
# Driving Business Value with

# Large Language Models (LLMs) in The Enterprise

## A Whitepaper for Business Decision Makers

Patrick Ward

Richard Jones

Mitko Vasilev

Version 1.0

September 2023

aitheriapartners.com

# CONTENTS

# EXECUTIVE SUMMARY

Generative AI 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 LLMs 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 LLM capabilities. Software companies are building LLM-based features into their packaged software, generally 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 make it all the more important that they are assessed, procured, integrated, supported and managed using solid governance principles and processes.

Aligned with good governance, the use of LLM 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 LLMs.
2. Outline options to LLM adoption, 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 and LLMs, 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.

# LARGE LANGUAGE MODELS: A BUSINESS BRIEFING

## An introduction to Large Language Models

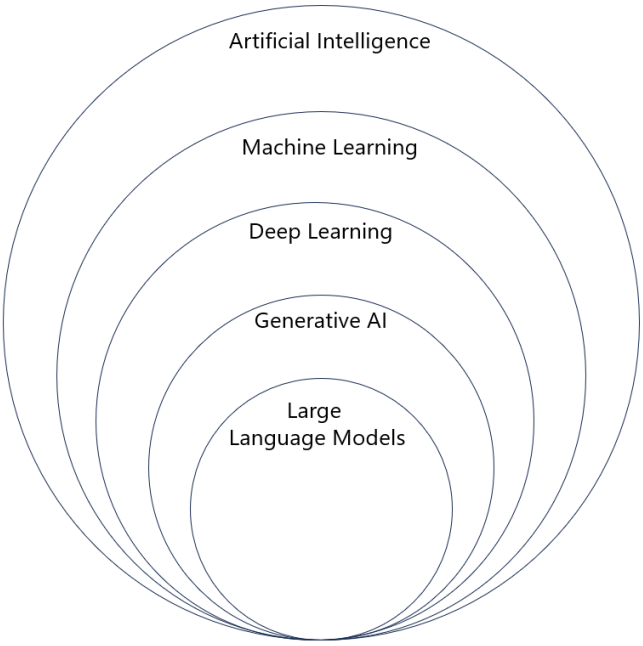


*Figure 1 Partial AI Hierarchy shows the classifications of AI into which LLMs fit*

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 an LLM.

## 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 Neural Networks

To create an LLM, 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.

## Foundation Models

Foundation Models are large-scale pre-trained models that serve as a starting point for various downstream tasks and applications. As noted above, these models are trained on massive amounts of text data which enables them to generate coherent and contextually relevant text in response to a prompt.

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

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

## Fine-Tuning

Once an LLM is trained (or often said to be "pre-trained"), it may be referred to as a "Foundation Model" or "base model". The model can be further "fine-tuned" to specific tasks or using information from specific domains.

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 the company, including the terminology and acronyms commonly used across the company, and content aligned to the brand's tone of voice.*

So, it begins to "speak the company's language", and provide responses that includes information buried deep in company documentation.

As part of Fine-Tuning, **Reinforcement Learning from Human Feedback (RLHF)** can be used to prompt the end-user to rate the 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.

## 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

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

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 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

Our focus in much of this document is on text-based models. 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 an image of a cat").

## 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:

- Good: "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."

## 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, LLM solutions can be implemented using one or a combination of closed-source and open-source solutions. Let's define those terms, then look at their application to LLMs.

### Closed-Source

Closed-Source solutions are created by a software company who do not release the source-code to the general public. The software is developed and maintained internally and released and licensed for use (typically accompanied with a monetary charge) by organisations and individuals.

**Example**: Microsoft Windows is closed-source, users must pay a licence fee, and all updates to the product are implemented by Microsoft.

## Open-Source

Open-Source solutions provide the ability of the general public to view and add to the source-code base, controlled by a set of formal processes. Open-Source solutions are typically monetized through the provision of support services.

**Example**: Ubuntu is a distribution of Linux that is free to use, but organisations may choose to purchase support contracts with Canonical, a UK-based company that provides commercial support and related services for Ubuntu.

## Open and Closed Source LLM technologies

Despite the "Open AI" name, Open AI's GPT-3 (and later) 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 subscription) to include interacting with those models as a component in their solutions.

It is 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.

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 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.

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

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 LLMs: Why now?

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

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 LLM 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 LLM (or indeed any neural network).
   The ability to distribute the training process across thousands of GPUs has meant that larger LLMs 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 an LLM models. Generated responses from foundational models may not provide desired results for organisations. 'fine-tuning' allows model outputs to align more closely to organisational standards. Other techniques such as using vector databases have further increased the capability to include well-structured and canonical content in the responses.

