Publication date:
April 2025
Author:
Eden Zoller
Bradley Shimmin

Open Source: The Fast-track to AI Success

Use open source to accelerate time-to-value and minimize future costs for both predictive and generative AI

Commissioned by:

Contents

Executive summary

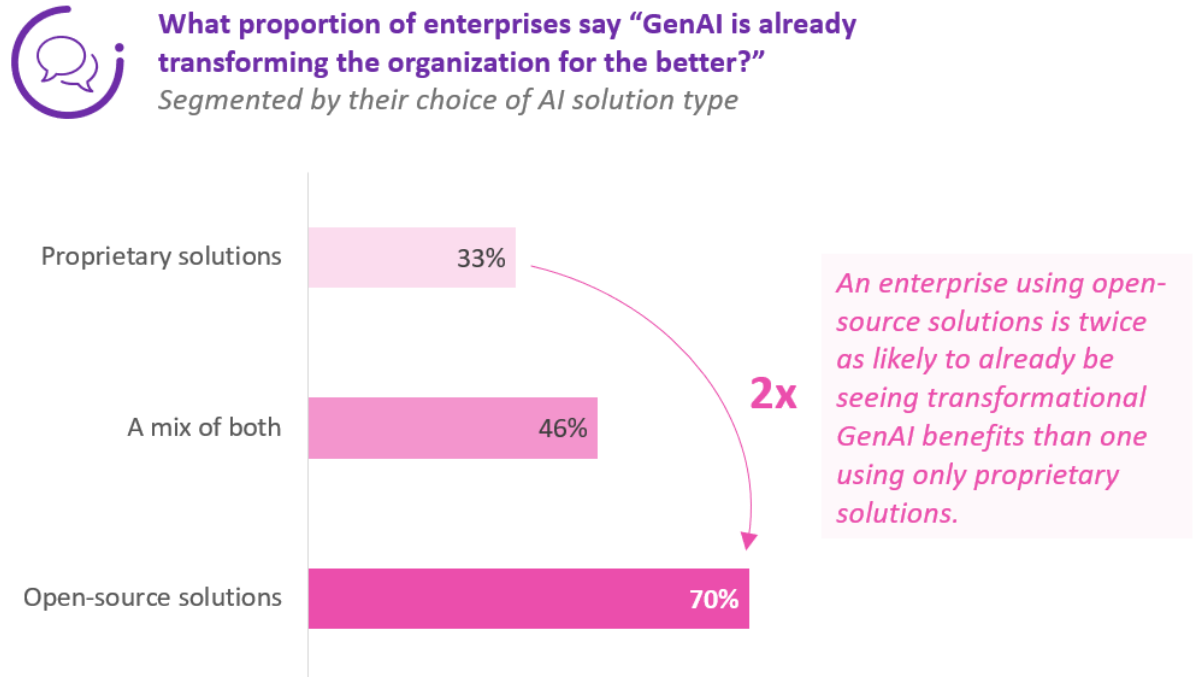
By adopting an open-source AI platform, organizations can accelerate development efficiency and deployment speed, as well as reduce costs, and mitigate exposure to risk.

Building AI in the enterprise is now more accessible than ever, thanks to affordable and sometimes free APIs backed by state-of-the-art Generative AI (GenAI) models. Large, powerful, proprietary models have played an important role in pushing boundaries and creating excitement, including offerings from OpenAI (GPT family) and Anthropic (Claude) and Google (Gemini).

However, AI solutions vary significantly in their ability to drive business impact and deliver value, and so enterprises need to invest with care. Proprietary foundation models can be very useful in certain scenarios but there are alternatives that can provide greater agility, scalability, and cost-efficiency. This is where open-source software comes into play, serving as an often hidden yet transformational force for the better, helping drive AI success within the enterprise.

This assumption is strongly evidenced by Omdia's 2024 Generative AI Enterprise Survey, which focuses on early adopters, where the highest response rate for the transformational impact of GenAI (70%) was among enterprises that had adopted open-source solutions (see Figure 1).

