

Publication date:

March 2024

Author:

Bradley Shimmin

# Building a strong operational foundation for Generative AI and beyond



Brought to you by Informa Tech

Commissioned by:



---

# Contents

---

Introduction	2
As simple as hitting the generative AI easy button?	3
Exploring the current AI platform landscape	6
Taking action: Moving from artisanal to operational AI	10
Conclusion: Putting it all together	16
Appendix	17

---

---

# Introduction

---



The generative AI hype of 2023 is giving way to the year of AI operationalization in 2024 – for both generative and predictive AI. With more stakeholders involved than ever, more opportunities, more risks and more choice (both good and bad) – it’s never been a more important time for companies to therefore invest a solid AI data science platform, preferably while working with a trusted vendor partner capable of conveying deep expertise in operationalizing AI.

That means adopting an ops-centric approach to AI development; embracing open, flexible technology solutions built to handle both traditional and Gen AI, or indeed whatever the future may bring.

For example, Red Hat OpenShift AI Red Hat’s integrated MLOps platform for building, training, deploying, and monitoring AI-enabled applications and models at scale across hybrid cloud environments, aims to provide this – with particular focus on:



- Simplifying AI adoption through access to the latest open source innovation
- Driving AI/ML operational consistency to speed up moving models from experiments to production
- Providing hybrid cloud flexibility to deploy models across on-premise, cloud, and disconnected edge environments
- Enabling data science and application developer teams to collaborate on a common, extensible platform enterprise-level support

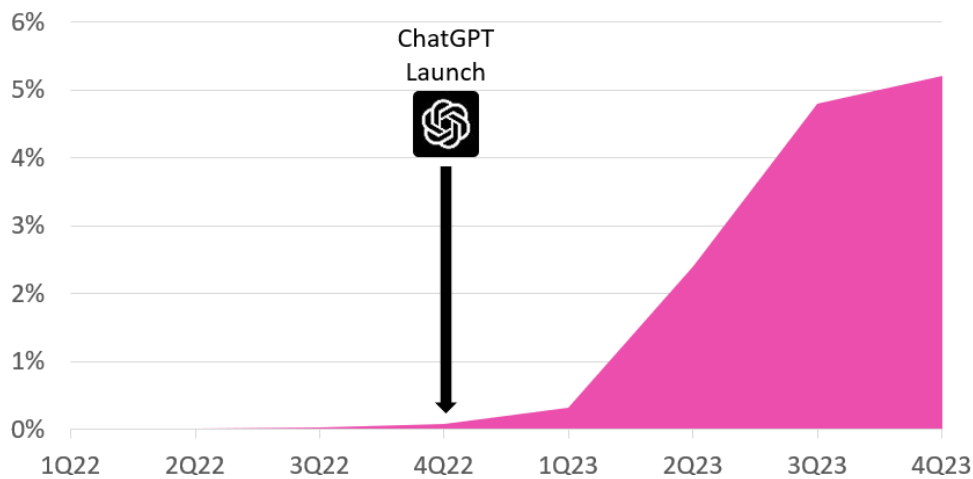
This Omdia whitepaper explores the background trends, challenges and future needs of the enterprise in successfully navigating this AI adoption path, including through such open-source platforms.

# As simple as hitting the generative AI easy button?



The tectonic aftershocks of innovation and disruption that continue to follow OpenAI's earthshaking introduction of ChatGPT in late 2022 have forever changed the way companies think about artificial intelligence (AI) and about how they build enterprise software in general. As borne out by current hiring practices (See figure 1), not only are Generative AI (GenAI) jobs on the rise, but of those, there are 12.6% more openings concerned with engineering expertise rather than traditional data science know-how.

**Figure 1: Percentage of AI jobs focused on GenAI technologies, 2022-2023**



Source: Omdia AI Skills Tracker 2H, 2023

Thanks in part to readily available, highly affordable, and remarkably capable large language model (LLM) hosted API services from AI leaders OpenAI, Microsoft, Google, Anthropic, Cohere, and more, enterprise practitioners are now "assembling" AI outcomes in short order. Equipped with feature-rich development frameworks, low-code model playgrounds, and increasingly GenAI-driven no-code tools, enterprise practitioners can now engineer advanced proof of concept (PoC) solutions across a wide spectrum of use cases, all within a matter of days -- something unheard of before 2022.

Unfortunately, this "easy button" aspect of GenAI has created a false sense of security among early adopters, blinding them to a myriad of operational risks that lie somewhere between PoC and production, risks ranging from unanticipated model hosting costs to unforeseeable privacy and security vulnerabilities.

## The promise of DevOps practices

To avoid these and many other potential pitfalls, enterprise practitioners need to envision GenAI solutions as a living element within the larger context of modern IT DevOps and machine learning operations (MLOps) practices. To do that enterprise AI practitioners simply need to invest in an AI and data science platform capable of operationalization of those AI assets using a solid MLOps foundation, a platform capable of taking IT concepts such as DevOps, as well as continuous integration and continuous deployment (CI/CD) and applying them to any and all AI endeavors.