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

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

## Key Features and Capabilities of LLMs

The following list, generated by ChatGPT and subsequently verified, summarises the key capabilities of an LLM:

1. **Language understanding**: understanding the meaning of words and sentences in context. This allows LLMs to accept prompts in "natural language".
2. **Text generation**: generating new text that is coherent and relevant to the given context, based on a given prompt.
3. **Question answering**: answering questions based on the given context.
4. **Translation**: translation of text from one language to another.
5. **Summarization**: summarising long pieces of text into shorter ones, retaining the most important information.
6. **Sentiment analysis**: analysing the sentiment of a piece of text and determine whether it is positive, negative or neutral.
7. **Text classification**: classifying text into different categories based on its content.
8. **Named entity recognition**: identify and classify named entities in text such as people, organizations, and locations.
9. **Text completion**: completing a sentence or paragraph based on the given context.
10. **Text correction**: correct spelling and grammar errors in text.
11. **Text summarization**: summarizing long pieces of text into shorter ones while retaining the most important information.

# FOUR KEY USES OF LLM'S 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]. LLMs fine-tuned on the Enterprise's own documentation and content can make high-value summarised content (e.g. Product Roadmap or Support information) easier and quicker to retrieve.

Information workers can use LLMs to:

- 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.
- Draft document content.
- Identify real-world objects such as Persons, Locations, Organisations, etc. to assist with information classification. This is a Natural Language Processing function known as Entity Extraction, or Named Entity Recognition (NER).

### Key considerations

- 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 LLM, 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 LLM output.
- Duplicate categorise of information can be problematic. For example, where two copies of annual leave policy exist, one of which is out of date, the use of LLMs 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, LLMs do not inherently support this model. It is therefore

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

critical to consider how Role-Based Access will be managed as part of the LLM-based Knowledge Management solution.
- Where an LLM is used to create content, there is a risk of IP issues arising. See Section "Key Risks – And How to Mitigate" below.

### Benefits

- LLMs 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.
- An LLM 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 retirement, years of accumulated knowledge leaves with them. Less experienced workers may have to duplicate old mistakes in order to learn efficient and effective working practices. LLMs can form part of a Knowledge Management solution that captures the knowledge of experienced workers and makes it available to all.

*LLMs 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.*

## 2. Guided Software Development

LLMs 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 LLMs can also be used to review existing software modules to detect bugs or identify opportunities for optimisation of the code.

### Key considerations

- The LLM 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 LLM based tools: it should be clear that accountability rests with the Developer.

*In Software Development, the LLM 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 a function in a desired programming language using a natural language description into.
- Reduction in development time.
- Increase in code quality 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."

## 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 LLM-based technology to support them.
- There may be options around integrating the LLM 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 LLM 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 "

- 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 LLM 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 LLM, 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 through LLM-engaged interactions.

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

- There is a Risk that an LLM divulges sensitive internal company information, information about other customers or other information protected by privacy regulation. Human review of content provided 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.
- Where an LLM 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

- LLMs 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 LLM increasing the throughput of customer interactions per CSA.
- LLMs fine-tuning 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 LLMs as a key tool to assist Marketers with their craft. LLMs 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."[7]

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.

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

An LMM 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 LMM to the Brand ToV. The LLM 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.

An LMM 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 LMM on the objective, style and audience of the message.

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

[7] 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 LMM 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 LLM is the same as or similar to text from a Copyright-protected source. See "Key Risk 5: Protection" in the "Key Risks – And How to Mitigate Them" section below.

## Benefits

- LLMs 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 Midjourney and Runway can be used as a rapid, low cost source of photography, illustration, video and graphics which for use in marketing communications.

# KEY 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 LLMs, 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

### The Risks

There is a risk that he use of AI generally, and LLMs specifically, 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 an AI, or is not aware that they are interacting with an AI.
- Duplication and inefficiencies – through the use of siloed approaches in different units across the organisation.
- IP Protection – where internal information gets divulged externally , or an AI generates content 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 by an AI.
- Business Continuity – where the execution of a key process becomes reliant on an AI capability, and that AI becomes unavailable for some reason.
- Retention of Employees – where employees are not clear on the boundaries of AI use, for example, or feel that their role will ultimately be replaced by the AI capability that is being used.

### 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. Be specific on which Functions in the organisation will use AI capabilities, and for what purposes. Ensure regular communication of the Policy, and key changes as it inevitably evolves over time.

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

- AI Policy should address the ethical concerns of AI, and measures the organisation will take to ensure ethical use. This 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
- 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. 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

LLMs 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 LLM 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 LLM-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. LLMs 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 LLM has a high degree of certainty of the response.

Establish a process to review LLM-generated 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[8] of a legal case in Ney York has highlighted the risks associated with LLM 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 an LLM and a search engine might return for a given request.

