by the European Parliament in March 2024. It regulates AI applications based on 4 levels of risk:

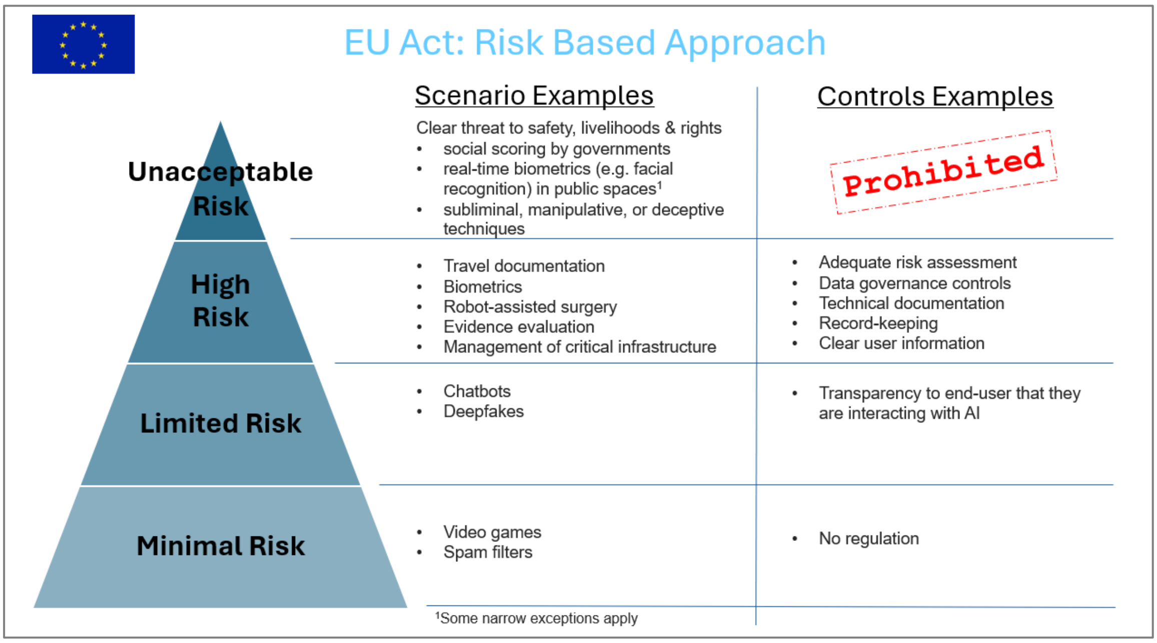
AI applications deemed to represent "**unacceptable risk**" are banned.

For applications deemed to present "**high risk**", it imposes controls around security, transparency and quality. These typically arise in areas like Health, Education, Employment, Justice and Law Enforcement. They consider applications of AI that may give rise to a significant impact on people's health and safety (e.g. the use of robot-assisted surgery) or fundamental rights (e.g. the use of AI in evidence evaluation).

> *Organisations operating in "High Risk" AI categories under the EU AI Act will have significant work to do to ensure compliance with the Act.*

"**Limited-risk**" applications must comply with end-user transparency requirements.

"**Minimal-risk**" applications will remain unregulated.



EU Act: Risk Based Approach

	Scenario Examples	Controls Examples
Unacceptable Risk	Clear threat to safety, livelihoods & rights • social scoring by governments • real-time biometrics (e.g. facial recognition) in public spaces[1] • subliminal, manipulative, or deceptive techniques	Prohibited
High Risk	• Travel documentation • Biometrics • Robot-assisted surgery • Evidence evaluation • Management of critical infrastructure	• Adequate risk assessment • Data governance controls • Technical documentation • Record-keeping • Clear user information
Limited Risk	• Chatbots • Deepfakes	• Transparency to end-user that they are interacting with AI
Minimal Risk	• Video games • Spam filters	• No regulation

[1] Some narrow exceptions apply

Figure 4 EU AI Act summary

Similar to the EU's GDPR regulations, the AI Act applies to service providers outside the EU if they offer services or products within the EU, or the AI system's output is used in the EU.

US Executive Order on AI

On 30 October, President Biden issued an Executive Order on AI to promote the "safe, secure, and trustworthy development and use of artificial intelligence[21]."

Extending to almost 20,000 words, the Executive Order covers eight areas: Safety and Security, Privacy, Equity and Civil Rights, Healthcare and Education, Workers Rights, Innovation and Competition, International Cooperation and Responsible Use by Government.

A White House fact sheet on the order can be downloaded here.

