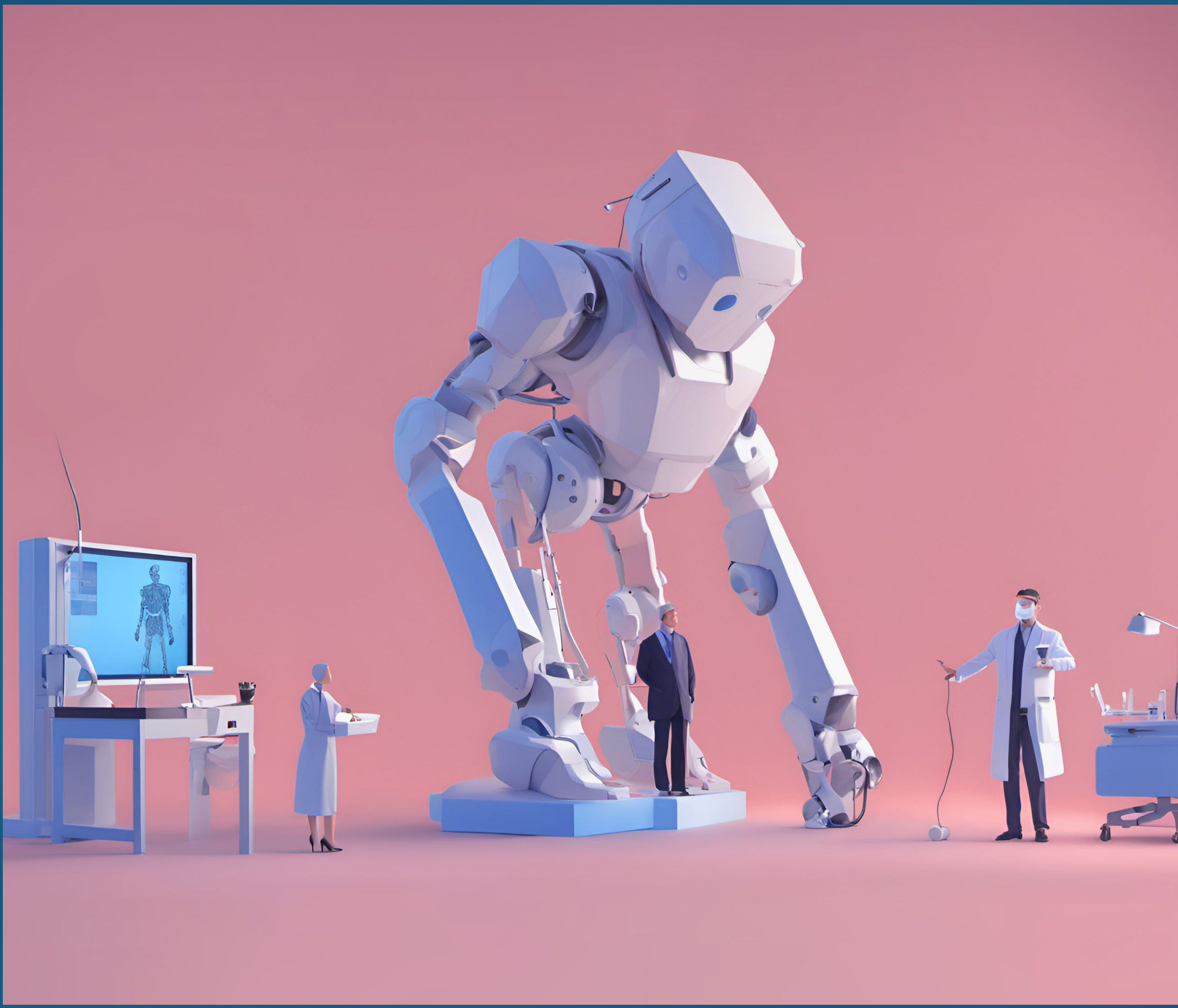
Mantel group

Whitepaper

Walk before you run: Getting smart when deploying Gen AI



The power of Gen AI is real.

But realising that power is far from simple.

All images crafted alongside Generative AI.

We know what you're thinking: "Not another document droning on about how Gen AI is going to change the world."