Figure 1: The transformative impact of open-source



Omdia Generative AI Enterprise Survey: Early Adopters – 2024 (356 respondents)

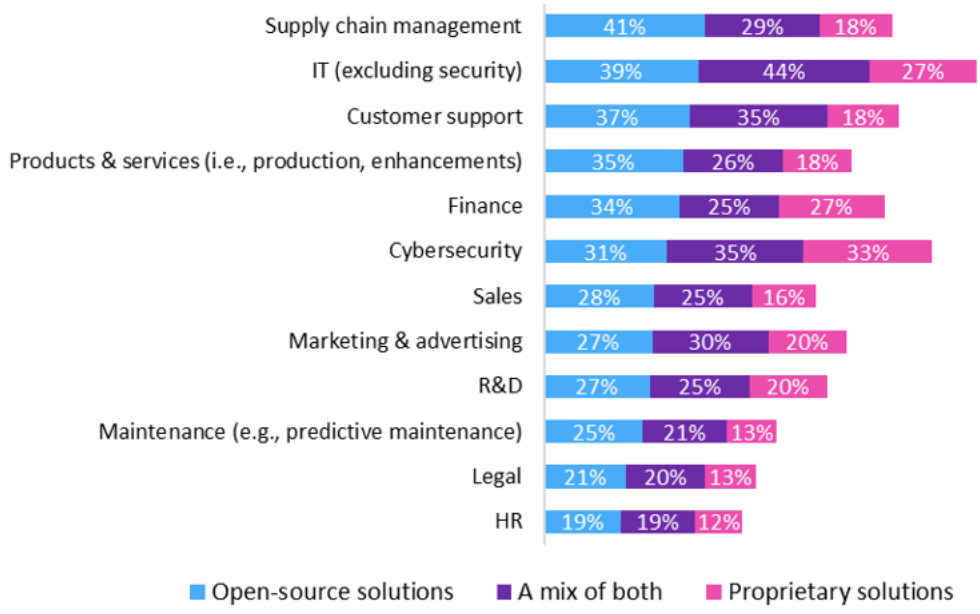
Through primary research and real-world examples, this paper will illustrate how open-source solutions can enable rapid model customization, streamline AI deployment complexities, and minimize future expense through an operational approach to AI.

Deploying AI may be easier, but measuring value is hard

The excitement surrounding GenAI and proliferation of GenAI models and tools has driven rapid adoption across multiple business functions and use cases. Omdia's 2024 Generative AI Enterprise Survey shows that in most cases open-source and/or mixed solutions have stronger traction than proprietary solutions (see Figure 2).

Figure 2: Open-source solutions enable rapid GenAI deployment

What proportion of enterprises say they have "Widespread adoption?" of the following core enterprise solutions?
Segmented by their choice of AI solution type

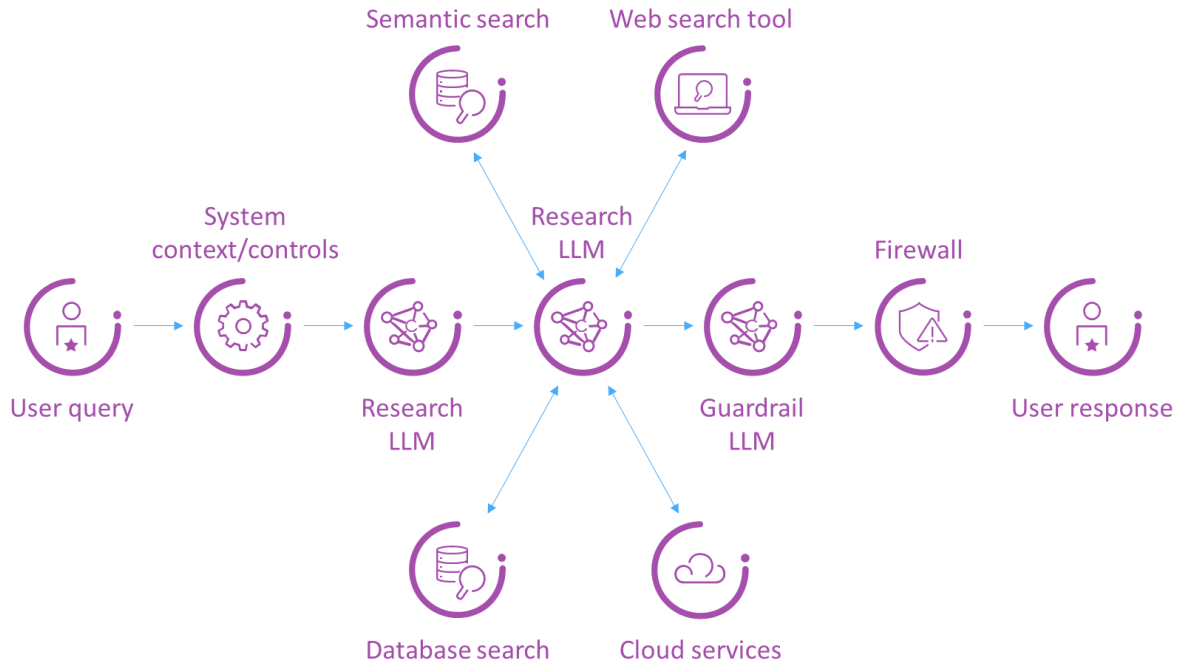


Omdia Generative AI Enterprise Survey: Early Adopters – 2024 (356 respondents)

However, the increased ease of AI deployment starkly contrasts with the difficulty of measuring its actual value. While traditional software and even predictive AI models offer relatively straightforward metrics like accuracy and performance, GenAI presents many unique challenges.