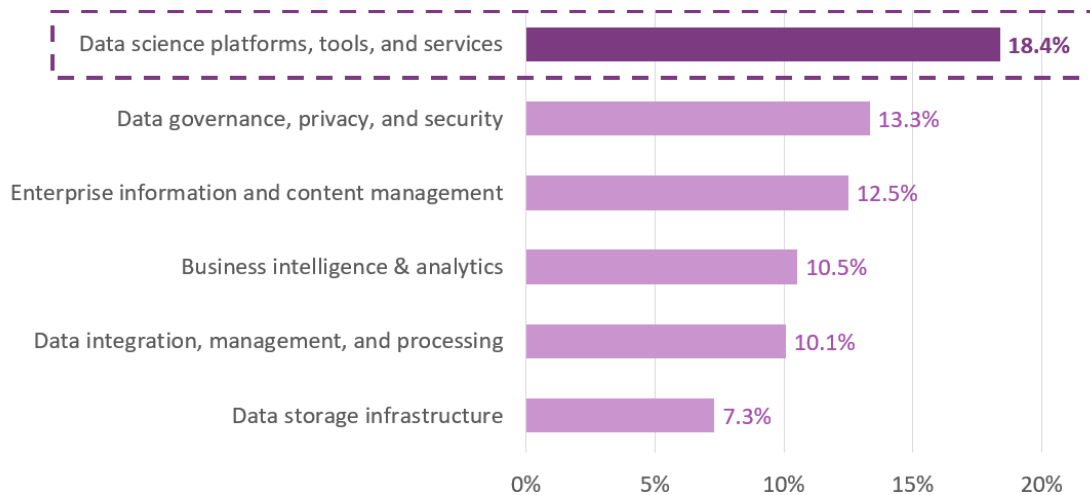
Within the broader software market for data and analytics infrastructure, Omdia views these platforms as a crucial, "must have" area of investment, one that is necessary to build successful

AI outcomes in the enterprise. Enterprise buyers agree. According to Omdia research, the market for AI and data science platforms, tools, and services represents the fastest growing analytics and data management market segment year-over-year, eclipsing traditionally dominant centers of investment such data governance, privacy and security (see Figure 2). Why are companies investing in these operational platforms? In a nutshell, the platform enables companies to streamline and even automate AI model development and deployment workflows; they also make it possible for companies to rapidly grow and adapt to both planned and unforeseen resource requirements; and they mitigate risk through improved governance and compliance, transparency, and explainability.

*"We had difficulties in getting AI/ML models into production. Those difficulties were created due to misalignment and 'silos' created along the way, creating an inefficient model development and a prevention of AI workloads getting into production for business decision making."*

*Director of Operations, Government Agency, Middle East*

**Figure 2: Analytics and Data Management market growth, 2022 - 2027**



Source: Analytics and Data Management Market Forecast 2023

Beyond these and many other benefits, which we'll explore more deeply in this paper, AI and data science platforms perform one ultra-critical task: They align AI development with enterprise software development. Before the rise of GenAI with its emphasis on engineering over traditional data science, it was moderately acceptable for companies to view data science projects as something separate from IT. Although it may have taken a few months to convert data science projects from experiment into deployable, enterprise-grade software, that was nothing compared to the amount of time it took to build an acceptable experiment in the first place.

Those days are over. The rise of GenAI and its use of pre-trained, foundation models (FMs) as fundamental building blocks that can be rapidly "engineered" into a working PoC, has opened up AI to an exceptionally wide audience of practitioners, be those business owners, data professionals, or application developers. With AI so close at hand for so many, it becomes an absolute imperative for companies to invest in a platform adept at operationalizing the entire spectrum of AI. That means the development of both GenAI and traditional, predictive AI; deployment across cloud, on-premises, and edge; and perhaps most importantly, unifying development across disparate technologies, be those closed or open source.

# Exploring the current AI platform landscape



Fortunately for enterprise buyers, there is no shortage of capable AI and data science platforms from which to choose. Omdia currently tracks more than 30 Ops-oriented such platforms from a wide range of vendors beginning with smaller, independent open source projects and ranging up to global hyperscale cloud providers and frontier GenAI model creators (see Figure 3).