Trust us, we know the feeling. We're not here to dazzle you with AI thought bubbles and jargon. We're here to cut through the noise, not add to it.

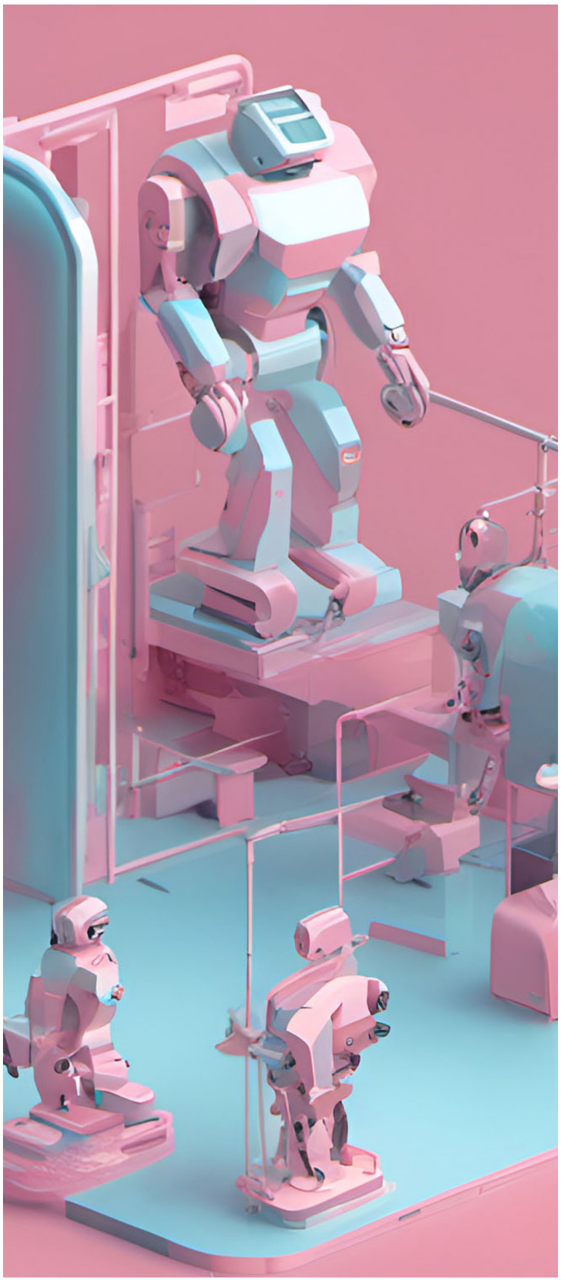
The excitement around Gen AI is justified. As the technology matures, it will have a profound effect on the way products and services are delivered. But to truly make use of Gen AI, organisations need to understand how the technology operates and how it can be useful to their business.

Before setting out on your Gen AI journey, companies need to ensure the fundamentals for implementing AI are in place. It means your AI tech stack needs to establish foundations that include people, process and governance.

From the start, a business needs to carefully consider what unique Gen AI use cases are relevant and useful within a particular company setting while also being critical about how (or, indeed, if) they can be put into production. This is not a time to be seduced by hundreds of pre-defined hypothetical use cases – the path to value comes from your unique requirements and how to leverage the strengths of Gen AI, not just what's easiest.

A business will also need to wrap its head around terms that are far from sexy, such as Machine Learning Operations and Large Language Models. In fact, we believe that the key to successful Gen AI models come through successful machine learning operations.

Building effective operational value in machine learning will generate the core levels of trust with your executives, your customers and your team, to ultimately deliver a successful Gen AI program.

