Freeform Dynamics in association FUJITSU

Generative AI Without Strings

A mid-size business perspective on the growing importance of private AI platforms

Time to demystify the 'magic' and get real

If your organization is like most others, you will have flirted with Generative AI (GenAI) through generic cloud services such as ChatGPT, or by switching on GenAI features in SaaS applications. Despite some success, the chances are that results have been mixed, and that you're still figuring out where the true value lies.

If you're further along in your GenAI journey, you may have explored the use of APIs to build your own applications that exploit cloud-based large language models (LLMs), perhaps leveraging some of your own data. Along the way, however, a range of control, cost and risk related issues have likely surfaced, together with a need to pay serious attention to data quality and integrity.

Zoom out to experiences across the whole business community, and there's an increasing realization that while GenAI shows huge potential, the truth is that it doesn't 'change everything' as many claim. When it comes to implementing it in a business context, the same overall approach and set of disciplines are required that you will have applied many times before in relation to previous data-driven solutions.

Against this background, our aim is to demystify the GenAI 'magic', and take a practical look at how mid-sized businesses can take advantage of what GenA has to offer.

Who is this paper for?

This paper has been written for IT or engineering professionals looking for a business-centric perspective on generative AI deployment, and tech-savvy business leaders seeking a better understanding of where and how generative AI might fit into their plans and activities.

Why this discussion, and why now?

When ChatGPT burst onto the scene in late 2022, the media painted a picture of imminent, radical change across all aspects of work and life. Investors scrambled to get a piece of the action, while technology providers rushed to incorporate GenAI features across their existing offerings.

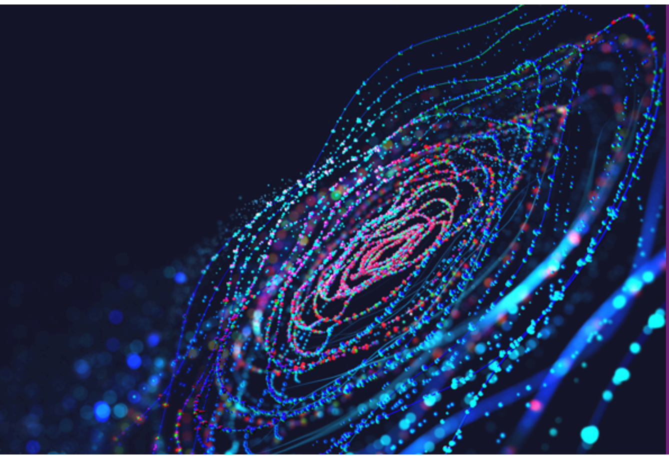
Wind the clock forward, and the sentiment today is quite different. Early pilots and deployments in business have shown promise, but most organizations largely remain at the exploration or pilot stage.

So why haven't things moved as quickly as anticipated?

Well early experiences have demonstrated that effective use of GenAI for most business purposes relies on incorporating your own data into the mix. The generic data sets used to train the large language models (LLMs) that underpin familiar GenAI cloud services, are inadequate to drive business decision making and customer interactions. Responses need to be grounded in accurate and up to date business information.

This in turn means you need to address data quality, security and access related concerns, which take on a new dimension with GenAI, especially if you're sending sensitive data to to a public cloud service.

Add lingering uncertainty about which use cases will deliver tangible real-world ROI, and it's not surprising to see many GenAI initiatives being paused or slowed down.



Most organizations still at the experimental or early pilot stage, it's better to take time to do things properly rather than rushing for fear of missing out.

It's time for a reset

Against this background, we need to get things onto a firmer footing and reset initial expectations. An understanding that GenAI is as much about data as it is about technology is an important part of this.

In the remainder of this paper, we'll review some of the key principles and imperatives in the AI platform space, and why cloud-based AI may not always be the best answer.

Along the way, we will look at an example of a real world solution - Fujitsu's Private GPT - to illustrate how modern platforms can ease the path to safe, secure and effective GenAI deployments.

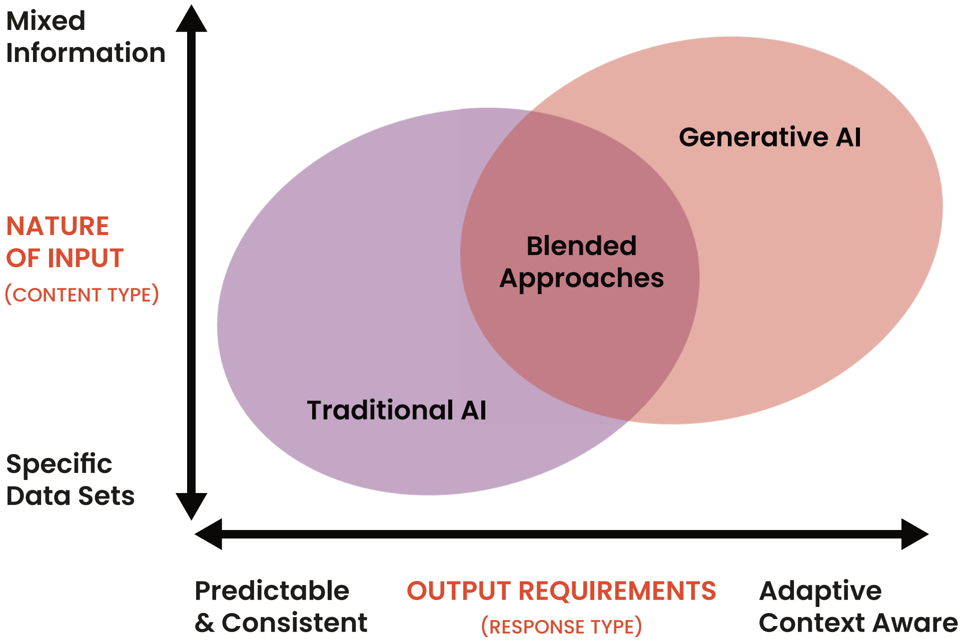
Let's start our discussion with a review of some fundamental principles and concepts that will help set the scene for our discussion of GenAI platforms and where they fit.