GenAI solutions integrate multiple systems and produce inherently probabilistic outcomes, requiring more safeguards and maintenance to sustain consistent user satisfaction—a critical measure of success (see Figure 3).

Figure 3: A typical GenAI solution



Source: Omdia

Traditional ROI calculations often fall short when assessing GenAI's impact. The dynamic nature of these models makes it difficult to predict long-term performance and quantify the return on investment with certainty. This uncertainty is made worse by several factors:

- **Delayed or stalled solution impact:** Realizing the full potential of GenAI projects can take considerable time and involve several dead ends, making it challenging to demonstrate immediate ROI.

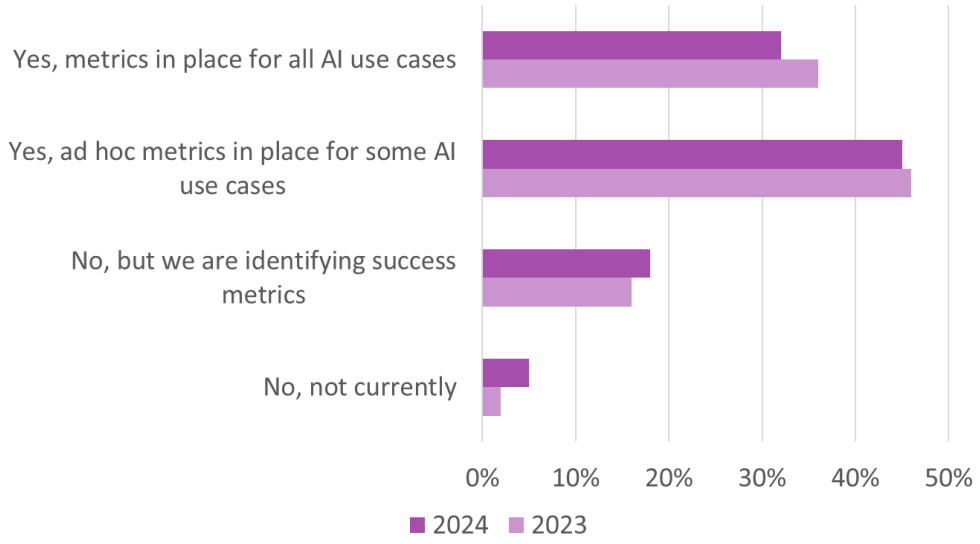
- **Hidden operational costs:** When deploying generative AI solutions, organizations face a distinct cost dimension that goes beyond traditional IT metrics like CPU or memory usage. Token consumption—the fundamental unit of LLM processing—represents an often misunderstood cost driver that can significantly impact budgets. Inefficiently designed applications may generate excessive token usage through redundant API calls or poor context management, silently inflating expenses. While cloud-hosted GenAI models offer scalability advantages, they introduce a deceptive cost structure where token consumption becomes the primary expense lever. This opaque pricing model can result in unexpected cost escalations, particularly when

processing large datasets or generating high-resolution content, as the relationship between workload and expense isn't as transparent as with conventional computing resources.

- **Future technical debt:** Inflexibly designed or rushed GenAI deployments can lead to accumulating technical debt, requiring costly rework later.

Recent Omdia research highlights this challenge. While 77% of enterprises claim to have metrics for assessing AI performance, only a third apply these metrics consistently across all their AI use cases (see Figure 4). This gap underscores the need for more sophisticated and comprehensive approaches to measuring the true value of AI investments, a difficult task given the rapidly evolving GenAI landscape.

Figure 4: Enterprises are struggling to get to grips with AI metrics



Source: Omdia AI Market Maturity Survey (2023 = 368 respondents; 2024 = 478 respondents)

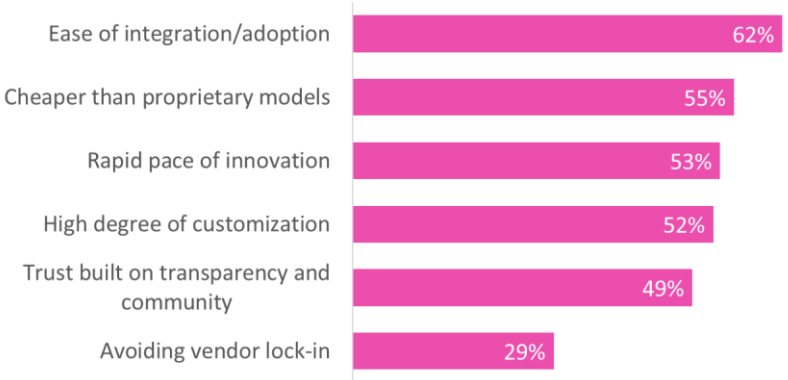
Addressing the ROI challenge: the open-source advantage