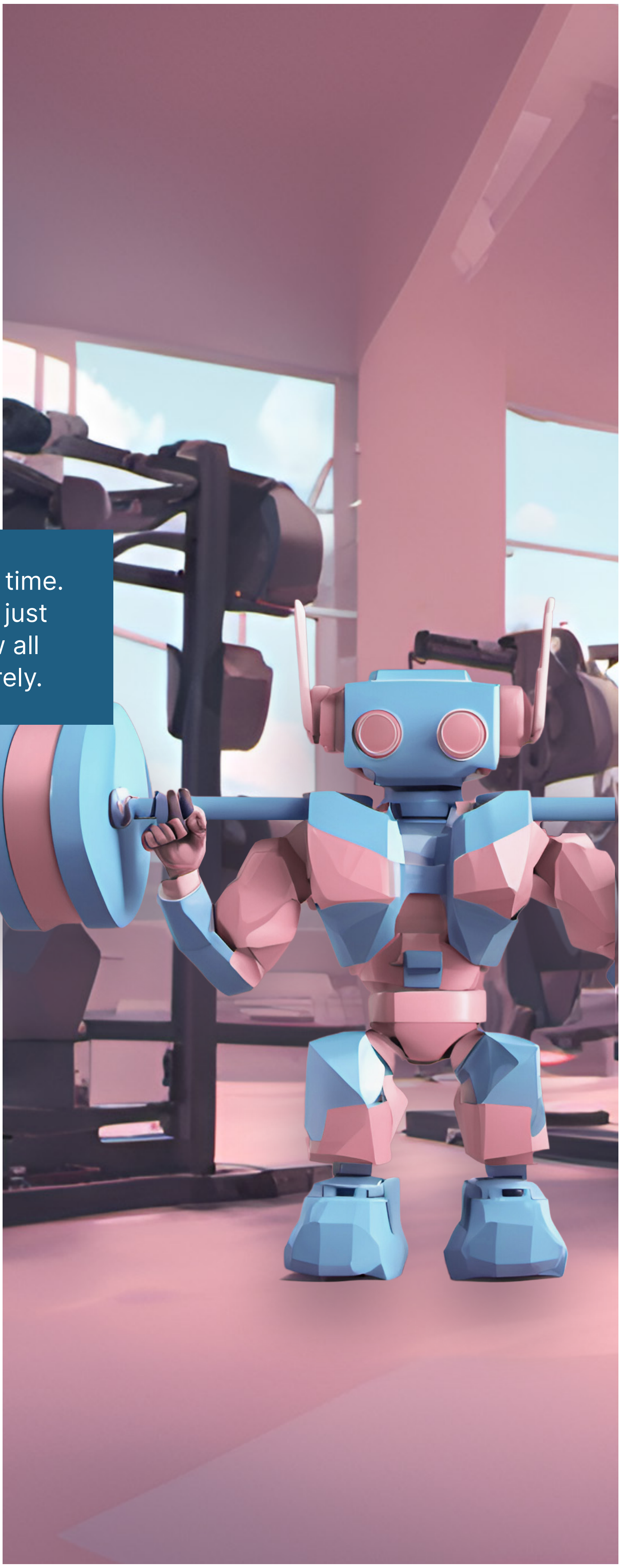


Gen AI is not an end-to-end system. Nor is it a case of simply utilising a publically available Gen AI API service, such as GPT 4, and pushing it onto existing processes. Instead, now is the time to get smart about creating the environment for making Gen AI work in the real world, because the business – from the C-suite to the factory floor – is finally listening. Best of all they're not only listening but, increasingly, they're demanding change.

Before you get AI Fit, you need an AI Workout.

The AI world is accelerating at a faster pace than many have anticipated. Unlike most other technological advancements where one part of the process lags another, the technology, research, people and platforms underpinning Gen AI are all rapidly developing at the same time. It means that your initial encounters with Gen AI will often be overwhelming and confusing, and by the time you learn all the lingo, there might be a whole set of new things to learn about.

It's not unlike walking into a gym for the first time. While you might enter having a clear view of just how buff you want to be, understanding how all the equipment will help is another thing entirely.



In the Gen AI gym of today, you'll immediately encounter concepts such as Top K and be faced with the choice of self-hosted models versus API driven models. You'll turn around and see tokens attached to big, muscular foundation models with 70 billion parameters. If this isn't intimidating enough, those fine-tuning methods and prompts that seem simple enough are available only if you're signed up to a special class that's run at dawn on Mondays.

Inevitably, you'll dive headlong into building Gen AI operational muscle without stretching first, and end up in a worse state than you were at the start.

That's where a 'personal AI trainer' comes in. Someone who can help you balance risk and reward. They'll match your error tolerance to the maturity of the tech. They'll help you find those use cases that deliver immediate value and know when to push the envelope. But what are the things you should ask them for help with?