A level-set on some key fundamentals

GenAI complements established AI models

While GenAI has captured imaginations and is currently receiving the lion's share of attention, it's important to understand that other forms of AI have existed for decades. Indeed, many traditional approaches, such as supervised machine learning, neural networks, and rule-based expert systems, are very well proven and still represent a better option than GenAI in some areas. The choice between traditional and generative AI - or a blend of both - often depends on the nature of your input and output data requirements.

Selecting the right AI approach



Traditional AI excels when working with specific data types and delivering predictable, consistent results. Consider fraud detection, where models analyze structured transaction data against established patterns, or quality control systems that process sensor data to make reliable pass/fail decisions.

Generative AI, by contrast, is particularly well suited to handling diverse information sources (including unstructured data) and delivering more contextual, adaptive outputs. This makes it ideal for tasks like mining documentation to answer user questions, or analyzing complex scenarios with a diverse set of variables to recommend an action.

We'll explore use cases more fully in the last section of this paper, but for now we can say that GenAI's ability to understand context and generate appropriate responses sets it apart from traditional approaches.

But traditional and generative AI are not mutually exclusive and are increasingly used together. As an example, you might use traditional AI to classify and route incoming customer inquiries, then GenAI to draft contextually appropriate responses.

Implicit in everything we have mentioned is the incorporation of your own data into AI solutions. It's worth taking some time to explore this further, at least at a high level.

GenAI applications must be business aware

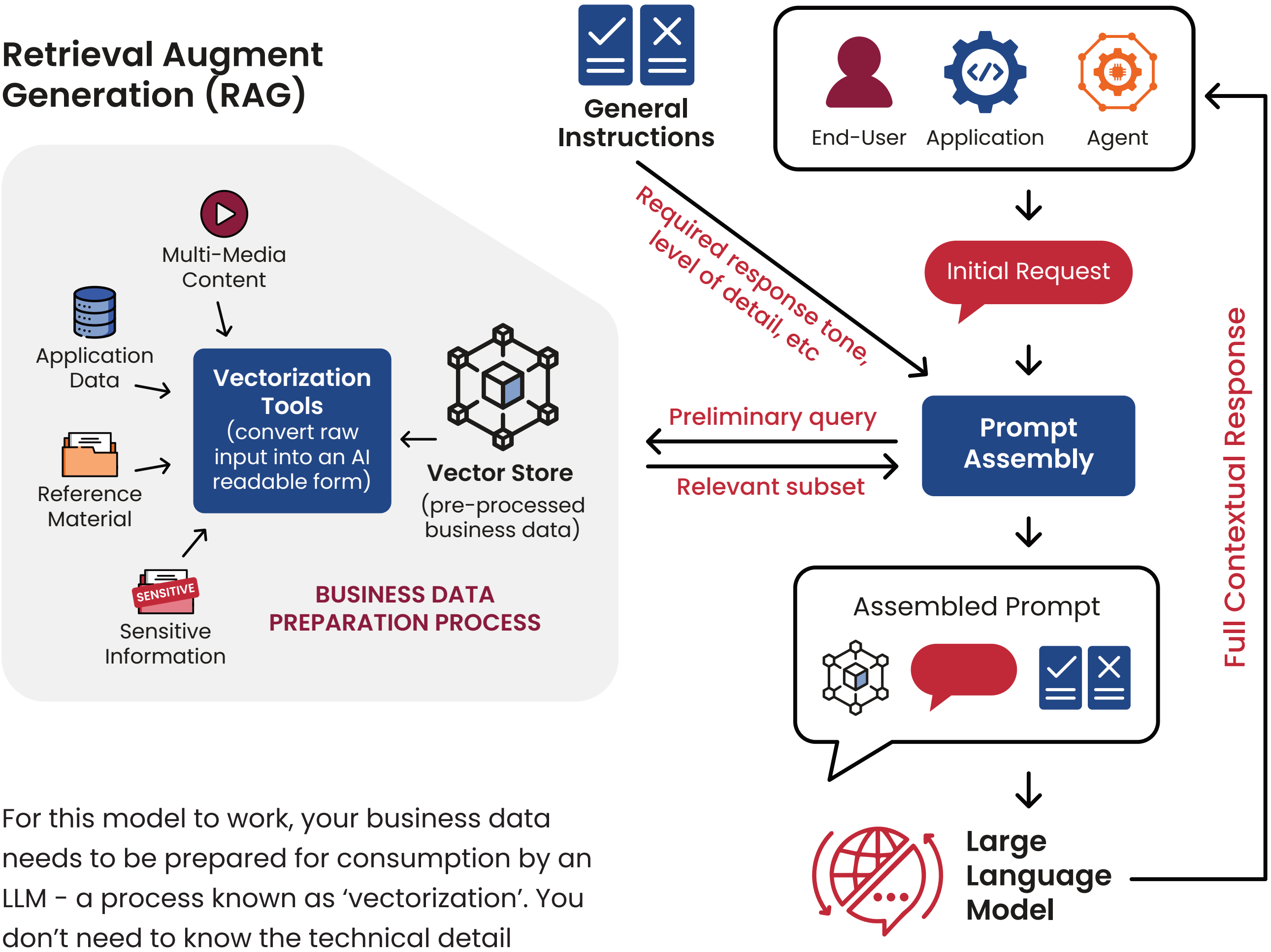
Requests sent to a general purpose GenAI platform through a chat interface (such as ChatGPT) are often referred to as 'prompts'. Prompts can be simple, e.g. "Why is the sky blue?", or can include more extensive instructions, e.g. "Please respond as Albert Einstein would when giving a lecture to post-graduate students", or "I need an explanation that my 5 year old child will be able to understand".

