EXCENTIUS

Implementing

Generative AI in portfolio companies

March 2024

eXcentius services

Value creation (cost reduction & revenue uplift)

Exit preparation

Hardware and software technology advisory

Strategic technology due diligence

Delivery assurance

Deal origination

Sector strengths

Technology-enabled Managed Services.

Cybersecurity.

Healthcare.

Life Sciences.

Business services.

Satellite communications and equipment.

Telecoms.

Edtech.

Fintech.

Cross-sector strengths

Data management for cross-border compliance.

Health services technology regulatory compliance.

Medical devices regulation.

Contents

Generative AI in a nutshell

Generative AI is a business issue, not an IT issue.

Can materially increase Enterprise Value when used to augment a business process.

Generative AI poses an existential threat to some industries and current investments.

Now a proven technology in the growth phase, and capability continues to advance at pace.

Relatively easy to adopt – it is does not require transformation of existing technology systems.

Proprietary enterprise data increases the ROI.

There are risks in implementation within a portfolio company.

December 2023

Using Generative AI models

From a user's perspective, the core of Generative AI is the prompt, which has three elements:

- What you 'ask' of the model (the query),
- Contextual information to focus the model's answer, and
- Output settings (what you want the answer to look like).

Unlike Machine Learning which asks the same question of different data, Generative AI can be used to ask different questions of the same data.

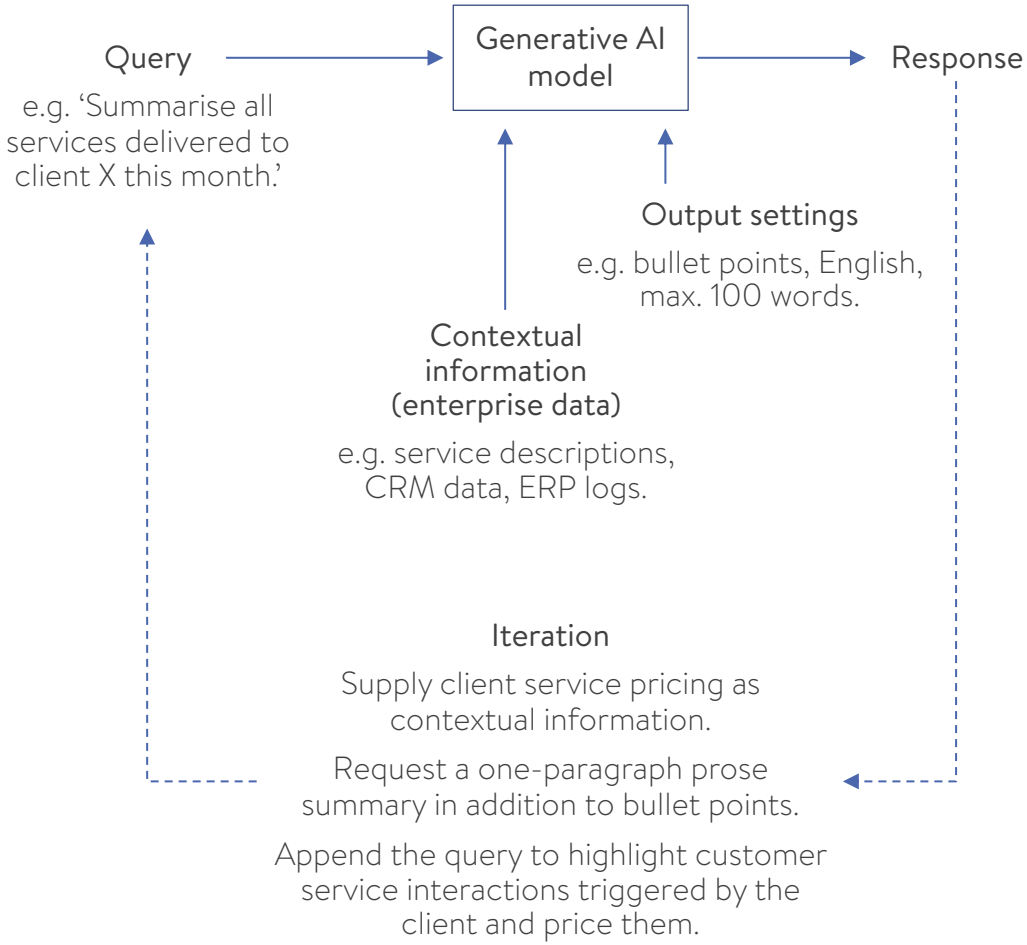
To avoid a generic answer, the user provides contextual information in the format of free text, written documents, video and image files, application log data, … for most commercial models this data can be kept private.