Gen AI is incredibly simple and incredibly powerful BUT it's also incredibly complex to manage.

The key component to success in Gen AI, just as it is for machine learning models, is confidence. A business needs to have confidence that the outputs that are being generated are the correct outputs, and customers have to have confidence that your use of AI is doing the right thing by them. There also needs to be a level of confidence within your executive team that your AI implementation won't pose any risk to the business, and only offers value. This could extend into your decisions around the selection and implementation of that technology, and an understanding of what your risk and rewards tolerances as a business are.

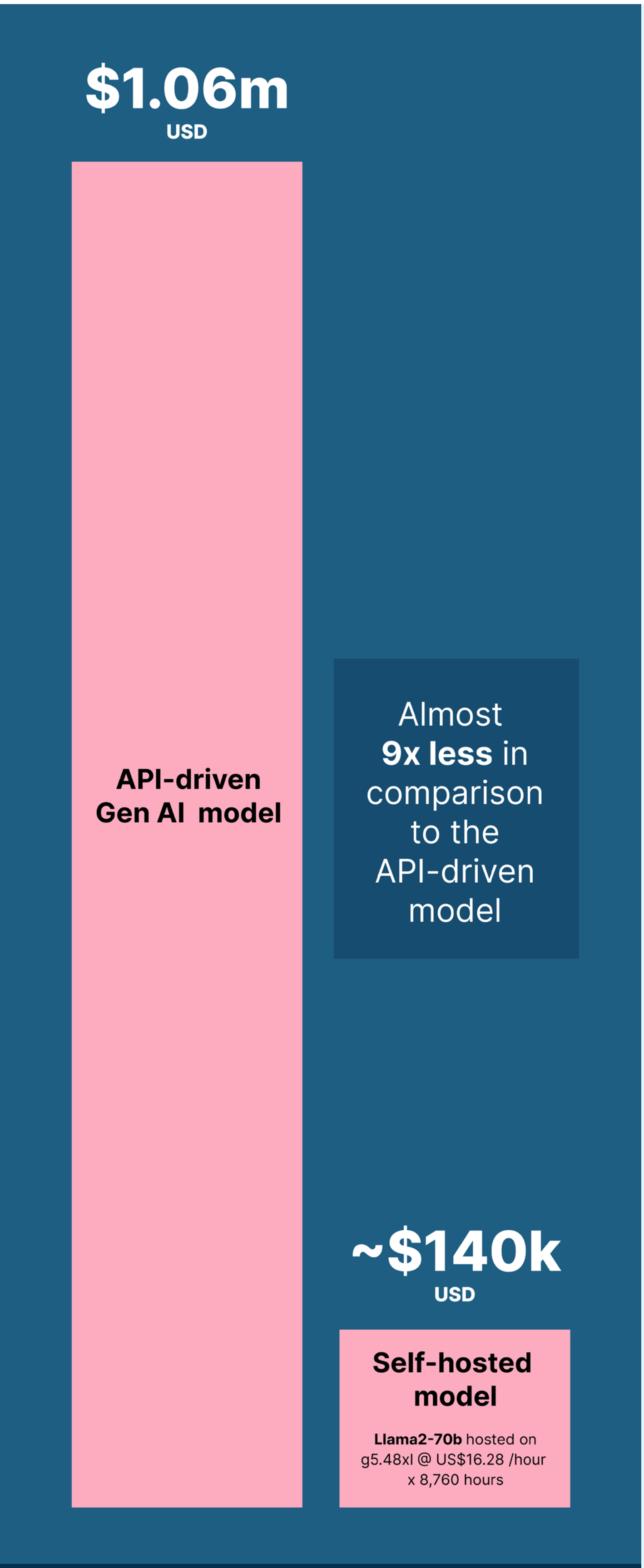
One of the most important, and fundamental concepts to grasp with this technology is that Gen AI models are not knowledge engines, instead they are rudimentary inference engines. They merely develop natural language responses based on the sequence of data that is ingested and what they determine is the most likely sequence of words or pixels to come next for a particular question or input.

As a business, the value of Gen AI doesn't exist in applying hundreds of different use cases. Instead, the value lies in knowing your unique use case, and building the foundations for Gen AI to be deployed and managed with maximum sophistication, efficiency and efficacy.