The latest incarnations of cloud-based GenAI services will also allow you to upload reference data, e.g. in the form of your own PDFs or image files. If you are familiar with this, you'll know how useful that can be.

The same basic concepts apply when building GenAI applications for your business, but the mechanism is a little different, typically based on a technique known as Retrieval Augmented Generation (RAG).

A RAG 'engine' firstly captures the initial request from a user, application or agent. Behind the scenes, it then adds further instructions plus relevant business data to construct an enhanced (augmented) prompt to forward to the LLM. This provides everything the LLM needs to generate a response grounded in your company's data and in line with relevant policies and preferences (see the right hand side of the graphic below).

Retrieval Augment Generation (RAG)



For this model to work, your business data needs to be prepared for consumption by an LLM - a process known as 'vectorization'. You don't need to know the technical detail behind this process, just think of it as converting your data into a form that a GenAI system can access in an efficient, high speed manner. The resulting vectorized data is held in a Vector Store that can be refreshed as often as your business needs require.

RAG transforms a general purpose AI system into a fully business-aware platform. With your potentially sensitive data now integral to the GenAI solution, however, this raises the question of control and protection.

Sovereignty and control really matter

The word 'sovereignty' comes up frequently nowadays in conversations about the public cloud. Sometimes it's associated with the location of data - e.g. whether a provider physically stores your data in a geography that's compliant with relevant regulations.

Increasingly, however, sovereignty is also used to refer to the level of control you retain over important decisions - e.g. how data is secured and protected, your ability to switch platforms, or the degree to which you have freedom to make changes at your own pace according your own priorities and timescales.

Both interpretations of sovereignty matter when considering the GenAI choices you make, and this brings us to the pros and cons of using hyper-scale cloud-based AI services from companies like OpenAI, Anthropic, Microsoft and Google.

At this early stage of the market, it's doubly important to properly think through supplier commitments before you make them. This will minimize the risk of getting caught out by changes to models, services, pricing, contract terms, development frameworks and so on.

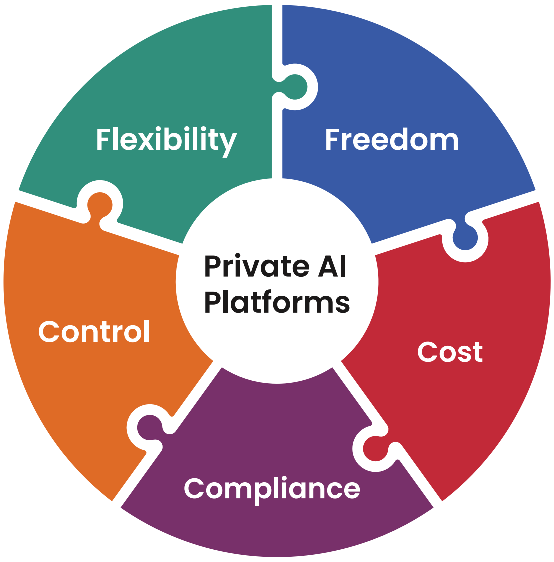
The low barrier to entry into the cloud world has its advantages, but it's also often accompanied by a high barrier to exit. This is a problem if the platform you choose evolves in a direction that deviates from your needs or starts to impose unexpectedly high costs.

For these reasons, we have recently seen the emergence of private GenAI platforms that run in your own data center or private hosted environment. These potentially provide a lot more control, freedom and flexibility, and address some of the cost and risk concerns.

Hyper-scaler Risks

Unpredictable costs
Volatile functionality
Private data exposure
Technical lock-in
Commercial lock-in
Slave to supplier agenda

Low cost of entry, but high cost of exit



The benefits of LLM freedom of choice

A particular aspect of private AI platforms worth highlighting is the freedom they allow you to design an environment specifically tuned to your business needs.

It's beyond the remit of this paper to go into model choices in detail, but suffice it to say that we've recently seen an explosion of both commercial and open source LLM options. This reflects the fact that with RAG in the mix, business needs are often better served by smaller models trained on narrower and/or more specialized data sets. They may still be general purpose, but could also be tuned to specific business domains or disciplines, e.g. healthcare, legal, analytics, coding, etc. This minimizes the impact extraneous or biased information, reducing both accuracy and consistency issues, Including hallucinations.

