cognizant®

Risk-based quality assurance of generative AI solutions

Introduction

From many conversations we have with our clients, it becomes clear that the most important challenge they see with the current state of generative AI (gen AI) is building trust in this technology. Gartner research from August 8, 2023 also confirms this by stating that the mass availability of gen AI has become a top concern for risk executives in the second quarter of 2023. The organizations need to ensure that they are acting ethically and managing client's information responsibly and in accordance with the regulatory norms. Apart from executives, users are also looking to build confidence in this new technology. To address these challenges, organizations are trying to make use of the same approaches they know from other information technology areas. But those approaches don't often take them further from the evaluation and learning stage. For organizations to move forward and start leveraging the benefits of gen AI at full scale, we believe they need to understand how it is different and how to deal with new risks properly.

The challenges of gen AI: Systems quality and security assurance

With gen AI, we have a situation where the technology develops faster than the law and existing risks and quality management frameworks of organizations. This limits broad, secure gen AI adoption and leads to 'shadow AI' usage (unofficial AI solutions not under the control of the IT or compliance departments) since many workers are willing to leverage the new productivity booster right away. The challenges include (but not limited to):

- Stochastic nature of outputs: The content generated by gen AI is stochastic with very limited ability to explain and repeat.
- Sensitive information exposure: Gen AI models may include and expose personal data and other sensitive information. The existence of sensitive information in the training data will likely result in models that may extract this information to the user.
- Hallucination tendency: Despite the convincing and realistic nature of generated output, a concern with gen AI models is their tendency to hallucinate facts and make up information. These cannot be eliminated by simple system configuration tweaks.
- Biases: The pretraining datasets of gen AI can contain political discourse, hate speech, discrimination, and other biases. Obviously, these might get into the model and then become a part of the generated output.
- Outdated information: Facts learned during pretraining can become outdated with time. However, retraining the model with updated pretraining data is expensive and fine-tuning is also challenging. Unlearning old facts and learning new ones are challenging.
- Intellectual property issues: Currently, the legal status of gen AI outputs is still unclear. To avoid any potential legal issue, in some cases, the organizations may want to control what is allowed to be generated and what not.
- Inference latencies: This may limit the approaches available for quality assurance.

To mitigate the related risks, organizations started applying the same approaches they know from data and software quality assurance. However, they must realize that it's not possible to directly remove bad or outdated data from the model and that cleaning up data won't mitigate hallucinations. In general, it's not possible to explain/interpret the outcomes of large gen AI models. New quality and security assurance methods and frameworks are required.

Firstly, organizations need to develop quality evaluation and risk mitigation strategies towards gen AI.

Gen AI quality evaluation strategy and risks management

ISO 25010 quality model defines quality characteristics specific to machine learning models and systems as robustness and the ability to explain and interpret. These correlate with challenges we mentioned, and we would use them for evaluations of gen AI, if not for the much greater complexity of gen AI models. The complexity of modern gen AI models (and of datasets that are used for their training) is impractical for any individual to explore and explain their logic. Since we cannot simply explore and explain that logic, we also cannot be sure that the outcome obtained using gen AI is 100 percent correct. With that, the most basic quality evaluation is to get some outputs and estimate how many correct or near-correct outcomes are generated, which will be measured for accuracy or truthfulness of the solution.

That type of testing approach is called adversarial testing. In this, the model's ability to separate fact from an adversarially selected set of incorrect statements is estimated. We can use public benchmarks or prepare our own set of golden questions to which the correct answers are known. Depending on how well the questions are answered (and this is measured by human testers) we will be able to analyze the accuracy of the model. To tell if the accuracy of the model needs to be improved, we need to know in advance, what level of risks the stakeholders and organizations can tolerate. For example, Med-PaLM2 model of Google can answer 85% of the US medical licensing exam questions, but this is still considered too risky to be adopted in clinical practice.

Accuracy will help evaluating characteristics like hallucinations, but not the cultural and social aspects. This is why the gen AI quality metrics are more variative in nature compared to traditional IT-systems. Additionally, we will need to measure some or all the characteristics, like fairness, bias, toxicity, privacy, transparency, accountability, robustness—and perhaps more—depending on the use case. With that we will potentially include several different quality metrics into our evaluation strategy for measuring the potential risks we have identified.