By thinking about your unique use cases, you will begin to understand the implications for cost and scale. The way that you use Gen AI has strong impacts on the cost of the implementation. If, for example, you've got a large amount of data that the model needs to ingest, there's a good chance an API-driven model will blow out your costs.

On the other hand, using an open source model such as Llama2-70b, requires you to have the right infrastructure, the right data, and the right people and governance structures to support the model. For businesses that don't have all of these necessary components, signing up to an API-driven model might be per token more expensive, but it might also be the most cost-effective option in the long run.

Regardless of the model and implementation approach your business adopts, a full reckoning of your use cases will help you understand the most critical aspect underpinning your Gen AI efforts: Machine Learning foundations. And what is the basis for these foundations? Ensuring your Machine Learning operations are in order.



Comparison of the cost of 60 million unique visitors/year to a homepage.

Use case

Using AI to personalise your home page

Everyone wants to feel seen. Imagine generating a hyper-personalised home page each time a visitor logs on to your website, delivering content directly relevant to their interests. For example, if a user logs on and reads an article on mergers of telcos, the next time they came to your homepage, they could instantly be delivered an AI-generated home page related to mergers, acquisitions, technology updates and telecommunication policy.

This sort of personalisation is possible with both API and open-source Gen AI models. However, understanding the hosting costs and enablement implications of each option is crucial.

If, for example, you had 60 million unique visitors a year to your homepage an API-driven Gen AI model would cost around US$1.06 million to run. The same content generated by a self-hosted model, on the other hand, would only cost around ~US$140k, almost nine times less than than the api-driven model (Llama2-70b hosted on g5.48xl @ US$16.28 /hour * 8,760 hours).

So why the disparity? The self-hosted model is priced on a per-instance approach, whereas the API-driven model is per-token. For large amounts of data per-token can end up costing a lot. However, it is far easier to support and manage – as the self-hosted pricing does not factor in the people costs you would have to have to support the running of that infrastructure.

MLOps is the catalyst for delivering Gen AI

Machine Learning Operations (MLOps) is how you deliver end-to-end machine learning in a productive way. That might sound technical – and it is! – but if you're thinking about starting a Gen AI program, you need to start with an MLOps program that is absolutely humming.

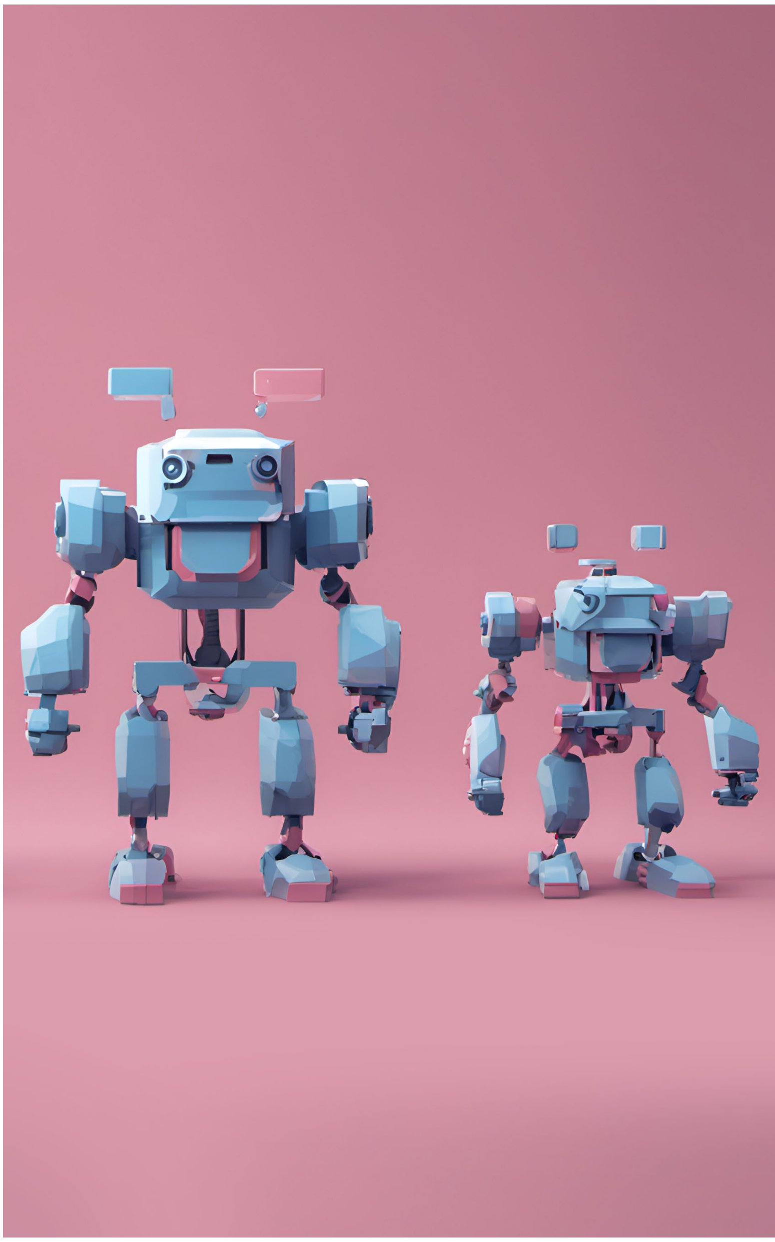
Why do you need a finely-tuned MLOps program before you start on Gen AI? In short, an MLOps program means you've done the work to make your machine learning (ML) projects functionable, reliable, repeatable and with a value-based mindset. A structured MLOps program is one that takes your organisation's idea through the ML lifecycle starting with data engineering, to feature engineering, to ML experimentation, training and model build and finally to deployment.