Smaller LLMs also typically need less computing power and are easier to monitor, manage and govern, reducing both infrastructure requirements and operational complexity - obviously a significant benefit in a mid-sized business environment.

Agents accentuate the need for robust controls

So far, we've only mentioned AI agents very briefly, but when we take a closer look at what they are about, it's clear that they reinforce the control imperative.

At their core, agents are like autonomous digital assistants that can work in the background and leverage AI systems to guide their operation. As part of this, they can follow multi-step processes, and have persistent 'memory', i.e. they retain context over time. Agents can also 'collaborate' with other agents and integrate with other systems resources and applications, e.g. CRM, ERP, logistics management, and so on.

As an example, imagine a 'bot' that works in the background, constantly monitoring customer input and either responding directly or routing requests to other bots, systems, processes or human operators.

While agent technology promises another leap forward in AI, it also reinforces the need for robust controls and foundations. Without proper grounding in accurate business data and careful control over its actions, an agent working in a customer service context, for example, could create significant problems - from providing incorrect information to taking inappropriate actions on customer accounts.

The upshot is that agent deployments need the same careful attention to data sovereignty and control as other GenAI applications - if not more.

And if agents require access to on-premises data and systems to operate effectively, there are clear advantages to running them in a private rather than cloud-based environment, which brings us back to the relevance of on-premises systems.

The anticipated rise in adoption of AI agents shines a spotlight on the need for robust security and data protection, in turn strengthening the case for considering private platforms.

Deploying on premises is now much easier

While we don't expect all GenAI deployments to go down the private route, the emergence of the latest pre-integrated platform options increasingly makes private GenAI an viable alternative to hyper-scaler offerings.

What once required complex integration projects and substantial technical resources is now accessible even to mid-sized organizations through solutions that combine optimized hardware, proven software stacks, and comprehensive management tools.

These modern platforms can dramatically simplify deployment while providing the commercial flexibility needed to start small and grow. You no longer need an army of AI specialists or massive infrastructure budgets - implementations can be operational within days or weeks rather than months, with initial costs frequently justified by a single well-chosen use case.

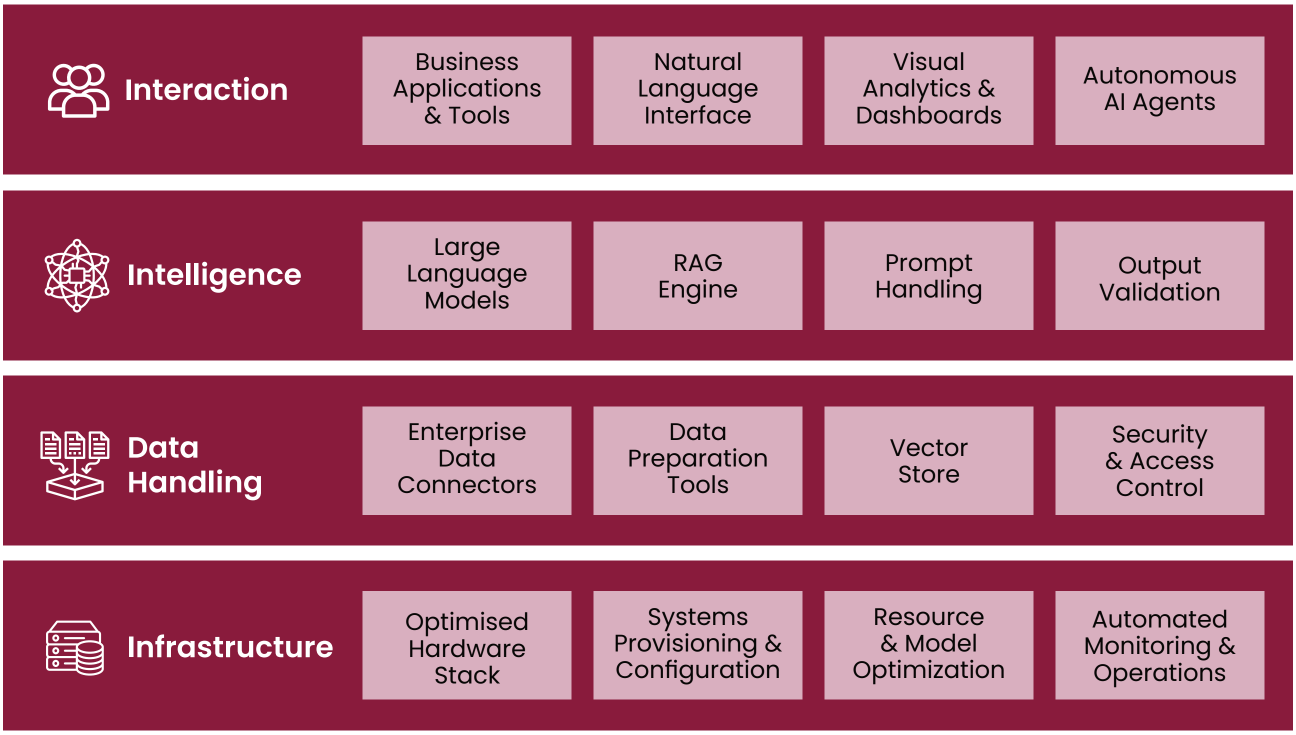
So what does a private AI platform actually look like?

