NTT DATA

Designed
with the power
of **GenAI**

Not all hallucinations are bad: acknowledging GenAI’s constraints and benefits

The AI daze

From crafting email responses to recommending a fitness routine and generating computer code, the list of what GenAI can do is seemingly endless.

Imagine telling a GenAI model to "draw a cool scene where dragons are flying over New York, leaving shiny trails behind them". The model creates a fantastic picture of dragons soaring over the city, adding a touch of magic to the skyline.

This ability to turn unusual ideas into "real" content in a matter of seconds is what makes GenAI so intriguing. But it also raises questions about the reliability of these advanced large language models (LLMs).

Experts are working to reduce the occurrence of hallucination, which is remarkably extensive at present. However, there is a parallel narrative: this creative aspect of GenAI also has the potential to foster innovation and open doors to new, previously unthought-of, possibilities.

The challenge for AI researchers is to find the right balance between responsible AI use and fostering creativity that enriches our world.

> "Hallucinations are completely fabricated outputs from large language models. Even though they represent completely made-up facts, the LLM output presents them with confidence and authority[1].

Gen AI hallucination in the headlines[2]

Can we trust everything that GenAI says?

Fake case law

When a lawyer relied on ChatGPT to prepare a filing on behalf of a man suing "Avianca Airlines", ChatGPT fabricated three (authentic-sounding) cases to support the argument: Martinez vs. Delta Air Lines, Zicherman vs.Korean Air Lines and Varghese vs. China Southern Airlines.

The fabrications were revealed when Avianca's lawyers approached the case's judge, saying they couldn't locate the cases cited in legal databases. As a result, the federal judge imposed a $500,000 fine on two lawyers and a law firm for submitting fictitious legal research.

Defamation

OpenAI's first defamation lawsuit involved an incorrect summary of the Second Amendment Foundation (SAF) vs. Ferguson case.

Mark Walters, a Georgia radio host, sued OpenAI after ChatGPT generated a false summary of the case. The summary incorrectly stated that Alan Gottlieb accused Walters of defrauding and embezzling funds from SAF.

[1] Tim Keary, Techopedia (3 September 2024). AI Hallucinations. Accessed on 4 September 2024: https://www.techopedia.com/definition/ai-hallucination

[2] Molly Bohannon, Forbes (8 June 2023) Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions.

Accessed on 4 September 2024: https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=5f0c9f937c7f

Generative AI: foundations and technology

GenAI acts as a creative mind for machines, empowering them to generate distinct and unique content such as images and text that closely mirrors human creation. To fully understand the phenomenon of hallucination within GenAI, we need to first understand the foundational technologies that drive these capabilities.

A key technique is the **generative adversarial network (GAN)**, which involves a dynamic interplay between a generator and a discriminator. The generator attempts to create content that resembles reality, while the discriminator distinguishes between real and generated content. This back-and-forth dynamic improves the generator's ability to create realistic outputs. The generator can sometimes push the boundaries of the ordinary and venture into the surreal. This divergence from the expected, known as "hallucination", showcases the intriguing potential of GenAI.

In addition to GANs, there are other significant techniques such as **variational autoencoders** (VAEs) and **autoregressive models**. VAEs can be compared with an artist honing their skills, learning and refining through imaginative processes. Autoregressive models resemble storytellers, carefully crafting narratives one step at a time.

A significant development in the field is the introduction of **transformer architecture**. Initially designed for natural language processing (NLP) tasks, this architecture led to revolutionary attention mechanisms that allow models to focus on different parts of input sequences, enabling parallelized processing and capturing long-range dependencies effectively.

In the context of GenAI, especially when integrated with GANs, the transformer architecture has significantly advanced the capabilities of GenAI.

Neural networks, inspired by the interconnected nodes of the human brain, learn patterns and structures from extensive data, enabling the generation of content that simulates human creativity. The learning process involves repetitive training and/or reinforcement learning from human feedback (RLHF) that combines reinforcement learning techniques, such as rewards and comparisons, with human guidance to train the model to produce and real-world examples.