Generative AI models can return a response in a myriad formats, styles and structures, and extrapolate to the n^{th} degree. This can be controlled by specifying the output parameters, for example 'written in the style of the Economist, in Spanish, in 300 words with an additional summary presented in bullet points.'

A key advantage of Generative AI services is that they enable sessions that 'remember' contextual data and previous responses. This allows for iterative prompting to fine-tune the final answer.

When starting out, prompting is an iterative process – it can take several weeks to identify the right contextual information, query phrasing and output settings to obtain the best response for a specific business process. Once these settings have been determined, they can be pre-set to make model usage efficient.

Schematic of model interaction



The need for enterprise data

There are two sets of data required to use Generative AI models effectively: the training data and the prompt data.

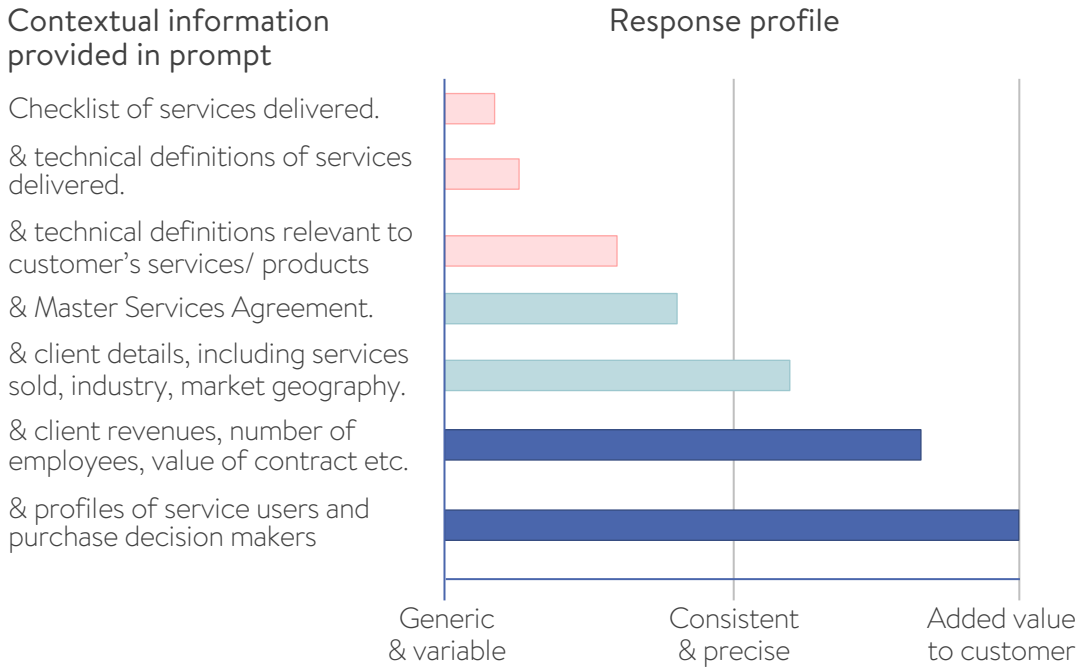
Generative AI models are Language Models trained on very large bodies of information[1], often obtained from a multitude of sources in multiple human and machine languages. As a result, the models have embedded knowledge that is far beyond the horizon of any company or employee. This enables the models to be excellent at finding similarities and summarising across large volumes of data. However, it also means that they give generic responses by default, which tends to be of little commercia value.

To increase the focus of the model's response, and thus the commercial value and ROI of deploying Generative AI, it is necessary to give the model contextual information when prompting. This additional 'prompt data' is not part of the model and can be varied with every request made to the model[2].

For commercial operations, the best prompt data tends to be the company's proprietary data that is spread across the entire enterprise (in all disciplines from sales to operations, in all formats including documents, chat messages and video calls) – it is a company's enterprise data that creates the commercial value[3].

Prompt contextual information & model response

Example prompt: Write a summary of the technical service delivered to client



1. Whilst there is no set definition, small Language Models tend to have about 10m parameters, large models have 1bn+ parameters, and very large models, such as OpenAI's ChatGPT version 4, have in excess of 1tn parameters. In contrast, AI models used for medical diagnosis, even in the most complex fields, typically have less than 1m parameters.
2. Some commercial services, including ChatGPT, retain prompt data within a session so that it can be built upon when asking a set of iterative questions.
3. Check the terms and conditions of the model service being used! Some services allow the model owner to use proprietary prompt data for subsequent retraining of the model and for fine tuning responses to other people's prompts – i.e., your proprietary data can become public information.