A closer look at the technology

Anatomy of a private GenAI platform

A modern GenAI platform combines multiple components in a carefully orchestrated stack designed to deliver business value quickly and reliably. While the underlying technology is sophisticated, the architecture is quite straightforward when viewed through a business lens. Here are some of the key layers and components within them, that will typically be included. You'll recognize some of the terms from earlier discussions.

Stylized Private GenAI Stack



Interaction	Business Applications & Tools	Natural Language Interface	Visual Analytics & Dashboards	Autonomous AI Agents
Intelligence	Large Language Models	RAG Engine	Prompt Handling	Output Validation
Data Handling	Enterprise Data Connectors	Data Preparation Tools	Vector Store	Security & Access Control
Infrastructure	Optimised Hardware Stack	Systems Provisioning & Configuration	Resource & Model Optimization	Automated Monitoring & Operations

At the foundation is the infrastructure layer, including the hardware and tools to optimize and operate it. The fact that data handling is a layer all of its own reinforces the principle that most GenAI applications succeed or fail based on how well enterprise data is integrated into the environment.

The intelligence layer contains the AI smarts, including the models and some of the control mechanisms we have been discussing. Lastly, at the top of the stack, we see the interaction layer, and the key takeaway here is that your private AI environment must be able to support different application and user types.

Moving from theory to real-world practice

From our discussion so far, you will hopefully have gained a good understanding of the key ideas and principles in the GenAI space, but it's always useful to look at a real world offering to get a feel for what's currently available on the market. We'll do this with a tour of Private GPT, a solution provided by the sponsor of this paper, Fujitsu. Please note, however, that this is for illustrative purposes only and should not in any way be taken as a supplier or solution recommendation by Freeform Dynamics.

A real world example: Fujitsu Private GPT

The challenge of making advanced GenAI capabilities accessible to mid-sized organizations has been central to Fujitsu's thinking in developing its Private GPT offering. Drawing on decades of experience in bringing enterprise-grade technology to broader markets, Fujitsu has taken a characteristically pragmatic approach to solving this challenge.

This starts with a fundamental philosophy that has served Fujitsu well across many technology transitions: identify best-of-breed components, integrate them with its own enterprise-class systems, and wrap the result with services that ensure customer success. In the case of Private GPT, this has meant carefully evaluating the rapidly evolving ecosystem of GenAI technologies, including both commercial and open-source options, to create a solution that's both powerful and practical.

What makes this approach particularly relevant for mid-sized organizations is the focus on reducing complexity without compromising capability. Rather than expecting customers to become AI experts, Fujitsu has packaged the technology in a way that emphasizes business outcomes. The solution can be deployed quickly with moderate resource requirements, yet provides a clear growth path as your needs grow and evolve.

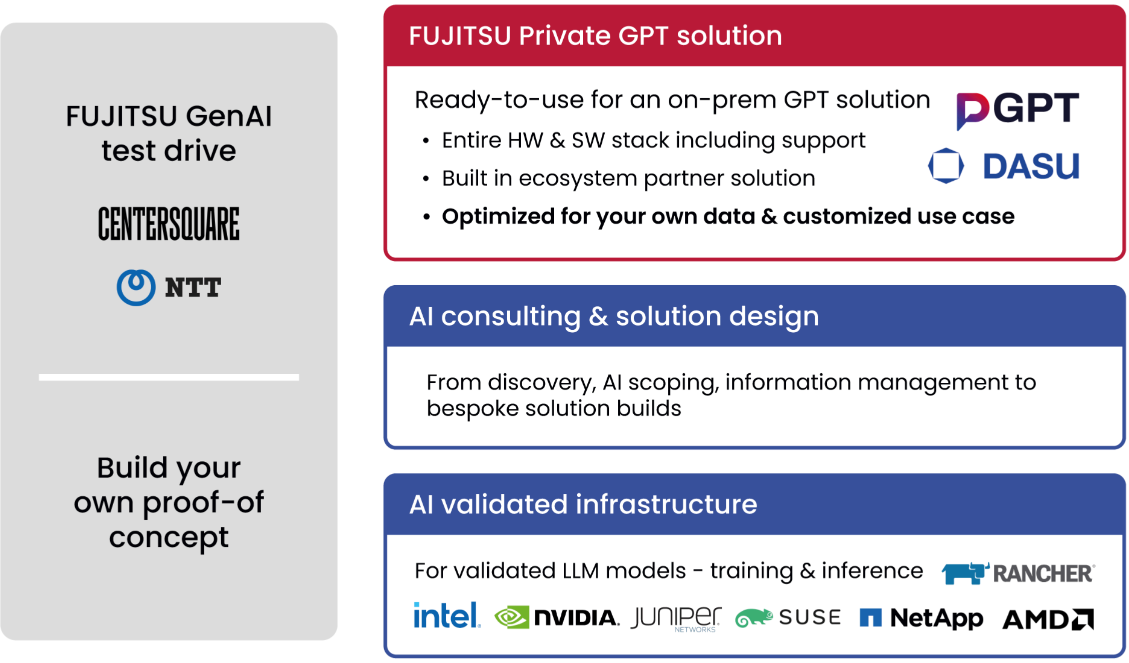
Let’s take a closer look.

