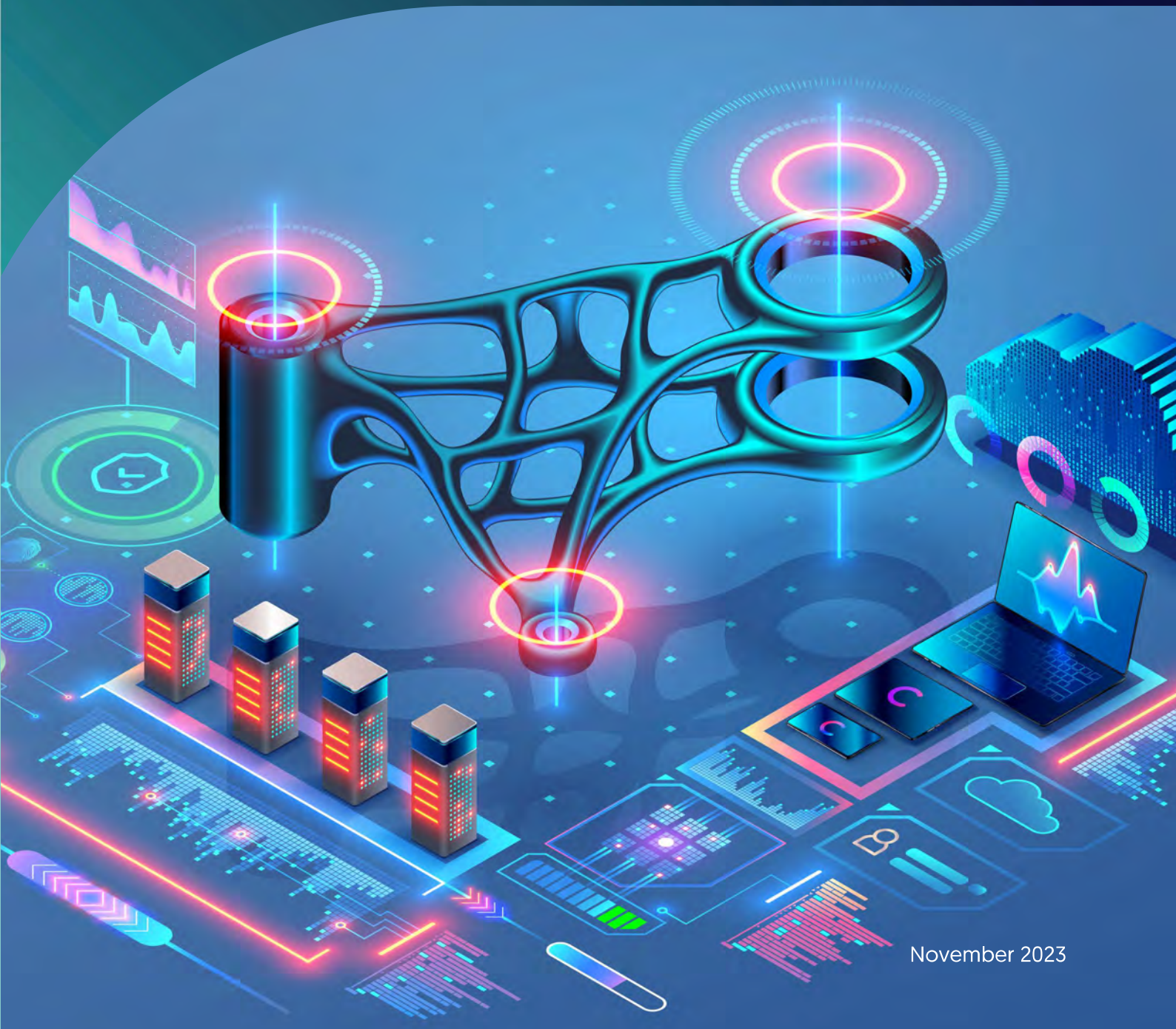
mtc
Manufacturing
Technology Centre

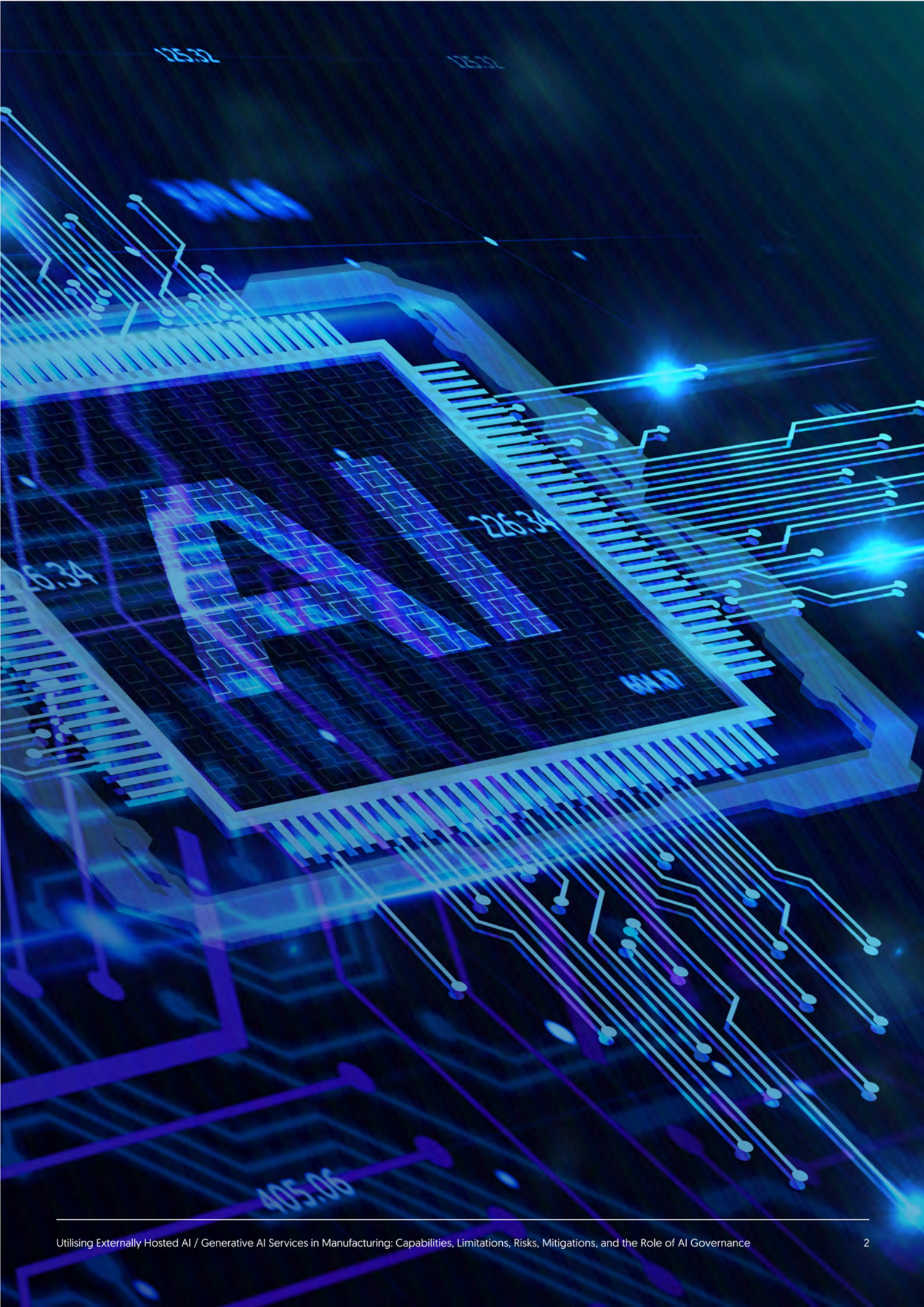
CATAPULT
High Value Manufacturing

UTILISING EXTERNALLY HOSTED AI/ GENERATIVE AI SERVICES IN MANUFACTURING:

Capabilities, Limitations, Risks, Mitigations, and the Role of AI Governance



November 2023


AI

Contents

List of Figures

List of Tables

Acknowledgements

Ula Hijawi
Advanced Research Engineer, Digital Engineering, The Manufacturing Technology Centre Ltd