Given the challenges and uncertainties surrounding AI metrics, it is easy to see why many companies take an ad hoc approach to measuring ROI. But this is not optimum and what companies need is a flexible, cost-effective approach by which they could achieve consistent, measurable AI value. Fortunately, the AI landscape has long relied upon and been driven by just such an approach, namely open-source software innovation – which offers a unique pathway to achieving measurable AI value by promoting transparency, cost savings, and adaptability across the entire AI lifecycle.

Open-source provides a unique pathway for enterprise adopters, one that is uniquely positioned to address the challenge of both measuring and delivering ROI.

Robust open-source AI assets, encompassing models, frameworks, tools, and platforms, can readily work in harmony to deliver tangible advantages. By reducing operational costs, simplifying integration, and accelerating model deployment, open-source helps create clearer ROI metrics. Omdia's survey reveals a strong preference for open-source adoption due to its positive impacts on cost and complexity (see Figure 5). The ability to readily integrate foundational libraries like PyTorch or frameworks like LangChain and LlamaStack, all without having to completely refactor existing code, can greatly speed innovation and delivery. This flexibility also extends to hybrid and multi-cloud environments, where software flexibility can likewise improve ROI. These factors all contribute to a significant reduction in the uncertainties associated with AI projects, paving the way for a more predictable and measurable ROI.

Figure 5: Why AI practitioners choose open source over proprietary software



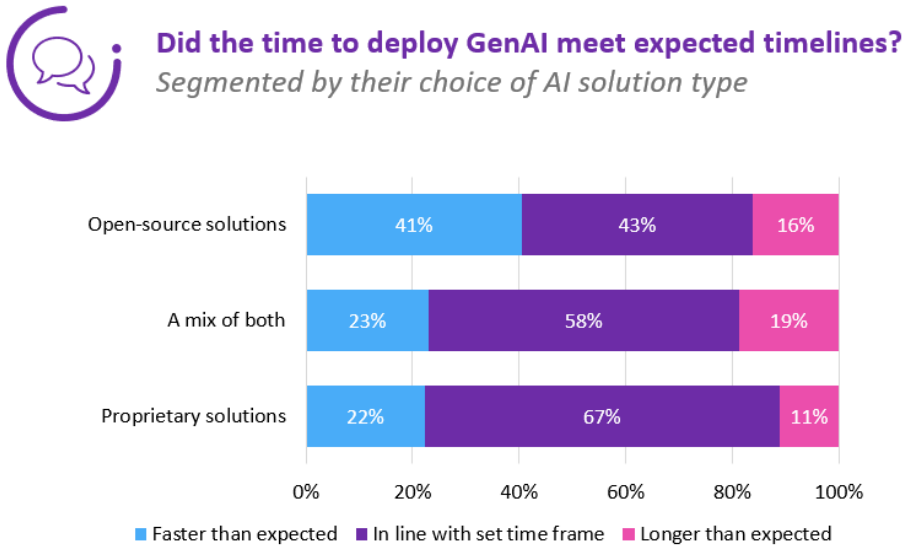
Source: Omdia Generative AI Enterprise Survey: Early Adopters – 2024 (291 respondents)

Accelerating time to value

Accelerating time-to-value is critical for enterprises looking to leverage AI effectively. By adopting an open AI development platform that prioritizes open-source technologies and practices, companies can bring AI-driven solutions to market faster than traditional proprietary systems allow.

Findings from Omdia's 2024 Generative AI Enterprise Survey confirm the assumption that open-source solutions can accelerate AI deployments, as shown in Figure 6. Forty-one percent of respondents adopting open-source GenAI solutions report that deployment times were faster than expected, compared to 33% using proprietary GenAI solutions.

Figure 6: Open-source solutions enable rapid GenAI deployment



Omdia Generative AI Enterprise Survey: Early Adopters – 2024 (356 respondents)

Consider the application of IT DevOps philosophies to AI development, a practice commonly referred to as Machine Learning operations or MLOps. This important practice helps companies streamline the entire AI lifecycle, managing everything from initial data preparation to model deployment and ongoing monitoring. Take, for example, the perennial challenge of optimizing resource allocation

and utilization spending. Support for popular open-source containerization technologies such as Kubernetes enables companies to tackle important AI infrastructure requirements.