Private GPT in more detail

Private GPT represents more than just an isolated technology product - it's a complete solution built to accelerate the journey from initial experimentation with GenAI to delivering tangible business value.

By tapping into the rich ecosystem in this space, including key open source projects, Fujitsu has created a platform that’s both open and fully supported.

At the heart of the solution is a carefully constructed technology stack that combines proven components into a coherent whole. Rather than reinventing the wheel, Fujitsu has selected and integrated established technologies that complement each other effectively. This includes everything from the underlying hardware through to selected LLMs and specialized management tools, all pre-tested and optimized to work together.



FUJITSU GenAI test drive

Build your own proof-of concept

FUJITSU Private GPT solution

Ready-to-use for an on-prem GPT solution

- Entire HW & SW stack including support
- Built in ecosystem partner solution
- **Optimized for your own data & customized use case**

AI consulting & solution design

From discovery, AI scoping, information management to bespoke solution builds

AI validated infrastructure

For validated LLM models - training & inference

Graphic reproduced with the permission of Fujitsu

The inclusion of comprehensive services to support use case discovery, scoping and information management is particularly significant. This addresses one of the most common stumbling blocks in GenAI adoption - the challenge of identifying and validating suitable applications of the technology. Fujitsu's structured approach helps organizations move quickly past initial uncertainty to focus on opportunities with clear business value.

Implementation support is equally well considered. Templates and service catalogs provide starting points that can be readily adapted to specific needs, while proven deployment methodologies help avoid common pitfalls. This combination of packaged knowledge and hands-on assistance significantly reduces both the time and risk involved in getting GenAI systems into full production.

Just as importantly, the solution is designed to evolve. Regular updates ensure you benefit from ongoing advances in the GenAI field, while the open architecture allows new capabilities to be incorporated as they come onto the scene and gain popularity.

This future-proofing is particularly valuable given the rapid pace of innovation in this space, including the emergence and ongoing evolution of agent-based AI applications.

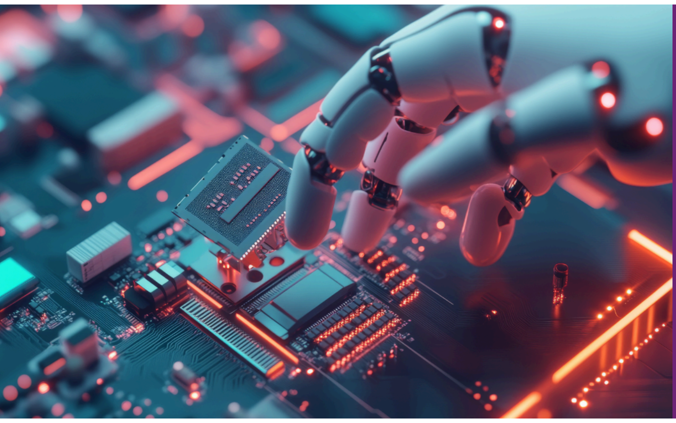
Right LLM, right data

Success with GenAI depends heavily on choosing the right model for your needs and feeding it the right business information. Fujitsu's experience in this area has directly informed how Private GPT approaches both challenges, making it significantly easier and faster to achieve clear results.

On the model side, Fujitsu helps you navigate the growing landscape of LLM options to find the best fit for your specific requirements. Rather than defaulting to the largest or most talked-about models, the aim is to select and implement the most effective and efficient LLMs based on the demands of your intended use cases and applications.

Equal attention is paid to data integration. Private GPT incorporates sophisticated Retrieval Augmented Generation (RAG) capabilities that allow the system to reference your business information when formulating responses. This ensures outputs are grounded in up-to-date facts rather than assumptions, while maintaining a clear chain of reference back to source.

Setting up and maintaining these data pipelines is made straightforward through pre-built connectors and automated tools. The idea is to help you identify relevant data, prepare it appropriately, and deploy update mechanisms to keep things current.



Fujitsu’s Private GPT offering provides LLM choice and incorporates the tools and services needed to seamlessly integrate your business data.

Efficient, optimized and scaleable stack

A common misconception is that GenAI always requires massive compute resources. Private GPT challenges this assumption by taking a more nuanced approach to infrastructure requirements, particularly valuable for mid-sized organizations watching their budgets.

The solution is architected to start relatively small for proof-of-concept work, then scale smoothly as applications move into production and demand grows. This might mean beginning with a configuration that supports a single department or use case, then expanding as value is proven. Fujitsu's uSCALE commercial model aligns perfectly with this approach, keeping costs in line with actual usage as you grow.

Efficiency is built in at every level. The infrastructure stack is pre-integrated and optimized to minimize both energy consumption and operational overhead. Sophisticated workload management ensures computing resources are used effectively, while automated monitoring helps identify opportunities for further optimization.

This efficiency-first approach doesn’t just lower running costs and reduce environmental impact, it also makes the solution more manageable for organizations with limited technical resources.

You get enterprise-grade capabilities with much less of the complexity and overhead normally associated with AI infrastructure.