Mohammed Begg
Research Engineer, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Mostafizur Rahman
Chief Technologist – Industrial AI, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Robert Munnoch
Enterprise Architect, Information Systems Management, The Manufacturing Technology Centre Ltd

Eleanor McCann
Paralegal, Legal team, The Manufacturing Technology Centre Ltd

Tom Winter
Technology Manager, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Yazan Qarout
Senior Research Engineer, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Stuart McLeod
Technology Manager, Digital Engineering, The Manufacturing Technology Centre Ltd

Dr. Nandini Chakravorti
Associate Director, Digital Engineering, The Manufacturing Technology Centre Ltd

Ginny Hawker
Governance and Planning Manager, Information Systems Management, The Manufacturing Technology Centre Ltd

Patrick Roxbee Cox
Chief Information Officer, The Manufacturing Technology Centre Ltd

Executive Summary

The rapid proliferation of AI systems, particularly Externally Hosted AI and Large Language Models (LLMs) like ChatGPT, have raised significant concerns regarding their responsible and secure use. This paper addresses these concerns, focusing on their application within the manufacturing sector. A comprehensive set is presented outlining a procedural risk analysis and AI governance framework at an organisational level to ensure the judicious deployment of these systems.

Central to the presented proposal is the establishment of an AI Governance Team, tasked with overseeing compliance with internal business policies and legal requirements and national/international policies/standards throughout the lifecycle of use-cases employing these technologies. This team plays a pivotal role in safeguarding data integrity and security.

To contextualise these guidelines, a case study is presented utilising OpenAI's ChatGPT and foundational models accessible through their API services. Within this study, a three-tiered risk analysis is demonstrated, spanning Generic, Categoric, and Use-Case specific risks. This analytical approach serves as a robust foundation for understanding and mitigating potential hazards associated with the integration of externally hosted AI systems in a business environment.

This framework aims to provide businesses in the manufacturing sector with a comprehensive toolkit to navigate the complexities of AI development and implementation, ensuring both efficiency and safety in their operations. The findings in the paper are based on research and data up to September 2023.

1. Background

The manufacturing sector is actively seeking innovative methods to optimise automation, gain operational insights, and speed up product and technology development. This necessitates manufacturing businesses to remain at the forefront of significant technical advancements. On the other hand, with the significant evolution of Artificial Intelligence (AI) technologies, third-party companies or providers and the open-source communities are making their AI services available for developers and organisations. Leveraging these AI services can lead to creating opportunities in manufacturing to build better solutions and to experiment with AI for various purposes.

Externally hosted AI services including Large Language Models (LLMs), such as ChatGPT, can offer great benefits to organisations in manufacturing when customised to their use-cases or integrated into their systems improving their efficiencies. However, these benefits come with several important concerns that need to be considered, as external AI services are usually hosted on third-party servers or clouds embedded into their infrastructure. Therefore, risks can emerge from third-party data, software, or hardware as their methodologies, business objectives and data sharing policies may not be aligned with the organisations' deploying the AI system. Data is considered one of the primary concerns, becoming a significant constraint for optimal LLM performance, as sharing sensitive data when submitting information to externally hosted AI models may lead to potential data breaches. Submitted data might be collected to help train the AI model and improve its performance or to customise the user experience. Another concern is the lack of transparency to ensure that the AI technology is developed and operated in a manner that inspires confidence by the developer and all associated stakeholders. Moreover, there are legal implications that can affect organisations as a result of integrating external AI services into their systems including Intellectual Property (IP) infringement and use of customers data.

Therefore, many organisations are looking into developing guidelines and policies and facilitating innovative approaches to leverage the capabilities of external AI technologies into their systems while addressing the associated risks and mitigations. This development is coupled with efforts of defining the areas of AI Governance at the organizational level.

The purpose of this paper is to address these concerns and provide effective and actionable guidance for AI developers, end-users and organisations utilising externally hosted AI services, such as ChatGPT, and other AI technologies. LLMs are chosen as a case-study of external AI services, which can serve as the foundational base for a wide variety of applications and tasks, showing an enormous potential in manufacturing. Moreover, the paper focuses on overcoming data constraints in AI systems that leverage external browser-based or Application Programming Interface (API) services to improve LLM effectiveness and protect the organisations' and their customers' data. It also addresses the importance of assessing terms and policies of external AI service vendors. To help achieve this, this paper introduces the role of the AI Governance Team. This Team is mainly focused on maintaining oversight of the AI Development Lifecycle, evaluating the methodology used in the AI solution and providing feedback on any compliance-related issues and governance protocols.

The scope of work is focused on the AI governance and the risk analysis framework addressing the development of the practical guidelines and processes for governing AI solutions utilising external AI services. This work does not fully cover the principles of responsible AI. Future publications are planned to expand on the scope of this work addressing the development of trustworthy and responsible AI frameworks for manufacturing.

2. Large Language Models (LLMs) as as a Case-study

LLMs are foundational models that utilise deep learning in Natural Language Processing (NLP) and Natural Language Generation (NLG) tasks, which are subsets of AI. The application of LLMs is one of the recent technical developments revolutionising the manufacturing industry. While LLMs may produce new and unique data based on patterns in existing data, they also go a step further by giving the capacity to analyse and organise complicated information, as well as provide human-like interaction. This leads to creating new opportunities in enhancing process automation and workflows and saving time and costs.

For example, modern manufacturing facilities have digital archives of reports, documents and manuals that are essential for the proper functioning of day to operations. Furthermore, many manufacturing assets come with the ability to ingest their various data sources. With the use of LLMs, automation of processes including generation of work instructions, reports, proposals, and design specifications can be made easier as well as carrying out causal analysis of machine or equipment anomalies. This has the potential to increase efficiency and quality of manufacturing output due to the inherent proactiveness in the application of LLMs. Moreover, R&D initiatives will have lower time to maturity as LLMs can be deployed as assistants in generating source code for developers or validating engineering models. Another example is the LLMs' ability to translate text-based prompts to fully fledged designs with integration with Computer-Aided Design (CAD) software [1], which can speed up design cycles and automate quality insurance processes.

In another direction, externally hosted LLMs, such as ChatGPT, can provide personalised engineering skill training [2]. Taking advantage of such models' vast pre-trained data and powerful text production capabilities, engineers can get up-to-speed with new manufacturing domains by gaining easy access to summarised core knowledge, case studies, and best practices.

Being foundational models, LLMs are first pre-trained to learn the basic language functions requiring computationally expensive resources and cutting-edge hardware. LLMs can then be further optimised, or finetuned, through transfer learning for other specific tasks or customised applications requiring less data and computational resources.

An example of the LLM architecture is a Transformer-based neural network as introduced in[3]. The number of parameters that a transformer has can inform about the sophistication and performance of the model. OpenAI have introduced the Generative Pre-trained Transformer (GPT models [4] that can be fine-tuned in customised manufacturing use-cases as discussed further in Section 2.3. Other renowned examples of LLMs are Google's Bard [5] and BERT [6] in addition to LaMDA [7].