The approach, which is now quite common for evaluation of these characteristics of gen AI, is called red teaming. Same as with the accuracy, in red teaming, the real-world adversaries will be emulated to identify risks, blind spots and potential harm. The members of the red team think like attackers and probe AI systems for failures. Like accuracy, there are public benchmarks for red teaming available, which can be reused for different cases.

Since achieving the highest quality levels in all characteristics might be very costly and economically inefficient, organizations should define up front what quality measures and ranges will be acceptable to mitigate the level of risks they defined for their use cases. With that, an up-front categorization of gen AI-based system risks and risk tolerance level identification is required. Risk tolerance will be defined by stakeholder's readiness or appetite to bear the risk to achieve its objectives. Legal and regulatory requirements should be considered and IT teams need to be involved as well. The criteria and methods for categorization should be defined at organizational levels in gen AI architecture blueprints.

For example, the risk levels might be defined in alignment with the European Union's proposal on AI regulation and they may look like this:

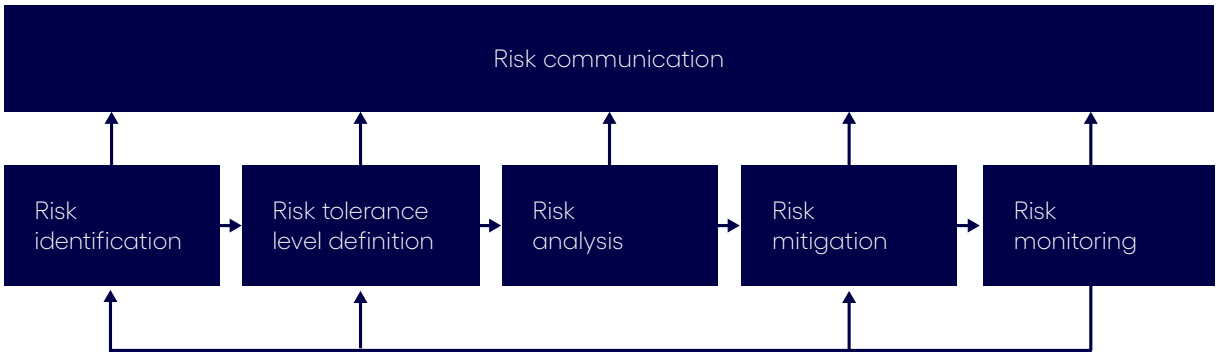

Unacceptable risk


High risk


Limited risk


Low risk

These are the first stages in our risk management process which look like this:

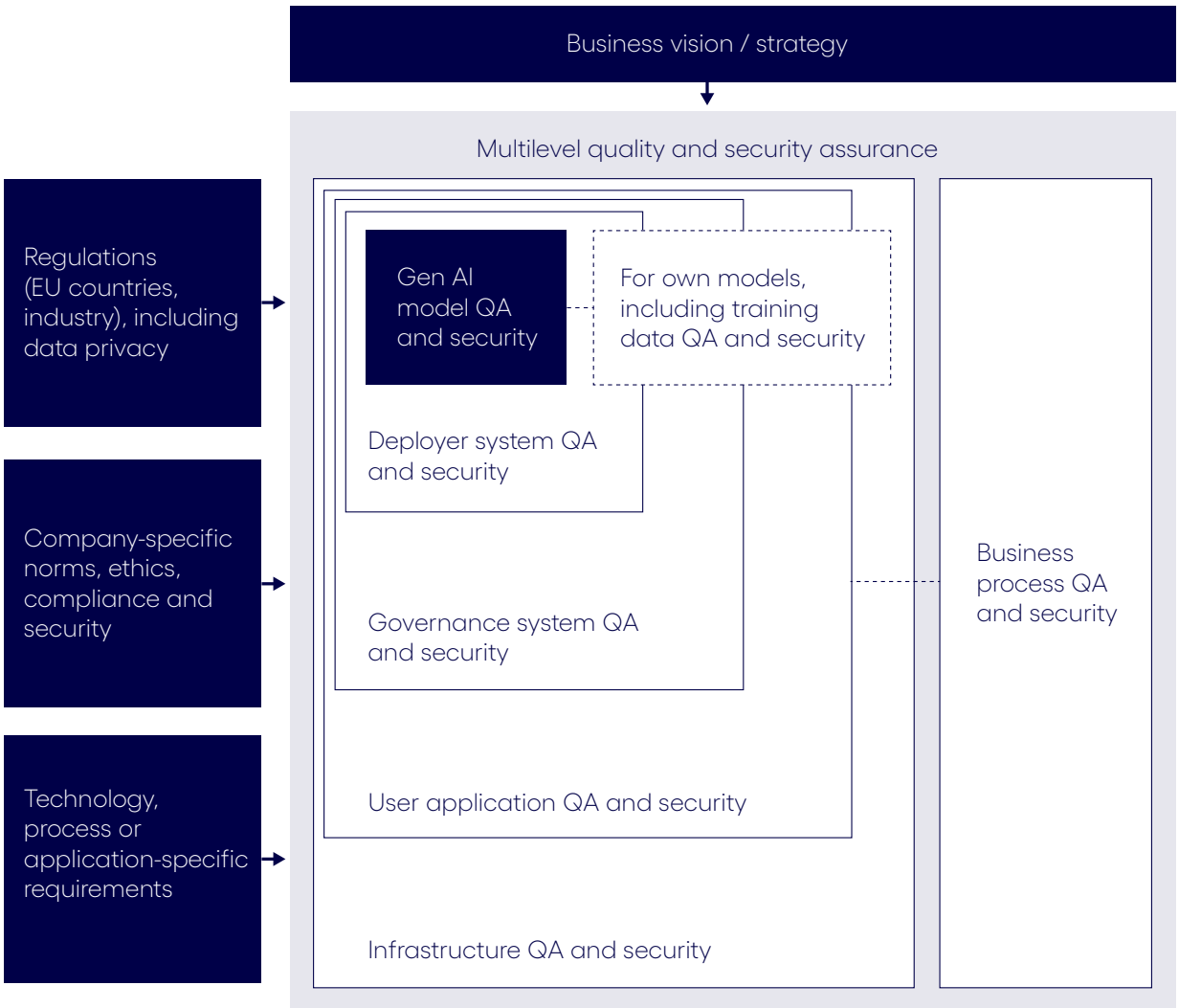


Now, we can move to the next step and identify at what levels of gen AI-based system the risk mitigation measures can be implemented.

Gen AI risk mitigation levels

A gen AI-based system can be represented by a multilayered model in which each of the levels impact the overall solution quality and security in its own way.

Organizations would usually use large gen AI models prebuilt by third-party vendors due to high costs of training their own models. While these models allow certain level of control of the output over so called system messages or prompts, they are still a kind of black boxes, which IT architects will naturally be willing to put into some controlled container. So, in the next level, there will be a system of a cloud provider (often called deployer), hosting gen AI. Both the gen AI model providers and deployer system owners bring their own safety systems. Next, gen AI systems may include several gen AI models and deployer systems. This is where the implementation service providers and independent software vendors bring third-party systems to govern multiple gen AI services of the organization (we will refer to this as governance system) with own risks management capabilities. Developers also have control over risks on application level, for instance, over metaprompts. Last but not the least, infrastructure and business processes represent other layers of gen AI quality and security assurance that bring us to the following model:



Understand gen AI model level quality risks mitigation options

When leveraging prebuilt gen AI models, the preference should be with ones providing transparent, independent, standardized evaluations of capabilities and safety. For instance, an industry body organized by Microsoft, Anthropic, Google, and OpenAI (and open to other vendors) called Frontier Model Forum promotes safe and responsible development of frontier AI systems, facilitating information sharing among policymakers and industry. It is expected that in the future, after regulations are settled, the vendors will also state the level of conformity of their models to this or other regulations.

Depending on solution type, different gen AI models (or combinations of models) and QA and security frameworks can be leveraged. For example, for financial question answering, BloombergGPT might be a better fit than some generic language model. For question answering in other industries, LLaMA or its specific fine-tuned variation might be a good option. For each specific model, there might already be a community of enthusiasts with appropriate frameworks for quality and security assurance as this is for instance the case with LLaMA. A model that is specialized for certain questions or domains specific to selected use cases will likely ensure better quality, but may perform poorly on more general tasks. In this case, additional configurations will be required on model or higher levels of the system.