From promise to real world value

Where to focus your efforts

The possibilities for GenAI to add value may not be limitless in a literal sense, but they might as well be when you're trying to assess the potential and figure out where to focus your efforts. The trick is to think in terms of problem domains or solution patterns rather than trying to define all possibilities. This will help you to structure your analysis and identify areas to explore further, before ultimately defining, evaluating and tackling specific use cases and applications.

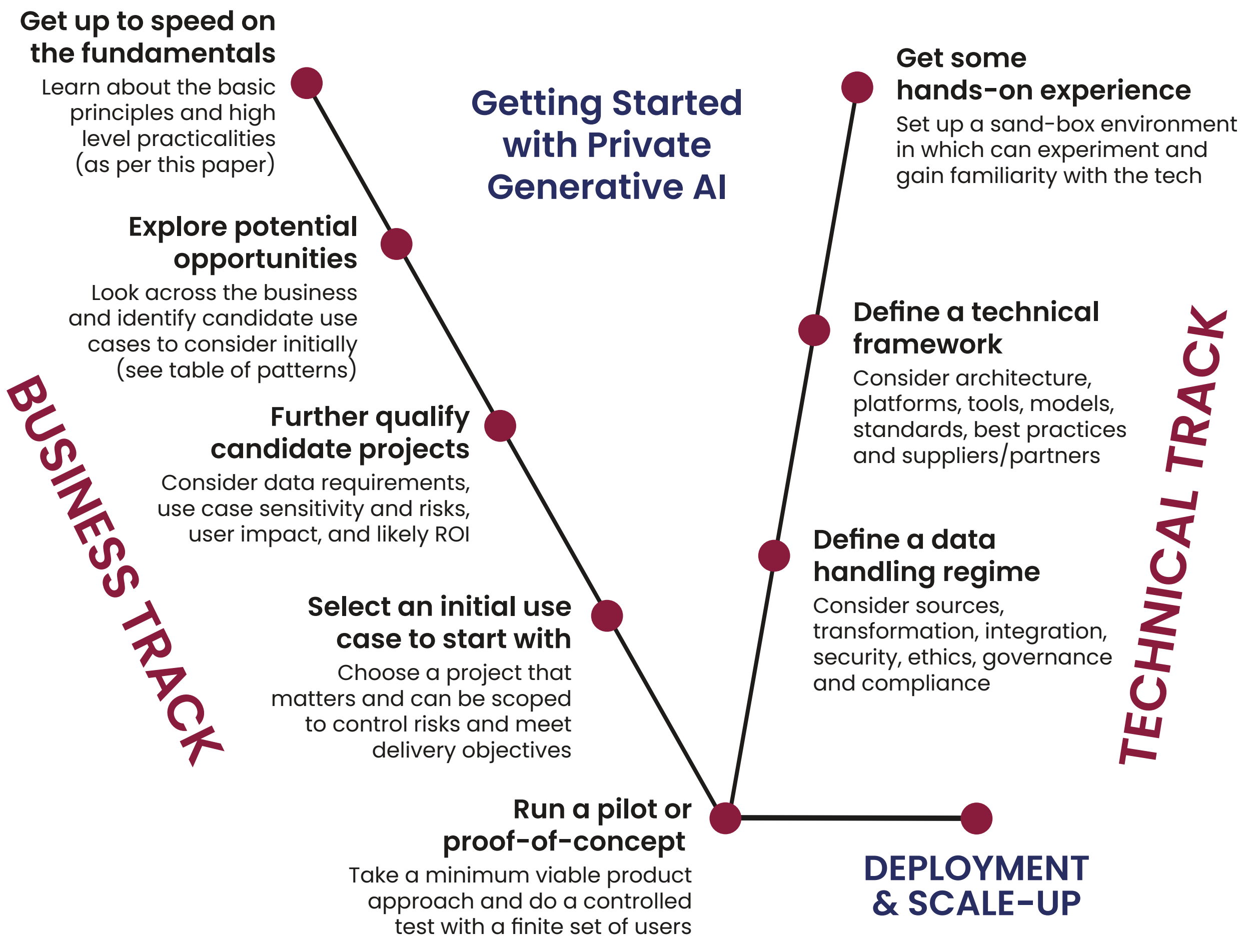
The following table is clearly not exhaustive, but it should be adequate to set you on the right path from a high level exploration and planning perspective.

Category	Domain/pattern	Examples
KNOWLEDGE WORK	Document synthesis & insight extraction	Legal document analysis, bid tender evaluation, complex proposal generation
	Research assistance, knowledge discovery	Product knowledge mining, IP analysis, market opportunity assessment
	Expertise capture & propagation	Engineering design, solution architecture, product documentation
	GenAI 'helpers' for developers & end users	Prompt-engineering assistants, response explainers, troubleshooting utilities
CUSTOMER EXPERIENCE	Automated customer response	Case handling, self-service sales & support, customer query resolution
	Intelligent support & troubleshooting	System diagnostics, complex fault analysis, manufacturing problem resolution
	Product & service management	Assisted design, automated deployment capabilities, technical spec generation
	Customer delivery enhancement	Requirements analysis, custom design & packaging, delivery planning
OPERATIONAL EXCELLENCE	Operations monitoring & control	Manufacturing execution, supply chain and logistics management, procurement systems
	Staff development & support	Training & competence management, on-demand guidance in context
	Smart monitoring & reporting	Regulatory compliance monitoring, anomaly detection and alerting
	Process mining & improvement	Critical process streamlining, safety SOP enhancement, quality system improvement

We should emphasize that the above table is not intended as a menu - our aim is to provide some examples to stimulate and inspire your thinking. Most of the ideas presented have broad cross-industry or cross-function relevance, and with some thoughtful word substitution you should be able to come up with specialized alternatives. As an example, smart monitoring and reporting could be a compelling use case in a manufacturing asset management and maintenance context.