2.1 CAPABILITIES AND APPLICATIONS OF LLMS IN MANUFACTURING

LLMs can be fine-tuned against specific use-cases, or tasks, of different capabilities. Examples of the functional categories of which state-of-the-art LLMs, such as OpenAI's GPT-4, can be utilised are summarised in Figure 1 and include, but not limited to [8]:

1. **Code Generation:** Enables software developers to generate scripts in various programming languages to efficiently resolve bugs/tickets of their system or introduce a new functionality in a shorter timeframe. Typically, the user will provide a prompt detailing what to script for and may provide a sample of their own code (e.g., if the code being generated is dependent on another piece of code) and the algorithm will generate a script accordingly. The scale of the output can vary from generating code for one specific functional requirement or even potentially creating an end-to-end application. A potential application within the manufacturing sector is the generation of ASCII (American Standard Code for Information Interchange) STL-format G-code for Additively Manufactured parts. A simple prompt can be provided to the LLM model to generate a model/part in the required format, which

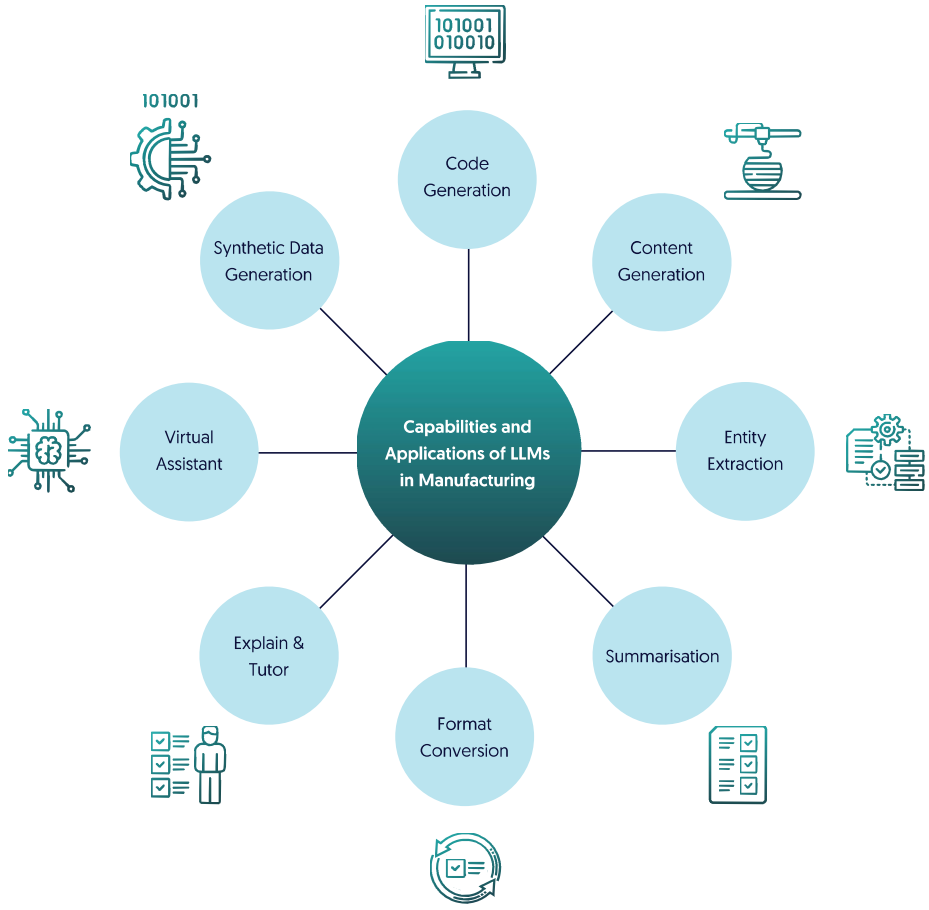


Figure 1: Illustration of LLM Applications in Manufacturing.

can be then manually saved as a STL file. Furthermore, the same model can be used to detect errors in G-code and provide corrections [9]. Therefore, having an LLM-based co-pilot that is trained on ASCII STL format G-code syntax can significantly free-up time of human operators and designers.

2. **Content Generation:** By providing the necessary prompts, human-readable information can be generated in different media formats and a variety of semantics, tenses, degrees of formality, and styles (E.g., email drafts, product descriptions/advertisements, reports, social media posts, presentations, etc.). A user would prompt with a request for a media format and a subject matter. The user can optionally provide keywords/subject matter to generate the intended content. In manufacturing, processes require a lot of documents and forms to be filled with data consuming a lot of time to be individually completed. An LLM can learn the various business templates and automate the data completion process, or even generate entire documents from scratch.

3. **Text Summarisation:** A human-readable block of text or paragraph of information can be provided as a prompt and the model can correspond to it with a summary. For example, researching through academic or a company's whitepapers can be made easier by having an LLM to sift through the text and summarise the contents, or present the key themes of the paper.

4. **Keyword Extraction:** By providing the necessary prompts, key information/keywords of interest can be extracted from a body of text (as a form of questions/answers). For example, when procuring new assets or onboarding software systems, they often come with lengthy instruction manuals, where an LLM can study these documents and provide a simplified and tailored set of instructions.

5. **Format Conversion:** The user provides a prompt of the original format, and the model generates the content in the targeted format. This includes converting information from a certain format to another (e.g., word, csv, etc. to tabular), or converting information from a certain presentation/modality to another (e.g., voice to text). For example, many manufacturing assets operate on various messaging protocols that as a pre-requisite require a certain format of the data which can be automated with an LLM upon learning the various format requirements for different protocols.

6. **Explain & Tutor:** When a subject is provided as a prompt, a concise explanation can be generated with varying levels of detail according to the user's requirement (e.g., explaining technical concepts, such as Machine Learning, Neural Networks, etc.). For example, LLMs can help new employees learning to operate machinery or use shopfloor software by acting as a virtual teacher providing real-time instructions and handling any queries from the new user. This would widely benefit newcomers in a business by reducing the steepness of the learning curve, which would otherwise rely on a human assistant.

7. **Virtual Assistant:** Consists of multiple of the aforementioned capabilities and can be deployed in business functions, such as Customer Support, Educational Applications, Personal Assistants, and Chatbots. For example, Business Management Systems (BMS) in the manufacturing industry contain a huge repository of forms, documents, policies, and templates for specific occasions. It can be quite daunting to find the right document for the right occasion without assistance. An AI-based chatbot/virtual assistant can be deployed to assist the user in finding the right documents by querying the users' circumstances and evaluating the best course of action. This bypasses the need to having to manually navigate across the BMS repository.

8. **Synthetic Data Generation:** By prompting the model to generate data samples of different modalities/representations (e.g., time-series or image-based data), synthetic samples can be generated to improve the size and quality of the dataset. For manufacturing vendors, utilising predictive maintenance models can cause a dilemma, as it may require sharing sensitive or potentially IP-based data for the model to

perform effectively. Moreover, a business may not have the required quantity or quality of data that is adequate for training a model. Therefore, an LLM can learn the nature of the dataset from a moderate sample and generate synthetic data to the required amounts excluding any sensitive business information or personally identifiable information. This reduces the barriers to deploying data-driven systems.

2.2 LIMITATIONS OF LLMS

While LLMs have shown a great potential in reproducing human-like language and can be adopted for an increasing number of use-cases, below are some of the key technical and non-technical limitations found in numbers of literatures [10], [11], [12].

1. **Training Set Limitation:** As LLMs are trained on massive amounts of data, the quality and relevance of the dataset used to train the model greatly determines its overall performance. With NLP technologies, such as GPT models, the training data, which is usually based on public information from the internet, are limited to a certain date. Therefore, with the passage of time into the future, the model can produce information that can be potentially outdated.

2. **Quality of Generated Content:** As LLMs rely on algorithms to process data, the AI-generated content may contain inaccuracy, misinformation, bias, or errors. For example, if a LLM is used to summarise a body of text, it may be possible for key bits of information to be excluded or provide misleading interpretations in the output due to a variety of factors, such as poor prompt engineering or unoptimized prompt parameters, and bias in the datasets. Therefore, verification is required to ensure accurate and unbiased output.

3. **Prompt Engineering:** Another key factor contributing to the generation of poor-quality content can be attributed to the improper use of prompts. With the increasing use of Large Language Models (LLMs), a sub-discipline known as 'Prompt Engineering' has emerged, which investigates how to query the model appropriately to achieve the desired results. Therefore, in the absence of training within this discipline, there is a risk that the model is not being used optimally, which may make it appear as though it is performing poorly.

4. **Trustworthiness:** Trustworthy AI refers to the development and deployment of AI systems that are secure, robust, transparent, fair, and aligned with governance values. The term often encompasses a set of principles and techniques aimed at ensuring that AI technology is developed and operated in a manner that inspires confidence by the user, developer, and all associated stakeholders. AI trust is a crucial topic for safeguarding against any harm that can come from the technology. Therefore, Trustworthy AI frameworks are being increasingly explored to eventually define 'trustworthiness' criteria for evaluating AI systems. Lack of trust acts as a barrier that is slowing AI adoption in many industries including manufacturing.