Once a gen AI model or a combination of models is selected, there might be additional risk mitigation activities required. Some of the options on this level include:

- **Improving outcome over API parameters:** The large commercial gen AI models come with several API parameters, which influence the outcome. One parameter, which is usually available for language models, is temperature. It defines the level of creativity of the responses. Reducing this towards zero will increase the chance of same response being generated for the same question, but will make these less natural for human and may lower the user experience.
- **Model retraining or fine-tuning:** As mentioned, these options require involvement of skilled data and AI experts, and are often associated with high costs. Also, some of the leading market models don't support these options. However, for very specialized and smaller models, the level of hallucinations decreases, repeatability improves, especially in a narrowed specialized application area with improved quality of new data used for retraining or fine-tuning.
- **Denoising language model corruptions with separate partner model:** We can add another model into the system which is trained on artificially noised statements and their clean counterpart. This technique will help avoid fine-tuning of the main model, which might be very costly because of its big size, and only fine-tune the much smaller partner model, which can be done in a frequency according to the risk levels identified.
- **Leveraging moderations endpoint:** Some gen AI service providers offer a moderation endpoint which is a tool to check whether the content complies with predefined policies. The moderation outcome is however normally not explainable as well and changes over time as it relies on another large model.

Deployer system-level gen AI risks mitigation

Hosting a gen AI model in environments under enterprise control allows organizations to better address security concerns (as compared to open SaaS platforms). There are many different techniques, methods and technical solutions emerging for improving the quality of the solution without changing the model itself. There is a bright variety of options like Safety System of Microsoft Azure or synchronous prompt modification related methods. The reasons for that kind of variety are high complexity and costs of model retraining, and the potential of addressing different kinds of use cases with the same model. Some of the mitigation options on this level are:

Content filtering: Deployers started equipping the gen AI solutions with content filtering systems put on top of gen AI models. For instance, the content filtering system integrated into Azure OpenAI Service runs alongside the core models and uses an ensemble of multiclass classification models to detect four categories of harmful content (violence, hate, sexual, and self-harm) at four severity levels respectively (safe, low, medium, and high). On top of that, there is an option to detect abuse cases (geo-specific) in asynchronous mode for both prompts and responses.

Moderation: Many deployers also offer separate moderation endpoints (like in the approach of gen AI service providers explained above) as well as monitoring features to filter out and log potential abuse cases.

Gen AI governance platform for organization-level risks mitigation and monitoring

Very soon, the number and types of gen AI models and sometimes their hosting platforms in organizations will increase, and companies will need an efficient way of governing all of them at once across those organizations. In other words, an additional layer for AI programmes across organizations will be required to operationalize and mitigate the risks on organizational level.

This can be implemented with the help of gen AI governance platform, which is independent from gen AI model providers and deployers. Governance platforms play an important role in risk management process, especially in the monitoring stage. Other than that, the governance platform can help assess, report, evaluate the risks and manage AI systems organizations build, helping to ensure they are compliant and safe, and understand where more attention is required.

Cognizant Neuro® AI is one of the advanced platforms which will help organizations enable and accelerate enterprise-grade AI adoption.

Application-level gen AI quality risks mitigation options

On top of the three levels we mentioned, the role of developers in application safety and quality is very significant. And this is not a straightforward approach since they have a broad variety of options on how to control what users are prompting or what they are prompting to get the desired output. They can, for instance, add system messages or metaprompts that are instructions provided to the model to guide its behaviour. Another protective measure can be to include detailed instructions in the prompt, instructing the AI not to answer certain type of questions. In this case, the response generated informs the user that the AI can only answer questions on specific topics and to avoid inappropriate questions.

Prompt modelling

The outcome of gen AI depends very much on how the prompts are formulated and what metaprompts/system messages are used. The task of a developer is to analyze how the variations in the prompts affect the safety and other quality metric levels in the input and output content, and what variations to go with in typical scenarios. While doing that, the developer will need to understand the specifics of quality assurance and the tools on other layers.

Retrieval augmentation

Solution architects and developers may decide to implement retrieval-augmented generation (RAG) approach in which they will create a retriever module which will retrieve the relevant documents (or passages) for a particular query from a large corpus of text. Then, they will feed these retrieved documents to the language model together with the initial prompt by reducing the hallucinations and improving some other quality characteristics.

Requesting gen AI to cite sources

Developers may add a request to cite sources which will force the model to rely on facts available in information sources. The relevance of this method however depends on application area since the source is not always available. For legal reasons, this might be very beneficial (besides training the model only on verified sources) since decision making based on made-up cases might have fatal consequences.