- Native support for hybrid-cloud deployments
- Automated resource provisioning and deployment
- Dynamic resource scaling to match user demand
- Seamless integration of disparate AI accelerator hardware
- Automate security and privacy compliance tasks

A commitment to open-source principles enables rapid iteration and adaptation. When deploying GenAI, teams can easily pivot to mature open-source frameworks for MLOps (e.g., Kubeflow) and open-source developer tools (e.g., LangChain or LlamaStack) without the complexities of extracting data and code from closed, proprietary platforms. This flexibility allows organizations to remain agile and responsive to evolving technological advancements and market demands. The result is faster innovation and a quicker return on investment.

Case Study: DenizBank transforms AI operations and time to market

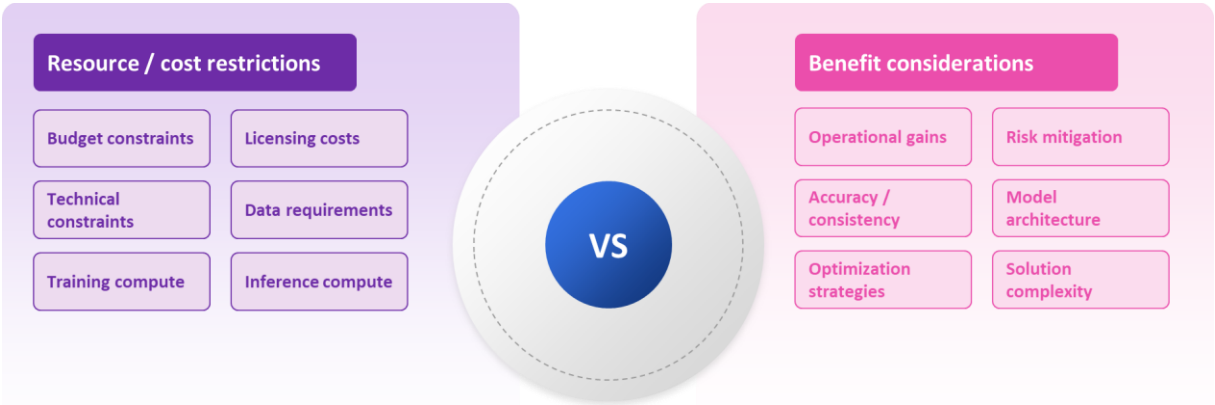
DenizBank is the fifth biggest private bank in Turkey, owned by Emirates NBD. The bank has around 120 data scientists that have developed over 100 AI and ML models to support critical banking activities, such as fraud prevention and credit application screening. Data scientists are supported by Intertech, the bank's IT subsidiary and a leading provider of financial service solutions in Turkey and beyond.

DenizBank tasked Intertech with transforming the model development environment (workbench) used by its data scientists. While useful, the existing workbench lacked standardization, was overly complex and resource-intensive, and required labor intensive setup for each individual model. Intertech wanted to address this by building a comprehensive, standardized workbench for data scientists that improves time-to-market while delivering cost savings. To achieve this, DenizBank and Intertech leaned into existing partner Red Hat, adopting OpenShift AI with support from Red Hat Consulting. Appealing attributes of OpenShift AI for the project included its use of Kubernetes for container orchestration, which helps simplify workflow deployment and management at scale. Moreover, OpenShift AI can be configured in ways that streamline and automate processes for data scientists, for example through the provision of standardized templates and pre-built cluster images, plug-and-play functionality for connecting data sources. These self-service capabilities help automate and accelerate model development. Intertech expects OpenShift AI to reduce the time to market for new models from around one week to just 10 minutes. The OpenShift AI solution also leverages GPU slicing to optimize GPU usage for model training and serving, drawing on NVIDIA's Multi-Instance GPU (MIG) technology. This maximizes resource utilization and increases flexibility.

Driving scalability and cost efficiencies

As AI initiatives expand, scalability and cost-effectiveness become paramount. Nowhere is this more evident than in building and serving AI models, particularly resource-hungry GenAI models that are capable of consuming tremendous amounts of storage and hardware. The bigger the GenAI model in terms of size (i.e., the number of parameters and size of its context window), the higher the training and inference cost -- and the longer it takes to deliver results (model latency) in production.

Figure 7: Balancing the scales of AI performance vs cost



Source: Omdia

In order to create impactful AI outcomes capable of driving value, companies must constantly balance these factors, matching available resources to solution requirements.

Proprietary AI models from providers like OpenAI, Anthropic, and Cohere can certainly take on complex tasks thanks to new innovations such as multi-modality, which allows a single model to take in and then generate text, images, audio, and video in a highly unified manner. More recently, model makers have begun to address more complex model use cases through in-model self-reflection, mixture of expert (MoE) model routing, and chain of thought (CoT) reasoning. However, these capabilities come with a hefty price tag in terms of storage and GPU compute costs, not to mention inference latency.