5. **Sustainability/Energy Consumption:** As LLMs are becoming more precise and accurate in generating text that is human-like (and even images that mimic real life), the underlying architecture is more complex, and the number of parameters increases exponentially. This has a causative effect of consuming more computing resources, and hence, more powerful Graphics/Central Processing Units (GPUs/CPUs) are needed to train and operate these models. With the rise of cloud computing technologies, it has become feasible to host such complex models, nevertheless, the concern on how much energy these models consume and their associated carbon footprint while in operation remain standing.

6. **Knowledge Retention:** As these models take an increasingly greater role in commercial and personal affairs, there will likely be a loss of knowledge on habits (e.g., report writing) that are often second nature to humans due to being reliant on these systems. Therefore, assurances are required that these models can reliably fulfil their functional obligations if human presence is to be continually diminishing. Furthermore, validation of AI models themselves will in the long term become an issue as the skills required

to perform this may be lost or require significant adaptation to validate the output of such systems.

7. **Privacy Concerns:** AI services are generally made available for users through the interaction with a browser-based application or Application Programming Interface (API) endpoints. When data within prompts are submitted to the AI service potentially containing sensitive information, data may be stored by the host of the AI service for training purposes as part of their continuous model improvement/development methodology. A particular concern is the inclusion of IP-sensitive data within prompts and being potentially exposed to the public.

2.3 OPENAI'S GPT AS AN EXAMPLE OF LLMS

Introduced by OpenAI, GPTs are family of neural network models that adopts the transformer architecture powering generative AI applications, such as ChatGPT, with the ability to produce human-like text and content. As GPT models are built on transformer neural networks, their transformer architecture follows an Encoder-Decoder structure. They analyse natural language queries, or prompts, summarise large sums of text, and use them to predict language patterns based on the GPT model's understanding of language. This can be achieved by training the GPT models with hundreds of billions of parameters on massive language datasets [13].

In a transformer, an Encoder maps an input sequence and converts it to a continuous vector, numerical representation. The encoded numerical representation holding the learnt features of that input (e.g., the encoded mathematical representations of a word) is then fed into a Decoder. The Decoder then generates an output sequence in an iterative procedure. At each step, the model regressively consumes the previously generated elements as additional input when generating the next [3].

This architecture requires a large amount of labelled data for specific use-cases making it difficult to configure. To address this issue, an alternative architecture in which a stack of encoders or decoders can be used. To reduce the requirement of labelled data, the transfer learning technique is utilised within the stacked transformers by initially training the model to get a foundational understanding of the language (hence referred to as foundational models) through a process known as 'self-supervised learning' [14]. Consequently, the model is further optimised through transfer learning for other specific applications (referred to as fine-tuning). GPT models can be fine-tuned in customised manufacturing use-cases across industries automating and improving a wide range of tasks.

OpenAI's GPT technology has progressed through several iterations to date with the latest version (as of 2023) illustrated in Figure 2 [4]. As newer versions are released, older versions will eventually be deprecated [15].

2.3.1 USAGE OF OPENAI'S CHATGPT AND EXTERNAL AI SERVICES

The user can interact with the GPT models through two main interfaces: OpenAI's browser-based application, ChatGPT, and OpenAI's API endpoints. The API service has been made available for use in custom applications, such that the base (i.e., foundational) models can be fine-tuned for the user's specific use-case.

The workflow process for fine-tuning OpenAI's base models starts with preparing the dataset following a particular 'prompt-completion' pair format. Each example in the training set needs to have a single input (i.e., prompt) and an associated output (i.e., completion). Following this, a fine-tuned model can be created by using an API endpoint to upload the training set and create a fine-tuning job customising the base model's name. The model training would start afterwards and can take time depending on the customised model and dataset size. Finally, once the training is complete, the fine-tuned model would be available for use via an API endpoint. The user can then start making requests to it by passing the model's name as the model parameter.

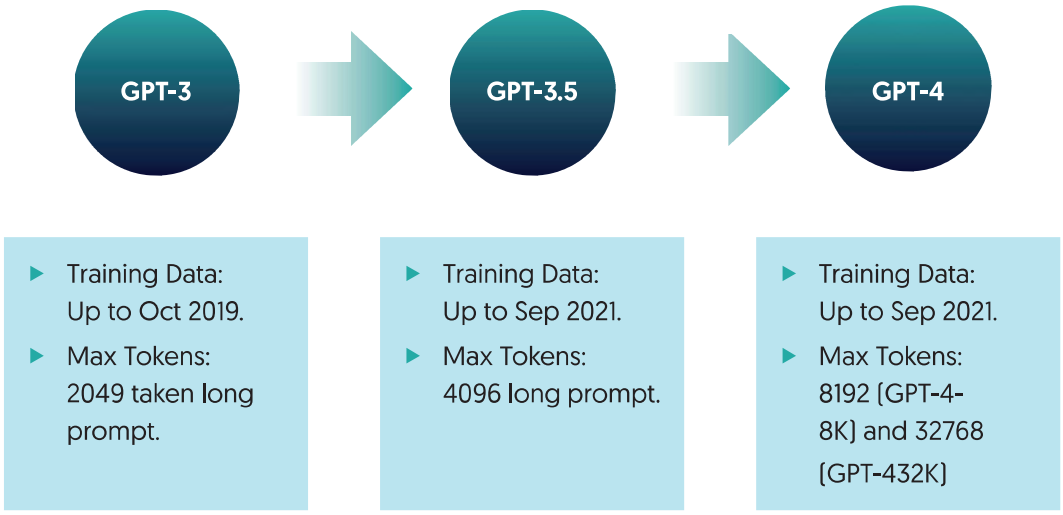


Figure 2: Active versions of OpenAI's GPT models as per September 2023.

2.3.2 UNDERSTANDING OPENAI'S TERMS AND POLICIES FOR API AND NON-API SERVICES

In general, there are three primary resources of data that are used to develop OpenAI's LLMs, including the models that power ChatGPT, as described in OpenAI's article [16]:

1. Information that is publicly available on the internet,
2. Information that OpenAI license from third parties, and;
3. Information that OpenAI's users or human trainers provide.

However, data from ChatGPT Enterprise and the API Platform (after March 1, 2023) is not used for training OpenAI's models [17]. User content is not shared with third parties for marketing purposes. A list of sub-processors that OpenAI has engaged with to provide processing activities can be found on OpenAI's Platform [18].

An illustration of the understanding of OpenAI's terms and policies of API and non-API services, particularly ChatGPT, API Services, and ChatGPT Enterprise, is depicted in Figure 3 with respect to data usage for training and data retention.

A. **Non-API Services (e.g., ChatGPT):** By default, when non-API consumer services ChatGPT or DALL-E are used, submitted data may be used to improve OpenAI's models [19]. However, users have the option of opting out of having their data used by switching off chat history in ChatGPT settings (under Data Controls) to turn off training for any conversations created while training is disabled [19], [20].

   When chat history is disabled, new conversations will be retained for 30 days, and can be reviewed by OpenAI only when needed to monitor for abuse, before permanently deleting [21].

B. **API Services:** OpenAI have released their API Platform [22], which gives developers access to powerful models like GPT-4 and GPT-3.5 Turbo. Data submitted through OpenAI's API is referred to as API data. By default, API data, inputs, and outputs are not used to train OpenAI's models or to improve OpenAI's service offering [17], [19]. However, users can decide opting in to share their data for API services to support the continuous improvement of their models [19], [20].

	Chat GTP	API Services	ChatGPT Enterprise
Data usage for training (by default)	✔	✘	✘
Data retention (30 day, by default)	-	✔	✔

Figure 3: OpenAI's Term and Policies for API and non-API Services as per September 2023.

Moreover, OpenAI may securely retain API inputs and outputs for up to 30 days to identify abuse. Users can also request zero data retention (ZDR) for eligible endpoints if they have a qualifying use-case [17]. Access to API business data stored on OpenAI's systems (i.e., stored API inputs, outputs, and fine-tuning data) is limited to [1] authorized employees that require access for engineering support, investigating potential platform abuse, and legal compliance and [2] specialized third-party contractors who are bound by confidentiality and security obligations, solely to review for abuse and misuse [17].