The need for a focused collaborative approach

As with all transformative digital initiatives, technical and business teams clearly need to work together on GenAI projects to achieve maximum impact. That said, it still makes sense to define two initial tracks of activity as you would normally - one primarily business led and the other primarily IT or engineering led. The graphic below illustrates how this might look in the lead-up to your GenAI pilot or proof-of-concept project.



Getting Started with Private Generative AI

BUSINESS TRACK

Get up to speed on the fundamentals
Learn about the basic principles and high level practicalities (as per this paper)

Explore potential opportunities
Look across the business and identify candidate use cases to consider initially (see table of patterns)

Further qualify candidate projects
Consider data requirements, use case sensitivity and risks, user impact, and likely ROI

Select an initial use case to start with
Choose a project that matters and can be scoped to control risks and meet delivery objectives

Run a pilot or proof-of-concept
Take a minimum viable product approach and do a controlled test with a finite set of users

TECHNICAL TRACK

Get some hands-on experience
Set up a sand-box environment in which can experiment and gain familiarity with the tech

Define a technical framework
Consider architecture, platforms, tools, models, standards, best practices and suppliers/partners

Define a data handling regime
Consider sources, transformation, integration, security, ethics, governance and compliance

DEPLOYMENT & SCALE-UP

In general terms, the kind of activity flow we see above will be familiar to anyone with experience of introducing an emerging technology in a business context. However, it's worth highlighting that the use case identification and qualification steps are particularly important with GenAI.

The problem is that so many users have formed their initial views of GenAI from relatively casual initial use. Whether through downloading a chat app onto their phone, or experimenting with AI features that just popped up in the office tools or SaaS services they use, this often provides a false sense of utility and business value. Indeed some IT leaders characterize early activity of this kind as simply representing a distraction.

With this mind, it's necessary to look past subjective feelings when assessing potential benefits, and focus on use cases that will deliver good, tangible business value. And note that these may not always be the most glamorous or intellectually interesting.

Some of the best early projects are actually not that challenging technically (if you have the right platform in place) and dependent on data sets that are simple, clean and readily available for incorporation via RAG. These are the ones to look for.

So, the old principle still applies of starting with projects that matter to the business but can delivered relatively quickly. It's that more rigorous qualification may be required.

Final Thoughts

The mid-market GenAI opportunity is real

Mid-sized organizations are often the last to benefit from transformative technology advances. If you work in that kind of environment, you're unlikely to have the resources and specialist expertise that allow large enterprises to embark on purely speculative or experimental exercises.

At the same time, you also don't have the same level of freedom as consumers and small businesses to just 'give it a go and see what happens'. This is especially the case with GenAI given the dependency on potentially sensitive and fragmented data.

This is why the emergence of private GenAI platforms is so significant. From both a commercial and technical perspective, you can now set up a modest-sized initial environment to gain experience and deliver some quick wins, with the ability to scale freely over time.

Platforms based on an open technology stack and running in your own controlled environment mean you can move forward confidently, without having to worry as much about data privacy, security, compliance, runaway costs or potential supplier lock in.

A well-designed and properly implemented private GenAI platform can represent a good long term investment.

Try to ignore the hype and move at your own pace

Our last piece of advice is to avoid getting caught up in all of the industry hype as vendors and service providers strive to create a sense of urgency.

Sure the opportunity is now, and it makes sense to move sooner rather than later. But early market volatility means rushing to the most obvious suppliers, namely the cloud hyper-scalers, might seem like the easy option now but could lead to undesirable and constraining costs and risks very quickly.

At the moment we are seeing many companies revisit their early public cloud commitments in a broader sense for a lot of the same reasons we've highlighted. 'Repatriation' is the watch word for some as they move applications back to their own data centers or private hosting environments.

As hyper-scalers' all grapple with how to monetize their huge AI infrastructure investments, it's unlikely that consumers will foot the bill. It's therefore just a matter of time before costs have be passed onto business customers, i.e. companies like yours.

In some cases it will be worth it, but as GenAI begins to pervade more aspects of your business, the majority of use cases will not need the huge level of resources media and pundits often suggest, so a well-designed and properly implemented private GenAI platform environment is likely to represent a good long term investment.

About

About Freeform Dynamics

Freeform Dynamics is an IT industry analyst firm. Through our research and insights, we help busy IT and business professionals get up to speed on the latest technology developments and make better-informed investment decisions.

For more information and access to our library of free research, please visit www.freeformdynamics.com.

About Fujitsu

At Fujitsu, we're passionate about using technology to create a more inclusive, sustainable and trusted future. It drives everything we do. Throughout our history, we've supported businesses and society through delivering robust and reliable IT systems.

Find out more about our business, our history, our philosophy and the countries we operate in, please visit www.fujitsu.com.

Terms of Service