Importantly, over the past twelve months, open-source models have quickly improved at matching larger, frontier models and have in many cases closed the capability gap that once existed between larger, frontier-scale models like OpenAI o1. The new wave of smaller, powerful open-source models includes Alibaba Qwen, IBM Granite, and DeepSeek-R1, Meta's Llama, and Microsoft Phi-3-mini. Some of the smallest models (e.g., the 3.8 billion parameters Phi-3-mini) are often small enough to

fit on a single consumer-grade GPU. The new wave of smaller open-source models can now keep pace with much larger models. R1 offers impressive performance, efficient use of resources, and low inference costs. According to DeepSeek, R1 beats OpenAI o1 on the benchmarks AIME, MATH-500, and SWE-bench Verified. The Qwen 2.5 Coder model from Alibaba is available in a 32 billion parameter configuration that performs within four percent of all larger models from Google, OpenAI, and Anthropic, according to third party benchmarks from Aider. IBM's new set of Granite Mixture-of-Experts (MoE) models can deliver performance across a wide range of use cases including RAG, classification, summarization, tool use, and entity extraction.

By integrating corporate data and expertise into smaller open-source models through fine-tuning, instruction tuning, and distillation, companies can develop models that surpass frontier models in both performance and functionality.

Moreover, when combined with retrieval augmented generation (RAG) techniques and prompt engineering, model fine-tuning represents the fastest and most efficient way for companies to ground models in contextual and timely facts while also expanding domain-specific capabilities. Low-code tools like Unsloth and Red Hat AI InstructLab can further simplify these techniques by automating complex tasks such as generating synthetic training data. These ideas are clearly on the minds of enterprises in Omdia's 2024 Generative AI Enterprise Survey, which found that 23% of North American enterprises currently follow fine-tune models to equip smaller models with company-specific skills and information.

Case Study: Red Hat empowers AGESIC to scale AI deployments and boost efficiency

Uruguay's Agency for Electronic Government and Information and Knowledge Society (AGESIC) oversees the implementation of the nation's digital strategies, including data, AI, cybersecurity and digital citizenship initiatives. It coordinates efforts across various government agencies, from local governments to ministries and its dependencies (approximately 50 in total) to ensure a consistent and effective approach. AGESIC is collaborating with Red Hat (a long-standing partner) to help government offices to test, deploy and scale AI solutions, using open-source platforms like Red Hat OpenShift and OpenShift AI.

The Uruguayan government strongly favors open-source software for public funded projects, and AGESIC notes this is because of potential cost savings and also because open-source solutions are flexible and well-suited to support modular architecture, sharing, and reuse. Red Hat solutions also support robust security, which is important to AGESIC's commitment to the provision of secure digital public Infrastructure (DPI). Projects supported by Red Hat include the development of a new AI enhanced platform to standardize, automate and improve the ticket classification process (i.e. for technical and associated issues) for multiple help desk teams supported by AGESIC. Prior to the introduction of the new platform, the ticket classification process was a manual, multi-step process. Besides being slow, the process could be inefficient, for example misclassified tickets that were not routed to the right technician. The Red Hat solution has produced significant improvements, with the automated ticket classification process taking a matter of seconds to complete where previously it could take up to an hour. AGESIC adds that help desk teams are pleased with the solution, which has freed them from a time-consuming mundane task to focus on higher value activities.

Reducing risk in AI investments

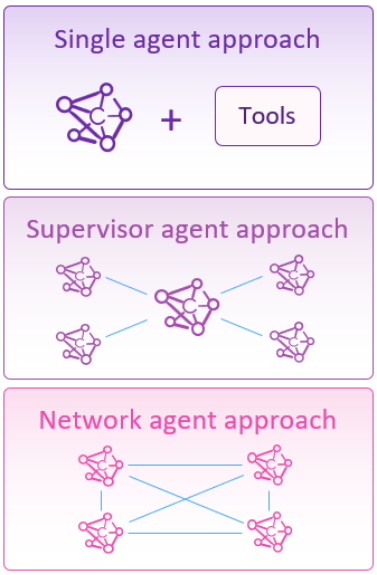
AI deployments can introduce significant risks, ranging from unpredictable costs to compliance concerns and security exposure. But investing in open-source solutions can mitigate many of these concerns by enabling greater transparency and accountability. This empowers organizations to minimize risk and maximize the long-term value and resilience of their AI investments. This contributes to a more predictable and manageable cost structure over the lifetime of the project, directly addressing the issue of future technical debt. There are many open-source model guardrail tools like Granite Guardian and NVIDIA NeMo Guardrails that automate responsible AI tasks such as model security, data privacy, and output consistency. Open-source communities like TrustyAI will further contribute to mitigating risk by fostering the development of tools focused on model explainability, model monitoring, and responsible model serving.