If you've done this work, there's a far greater chance that you'll be able to avoid the pitfalls of scaling up to a Gen AI program. You will also have a much clearer view of the benefits that your data can deliver. Conversely, without an MLOps program, there's far less chance of your data being in a state able to interact with Gen AI programs.

There are a number of key questions to ask yourself and the business when considering shifting from ML Ops to Gen AI:

1. **How well are your MLOps processes working – if they are established at all! – and how do you then integrate that with Gen AI models?**
2. **What programs and additional features do you need to scale across Gen AI and all other applications of machine learning?**
3. **What other dimensions, such as different cost structures, do you need to consider when moving from MLOps to Gen AI?**
4. **How can you overcome your fears of transitioning the unique needs of Gen AI into MLOps?**

This is the moment that businesses need to make MLOps the catalyst for building a successful Gen AI program.

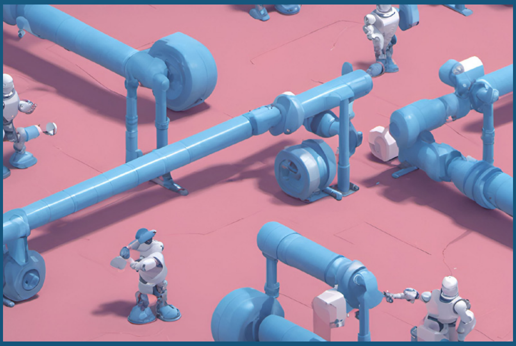


How is MLOps different to LLMOps?

Both terms are a bit of a mouthful but Large Language Model Operations (LLMOps) is, in very basic terms, Machine Learning Operations (MLOps) with a bit more added into it. Where MLOps supports the development of machine learning models from scratch, and has a more obvious path to defining what is performing well and what's doing not-so-well, LLMOps adds a layer of complexity to these governance steps where the solution doesn't seem as obvious out-of-the-box.

Imagine setting out on a road trip, where the route-finding tools at your disposal mirror the intricacies of MLOps and LLMOps. On one hand, think of MLOps as akin to using a modern GPS device. Enter your destination, and you're provided with clear, turn-by-turn directions. As you traverse your path, the GPS continuously monitors the route, providing alerts for traffic jams and effortlessly suggesting alternative routes for any obstacles. It's a straightforward, efficient, and largely automated experience. The journey is predictable, and there's minimal need to interpret or intervene, mirroring the streamlined approach of MLOps where monitoring and governance are simplified.