C. **ChatGPT Enterprise:** OpenAI have released their ChatGPT Enterprise Platform [17], which is built for business offering organisations the ability to use ChatGPT with controls, deployment tools, and speed required to make organizations more productive. ChatGPT Enterprise data, inputs, and outputs are not used for training OpenAI's models [17]. Within an organization, only end users can view their conversations. Workspace admins have control over workspaces and access. Authorized OpenAI employees will only ever access the organization's data for the purposes of resolving incidents, recovering end user conversations with the organization's explicit permission, or where required by applicable law [17].

Data is securely retained by ChatGPT Enterprise to enable features like conversation history. The organisation can control how long their data is retained. Any deleted conversations are removed from OpenAI's systems within 30 days. Note that shorter retention periods may compromise product experience [17].

Note that OpenAI's terms & policies [23] are subject to change over time. It is recommended to refer to the latest terms and policies.

3. Key Recommendations

Organisations in manufacturing are increasingly exploring possibilities of employing externally hosted AI services, such as LLMs, to leverage their potential as part of the AI solution development while considering the limitations. Recommendations to be considered are proposed as follows, which are meant to serve as guidelines, and not as a legal advice:

1. Identifying and validating a solid use-case for utilising the AI technology and justifying the use of the external AI service to achieve the solution's objectives in terms of its benefit.
2. Evaluating the sensitivity level of data required to be submitted to the external AI model.
3. Minimising the amount of submitted dataset where applicable, i.e., collecting and processing the minimum data necessary to achieve the objectives of the AI solution to avoid a larger scope of risks.
4. Thoroughly reviewing the use of LLMs and ensure that it does not infringe any third-party Intellectual Property (IP) rights.
5. Including a Human-in-the-Loop process to implement measures to validate AI-generated content, such as reviewing and filtering AI-generated content before publication or final use.
6. Incorporating data transparency measures within the AI solution development, such as, but not limited to, providing a clear understanding of the used data sources and the AI model's training processes, i.e., how the AI model is trained, and what datasets are used.
7. Analysing risks associated with the AI solution or the adopted use-case and their mitigation measures.
8. Including a declaration statement in case AI-generated content has been embedded within the output.
9. Developing awareness in the business of the potential implications of AI-generated content, its alignment with the organisation's values, and the associated risks and mitigations of utilising LLMs.
10. Development of an AI usage policy/AI governance practices within the business that guide users such that the above recommendations are considered.

4. Risks and Mitigations of Externally Hosted AI Services in Manufacturing

Analysing the risks and mitigations for utilising external AI services, such as LLMs, comes forward as a crucial step to leverage their potential capability while mitigating harm or loss to the business. This section provides a guideline for the type of risks to consider and their example mitigations.

4.1 GROUPS OF RISKS

Risks of using external AI services can be broadly grouped into three categories as illustrated in Figure 4.

1. **Generic Risks:** risks that are concerned with the business and its relationships with their customers, and the resulted impact that these risks can have if they were to materialise. They can be applicable to all application categories (e.g., applications shown in Figure 1) and should always be identified before assessing any further use-case specific risks and mitigations.

2. **Categoric Risks:** in this context, the noun 'category' refers to the types of the functionality that the use-case will be leveraging (e.g., applications shown in Figure 1). The list of application categories can be identified by the business and continuously developed based on the captured requirements along with the associated risks and mitigations for each application category.

3. **Use-case Specific Risks:** risks being directly related or specific to the adopted use-case/AI solution development. A use-case specific risk analysis would typically incorporate applicable Categoric and Generic risks.

All groups of risks are iterative and can be continuously developed as a live register considering lessons learnt from AI solutions or projects utilising external AI services over time. Risks related to ethical, trustworthy, and responsible AI are covered throughout the 3-teird risk and mitigation analysis, which will be expanded on further in future work.

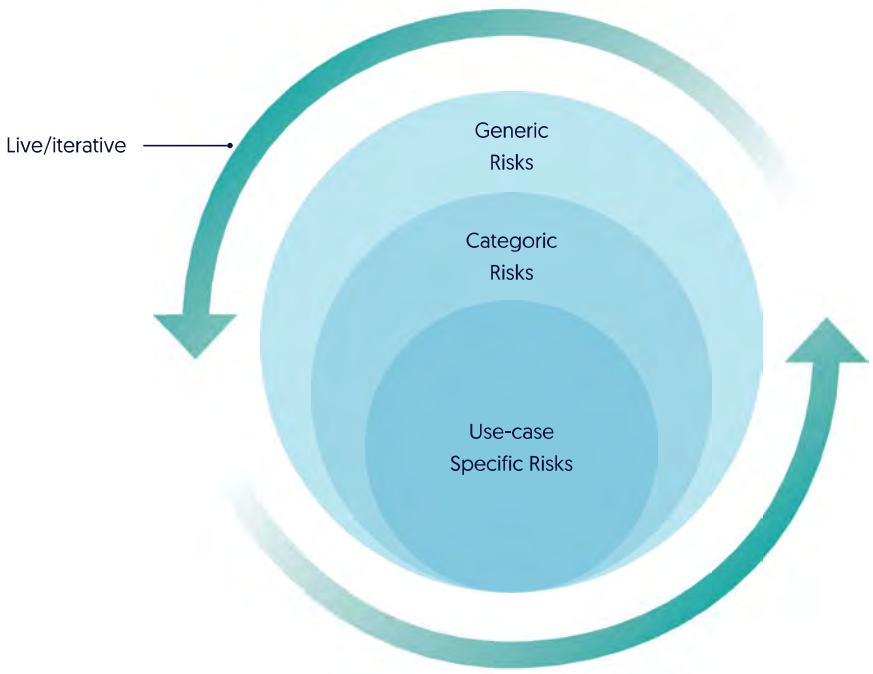


Figure 4: Risk categorisation of leveraging external AI services.

4.2 EXAMPLE GENERIC RISKS

Examples of Generic Risks surrounding the use of an externally hosted LLM are summarised in Table 1:

No.	Generic Risk	Description
1	Loss of or unauthorised use of data	Any confidential data or personal data submitted to an AI service could be at risk of loss or unauthorised use, which could result in claims or complaints by customers whose data is affected, interruption to business operations, loss of valuable data and/or fines by data protection regulators.
2	Accuracy and Reliability of AI Generated Content	Content generated by the system is not likely to be verified by the service provider and therefore poses a risk of inaccurate information or information that may infringe on third party intellectual property rights. This could lead to complains and/or claims by customers.
3	Use of External AI API Services	Once the solution has completed its term and is no longer in active use, in the absence of a decommissioning plan there is a risk of the service to continue being interacted with little or no over-sight from the necessary team/stakeholders. This can potentially lead to misuse of the system ultimately impacting the business's resources due to added costs.

Table 1: Generic Risks associated with using Externally Hosted LLMs.

4.3 EXAMPLE CATEGORIC RISKS AND MITIGATIONS

While Generic Risks are considered relevant to each business, 'Categoric Risks' observed as the middle circle in Figure 4 are investigated along with their mitigations and proposed as an example as shown in Table 2. The risks are analysed against the application categories shown in Figure 1 and described in Section 2.1, which can be of interest in the manufacturing sector.