Key points of note:

- Mandates AI developers to share safety data, training information and reports with the U.S. government prior to publicly releasing new or updated large AI models
- Mandates the National Institute of Standards and Technology (NIST) to create federal standards and tests to ensure AI doesn't create a threat for national security.
- Establishes a range of federal task forces and advisory committees related to various aspects of AI safety and security
- Directs federal agencies to issue their guidelines for AI within a year

[21] Executive Order on AI – Whitehouse Briefing 30 October 2024

AITHERIA PARTNERS - HOW WE CAN HELP

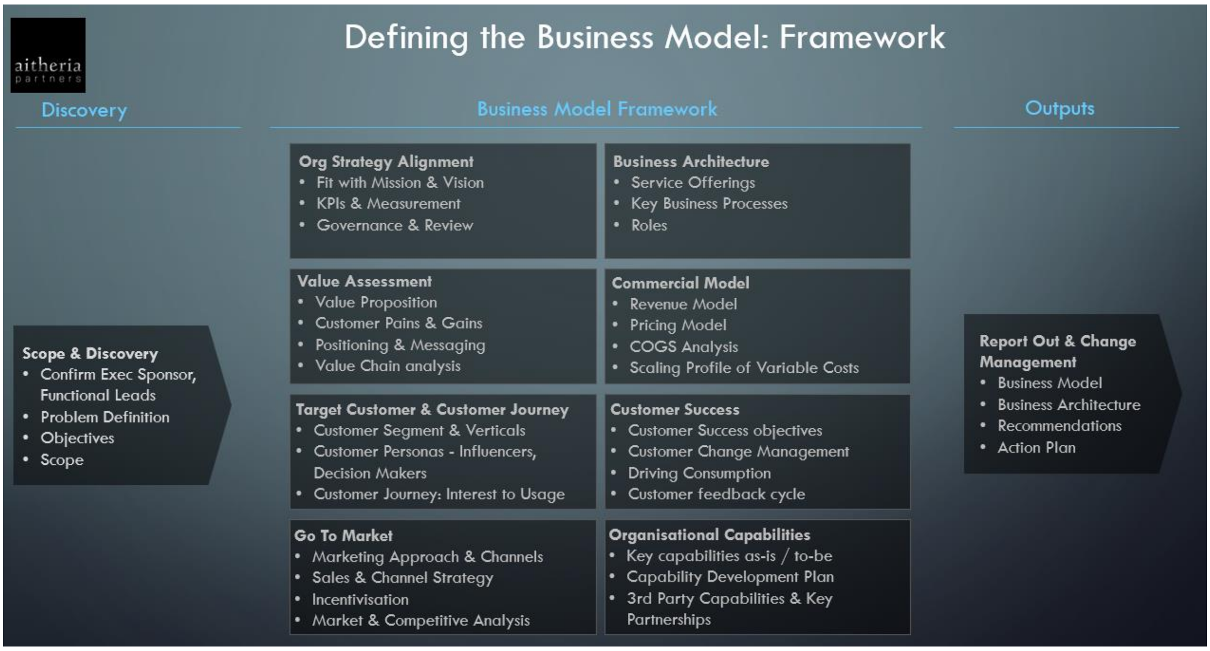
We are specialists in guiding organisations to define their business-led Digital Transformation. We help organisations define their Business Model, Go-To-Market Plan and Solution Architecture for their digital products and services.

Generative AI introduces new concepts, new capabilities, new opportunities and new risks. However, its evaluation, procurement and deployment should follow similar Governance and Change Management to those of most technical toolsets and capabilities.

With highly innovative and novel technology like Generative AI, we generally start with an overview of its capabilities and common Use Cases – in the context of the vertical industry in which the client is operating. One of our objectives in creating this White Paper is to establish a good starting point to achieve this overview.

Then, we seek to clearly define what the organisation is seeking to achieve, and how it will achieve it.

Where AI (or indeed any technology) is being applied to create a new or enhanced market offering – a new product or service – we seek to comprehensively define the Business Model. We facilitate a workshop with the Leadership Team – the custodians of the Business Model – to get really clear on the value proposition, the target end-customer and the optimal sales channel to the customer. We help define the commercial model, considering costs, pricing and revenues. And we look at the organisational capability required to deliver the proposition to customers at scale.



Defining the Business Model: Framework

Discovery

Scope & Discovery
- Confirm Exec Sponsor, Functional Leads
- Problem Definition
- Objectives
- Scope

Business Model Framework

Org Strategy Alignment
- Fit with Mission & Vision
- KPIs & Measurement
- Governance & Review

Business Architecture
- Service Offerings
- Key Business Processes
- Roles

Value Assessment
- Value Proposition
- Customer Pains & Gains
- Positioning & Messaging
- Value Chain analysis

Commercial Model
- Revenue Model
- Pricing Model
- COGS Analysis
- Scaling Profile of Variable Costs

Target Customer & Customer Journey
- Customer Segment & Verticals
- Customer Personas - Influencers, Decision Makers
- Customer Journey: Interest to Usage