On the other hand, LLMOps is like navigating with a detailed paper map and a compass. This method immerses you in rich details, from topography to scenic routes. Instead of simply following voice prompts, you're engaged in understanding the map's nuances and intricacies. When an unexpected challenge arises, you find yourself interpreting the map, consulting the compass, and making judgement calls. This journey, while potentially more enriching and scenic, demands deeper engagement, understanding, and interpretation, echoing the intricate governance steps inherent in LLMOps. But unlike a paper map, ultimately you want to automate the process of LLMOps just like MLOps.



Making AI work for you, not the other way around

Once you've got your ML foundations in place, how do you make the step from MLOps to Gen AI? It all starts with something that's hardly sexy but is absolutely necessary: governance.

Governance is possibly the most important yet overlooked part of a Gen AI program. When you turn on your Gen AI model and it starts generating answers, how do you know that those answers are what you want? Or need? Do the answers, for example, have repeatability, explainability, reliability? These are all questions that good governance oversight allows you to answer.

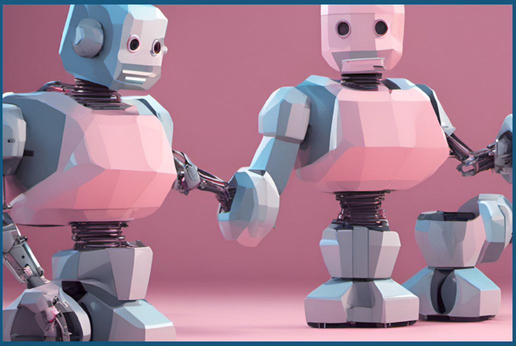
Baking in responsible AI at the beginning not only gives you internal confidence in the model's performance, it will help lead to better outcomes and value realisation. These outcomes will be enabled via a range of factors from prompt validation, prompt guardrails, monitoring, access control and data governance.

But governance shouldn't be set up as a series of roadblocks. Businesses need to iterate fast and innovate. They need to learn and increase value. And they need to continually improve. You still need to make sure you treat your Gen AI project and the subsequent model like any other machine learning use case.

Ride the wave, don’t drown in hype.

We don’t need to tell you that there’s a lot of noise around about Gen AI. As the technology continues to improve, you need to carefully sift through the options and understand exactly how the technology’s current maturity levels can apply to your use cases. Don’t be swept along by the new, shiny toys – starting with strong technological foundations will often be the best place to start.