**Figure 3: AI and data science platform marketplace**



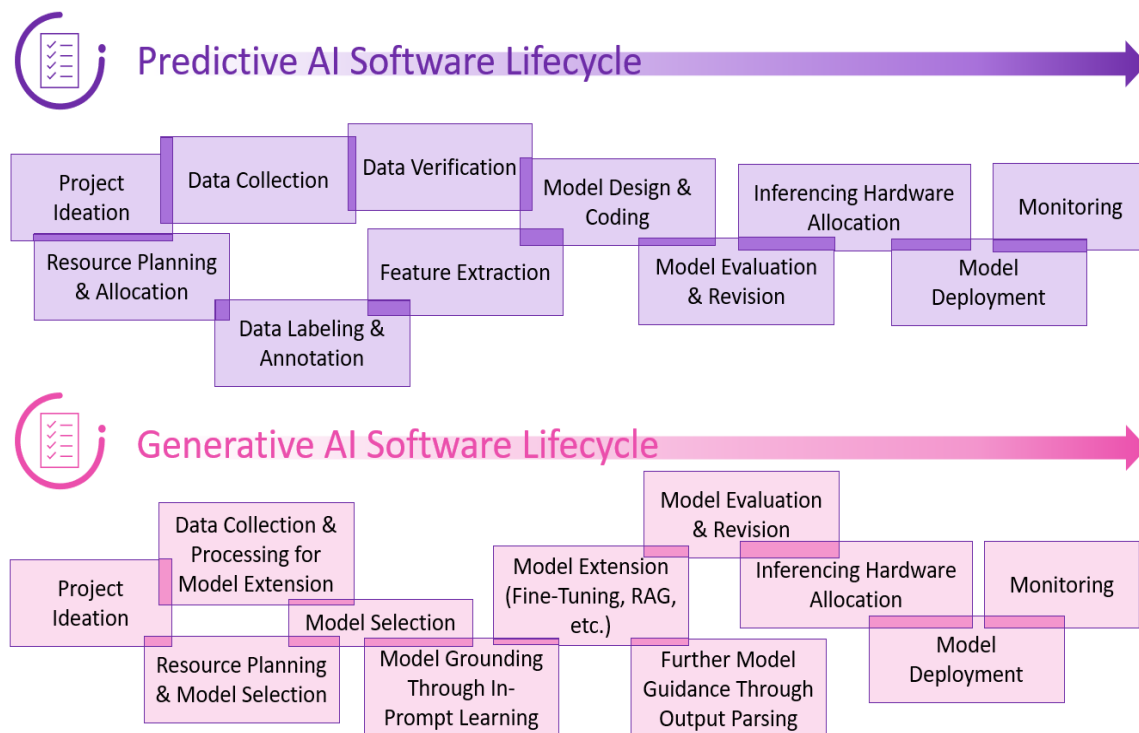
Source: Omdia

The trick, of course, is selecting the right platform. To that end, Omdia recommends first ensuring that the platform speaks both predictive and GenAI with equal aplomb. That done, buyers then need only ensure the platform aligns with their company in terms of AI maturity level, internal technology expertise, standing cloud/premises investments, and philosophical approach to IT and AI operations (e.g., level of commitment to DevOps principles) -- a sizable task to be sure, but one well worth the investment. Very soon if not already, companies that cannot rapidly and affordably build and scale solutions that span (and often combine) both predictive and GenAI use cases, will fall behind those able to do so.

## Why operationalize when AI seems so easy?

While there are many benefits to be found in MLOps-savvy AI and data science platforms, they all aim at controlling technical debt. In short, technical debt is a simple way of referencing the cost of dealing with the ongoing complexities inherent in a project's processes, procedures, technologies, and assets (e.g., Python code, LLMs, data pipelines) over time (see Figure 4). All projects generate technical debt, even seemingly simple GenAI projects.

Figure 4: Centers of technical debt for both predictive and GenAI lifecycles



Source: Omdia

It may take only a few weeks to stand up a working chatbot fine-tuned to answer employee human resources (HR) questions, for instance, but it may cost the company a tremendous amount of time and energy to maintain the functionality and performance of that project going forward. Now triple those costs for three similar AI projects enacted independently across sales and marketing. With each project lifecycle there are several steps and within each step, there are several opportunities to accumulate technical debt:

- Project ideation and technology selection:** Poorly defined or misaligned project goals coupled with a lack of understanding with regards to what the selected technology (not just an LLM but all supportive hardware and software) can and cannot do.



- **Data collection and processing:** Not enough data, inaccurate, or low-quality data, or worse the wrong data.
- **Model selection and evaluation:** Choosing the wrong model or one that's ill-suited to the project or misaligned with the data (e.g., performs well during testing but poorly in production).
- **Model evaluation and integration:** A lack of sufficient outcome validation coupled with poor project documentation and version control.
- **Model deployment:** Unanticipated resource requirements, a lack of compatibility with existing systems, a lack of automation pipelines and overall lack of scalability.
- **Project Monitoring:** Inadequate facilities to detect immediate problems such as data privacy/security breaches as well as cumulative issues such as data or model drift, and AI biases.

*“People may say 'make things easier' but they like when brands are honest about how hard things actually are. Complexity stands out, because you're not selling that it's going to be a walk in the park. You're saying that there are hard parts. You understand us and the struggles we have here.”*

*Software Developer, Automotive*

Unfortunately, there is no easy button that can foresee, observe, and mitigate these and many other forms of technical debt. Even the most advanced AI and data science platform is not enough on its own unless it is coupled with equally advanced best practices and staffed by appropriate practitioners. Such an investment in people, process, and platform, including technology partners who can act as trusted advisors, is exactly what it takes for companies seeking to leap from isolated artisanal efforts to enterprise-class AI products.