Non-integrated & integrated implementation

Non-integrated implementation

Model operators such as Amazon, Google and Microsoft each provide a generic prompt interface (a web page) which can be used manually as part of a business process. A proliferation of third-party interfaces also exist. They are designed to make it easier to specify more complex query and output settings, and they provide pre-loaded contextual information to increase the utility of the model for specific use cases in specific industries.

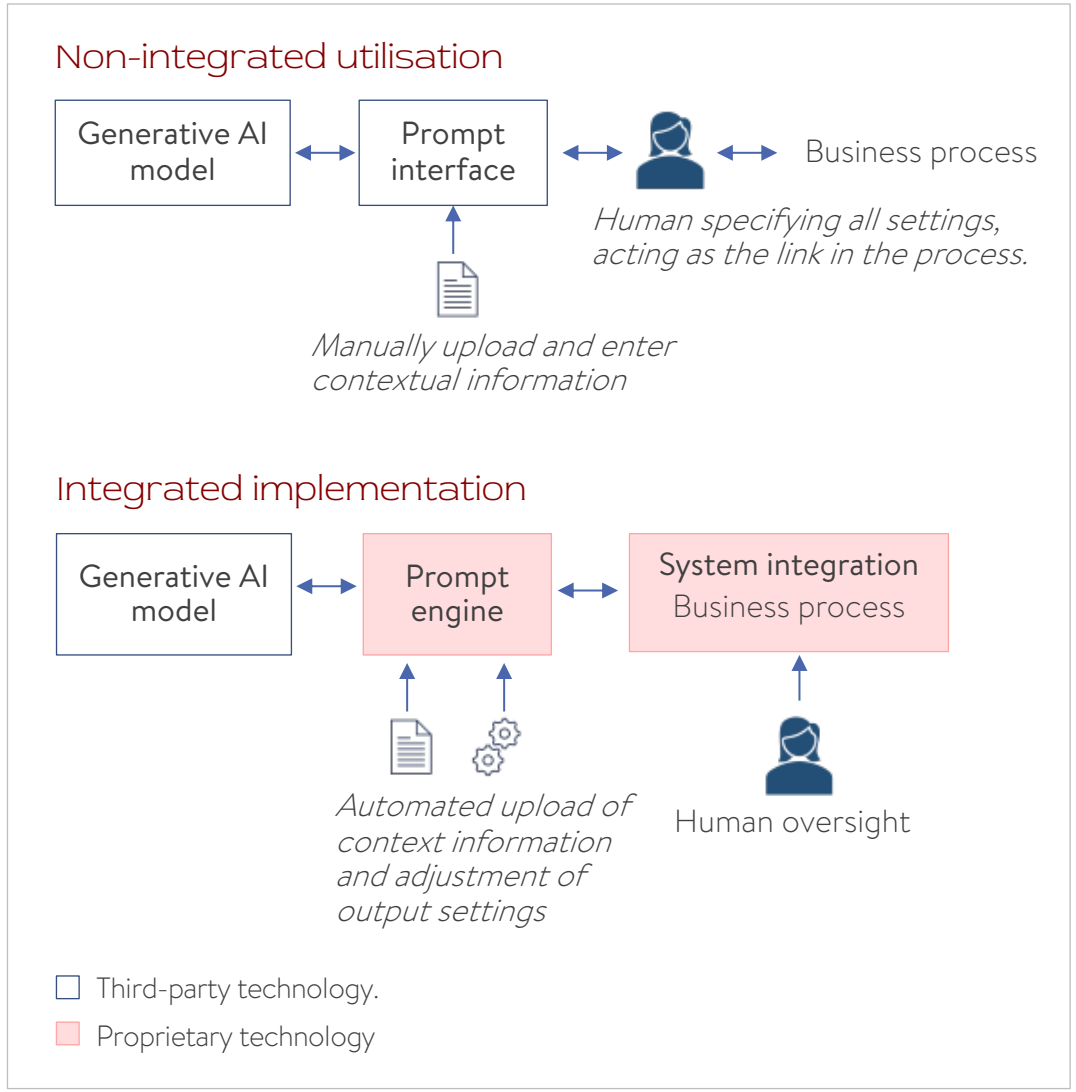
The prompt interface is used independently of a company's technology systems – the user manually uploads documents, types in free text and settings, then acts upon the response. This process requires tools embedded within the model service, such as the ability to read Microsoft Office documents.

Non-integrated implementation requires the user to have access to the needed contextual data and it relies on that user to enforce confidentiality.