Application Category	Categoric Risk	Resulted Impact	Controls and Mitigations
Code Generation	Code extracts being stored on an external server.	▸ IP/Sensitive data that gives the business its competitive edge is outside the business systems.	▸ Provide an abstract prompt that does not contain sensitive information, and then modify the generated output to be configured for the intended uses. ▸ Ensure to review the AI service's vendor's terms and privacy policy. ▸ Ensure to use an organisational account.
	The AI model generates incorrect syntax/code.	▸ Creates unreliability in the system. ▸ Low quality of generated code as faulty scripts being developed. ▸ Reduced maintainability of the code. ▸ Knowledge retention issues as a result of reliance on AI-generated code.	▸ Include a Human-In-The-Loop process, such as a code-review gate in the development process, prior to accepting AI-generated code. ▸ Refer to the Organisation's in-house software development guidelines.
	AI-generated code does not meet industry standards.	▸ Low quality of AI-generated code. ▸ Non-conformance of software application code	▸ Review and address advised practices for proper prompt usage. ▸ Ensure a Human-In-The-Loop process, such that a code-review gate in the development process is included prior to accepting AI-generated code.
	AI-generated code contains security vulnerabilities.	▸ Susceptible to malicious attacks caused by cybersecurity vulnerabilities leading to compromising the system.	▸ Include a Human-In-The-Loop process, such as a code-review gate in the development process, prior to accepting AI-generated code. ▸ Refer to the Organisation's in-house software development guidelines
Content Creation	AI-generated content contains correct information without reference.	▸ Prior work is embedded in AI-generated content without an indicating reference. ▸ Possibility of IP infringement or plagiarism issues in AI-generated content for academic articles or technical reports.	▸ Proper prompt engineering to ensure that references are included in the response. ▸ Usage of plagiarism detector tools (a database software to scan for matches between the generated text and existing texts) to check whether AI-generated content contains third-party content or IP infringement.
	AI-generated content is biased or discriminatory.	▸ AI-generated content contains information that is potentially biased or discriminatory. ▸ Reputational harm if the content is publicised or if its traceable back to its origin.	▸ Include a Human-In-The-Loop process throughout the development to evaluate AI-generated content not to include bias. ▸ Usage of moderation tools (e.g., software tools that help identify & remove harmful content), where possible, to check whether generated content complies with pre-defined usage policies.

Table 2: Categorical Risks and Mitigations. (continued on page 20)

Text Summarisation	Incomplete/inaccurate information summarisation.	▸ Innacurate response by missing parts of information and generating an incomplete summary yielding misleading or low-quality information.	▸ Domain experts are included in the process to review accuracy. ▸ Enhance prompt usage/engineering (see 2.2).
Entity Extraction	Incomplete/inaccurate information extraction.	▸ Incomplete extraction yielding misleading or low-quality information. ▸ Incorrect information extraction based on the prompt provided.	▸ Domain experts are included in the process to review accuracy. ▸ Proper prompt usage/engineering.
Format Conversion	Format conversion errors. AI-generated content does not comply with the correct targeted output format.	▸ Inaccurate or incomplete output leading to low quality generated content impacting the business's reputation in addition to rework efforts increasing costs and resources.	▸ Include a Human-In-The-Loop process throughout the development to identify errors or missing information. Rework to be concluded if required.
Explain and Tutor	Technical Concepts being explained without reference.	▸ Prior work is embedded in AI-generated content without an indicating reference. ▸ Possibility of IP infringement or plagiarism issues in AI-generated content for academic articles or technical reports.	▸ Proper prompt engineering to ensure that references are included in the response. ▸ Usage of plagiarism detector tools to check whether AI-generated content contains third-party content or IP infringement.
	AI-generated content lacks context, and therefore, being intrinsically ambiguous for audience of different technical backgrounds.	▸ Effectiveness of the AI tool does not meet expectations.	▸ Proper prompt usage/engineering to address level of detail and targeted audience.
	Incomplete/inaccurate explanation.	▸ Incomplete explanation yielding misleading or low-quality information. ▸ Incorrect information explanation based on the prompt provided.	▸ Domain experts are included in the process to review accuracy. ▸ Proper prompt usage/engineering.
Synthetic Data Generation	AI-generated synthetic data is not realistic, i.e., does not represent realistic features.	▸ Creates inaccurate representations in the dataset and any subsequent analysis performance will be degrading. E.g., inaccurate training data can lead to low performance of AI models trained using AI-generated synthetic data.	▸ Domain experts are included in the process to evaluate the quality of AI-generated synthetic data.
	Lack of class diversity of represented real-world data.	▸ Can lead to creating bias within the dataset producing a restricted set of scenarios or uses. For example, AI models trained on unequally distributed AI-generated synthetic data can lead to a degrading generalisation capability, hence lower performance. Inaccurate synthetic data generation can also produce misleading representations of real-world scenarios in training and education programs, which can result in learners not being adequately prepared for real situations.	▸ Domain experts are included in the process to evaluate the distribution (i.e., diversity) of AI-generated synthetic data.

Table 2: Categorical Risks and Mitigations.

5. Developing the Process of Using Externally Hosted AI Services and the Role of AI Governance

5.1 AI DEVELOPMENT LIFECYCLE

Figure 5 below illustrates the AI Development Lifecycle typically adopted by companies that create their own AI solutions. It has similarities to the CRISP-DM [24] methodology and the CDEI portfolio of AI assurance techniques [25]. Each stage involves different team roles, and while the general flow is cyclic, it is common for the process to become iterative between two steps, such that it arrives at an optimal state.

The different team roles are explained as below:

- **Domain Expert:** Provides context to the development team for the AI solution to be tailored for their application.
- **Data Scientist:** Translates the objectives put forward by the Domain Expert into data driven tasks.
- **AI Engineer:** Specialised in the AI/ML subset of Data Science, they are responsible for training the model and evaluating its performance and feeding back necessary adjustments.
- **Data Engineer:** Builds the systems/pipelines for the collection and storage of data. For an AI development that requires the use of large datasets, they can be particularly useful in managing the 'Data Gathering & Preparation' step.
- **Software Engineer:** Building applications around the AI model, and/or integrating the AI model with a pre-existing one.
- **Machine Learning Operations (MLOps) Engineer:** With the model deployed into production, the MLOps Engineer monitors and maintains the model pipeline, and ensures compliance with the MLOps framework [26].
- **AI Governor:** Maintains the oversight of the AI Development Lifecycle from scoping till retirement, evaluates the methodology used in the AI solution and provides feedback on any compliance-related issues and governance protocols against the relevant organisation's policies and standards. The AI Governor has in-depth knowledge of AI and are responsible for overseeing the development and use of AI systems. The scope of their oversight can vary from technical/methodology to bias and ethical concerns depending on the circumstance.

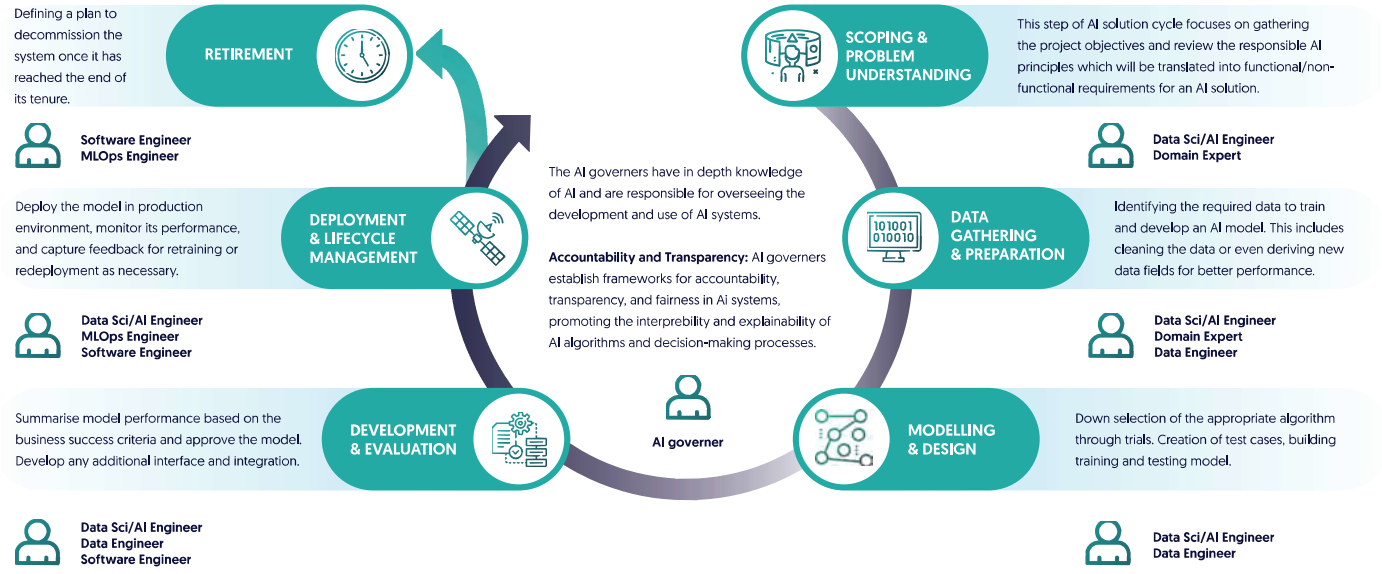


Figure 5: AI Development Lifecycle.