## Addressing the supposed rift between GenAI vs "plain old" AI

Given such demands, it can be quite tempting to think of GenAI solutions as an entirely new form of AI, one that doesn't have to play by the many operational rules listed above. Moreover, built on top of current model architectures using neural networks, GenAI solutions can readily appear capable of taking on seemingly any use case or vertical requirement involving the manipulation of language, code, and images.

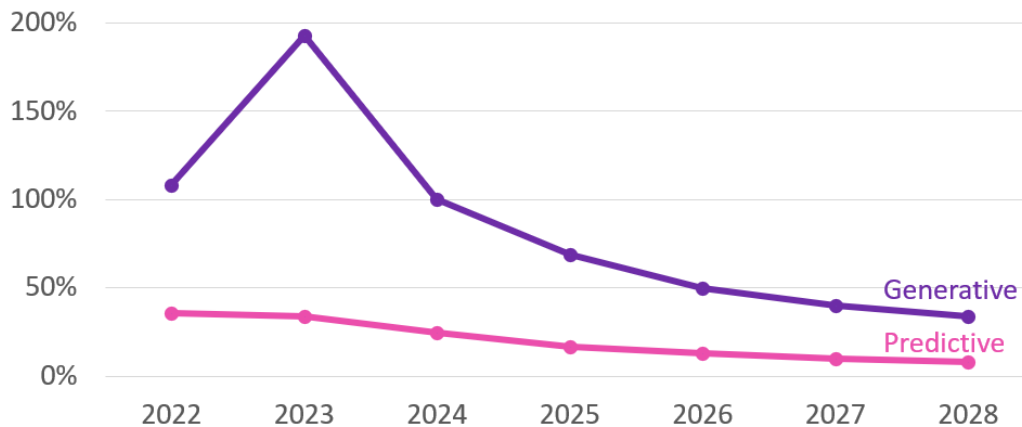
This impression has been fueled by highly public frontier models such as OpenAI GPT and Cohere

*“Many customers have challenges operationalizing, their AI experiments, whether it's classic AI or new generative AI projects with large language models, or RAG patterns. Providing a consistent AI platform as a foundation to handle the ever-changing AI landscape will help these customers grow with the innovation that happens in open source”*

*Steven Huels, General Manager, Red Hat AI Business Uni*

Claude, which show impressive aptitudes across many different human endeavors. Even smaller efforts such as the popular Meta Llama family of models are beginning to show similar levels of capability, especially when combined with techniques like model fine-tuning using corporate data or enriched model responses using in-prompt contextual knowledge techniques like retrieval augmented generation (RAG) data pipelines. Even more, with the recent influx of multimodal models (which incorporate various types of data, e.g. text, video, images, etc.) and other recent developments, companies are more inclined to view LLMs as a one-stop, "do-anything" solution that can sweep away traditional predictive AI technologies and techniques (see Figure 5).

**Figure 5: Predictive and Generative AI software growth rates: 2023-2028**



Source: Omdia AI Software Market Forecast 2023

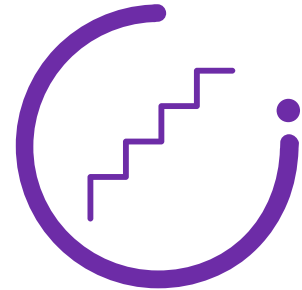
The truth, however, is quite the opposite. Whilst GenAI indeed presents a number of unique opportunities and challenges, at its core, it's no different than traditional, predictive AI when it comes to building, deploying, and governing a given AI solution at scale in the enterprise. From an operational perspective, a GenAI transformer model architecture looks no different than a traditional AI and data science approaches. They are both built and trained using virtually the same machine learning (ML) frameworks such as PyTorch, JAX, and DeepSpeed.

Both require the same careful attention in terms of planning and building training and inferencing resources. And in truth, this area of technology is moving so fast and that innovations are released daily. Therefore, relying on an AI and data science platform that supports both use cases and focuses on incorporating or supporting such innovations quickly is the best choice when looking how to support a differing variety of AI use cases.

---

# Taking action: Moving from artisanal to operational AI

---



No matter the form that AI takes going forward, one constant will remain: AI must be built on top of a solid MLOPs-savvy AI and data science platform, one steeped in predictive AI that's also capable of operationalizing whatever new technologies, development practices, or deployment patterns may come its way. A smart investment in a platform that aligns with corporate requirements can turn even the most meager of experiments into an impactful piece of software by not only eradicating many of the tactical pitfalls mentioned previously, but by shifting the manner in which AI software is built from experimental to transformational. Below are a few examples of how this shift can benefit enterprise practitioners:

- **Accelerating time-to-value:** Automating and securing access to relevant supporting data, as well as proper documentation (increasingly generated by LLMs), and automated testing and validation of all AI assets can greatly shorten the time it takes for IT to integrate, refactor, and deploy projects.
- **Bridging the gap between all stakeholders:** With built-in collaborative tools, GenAI-powered assistants, and role-based user experiences as seen with GenAI playgrounds, companies rapidly iterate without incurring risk. Using a shared platform between app developer and data scientist, capable for both GenAI and Predictive AI can help.
- **Engendering both scalability and flexibility:** An AI and data science platform capable of managing assets both self-hosted, managed-hosted, and on-premises, enables customers to prioritize both security and performance over the long term.
- **Optimizing resource utilization:** By automating many operational tasks such as allocating and managing the most efficient project resources across both model training and inferencing (i.e., distributed training and access to right-sized AI hardware acceleration), AI and data science platforms can efficiently balance performance and cost.
- **Enabling continuous innovation:** With the power to identify even the smallest of interdependency issues across a multitude of development frameworks, AI and data science platforms can greatly speed up project and experiment initiation. Bonus points for platforms that can compile optimal experiment stacks using best practices specific to a given use case.

- **Supporting governance and compliance:** Regardless of the model architecture and underlying technology selected for use, an AI and data science platform delivers transparency, accountability, and control. They accomplish this through several paths, including audit trails, guardrails specific to corporate intellectual property (IP) and customer personally identifiable information (PII), and measures against established regulatory requirements.

In this way, AI and data science platforms can function to make life easier, equipping customers with ready-made tools designed to make AI an integral component of enterprise software. This is the true key to unlocking the potential of AI adoption at scale across the business.

## To build, buy, or subscribe: which way forward?

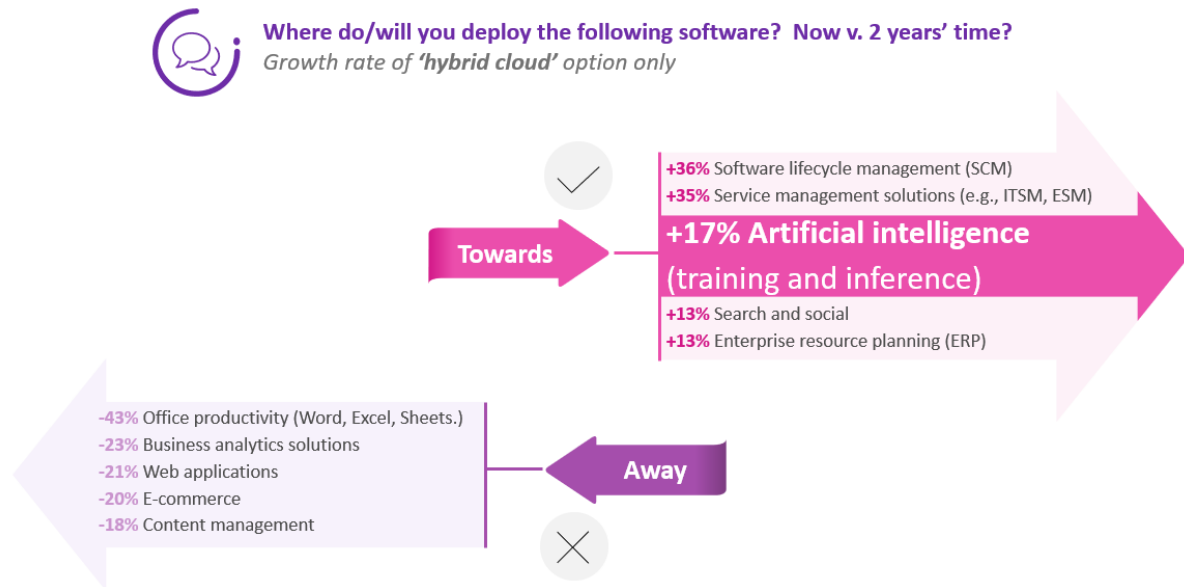
Just as every enterprise is unique, so too are an enterprises' requirements when it comes to building an AI solution. As mentioned above, there are many MLOps solutions in the marketplace, some geared toward adoption as a fully managed cloud service, and others tailored for self-assembled and self-hosted deployment on premises. The same applies to both predictive and GenAI models themselves. Practitioners have at their fingertips a wealth of options. They can very easily sidestep having to manage infrastructure resources and subscribe to proprietary GenAI models via managed APIs from hyperscale cloud providers Google, Microsoft, IBM, AWS, etc. Conversely, they can opt to tune or tailor existing models – either hosted through a hyperscale platform or host on-premises as a self-managed cloud service behind the corporate firewall. Alternatively, practitioners can also adopt an open source model, training or fine-tuning the model to match business requirements, with the option of deploying it either through on-premise data centers or hosting it on a hyperscale platform for inferencing.