Integrated implementation

Alternatively, you can build a proprietary 'prompt engine'– a small software product that programmatically interacts with the model to automate queries, prompt settings and the provision of contextual data. The response can also be integrated to trigger actions in other business systems, for example capturing the written response as a PDF document in the CRM. In either instance, human oversight is essential to catch inappropriate or incorrect model responses.

Back-office systems providers, such as Salesforce and Microsoft, are already providing model integrations so that you can integrate Generative AI into your business processes without building your own prompt engines. Adoption still requires some degree of systems engineering, but significantly less proprietary development. Of course, systems providers charging for the privilege...



Ensuring optimal ROI: the right project people

Implementing Generative AI in a business process is a project of two parts: identify the opportunities and prove the business case, then build a product (if required). The purpose of the project is business, not technology, and IT and (software) Engineering are only required at the end of the project if the business case is proven to proceed. Therefore, we recommend the following team structure to ensure success:

Project role	Ideal candidate	Rationale
Project sponsor	Member of the board	The objective of implementing Generative Ai is either to shift EBITDA or strategically protect revenues or exit multiple, and implementation requires additional investment. The board have this expertise and authority.
Project owner(s)	Member of the Senior Management Team (SMT), responsible for an operational area where Generative AI is being considered.	Successful implementation will require access to enterprise data that is typically under the control of members of the SMT, and may also require changes in operational metrics, personal performance metrics, personal objectives, and staff allocation. In addition, the junior staff who should be tasked with exploration (see 'Use case explorer' role) will need protecting from middle management. Only the SMT have this authority and capability. In addition, the Project Owners will be responsible for the quality of the business cases, and ultimately for financial success if implemented.
Project Manager	Product Manager or senior project manager.	The candidate(s) need to be senior because they will be navigating across organisational functions and around interests that are vested in the status quo. The candidates also need a solid understanding of P&Ls and business metrics. If the candidate is from the Product Group, they must be commercial and not technical (hence a Product Owner is not suitable).
Business case analyst	Analyst with experience of business P&L modelling.	Scenario modelling of the P&L is required to produce a realistic business case because the impact and utilisation of Generative AI is not certain. Whilst the model inputs will come from others, experience is required to deliver realistic forecasts.
Use case explorer	A person involved day-to-day in operational processes that involve a creative element. Junior to below middle management.	The purpose of this role is to discover potential opportunities for Generative AI in the business, without filtering. A person who works within a process on a daily basis has an intuitive understanding of the challenges, inefficiencies and process opportunities. A junior person is more likely to play with Generative AI and be more inquisitive, and thus discover more opportunities. Middle Management tend to have vested interests not to change.

Ensuring optimal ROI: finding the opportunity

Where to start: map the opportunities

The map opposite can be used to identify areas in the business that may yield commercially viable opportunities for Generative AI. Each area can be narrowed using the table 'Where Generative AI can deliver the most investor value' of page 8.

The use cases of Generative AI are myriad within these areas, even for specialist smaller models, and the ideal prompt settings, input data and response parameters will vary by individual process. Therefore, the best way to identify opportunities is to explore by trial and error.

Explore by trial and error

As highlighted previously, a junior person involved day-to-day in the processes is the the best type of explorer. To be successful they need to be given time, a non-judgemental space, and protection from interference by others.