5.2 METHODOLOGY OF DEFINING ORGANISATIONAL AI USAGE POLICIES

There are no universal regulations applied to the use of externally hosted AI services, however, organisations are forming their internal frameworks and processes to analyse and mitigate the associated risks and maximise the potential of these services. The MTC has followed an approach that is similar to the one introduced in Figure 6 for developing the policy and process of using externally hosted AI services.

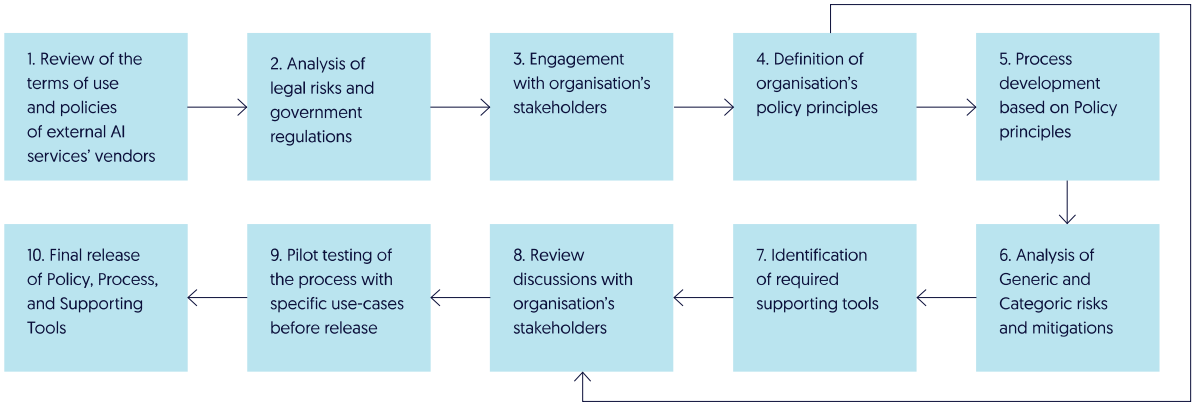


Figure 6: Approach for forming organisational processes of using externally hosted AI services.

The approach consists of 10 main steps as described below:

1. Reviewing the terms of use and privacy policies of different vendors providing externally hosted AI services facilitates a better understanding of the legal and technical risks and contributes to shaping the main guidelines and principles of the organisation's AI policy.

2. Risks and mitigations are analysed from a legal perspective and against the governmental regulations and policies. This analysis contributes to the organisation's AI policy as well as the Generic Risks and Mitigation list that is considered throughout the development of each use-case following the process of using externally hosted AI services.

3. Engagement with the organisation's different stakeholders on a regular basis throughout the development of the policy and process. Stakeholders include, but are not limited to, the Legal team, the IT team, and the Engineering team, end users, developers, regulators, policymakers, partners, etc. This is important to ensure that their requirements are considered and that their input is contributing to the policy and process development.

4. The policy of using externally hosted AI services acts as the base charter, or principles, which are embedded throughout the process development in the following step. The policy's clauses are defined in alignment with several factors including, but not limited to, the organisation's values, pre-existing policies of the organisation's and customers' data usage, legal risks and mitigations, and stakeholders input.

5. The process acts as a practical guidance for developing solutions involving a use-case that leverages an externally hosted AI service. Throughout its development, the process shall adhere to the principles/clauses introduced by the policy. It also clearly outlines the interactions between the relevant stakeholders and when their action is required. The process consists of a set of steps to be followed and supporting tools to be used throughout the full lifecycle of a project or an AI solution and until the end of its course.

6. The Generic and Categoric risks and mitigations are analysed following the structure introduced in Section 4. The analysis is recorded in a live register that is continuously monitored by the organisation's AI Governance team, such that it incorporates emerging risks and mitigations from new AI solutions over time. The register is also considered as a starting input to the use-case's specific risks and mitigation analysis.

7. To aid developers throughout the process, supporting tools are provided and referred to within the process's steps. Examples of such tools include, but are not limited to, risks and mitigations assessment templates, transparency report forms, IT forms, and checklists for providing information about the use-case and requirements needed to facilitate access to the external AI service.

8. Steps 4-7 above are reviewed and discussed with the involved stakeholders identified in Step 3 to ensure that their input and concerns are addressed. The review discussions are concluded iteratively and feedback to the policy and process development.

9. Upon the finalisation of the policy, process, and supporting tools, defined use-cases can be adopted for pilot-testing or implementing the process before its release to the wider organisation. This also enables other teams to offer their points-of-view and highlight gaps or modifications that need to be addressed in the process.

10. Once the pilot-testing is concluded, the policy, process, and supporting tools are made available for the organisation's teams' access.

5.3 PROPOSED PROCESS FOR DEVELOPING AI SOLUTIONS UTILISING EXTERNALLY HOSTED AI SERVICES

Developing policies and approaches in organisations for leveraging external AI services plays a significant role in governing and maintaining AI solutions utilising these technologies. In addition, a defined process ensures that an organisation incorporates the mitigations highlighted in Section 4. A proposed flowchart that depicts this development is illustrated in Figure 7.

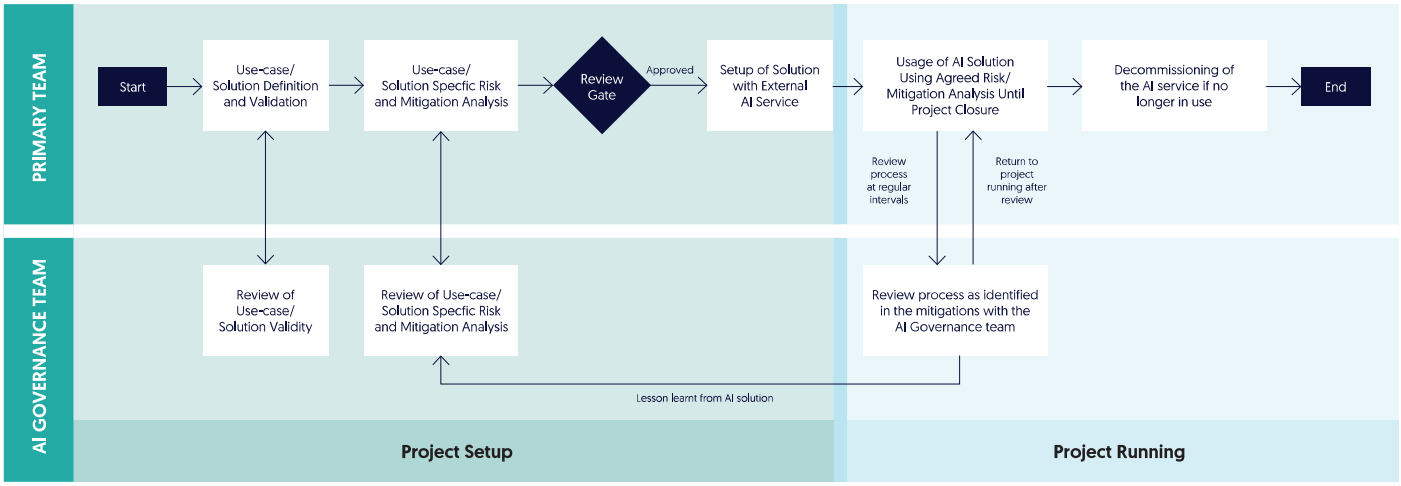


Figure 7: Process of Using External AI Services.