*“The challenge of moving from local model development to deploying models in production is one of the harder problems in ML operations. So anything that makes that easier is very relevant. This will help when your infra team is not very well versed in ML.”*

*Director of ML, Financial Services*

Truly, the options are endless, as customers can combine any number of build, buy, subscribe approaches, applying one or another at any point in the solution lifecycle. For example, practitioners can prototype an entire solution using hosted APIs and then reconstruct that same solution in a bespoke manner locally. Alternatively, they can also adopt an on-premises model just for select tasks such as data preparation (i.e., data labeling) or vector embeddings. Selecting the right approach at each step in the project lifecycle can be a daunting task that often comes down to picking the most optimal ratio of cost, performance, security, and control for each project. For this reason, companies building AI outcomes are overwhelmingly opting to build a truly hybrid AI platform that performs equally well across both public and private cloud. To illustrate, a recent Omdia survey of more than 5,000 enterprise IT practitioners found that companies look upon AI training and inferencing workloads as a core IT operational asset. When asked how they would look to deploy several different types of enterprise software, the answer shows AI inferencing and training is one that must increasingly be managed across both cloud and premises (see Figure 6).

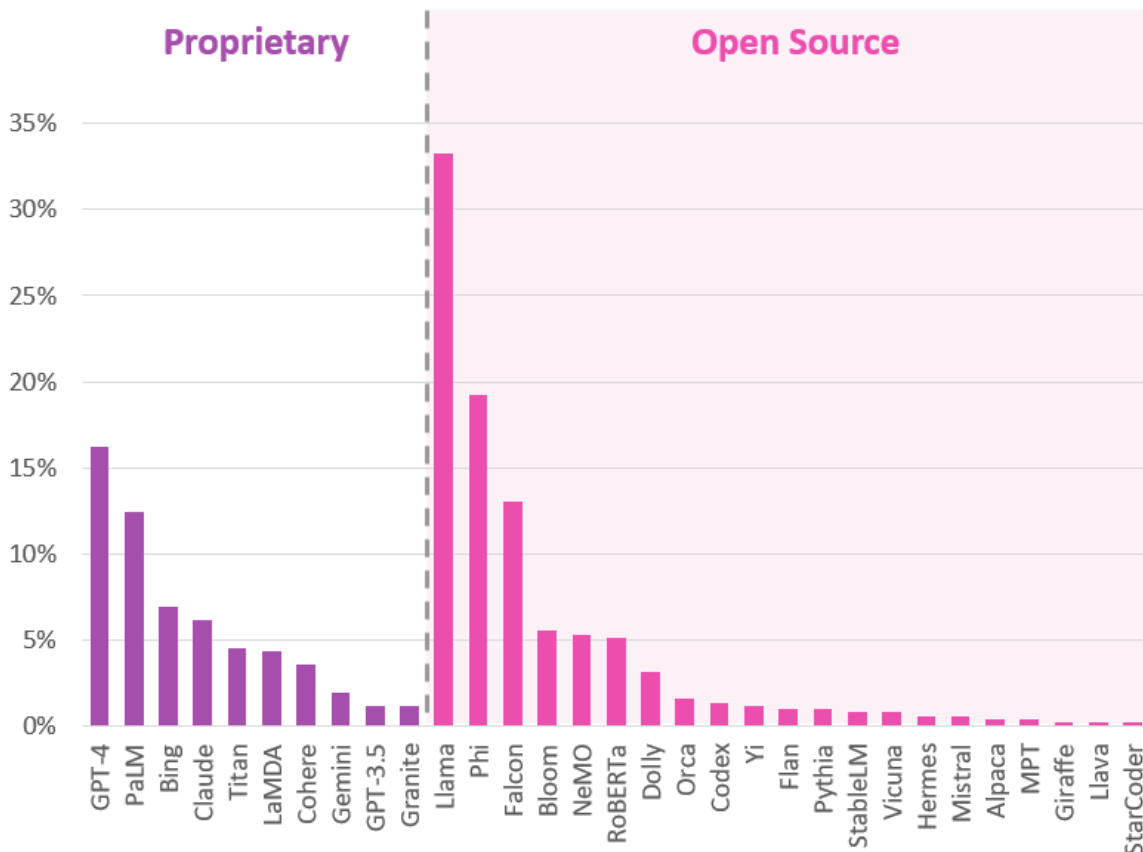
Figure 6: The changing approach to hybrid cloud/premises software deployment



Source: Omdia IT Enterprise Insights 2023

## The increasingly important role of open source software

AI in the enterprise has enjoyed a long-lived collaboration with the open source community. Most AI solutions running in the enterprise today rely to one degree or another upon open source projects, be those tools, development libraries, toolkits, models, or even data science platforms themselves. As expected, at the very foundation of ML development, enterprises depend upon well-trodden libraries like TensorFlow, PyTorch, and Keras for both predictive and GenAI solutions. Looking at GenAI solutions in particular, several toolkits like NVIDIA NeMO, vector databases like Milvus, LLM models like Meta Llama 2, and orchestration frameworks like Kubeflow are finding their way into production environments at a rapid pace. As a matter of fact, in looking at job postings for the 2nd half of 2023, Omdia found that twice as many job posts specific to GenAI called for expertise in Meta's Llama family of open source models as compared with the consumer juggernaut GPT-4 from OpenAI (see Figure 7).