Fact verification methods

In this case, the evidence is retrieved from an external database to assess the veracity of a claim. The inference costs required for verification should be considered and weighted against the risks.

Gen AI infrastructure and information security governance

Gen AI introduces new security risks, and they won't be mitigated by traditional cybersecurity methods. With that, the security team will need to handle security of two types of components:

- Security of underlying infrastructure
- Security of the AI system

While the first part is familiar to most security professionals, the second part requires new ways of protection with new approaches, new governance elements and frameworks. These are some new types of risks to be addressed:

- AI behavioral vulnerabilities; the attackers may try to bypass expected AI behavior or make AI systems perform unexpected jobs
- Additional legal and regulatory risks, including copyright and ownership-related risks
- Risks in content being generated by gen AI; for instance, insecure code generation
- Social, including bias, discrimination, reputation and ethics

This requires reassessment of risks and their impacts to organizations, development of gen AI best practices and frameworks, actualization of security awareness trainings. We believe, a gen AI-specific policy is also required.

Keep in mind, the technology and its understanding are still in the initial phases and the full roster of risks is still unknown, and that continuous learning and reassessment is required.

Business process quality and security assurance

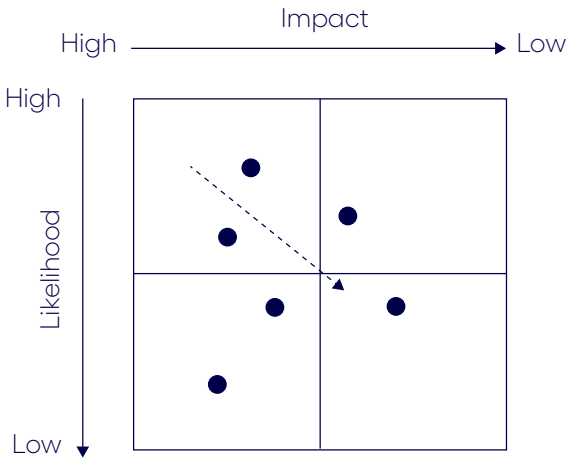
Countries and country unions are actively working on their own AI regulation. For instance, on June 14, 2023, the European Parliament adopted the latest draft of EU's Artificial Intelligence Act. Gen AI solution vendors and organizations will need to follow these to conform and add another level of trust. In general, regulators are following a risk-based, principle-driven approach that can be summarized as:

- Users should be made aware when they are interacting with AI
- Generating illegal content is not allowed
- Copyrighted data used for training should be respected

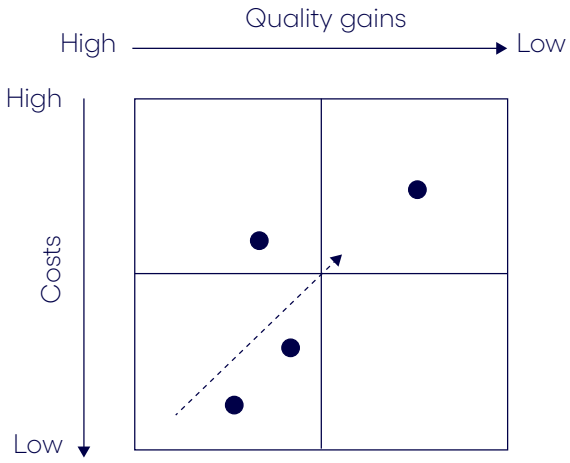
EU also proposes to ban any attempts to use AI to manipulate people, groups, do social scoring or biometric identification.

Combined risks mitigation approach

Since there is no method that would address and mitigate 100% of the challenges of gen AI, we need to follow a certain approach for cost-effective quality assurance. We spoke about an up-front categorization of gen AI-based system risks and we now understand what kind of mitigations may be applied. With that we can distribute the risks according to their likelihood, and impact and distribute the mitigation measures according to their costs and potential quality gains, as in the pictures below.



Risks with likelihood and potential impact—start from highly likely risks with potentially big impact.



Risks mitigation measures—start from low-cost measures with potentially high-quality gains.

Now, in our risk management process, we will address the risks and implement mitigations until the desired level of quality as per preselected metrics is reached.