The process flowchart consists of a set of practical steps to be followed throughout the setup and running phases of a project or AI solution involving a use-case that leverages external AI services and until the end of the AI solution's course. There are three parties interacting with each other throughout the process: the Primary team, AI Governance team and Organisation Enterprise IT team as defined below:

1. **The Primary team:** organisation personnel using the externally hosted AI service. Typically, the Primary team would be the project team driving the development of the AI solution.

2. **The AI Governance team:** involve AI Governors monitoring all AI solutions. Generally, they would evaluate the methodology used by the Primary team and provide feedback on any compliance-related issues and governance protocols. Please see Section 5.1 for the description of an AI Governor.

3. **Enterprise IT team:** organisation personnel ensuring that the required resources are available to the primary team to interact with the external AI service as part of their solution. They will be allocating and managing the infrastructure and networking requirements in with respect to the organisational-specific IT policies with responsibility based on the Primary controlling the risks using the agreed mitigations and the review process.

The Process steps are described as follows:

A. Project Setup Phase:

1. **Use-case/Solution Definition and Validation:** The Primary team defines the AI Solution adopting a use-case using external AI services and assesses the use-case's validity against pre-defined guidelines provided by the AI Governance team.

2. **Use-case / Solution Specific Risk and Mitigation Analysis:** the use-case specific risks and mitigations are analysed by the Primary team taking into consideration the Generic and Categoric risks and mitigations, where applicable. The conducted risks/mitigations analysis is then reviewed with the AI Governance team.

3. **Review Gate:** With the Use-Case Definition and Validation and Risk and Mitigation Analysis steps complete, a final review is carried out on the process followed thus far and a decision is made as to whether the AI solution is appropriate.

4. **Setup of Solution with External AI Service:** The setup of the AI solution is facilitated by providing the required access to the AI Service with any other necessary Information Systems (IS)-based infrastructure with the support of the Enterprise IT team.

B. Project Running Phase:

1. **Usage of AI Solution:** the project is active, and the AI solution is deployed using external AI services. The analysed mitigations are applied by the Primary team until the project's closure.

2. **Review Applied Mitigations:** Within regular time intervals, the Primary team reviews the applied mitigations with the AI Governance team for their feedback. Lessons learnt in this review process are feedback to the Generic and Categoric risks and mitigation analysis to be considered in future AI solutions.

3. **Decommissioning of the AI service:** By the end of the project's/solution's running duration, if the AI service is no longer required then it is decommissioned.

6. Conclusion

This paper has tackled the pressing concerns surrounding the proliferation of Externally Hosted AI systems and their seamless integration with business operations. It has honed in on critical aspects such as data privacy, security, potential intellectual property infringements, and legal ramifications. In response to these challenges, a comprehensive guidance framework has been developed, exemplified through the characterisation of externally hosted Large Language Models (LLMs) like OpenAI's ChatGPT.

The framework comprises key recommendations, including:

- Establishing an AI Governance Practice
- Clearly defined use-cases, data sensitivity and volume.
- Thorough review of the terms of use and data policy for the externally hosted LLM services.
- Importance of human in the loop (HIL).
- Implementation of a multi-tiered risk analysis (See Section 4.1), encompassing:

    - Generic (See Section 4.2),
    - Categoric (See Section 4.3), and
    - Use-Case specific risks.

Moreover, a procedural flow (see Figure 7) has been proposed to integrate essential checks and balances, ensuring the appropriate deployment of these systems within business applications. Central to this process flow is the establishment of an indispensable AI Governance Team along with polices. This team assumes the pivotal role of overseeing the entire process, ensuring strict adherence to internal policies and legal requisites, and acting as a learning focus for future development. Its members, possessing extensive experience in AI systems, form the bedrock of this governance structure and leveraging AI for future opportunities.

By implementing this framework and process flow, businesses operating in the manufacturing sector can confidently navigate the intricate landscape of AI integration, simultaneously enhancing operational efficiency and ensuring the highest standards of safety and compliance. Future publications are planned to expand this work addressing the development of trustworthy and responsible AI frameworks for manufacturing.

7. References

[1]. L. Makatura, “How Can Large Language Models Help Humans in Design and Manufacturing?,” arXiv.org, 2023.

[2]. X. Wang, N. Anwer , Y. Dai and A. Liu, “ChatGPT for design, manufacturing, and education,” Procedia CIRP, vol. 119, pp. 7-14, 1 1 2023.

[3]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention is All you Need. In: Advances in Neural Information Processing Systems,” 2017.

[4]. OpenAI, “OpenAI Platform,” [Online]. Available: https://platform.openai.com/docs/introduction/overview. [Accessed 14 August 2023].

[5]. Google, “Bard,” [Online]. Available: https://bard.google.com/. [Accessed 24 8 2023].

[6]. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 24 May 2019. [Online]. Available: 10.48550/arXiv.1810.04805. [Accessed 14 August 2023].

[7]. R. Thoppilan and et al., “LaMDA: Language Models for Dialog Applications,” 10 February 2022. [Online]. Available: 10.48550/arXiv.2201.08239. [Accessed 14 August 2023].

[8]. T. Eloundou, S. Manning, P. Mishkin and D. Rock, “GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models,” 23 March 2023. [Online]. Available: 10.48550/arXiv.2303.10130. [Accessed 14 August 2023].

[9]. S. Ekaran, “Tom's Hardware,” 2023. [Online]. Available: https://www.tomshardware.com/how-to/use-chatgpt-for-3d-printing-g-code-stl. [Accessed 07 09 2023].

[10]. M. Chen et al., “Evaluating Large Language Models Trained on Code,” 2021. [Online]. Available: http://arxiv.org/abs/2107.03374. [Accessed 14 August 2023].

[11]. E. L. Rimban, “Challenges and limitations of ChatGPT and other large language models,” International Journal of Arts and Humanities, vol. 4, no. 1, pp. 147-152, 21 June 2023.

[12]. A. Tamkin, M. Brundage, J. Clark and D. Gang, “Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models,” 4 February 2021. [Online]. Available: http://arxiv.org/abs/2102.02503. [Accessed 14 August 2023].

[13]. AWS, “Amazon Web Services,” [Online]. Available: https://aws.amazon.com/what-is/gpt/. [Accessed 12 9 2023].

[14]. K. N. T. S. I. S. Alec Radford, “Improving Language Understanding by Generative Pre-Training,” OpenAI, 2018.

[15]. OpenAI, “GPT-4 API general availability and deprecation of older models in the Completions API,” 6 July 2023. [Online]. Available: https://openai.com/blog/gpt-4-api-general-availability. [Accessed 14 August 2023].

[16]. M. Schade, “How ChatGPT and Our Language Models Are Developed,” [Online]. Available: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed. [Accessed 15 August 2023].

[17]. OpenAI, “Enterprise privacy at OpenAI,” [Online]. Available: https://openai.com/enterprise-privacy.

[18]. OpenAI, “OpenAI Subprocessor List,” [Online]. Available: https://platform.openai.com/subprocessors. [Accessed August 2023].

[19]. M. Schade, “How your data is used to improve model performance,” [Online]. Available: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance. [Accessed 15 August 2023].

[20]. OpenAI, “Security & Privacy,” [Online]. Available: https://openai.com/security. [Accessed August 2023].

[21]. OpenAI, “New ways to manage your data in chatgpt,” 2023. [Online]. Available: https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt. [Accessed 2023].

[22]. OpenAI, “API Plaform,” [Online]. Available: https://platform.openai.com/docs/introduction.

[23]. OpenAI, “Terms & policies,” [Online]. Available: https://openai.com/policies. [Accessed October 2023].

[24]. IBM, “IBM Documentation,” 2021. [Online]. Available: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview. [Accessed 27 9 2023].

[25]. UK Government, “CDEI portfolio of AI assurance techniques.,” 7 June 2023. [Online]. Available: https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques. [Accessed 8 November 2023].

[26]. Databricks, “What is MLOps?,” 2021. [Online]. Available: https://www.databricks.com/glossary/mlops. [Accessed 27 9 2023].


LAN LPN
PLANNING
SERVICE
REPAIR

mtc
Manufacturing
Technology Centre

CATAPULT
High Value Manufacturing

the-mtc.org