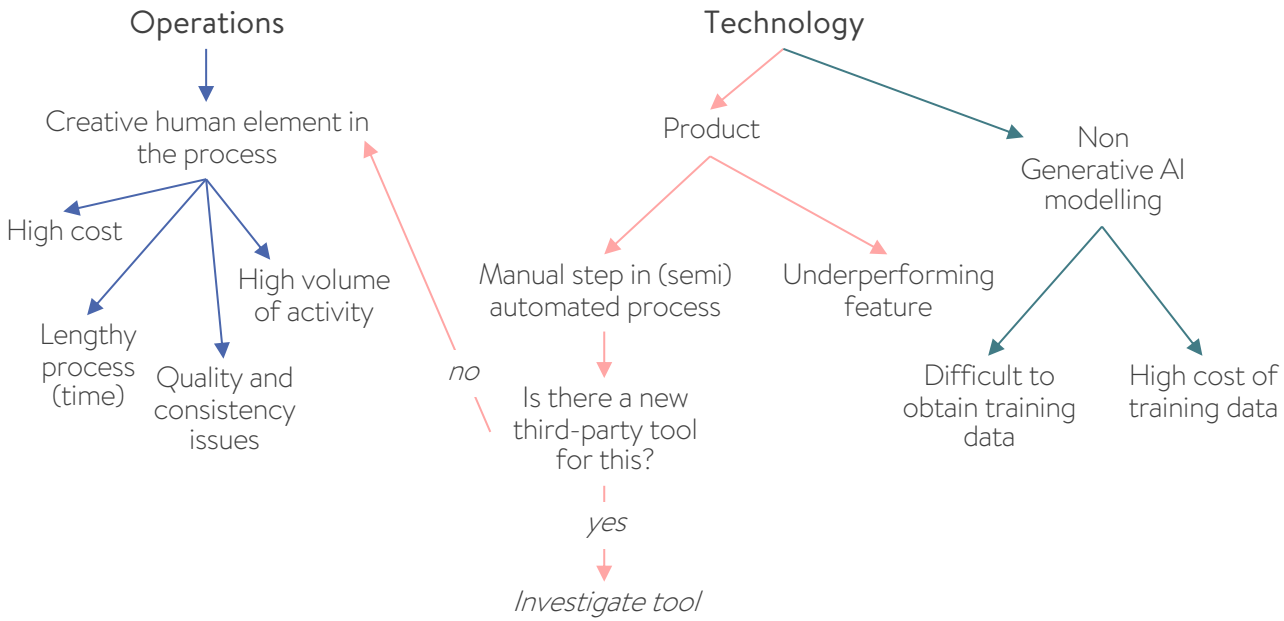
Exploration also requires decent tools. We recommend a paid-for web-access account to the model service, which is designed to enable exploration. Using a free version will probably be inefficient due to limited features.

Set a fixed time frame for initial exploration.

Use a generalist model

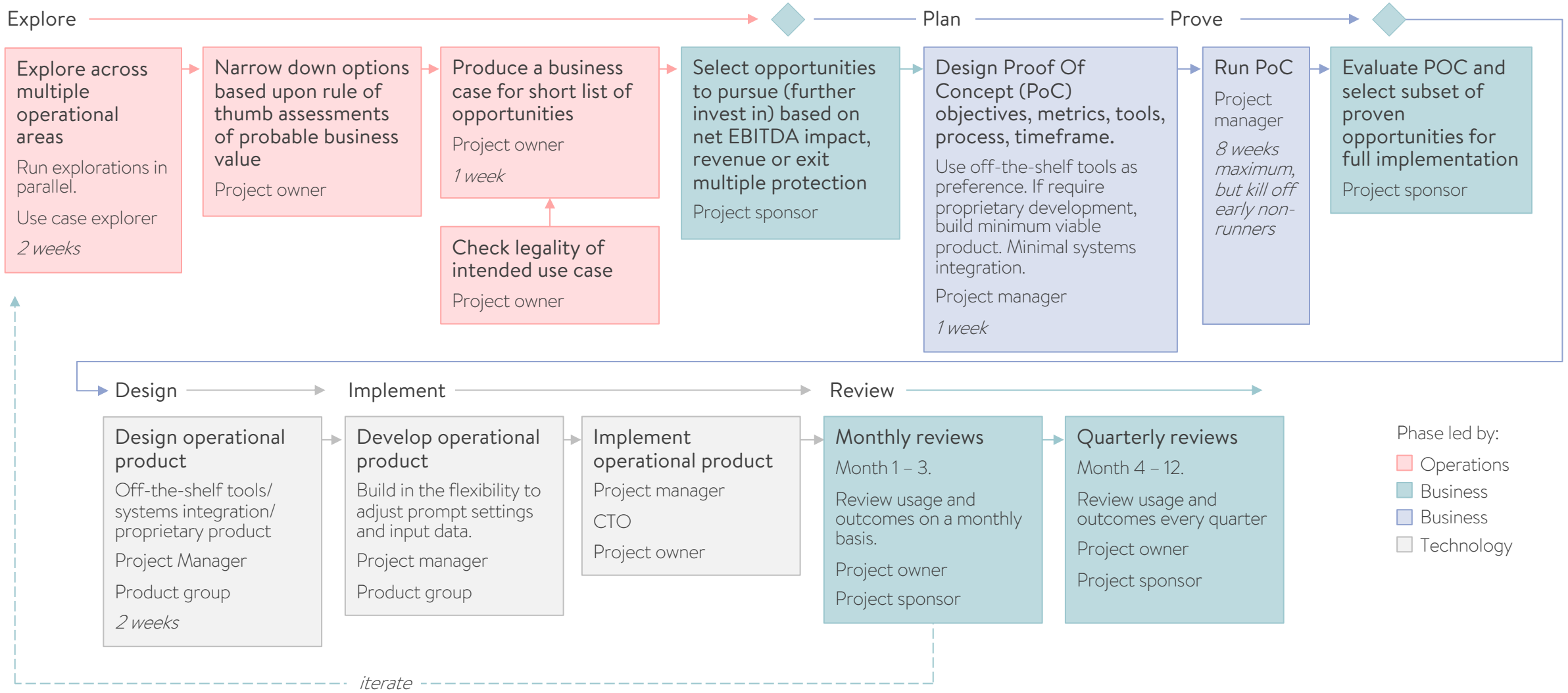
Start with a generalist, large model for the exploration – they present more opportunities because they are generalist, but they take a bit longer to fine tune.