It also emphasises how essential human review and fact-checking is, especially when LLM-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*".

## 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 Generative AI spreads organically without a stated Policy and in the absence of managing the adoption of Generative AI toolsets in line with good governance.

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

### 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 AI 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[9] 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.

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

- Providers of AI services may be willing to underwrite the risk. For example, Microsoft recently announced a new 'Copilot Copyright Commitment for customers'[10] providing an undertaking to 'assume responsibility for the potential legal risks involved'.

*Microsoft recently announced 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 LLM based solution in an Enterprise, the information given to it in the form of prompts may be used by the providing party 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.

A recent example where employees at Samsung entered company proprietary information into ChatGPT[11] got significant attention. Samsung is said to have subsequently banned use of public generative AI 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.

[10] Microsoft announces new Copilot Copyright Commitment for customers – Microsoft Blog, Sept 2023
[11] Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak – Bloomberg, May 2023

- 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"[12].
- 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.

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

# ADOPTING LLMS IN BUSINESSES: 5 APPROACHES

In this section, we guide on how to adopt an LLM 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.

*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.*

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.

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 "Key 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:

- Easy to implement.
- Intuitive to use with little or no training for most employees.
- Low or zero cost.

Cons:

- IP Risk, as outlined above.
- Where some employees adopt Generative AI Services themselves, other employees with little understanding of Generative AI 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.

## 2. LLM-Powered Enterprise SaaS Assistant Tools

Many commercial software products now come with AI 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 AI: 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 AI: 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 will cost $30 per user per month for Microsoft 365 E3, E5, Business Standard, and Business Premium customers[13], once Generally Available (timing to be announced). This will almost double the cost of an E3 licence which is currently $36.

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:

[13] "Announcing Bing Chat Enterprise and Microsoft 365 Copilot pricing" – Microsoft Blog 18 July 2023

- 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.

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 LLM capability is limited to the specific application suite.

## 3. Enable In-House Applications to Access LLM 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 LLM-based components to unlock the benefits of Generative AI as part of an application's capability.

This can be particularly valuable where the LLM 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 LLM 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 LLMs created and/or hosted internally (see Options 4 and 5 below).

### Pros:

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

### Cons:

- 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 Large Language 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.

These options require technical capability which may be 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.*

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 LLM-based solution. Where a partner is used to implement or host the LLM, 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 LLM 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 LLM.
- Allows the organisation to quickly adopt ground-breaking LLM 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

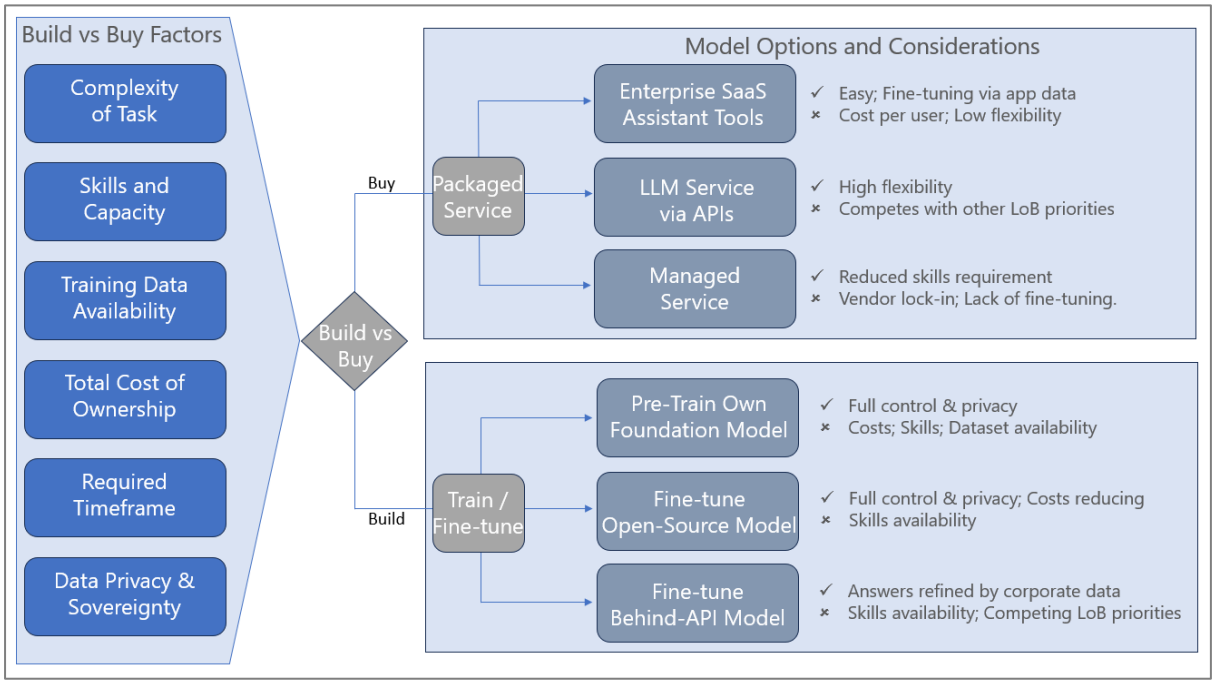


*Fig. 2 Build vs Buy and Model Considerations*

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 LLM 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 LLM 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 LLM 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 LLM 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 LLM 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[14]. 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.