**Figure 7: Comparing open source and proprietary GenAI skills in job postings**


Source: Omdia AI Skills Tracker 2H, 2023

There are some very good reasons for the popularity of open source models like Meta Llama and Microsoft Phi, which itself ranks second to Llama on team open source in job postings. First and foremost, open source models can be pulled apart and evaluated to a much greater degree than proprietary model services like OpenAI GPT-4, providing companies with greater transparency into the way the model arrives at its output. Note also that many open source models are beginning to incorporate training data that is itself open source, which can serve as a tremendous boost for companies seeking to gain better control over model bias concerns. Second, they are cost effective with no licensing fees to pay either in full or by transaction/token. Companies can deploy these models anywhere they want, on premises or cloud, at the edge, or even on-device, depending upon model size. Third, over the past year, a tremendous amount of innovation has emerged from the GenAI open source community of enterprise developers, researchers, and technology companies. Innovations such as model quantization, which allows models to run on less hardware without sacrificing too much accuracy have quickly become a standard practice. And new ways of aligning models with corporate expectations using LLMs themselves using techniques like direct priority optimization (DPO) can greatly simplify what was before a complex and expensive process, available only to highly funded model builders.

Despite these benefits, enterprise buyers must approach open source technologies with open eyes and careful consideration. While providing the advantage of allowing these technologies to be deployed on-premise, open source technologies require a higher degree of technical expertise to integrate and manage than hosted public cloud services. And open source technologies can introduce security concerns such as malicious actors and present compliance risks within highly regulated industries or countries bound by unfavorable legislation. One option is therefore to work with trusted open source technology providers who can curate AI tooling before providing a commercial version of the software. And even despite some of the potential downsides, within the realm of AI and data science, open source software continues to thrive, showing no signs of slowing in terms of influence or adoption among enterprise practitioners.

## Avoid failure by measuring success

Unfortunately, IT-centric risks such as bias and data privacy makeup only one half of the value equation for enterprise AI practitioners looking to move from artisanal into transformational AI outcomes. The art of aligning business objectives with IT capabilities can make or break an AI project. Practitioners must set proper Key Performance Indicators (KPIs) for each AI project that go beyond infrastructure health and costs to also encompass AI-specific measures. Only in this way can project sponsors accurately judge the value and health of a project from inception to deployment and beyond. Here again, investing in a highly operationalized AI platform is vital, as AI and data science platforms aren't just built to track code changes. In managing such assets, these platforms make for a source of truth, accurately instrumentation and monitoring all of the many disparate assets that go into an AI solution, covering every dataset, Python script, and app container in context over time.

From these assets, AI and data science platforms can surface invaluable business-level reports, dashboards, and alerts concerning important measures, including the following examples:

- **Performance:** Model accuracy (precision, recall, etc.), generalization, consistency
- **Transparency:** Model documentation and output explainability
- **Data quality:** Data drift as well as bias, fairness, and inclusiveness detection/monitoring
- **Scalability:** Model computational requirements across both training and inferencing

*“Building and deploying models is hard. The connection between how they get built and how they get deployed is extremely important. Our chosen solution (Red Hat OpenShift AI) ties those two things together very, very closely. You don't have to worry about that mismatch when you deploy it. This takes a huge burden off your infrastructure team. Infrastructure people waste a lot of time on failed models. There's a really big value to highlight right at the beginning.”*