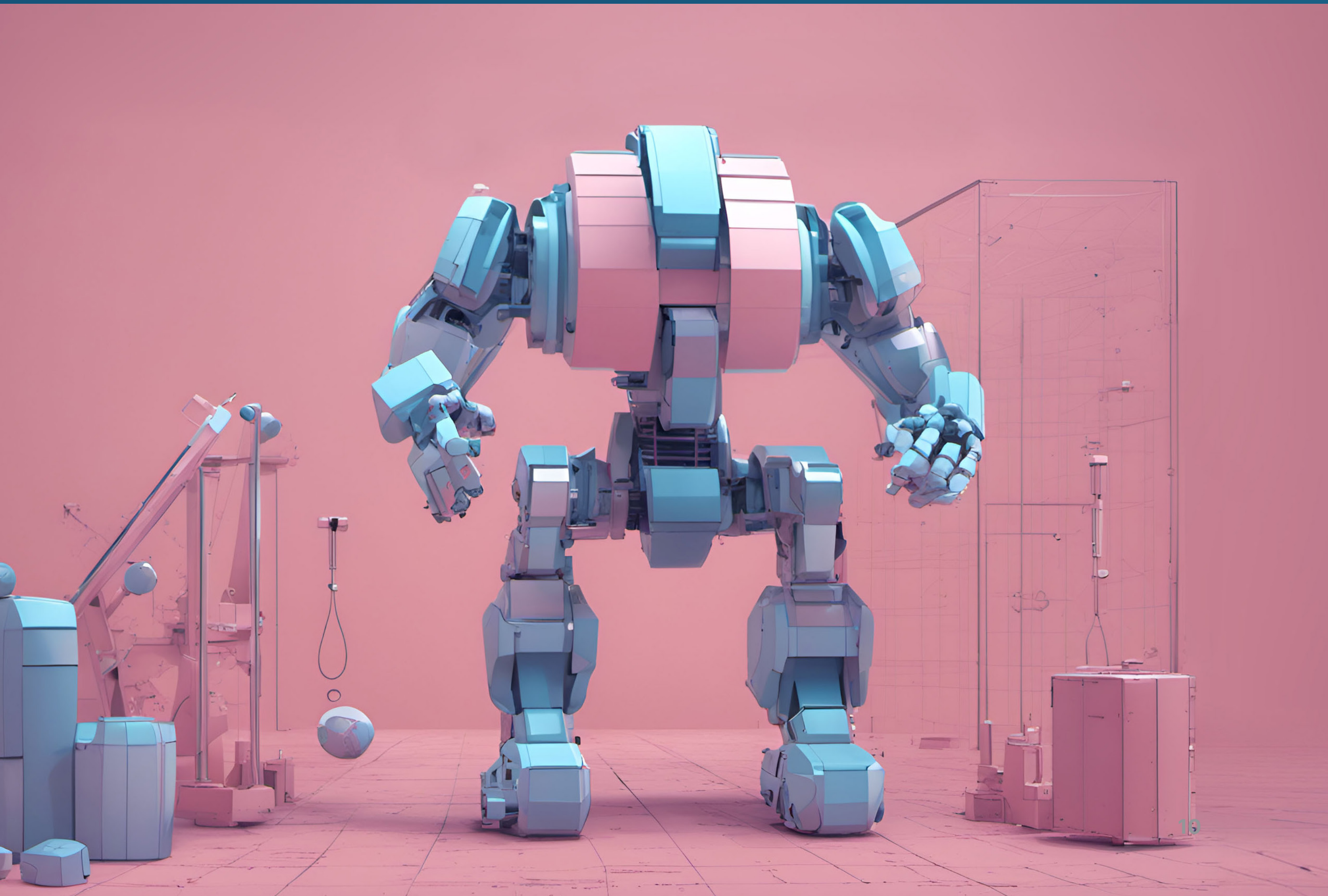
Businesses also need to remember that building a Gen AI use case can be a complex task, often requiring high levels of customisation and governance. This isn’t a simple plug and play situation. Because of the huge adoption and technology curve, by the time you build out your program, the technology may have advanced to or past that point. Accelerating ahead of the mature (or maturing) technology has the potential for cost blow-outs and suboptimal implementation.



For businesses ready to start using Gen AI, here’s a basic playbook to consider:

- Ensure your foundational data programs are in place via strong MLOps.
- Explore the value and pitfalls of Gen AI in your business through experimentation, but don’t expect this to directly convert to a production-ready output.
- Hone your MLOps before considering productionising your less complex Gen AI use cases.
- As you refine your governance, error tolerance, ‘hallucination’ and customisation of the Gen AI models, move to more and more complex tasks.
- Understand that fine tuning a model doesn’t always make a model more accurate – the law of unintended consequences applies.
- At each step, perform a full risk versus reward assessment.
- Just like the Gen AI model itself, learn from your mistakes.
- If you reach an impasse or the limits of your knowledge, don’t try to guess the next step. Make sure you have expert help and guidance.

Key takeaways



Navigating the world of Gen AI is like wading through treacle – the faster you try to move, the stickier the situation becomes.

It is imperative that heads of technology in any business convey the risks and opportunities involved with Gen AI and MLOps to the C-suite and executives. At the same time, they need to explain to the business how these models will operate in reality, and the limitations of the models, in order to continue to win support for organisational engagement in machine learning.

Ensuring that your team has the background and experience in implementing and scaling Gen AI solutions, though upskilling or in partnership with experts, means that you will have the tools to convince your business to invest in ML foundations and focused Gen AI solutions, instead of spending tens of thousands on a proof of concept that has no ability to scale or achieve real value.

Mantel Group has the best people in the business for creating a Gen AI journey that is productive and scalable. Not only do we have the tech smarts and the business nous but our work is always grounded in the reality of your needs, now and in the future.

At Mantel Group we're always keen to start new conversations on using technology to impact people in a positive way.

Reach out today

Speak with Brendan Wilkinolls about how your business can utilise the potential of Gen AI.

☎ 0490 003 350

✉ brendan.wilkinolls@mantelgroup.com.au

Mantel group

Unlock your Gen AI potential

We’re always keen to start new conversations on using technology to impact people in a positive way.

Contact Us