[14] 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 LLM 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 AI 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 LLM-based 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 LLMs are used and where they must not be used.
  2. The training users need to use them effectively and safely.

- Ensure it is clear to individuals using AI-based tools that they are accountable for AI-created content & responses. For example, an email generated by an AI tool must be proof-read before sending it, and all facts asserted should be checked; Code generated by AI-based 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 Model 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 LLMs 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 AI 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 Prosci15.
- Ensure the programme is user-centric and gives individuals the information they need to safely and effectively use LLM-based tools to bring maximum benefit to the organisation while mitigating the kinds of risks outlined in the "Key Risks – And How to Mitigate" Section.

[15] The Prosci ADKAR Model

# 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 and the 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. 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 will 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.


*Fig. 3 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.

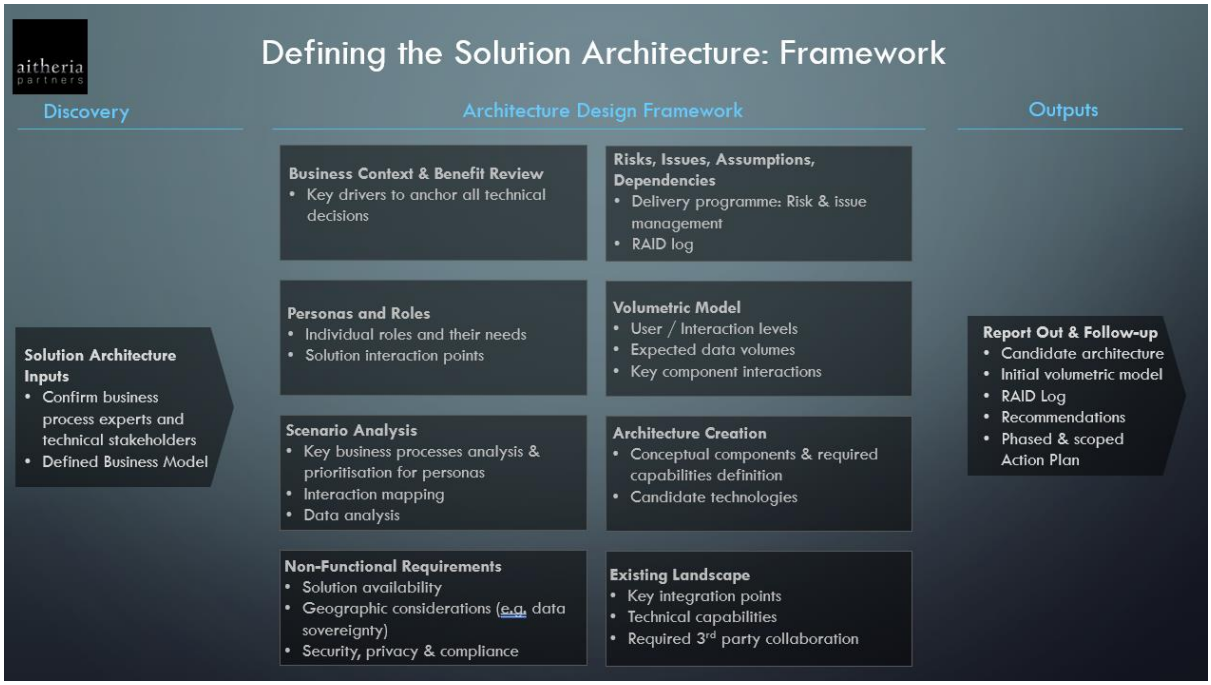


*Fig. 4 Aitheria Partners Solution Architecture Framework*

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. |

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. Generative AI: The creation of new content, based on a prompt. Pre-trained: Using massive amounts of text data to enable it to generate coherent and contextually relevant text responses to a prompt. Transformer: 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

- **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

## Appendix 5: The Use of Generative AI in the creation of this document

From a blank page, we used ChatGPT to "Propose the headings for a White Paper entitled 'Driving Business Value with LMMs in Enterprise'". We kept many headings, deleted some, changed some and added others.
The section "

Key Features and Capabilities of LLMs" was created using ChatGPT.