Open-source projects lower security risks such as hidden backdoors or unpatched vulnerabilities because the project code is open to community scrutiny. In addition, companies can inspect all aspects of an open-source project, which gives greater control over AI outcomes - a difficult task for GenAI in particular, where each word or token returned by a model hinge on a probability curve. At the same time, end-to-end visibility into a model's architecture, training data, and alignment techniques enables companies to significantly enhance trust in its outcomes. This is essential for meeting increasingly stringent regulatory requirements and ethical concerns.

Going small and agentic

While massive LLMs garner significant attention, they present considerable challenges in deployment and cost-efficiency. Their sheer size requires substantial computational resources, leading to high infrastructure costs and lengthy spin-up times (even tens of minutes). This contrasts sharply with the agility and scalability found in many smaller LLMs. Companies do find tremendous value in leveraging hosted LLM APIs to quickly prototype solutions and explore new use cases. However, in order to realize corporate performance, security, governance, compliance requirements, AI practitioners are increasingly readying these proof of concepts (POCs) for production by moving to several smaller models that have been fine-tuned using corporate knowledge and skills and then hosting the final product on company-controlled infrastructure, be that on-premises or in the cloud.

Figure 8: Agentic LLM architectures



Source: Omdia

This modular approach pairs perfectly with the use of smaller, open-source models that have been tuned to target specific tasks using corporate knowledge and skills. Instead of over engineering numerous prompts in order to encourage a single model to behave in different ways, AI practitioners can instead architect highly optimized and flexible AI systems, systems that can take on more advanced capabilities such as multi-modal support for language, image, video, etc. or even the introduction of autonomy via agentic processes where multiple models can work in harmony to solve complex, multi-step processes (see Figure 8). These optimized AI systems can also include the latest compression techniques like quantization and sparsity available in modern open source inference servers like vLLM.

The orchestration of multiple, smaller models via an open-source AI platform can take a solution beyond what is possible with a single, large frontier model in terms of capability, cost, and performance. For example, a series of smaller, orchestrated models can be more readily scaled up or down as needed, mirroring cloud-native app architectures in controlling both inferencing costs and minimizing app latency. They can also be switched in and out with old models being demoted and new models being promoted as needed, even in real-time and in a fully autonomous manner. Furthermore, smaller LLMs significantly reduce the risk of managing a solution over the long term as they can be frozen in time, thereby avoiding break/fix updates, capability drift, and even outright inference refusals that are unfortunately common with fully managed frontier model offerings.

By taking full ownership of AI with smaller, open-source LLMs, companies can achieve faster iteration cycles, deploy updates more rapidly, and reduce operational overhead.

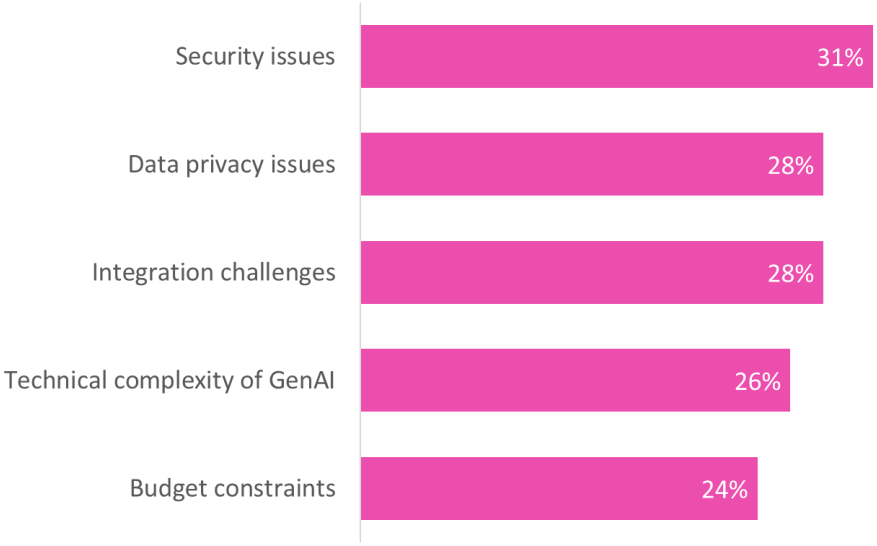
Mitigating concerns, removing misconceptions

There are many advantages inherent in using open-source for enterprise AI. For example, licensing costs can be taken off the board, leaving more room for high-value investments in the underlying infrastructure. But open-source is not without challenges, and concerns remain regarding security, stability, and support. The ability of open-source to match the capabilities of proprietary solutions is less of an issue as the feature and performance gap is closing, and indeed many open-source models are now on parity or outperform large proprietary models, as noted earlier in this paper.