Customer Success
- Customer Success objectives
- Customer Change Management
- Driving Consumption
- Customer feedback cycle

Go To Market
- Marketing Approach & Channels
- Sales & Channel Strategy
- Incentivisation
- Market & Competitive Analysis

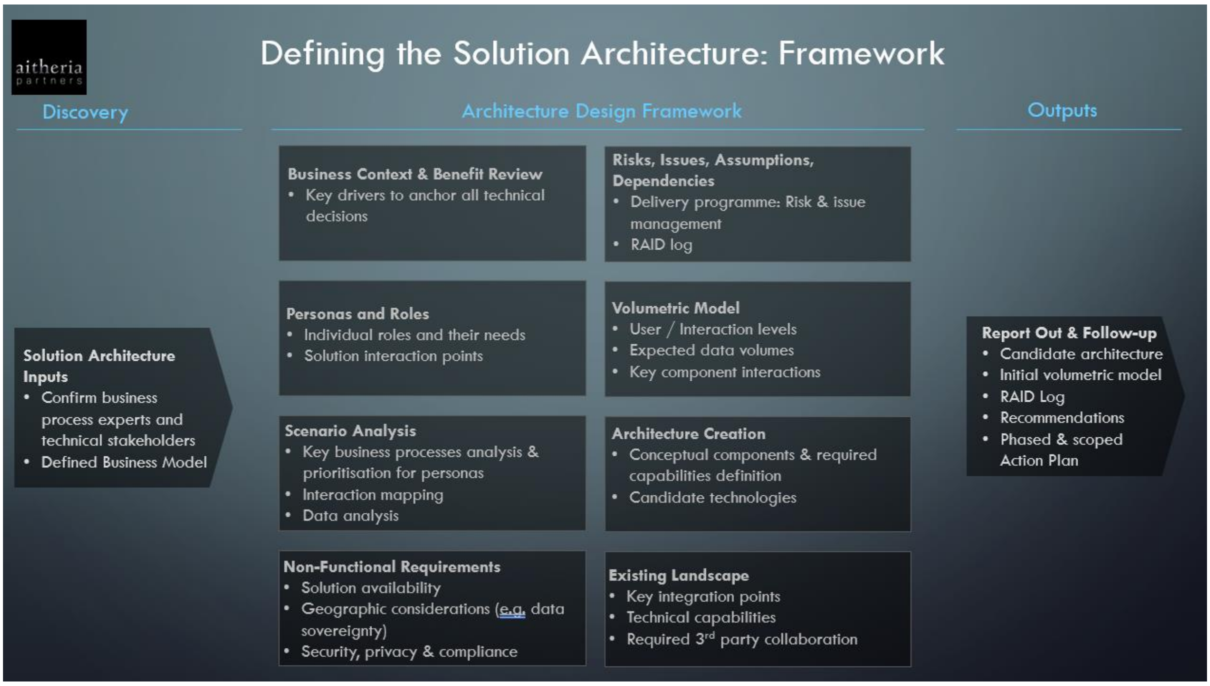
Organisational Capabilities
- Key capabilities as-is / to-be
- Capability Development Plan
- 3rd Party Capabilities & Key Partnerships

Outputs

Report Out & Change Management
- Business Model
- Business Architecture
- Recommendations
- Action Plan

Figure 5 Aitheria Partners Framework for defining the Business Model of a digital proposition

We believe in establishing clarity on the Business Model first, before delving deep into the technical solution. A comprehensively defined Business Model is the right starting point for defining the Solution Architecture. We use the Aitheria Partners Business Model Framework (Fig. 3) and Solution Architecture Framework (Fig. 4) to guide this definition in a structured, repeatable way.



Defining the Solution Architecture: Framework

Discovery

Solution Architecture Inputs
- Confirm business process experts and technical stakeholders
- Defined Business Model

Architecture Design Framework

Business Context & Benefit Review
- Key drivers to anchor all technical decisions

Risks, Issues, Assumptions, Dependencies
- Delivery programme: Risk & issue management
- RAID log

Personas and Roles
- Individual roles and their needs
- Solution interaction points

Volumetric Model
- User / Interaction levels
- Expected data volumes
- Key component interactions

Scenario Analysis
- Key business processes analysis & prioritisation for personas
- Interaction mapping
- Data analysis

Architecture Creation
- Conceptual components & required capabilities definition
- Candidate technologies

Non-Functional Requirements
- Solution availability
- Geographic considerations (e.g. data sovereignty)
- Security, privacy & compliance

Existing Landscape
- Key integration points
- Technical capabilities
- Required 3rd party collaboration

Outputs

Report Out & Follow-up
- Candidate architecture
- Initial volumetric model
- RAID Log
- Recommendations
- Phased & scoped Action Plan

Figure 6 Aitheria Partners Solution Architecture Framework

Over the last 10 years or so, we have been guiding organisations large and small across many verticals on the adoption of AI capabilities. Having worked with the customer to identify the right stakeholders in the organisation, we run a series of facilitated workshops using a structured framework to assess, analyse and define the use of AI, both from a business and a technical perspective.