Required gen AI-specific quality assurance capabilities

Not only businesses, IT and security departments are disrupted by gen AI as we can see that quality and security assurance of organizations need to keep up with changes and bring new capabilities. Some of the traditional capabilities are still required for gen AI, but they are not enough and new capabilities are required:

Area	*Traditional capabilities required*	*Additional gen AI-related capabilities required*
Data	• *Data testing* • *Manual test data creation* • *Automated test data generation*	• *Gen AI training data quality assurance* • *Gen AI fine-tuning data quality assurance* • *Prompt injection data quality assurance*
Gen AI model	• *N/A*	• *Truthfulness benchmarking Q&A repository* • *Red teaming benchmarking repository* • *Response sampling to detect hallucination* • *Small model testing and extrapolation techniques* • *Gen AI red teaming techniques*
Deployer/ governance system	• *Integration testing* • *Performance testing*	• *Repository of questions for adversarial testing* • *Content filtering quality assurance* • *Meta-prompt quality assurance* • *AI behavioral vulnerabilities testing (jailbreaking)* • *Augmented retrieval data sources quality assurance*
Application / user experience	• *E2E and other types of manual functional tests* • *Robotic or AI-infused test automation* • *Cross-channel testing* • *Integrated channel experience*	• *AI trust-level assurance* • *Application specific risk-level assurance (legal, healthcare, etc.)*
Business processes	• *Scenario repository of business processes* • *User journey testing* • *Regulatory compliance assurance*	• *Gen AI regulations* • *Company-specific norms, ethics, compliance and security (e.g., abuse monitoring)* • *Copyright and ownership related*
Security	• *Security tests* • *Infra-as-a-code artifacts quality assurance*	• *Gen AI-generated code security* • *Data Protection Impact Assessment (DPIA), Privacy by Design (PbD) assessments and balancing tests completed at the development stage by a compliance team*

Human centricity is key

Irrespective of the risk management frameworks and techniques, human centricity in gen AI implementation and adoption is key.

According to the European Union's proposal on gen AI regulation, for high-risk AI systems, human oversight throughout the AI systems' lifecycle is strictly necessary to mitigate the risks to fundamental rights and safety posed by AI. It means that human involvement is required in all stages of the risks management processes we have described above.

Also, when it comes to adoption, one of the key elements of building trust and gen AI success within an organization is transparency of advantages, limitations, and safety to the users. After risks are addressed and mitigated, quality tests are successfully completed. We recommend piloting the new system with limited number of key users for one to three months before it is made available to a broader user base. After that, trainings need to be offered to a broader user base which would explain:

- How gen AI helps in generating new content and be more productive
- How to create effective and safe prompts
- How to review the outcome and handle the new content in the context of their work inside and outside the organization

Users are also key in improving the quality of gen AI solutions with time. This is true with the quality of knowledge they will continue bringing to the system, and the assistance in evaluating the system and cooperating with the implementation team with new feedback and ideas.

Other elements of gen AI program success

Risk-based quality assurance of gen AI solutions which we have explored is one of the key elements of gen AI success in organizations and it goes though all stages of gen AI implementation lifecycle. In the series of gen AI-related whitepapers, we will explore other elements of gen AI success, including:

- Gen AI culture implementation
- Gen AI business strategy implementation
- Gen AI business value management
- Gen AI operating model implementation
- Building gen AI capabilities

We will also explore more on gen AI security and data privacy risks management, which is a very broad and important topic, discussing all the gen AI quality assurance layers that deserve a separate article.

Conclusion

Gen AI is still in the early maturity stage and its application in corporate environment comes with number of challenges. All parties (model owners, deployers, service providers, organizations and developers) address these challenges in their own ways which require a multilayered approach for quality assurance and risks mitigation with a wide variety of options. It will take some time before all regulations are in place and the industry standards and best practices are established. However, not acting is not an option due to potential competitive disadvantages of not using the benefits of gen AI in the long run. With that, organizations must evolve and act now to understand its benefits, as well as potential risks and ways of dealing with them. New policies, practices, frameworks and data quality improvement programs need to be implemented while user experience and building trust should remain in focus. We should also keep in mind things are changing very fast in this area and organizations should revisit their approach towards gen AI frequently over the next few years.

Authors

Dr. Ulrich Faisst,
Chief Technology Officer
Central Europe
(Ulrich.Faisst@cognizant.com)

Dr. Almir Mardanov,
Lead Enterprise Architect
Central Europe
(Almir.Mardanov@cognizant.com)