Map to identify where to start looking for financially viable Generative AI opportunities within the operating business.



Implementation process overview

The following process is designed to maximise investor ROI when implementing Generative AI in portfolio companies.



Legality of operation

There are two aspects to the legality of AI: the data used by the model, and what the model is used for.

Data

Like most models, Generative AI has training data that is used periodically, and input data that is provided each time the model is called. Unlike most AI, the large Generative AI models have been trained on very large volumes of data that includes unauthorised copyrighted material. At present this material is being used freely but there is a growing regulatory consensus that the model owners must respect copyright laws, or at least enable users to do so. It is likely that the model owners, not the model users, will be forced to provide remedies for this issue. If that is the case, it is reasonable to expect the model owners to make it as easy as possible for model users to comply to new regulation (and hence continue to use their paid-for services).

There are very clear regulations regarding the data used for prompting AI models, including general laws such as the EU's GDPR and industry regulations such as the USA's HIPAA. These regulations govern what data can be used, how it can be collected, and how it must be managed.

Model usage

At present there is little regulation governing what the output of an AI model can be used for. But this is changing rapidly, and the EU's new regulation of AI is expected to come into force in 2024 having now achieved provisional agreement. As with other EU regulation governing access to its markets, it is likely that this legislation will set the standard which portfolio companies choose to comply to (even if national regulation varies, as is expected with the United Kingdom for example).

The EU regulation assigns categories of risk to AI models, based upon capability regardless of the context within which it is used. Essentially, the higher the risk, the stricter the rules. But there are certain capabilities that will be banned outright, including cognitive behavioural manipulation, emotion recognition in the workplace and educational institutions, and social scoring.

Specialist intermediaries

A few model owners due to cost

Due to the complexities of design and the cost of training models, owning proprietary Generative AI models is currently beyond the reach of most businesses. We expect to see a fall in access costs as new processing GPUs are brought to market. But there are no indications that set-up costs will change materially in the foreseeable future, thus we expect Generative AI models to remain a service of the global infrastructure providers (Google, Microsoft, AWS etc.) and a small number of independent model owners who provide smaller models focused on specific industries.

A multitude of specialist intermediaries - unlikely to survive the next 12-24 months

There is a large and growing number of service providers who are selling access to 'specialist versions' of Generative AI models. These intermediaries provide a proprietary prompt engine to access the models owned and operated by others, which uses their own settings and proprietary data tailored to industry verticals. Explicitly, they do not have proprietary Generative AI models, only a proprietary way of prompting and parsing responses.

The large model service providers have already started to acquire and consolidate the specialist intermediaries because they add value to their model services.

The business has the true value, not the intermediaries

In many cases it is probably an illusion that using a pre-configured intermediary will make it easier and quicker to adopt Generative AI – only the business knows the best prompt settings and how to fine tune them, because only the business has the intimate knowledge of the process it is trying to augment. And in most cases, the business has the best prompt data.

Nor is it technically difficult to implement Generative AI into a business process - beware the specialist 'prompt engineer', who is actually a standard programmer who has simply read the instruction book of the model provider, something your own IT staff can do themselves.

Using intermediaries over the next 12 months

For the next 12 months we see material risks in using an intermediary to access a Generative AI model service such as ChatGPT, rather than using the model owner's service directly.

Using intermediaries might return value more quickly, but their additional costs can easily erode ROI over the longer term (particularly costs incurred in fine tuning the prompting). There is also a material risk that an embedded intermediary will cease to operate, resulting in expensive business interruption.

The value and risks of using intermediaries will be more transparent and certain after they have matured and the market has consolidated, which we expect to see in 12 - 24 months.

RAG, vector databases & accelerated digital transformation

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) is an AI framework for improving the quality of Generative AI responses. As with uploading contextual information in an unstructured format, RAG grounds the model on external sources of knowledge, ensuring that the model has access to the most current, reliable facts. It also ensures that users have access to the model's sources, allowing claims to be fact-checked and thus enabling the ability to build trust.