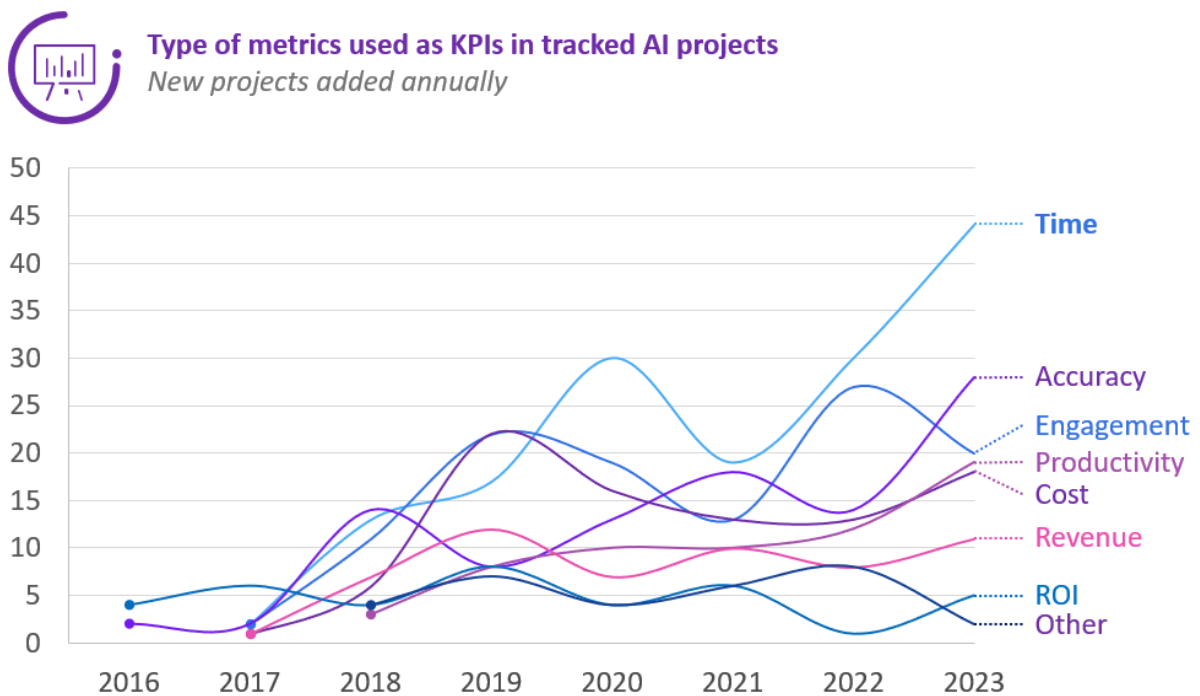
*Director of ML, Financial Services*

- **Robustness:** Security, stress and adversarial attack measures

Together these measures jointly paint a more complete picture of higher level business KPIs such as time to market along with solution accuracy, engagement, productivity.

Among these measures, Omdia has found that time to market and solution accuracy are currently in demand well above other measures (see Figure 8). These two KPIs align well with GenAI in particular, as they can go a long way toward establishing trust in a given implementation, proving, for example, that a model is delivering unbiased, accurate, and useful information; or that the model is saving customers time. The ability for an AI and data science platform to combine technology and business measures can also serve as an extra layer of protection against both seen and unseen risks. For example, a model may be delivering accurate results, but if it is not saving users any time, that situation bears further investigation. The same goes for assessing ongoing compliance with industry standards and governmental regulations (moving targets in their own right).

Figure 8: Evolving enterprise AI priorities, 2016-2023



Source: Omdia AI Business Performance Metrics Database 2H, 2023



---

# Conclusion: Putting it all together

---



While it is impossible to predict how the next AI shockwave will disrupt companies seeking to build business value around AI outcomes, one resounding truth is sure to resound in an unbroken call to arms: in order to maximize their investment in AI while simultaneously lowering associated risks, companies must invest in a solid AI data science platform, preferably while working with a trusted vendor partner capable of conveying deep expertise in operationalizing AI.

That means adopting an ops-centric approach to AI development; embracing open, flexible technology solutions built to handle both traditional and Gen AI, or indeed whatever the future may bring. These solutions should handle diametrically opposed ideologies such as proprietary vs open source or cloud vs premises deployment; regarding GenAI as the latest predictive AI innovation, one that will surely give way to an as-yet unknown technology; and embracing accountability through intertwined technical and business KPIs.

With this mindset, organizations can more readily take control of their investments, and in so doing meet today's demands while making room for tomorrow's innovations. This is the key to successfully navigating the complex landscape of AI development and unlocking new opportunities for growth and innovation.

# Appendix

---

## About Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers integrate new and existing IT applications, develop cloud-native applications, standardize on our industry-leading operating system, and automate, secure, and manage complex environments. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500. As a strategic partner to cloud providers, system integrators, application vendors, customers, and open source communities, Red Hat can help organizations prepare for the digital future.

## Further reading

[OpenShift AI solution brief](#)

[OpenShift AI demo video](#)

[Top 5 ways to implement MLOps successfully in your organization](#)

[Red Hat customer successes eBook](#)

## Author

**Bradley Shimmin**

Chief Analyst, AI & Data Analytics

Applied Intelligence

Bradley.Shimmin@Omdia.com

## Get in touch

[www.omnia.com](http://www.omnia.com)  
[askananalyst@omnia.com](mailto:askananalyst@omnia.com)

## Omdia consulting

Omdia is a market-leading data, research, and consulting business focused on helping digital service providers, technology companies, and enterprise decision-makers thrive in the connected digital economy. Through our global base of analysts, we offer expert analysis and strategic insight across the IT, telecoms, and media industries.

We create business advantage for our customers by providing actionable insight to support business planning, product development, and go-to-market initiatives.

Our unique combination of authoritative data, market analysis, and vertical industry expertise is designed to empower decision-making, helping our clients profit from new technologies and capitalize on evolving business models.

Omdia is part of Informa Tech, a B2B information services business serving the technology, media, and telecoms sector. The Informa group is listed on the London Stock Exchange.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help your company identify future trends and opportunities.

## Copyright notice and disclaimer

The Omdia research, data and information referenced herein (the “Omdia Materials”) are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together “Informa Tech”) or its third party data providers and represent data, research, opinions, or viewpoints published by Informa Tech, and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an “as-is” and “as-available” basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees, agents, and third party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.