Security

Numerous studies from Omdia show that security forms the single biggest barrier to AI adoption in the enterprise (see Figure 9). Bad actors actively seek out vulnerabilities in both open-source software and its supply chain, as evidenced in the recent discovery of malware within the Ultralytics YOLO11 vision model. A fully supported open-source platform such as Red Hat OpenShift AI, however, can help to mitigate this and many other security risks through transparency, accountability and supporting a secure software chain. The vendor constantly inspects any modification to all supported AI assets. Further, vendors like Red Hat provide numerous automated security compliance features, such as regular patching and updates.

Figure 9: Security is the biggest concern among early adopters of GenAI



Source: Omdia Generative AI Enterprise Survey: Early Adopters – 2024 (Top 5 answers only; 356 respondents)

Stability

The community-supported nature of open-source projects can lead to the perception that open-source projects are not as reliable as their proprietary counterparts. This is not the case with mature platforms rooted in open-source, as is the case with Red Hat OpenShift AI). Such platforms offer the stability of a robust, enterprise-grade support program that does not depend upon the end user organization working with the developer community to report bugs and patch code. This forms a collaborative advantage not found in proprietary solutions that can lock organizations into costly long-term commitments to both technology and support.

Conclusion: Open source is the compass for enterprise AI

Adopting open- source software is more than a technological choice—it's a strategic imperative. This approach empowers enterprises to build resilient, adaptive AI solutions that deliver value today while anticipating future advancements.

Proprietary, frontier models have their place and can add significant value as a tool for AI deployment, for example through the provision of readily available APIs. However, such benefits should not overshadow the importance of measurable value and long-term cost efficiency. Proprietary solutions often fall short in these areas, leading to hidden costs, technical debt, and hampered agility.

Open-source platforms, exemplified by Red Hat OpenShift AI, offer a compelling alternative. They streamline the AI lifecycle, from data preparation to deployment and monitoring, fostering collaboration and reducing bottlenecks. This translates to faster time-to-value and improved ROI through optimized resource utilization and more efficient workflows.

The inherent flexibility of open-source, combined with MLOps principles and hybrid/multi-cloud capabilities, enables greater scalability and adaptability. The transparency of open-source solutions mitigates risks associated with security, compliance, and intellectual property, allowing for greater control and customization. The shift to smaller, more manageable open-source LLMs contributes to cost reduction and improved efficiency.

Recommendations

For companies starting their AI journey:

- Strategic partnerships: Collaborate with a proven open-source provider offering comprehensive AI stack support. This ensures seamless integration and reduces the burden of managing disparate technologies.
- Open platform foundation: Begin with an open-source platform that readily accommodates diverse technologies and multiple data sources. Such flexibility is crucial for adapting to evolving AI requirements while avoiding vendor lock-in.
- Prioritize security and transparency: Implement security measures from the outset, ensuring compliance with relevant regulations and in-house ethical requirements.
- Phased implementation: Adopt a phased approach to avoid overwhelming your team and resources. Starting with a well-defined, manageable project allows for iterative learning and reduces the risk of project failure.
- Establish metrics from the start: Develop robust metrics to accurately measure the impact and ROI of your AI initiatives at all levels of abstraction from resources up through solutions.

For organizations with established AI initiatives:

- Optimization through open-source: Leverage open-source solutions to enhance both cost efficiency and performance.
- Hybrid/multi-cloud flexibility: Enhance agility and scalability through hybrid and multi-cloud capabilities. This allows you to distribute workloads optimally, minimizing costs and maximizing resource utilization.
- MLOps for lifecycle management: Embrace MLOps principles to streamline the entire AI lifecycle. Such automation reduces manual intervention (e.g., technical debt), resulting in faster iteration and lower operational costs over the long term.
- Future-proofing through open standards: Invest in solutions based on open standards and engage with numerous open-source communities. This ensures long-term compatibility, adaptability, and access to a rich pool of expertise 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.

Author

Bradley Shimmin
Chief Analyst, AI & Data Analytics
Bradley.Shimmin@omdia.com

Eden Zoller
Chief Analyst, AI
Eden.Zoller@Omdia.com

Get in touch

www.omdia.com
askananalyst@omdia.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 TechTarget, 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 TechTarget, Inc. and its subsidiaries or affiliates (together “Informa TechTarget”) or its third party data providers and represent data, research, opinions, or viewpoints published by Informa TechTarget, 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 TechTarget 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 TechTarget 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 TechTarget 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.