By systematically grounding through RAG, the model uses less of the information that was embedded into it in training, reducing the chances of leaking sensitive data and of 'hallucinating' incorrect or misleading information.

And as IBM notes, 'RAG also reduces the need for users to continuously train the model on new data and update its parameters as circumstances evolve. In this way, RAG can lower the computational and financial costs of running LLM-powered chatbots in an enterprise setting'.[1]

Whilst this technology is being adopted at pace, RAG is currently not perfect, can be difficult to set up correctly, and requires advanced IT and database skills to implement.

Vector databases

RAG works by presenting data as vectors, which requires the underlying information to be stored in a vector database. These databases are highly efficient at indexing, storing and retrieving information. This is because, unlike relational databases of rows and columns, data in a vector database is represented by vectors with a fixed number of dimensions that are clustered based on similarity. This clustering is a semantic understanding of the relationships between individual parts of the data stored within them, without an understanding of the meaning of the data.

Data in vector databases can also be labelled to ensure confidentiality, access restrictions, privacy and regulatory compliance.

Accelerated digital transformation?

In most organisations, by far the biggest problem faced in digital transformations is the categorisation of enterprise data based upon the meaning of the data.

Good categorisation is the cornerstone of success when using traditional relational databases because it enables fast and cost-efficient data retrieval. However, understanding the meaning and designing categories that can be used efficiently in multiple applications is both time consuming and expensive, at in many larger companies it is effectively an impossible task.

Deploying vector databases for Generative AI does not require meaning-based categorisation of enterprise data – thus the main bottleneck of digital transformation can be avoided, or at least materially mitigated.

1. Source; 'What is retrieval-augmented generation?', Kim Martineau, 22 August 2023. https://research.ibm.com/blog/retrieval-augmented-generation-RAG. Accessed on 10 December 2023.

Operational risks

1. Inadvertent use of confidential information

Most businesses have both confidential and non-confidential data sitting on the same IT systems (albeit partitioned with different access controls) and, by default, a generative AI model will read all information it has access to. This can result in confidential information being consumed by Generative AI models within the business when it is inappropriate or not permitted. The solution is to enforce access controls, label data and assign usage restrictions.

2. Biased decisions based on data

Processes enhanced by Generative AI can lead to biased decisions based on data due to AI hallucination and due to reduced human questioning of automated decisions. The impacts can be indirect and manifest over time, such as slower revenue growth due to increased customer churn, or direct and immediate such as incorrect assessments made in a regulated financial process. The solution is to ensure on-going human oversight and routine assessment of model / process outputs.

3. Breach of copyright

Inadvertent plagiarism is a real possibility because most of the commercial Generative AI models do not report sources unless instructed to do so. Models also paraphrase, of which the legality in as yet untested. Source checking is essential when using quoted information, and a simple Google search can be used to highlight any possible infringement.

4. Release of proprietary, confidential or regulated information

The terms of most commercial model services allow for the use by the model owner of any data uploaded by a user. This data is often used to augment the model, and thus can appear in model responses given to other users (without copyright protection).

Third-party use of private, confidential and regulated data can usually be avoided by labelling uploaded data in the prompt process – though we recommend close inspection of the service terms and conditions.

5. Model collapse through use of synthetic data

When instructed to create synthetic data to be used by other ML models, Generative AI will automatically produce data which focuses on a limited number of scenarios and which excludes the atypical (rare) scenarios found in the real world. Most ML models will collapse (cease to function accurately) if trained on a limited number of scenarios. Model collapse can be avoided by explicitly directing the Generative AI model to create atypical training data.

6. Insufficient transparency

In default mode, Language Models do not provide clarity on what data has been used and why interpretations have been made. This lack of transparency can be problematic, particularly where individual people receive negative service outcomes based on AI inputs. Generative AI can be instructed to cite information used in a response, but it remains difficult to obtain an explanation of the data interpretation in a format that a lay person can easily understand.

eXcentius case studies

Case study | Value creation & investment preparation

Portfolio company

US based provider of military and civilian satcoms equipment, including hardware and software for ground station hubs and modems for ships, planes and fixed sites.

Global revenue USD$0.5bn.

Value created

- Product portfolio strategy and roadmap that reversed $261m of declining revenues and increased total market share.
- Market research, customer research and product analysis to secure $100m investment raise for product portfolio transformation.
- Design of the technical architecture of next generation hardware & software products to increase long-term ARR.
- Map of operational transformations to transition Product Management and Engineering functions and reduce cost base.
- Identification and recruitment of off-shoring specialist technology partner.