The key outputs of this process are:

1. Business Strategy and Business Model
2. Solution Architecture
3. Recommendations and Action Plan

This aligns all functions in the organisation around a common strategy and plan. It creates the clarity required to build or buy the right solution, manage the organisational change and (where applicable) bring a new proposition to market. Customers typically seek to manage this process themselves, or seek the support of a Partner. Where required, we have trusted relationships with Partners that can help with managing business change and technical implementation.

To discuss how AI and Digital can create exceptional value for your customers, please get in touch.

APPENDICES

Appendix 1: Guiding Principles for the creation of this Whitepaper

- Target is a Business Audience
- Focus on Business Value - the "Why" of deploying this technology
- Plain speaking - minimise jargon (and explain it when it's necessary)
- Include Risks and Mitigation – and how we can help
- Balance "keeping it brief" (to maximise reader engagement) with "make it comprehensive" (to maximise the value delivered)
- Use diagrams / pictures to help the narrative

Appendix 2: Document Control, History and Feedback

Version	Date	Author(s)	Updates
1.0	15 Sept '23	Patrick Ward, Richard Jones, Mitko Vasilev	Creation of initial document focused on Large Language Models (LLMs)
2.0	Sept 2024	Patrick Ward, Richard Jones,	Broadened scope from LLM to multi-modal Generative AI. Addressed various points of feedback from multiple sources. Added new "Deep Dive on GenAI for Knowledge Management" Added new Risk around the use of "Deepfakes". Added new "Emerging Regulations and Standards" Section.

We strongly welcome feedback, insights, links to good content for inclusion in future versions of the document. Updates made, and names of those providing input, will be recorded in "Appendix 2: Document Control, History" above.

Feedback can be provided by clicking here.

Appendix 3: Terminology

- **BERT –** A foundation model from Google - Bidirectional Encoder Representations from Transformers.
- **Chat GPT** – **Chat**: natural language human computer interaction. **G**enerative: The use of AI for the creation of new content, based on a prompt. **P**re-trained: Using massive amounts of text data to enable it to generate coherent and contextually relevant text responses to a prompt. **T**ransformer: A type of neural network architecture proving breakthrough capability for computers to understand and generate human language.
- **Computer Vision** – a branch of AI enables computers to understand, interpret, and extract meaningful information from visual data, such as images and videos.
- **CRM – Customer Relationship Management**: a software system that helps organisations manage and maintain customer interactions, data, and relationships.
- **CSA - Customer Support Agent**: A person, typically in a Call Centre setting who is responsible for interacting with customers, addressing their inquiries, providing information, resolving issues, and facilitating transactions on behalf of their company.
- **ERP - Enterprise Resource Planning**: a software system that integrates and manages core business processes, functions, and data within an organization to enhance efficiency and streamline operations.
- **Foundation Models –** large-scale pre-trained models that serve as a starting point for various downstream tasks and applications. Foundation models provide a baseline of language understanding that can be fine-tuned and adapted to specific tasks in areas like natural language processing, chatbots and text generation. Examples: OpenAI's GPT (Generative Pre-trained Transformer) models, Google's BERT (Bidirectional Encoder Representations from Transformers), Meta's Llama, Stable Diffusion from Stability AI (images) and Runway ML (video).
- **GPU (Graphics Processing Unit)** – a specialised electronic circuit initially created for visual display via a monitor. Because of their abilities to perform intensive calculations, they are now also used for training AI models and in some cases designed specifically for this purpose.
- **LLM (Large Language Model)** – A type of Neural Network trained specifically on language and able to analyse and produce language text.
- **LoRA** (Low-Rank Adaptation of Large Language Models) – a technique that accelerates the fine-tuning of large models while consuming less memory
- **Machine Learning** - a branch of AI that uses algorithms and models that enable computers to learn from and make predictions or decisions based on data without being explicitly programmed.
- **Neural Networks –** types of Machine Learning models that mimic the structure of organic brains with layers of neurons interacting to produce the output

- **Plug-Ins –** An OpenAI specific term to describe a mechanism to extend the functionality of AI models
- **Retrieval-Augmented Generation (RAG) -** a technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources.
- **Sora –** A new video generation service from OpenAI. Announced February 2024. Release date not public at the time of writing.
- **Transformers –** A mechanism used to enhance the performance of an LLM by making contextual connections between words
- **QLoRA (Quantised LoRA) –** a fine-tuning technique based on **LoRA** bringing further efficiencies