Critical insights

Our forensic examination of the market technology assumptions in the 5-year P&L forecast showed that 61% of projected revenues were not achievable with the company's current technology strategy.

Case study | Value creation

Portfolio company

A global telecoms operator providing mobile voice and data services for civilian and defence, with a complex product and services portfolio due to little customer migration to newer products over a long trading history.

Global revenue USD$1.4bn.

Value created

- +9% EBITBA.
- 30% reduction in cost of Customer Support.
- Reduced complexity in Sales, Product and Engineering management.
- 37% reduction in technical debt.

Key deliverables

Rationalisation of the portfolio from 140 to 40 products, with ring-fenced critical products (military, blue-light etc).

Product retirement plans and product feature migration plans.

Frameworks to renegotiate contractual terms with suppliers and customers.

Architecture for customer outreach plans.

Case study | Technology due diligence of telemedicine service

Target company

UK-based remote dermatology services in secondary healthcare, with majority of revenues from health insurers.

Clinical diagnosis of skin conditions using a patient app and ‘out of hours’ assessment by physicians.

Proprietary technology of mobile app, case management software and image analysis AI.

Key findings

- 37% of investment period revenues unlikely to be achieved.
- Systematic reduction in Enterprise Value due to transfer of Intellectual Property

It would not be possible to deploy the required AI within the investment timeframe due to the unforeseen need for regulatory certification, and the product roadmap would not deliver the additional revenue-generating services on time.

The shareholding and licensing agreements, plus the technical requirements of scaling the proprietary AI, would lead to transfer of IP to the value of a loss in 3x multiple.

Additional deliverables

- Design of a realistic technology platform to deliver the investment thesis.
- Identification of an alternative investment.

We designed a technology platform that could deliver the investment objectives, articulating the technical architecture, costs, skills and schedule to implement, and the associated product roadmap.

And with the investor’s consent, we conducted a global search for an alternative target and identified a more mature company with better technology and a better fit for the investor’s existing portfolio businesses. We provided introductions to the board, a view of the quality of the technology, and an assessment of potential fit within the investor’s portfolio.

Legal disclaimer

This document was prepared for information purposes and should not form the basis of, or be relied on in connection with, any investment decision or any contract or commitment whatsoever with respect to a proposed transaction or otherwise.

Any recommendations contained herein are necessarily based on technical, financial, economic, market and other conditions prevailing as at the date that the due diligence report was authored by eXcentius (stated in the title section of this report). eXcentius is under no obligation to update the report (including the recommendations). Accordingly, the report does not take into account any information, events or financial, economic, market or other conditions that has (or have) become apparent or come into existence since the date of submission of the report.

The report was prepared by eXcentius based solely on information obtained from suppliers contracted to eXcentius, individual people in discussion with eXcentius and from public sources on or prior to the date of issue. eXcentius has assumed and relied upon without independent verification the accuracy and completeness of such information for the purposes of rendering the report. No representation or warranty, express or implied, is or will be made in relation to the accuracy or completeness of the report (including the recommendations) and no responsibility or liability is or will be accepted by eXcentius or by any of its officers, employees or agents in relation to it. eXcentius and its subsidiaries and associated companies and their respective officers, employees and agents expressly disclaims any and all liability which may be based on the report, and any errors therein or omissions therefrom.

The report does not constitute an offer or invitation for the sale or purchase of shares or any of the businesses or assets described herein and does not constitute any form of commitment or recommendation to prospective purchasers or any other person on the part of eXcentius or any of their respective subsidiaries or associated companies.

The distribution of the report in certain jurisdictions may be restricted by law and, accordingly, recipients of the report represent that they are able to receive this information without contravention of any unfulfilled registration requirements or other legal restrictions in the jurisdiction in which they reside or conduct business. In particular, the report is only being distributed in the United Kingdom to persons who (i) have professional experience in matters relating to investments or (ii) are persons falling within Article 49(2) (a) to (d) (“high net worth companies, unincorporated associations etc.”) of The Financial Services and Markets Act 2000 (Financial Promotion) Order 2001 (as amended) or to whom the report may otherwise be lawfully distributed (all such persons together being referred to as “relevant persons”). The report is directed only at relevant persons and must not be acted on or relied on by persons who are not relevant persons. Any investment or investment activity to which the report relates is available only to relevant persons and will be engaged in only with relevant persons.

Technology advisory | Strategic technology due diligence | Value creation & exit preparation | Deal origination

For further information, please contact Paul Oliver at paul.oliver@excentius.com.
www.excentius.com