The combination of these technologies, coupled with a sophisticated learning process, often pushes machines into new, imaginative territories, resulting in creative and, at times, hallucinatory or unreal outputs.

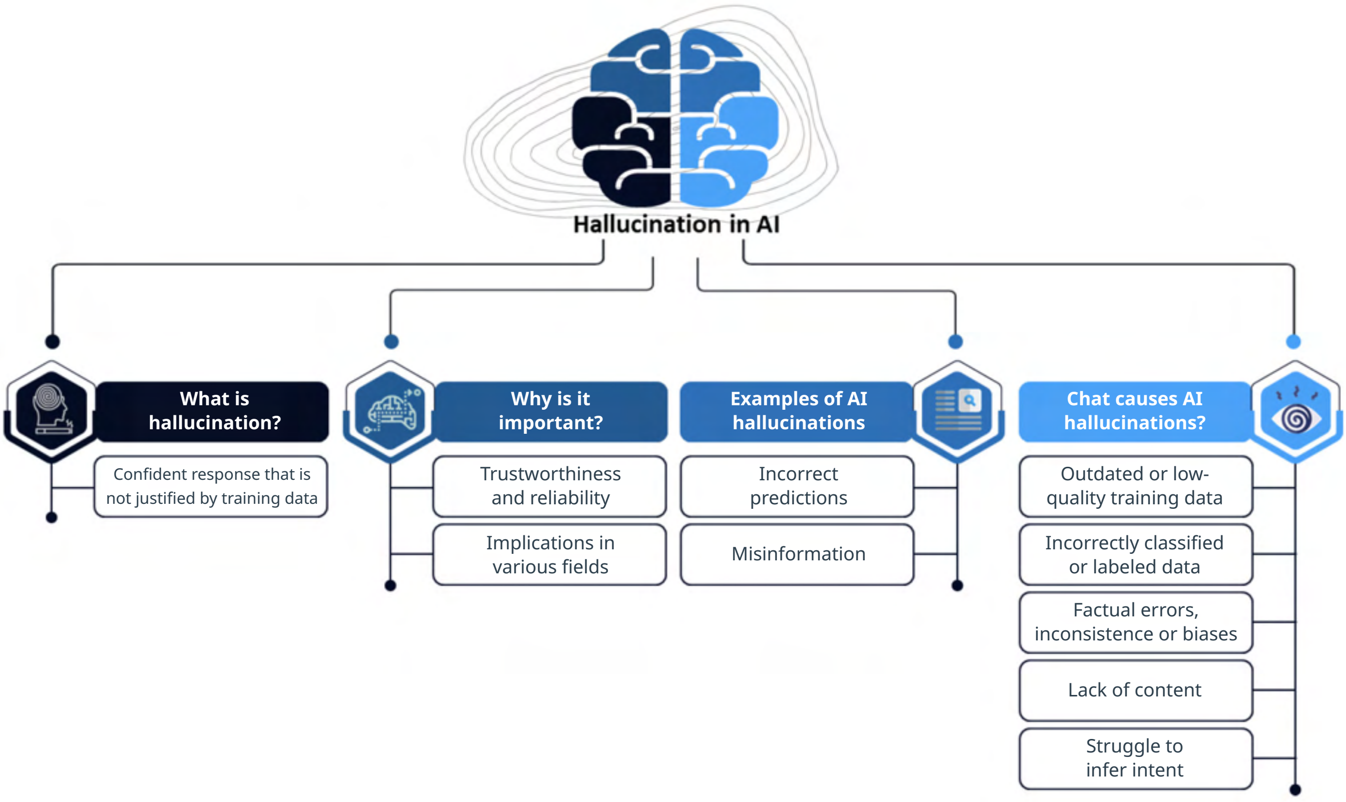
Understanding the mystery of data

In the world of AI, data plays a similar role as experiences do in the human world. It acts as both the canvas and the brush interchangeably, allowing machines to mimic and innovate, mirroring human cognition and creativity.

The term **"anthropomorphism"** is used to describe and explain hallucinations in GenAI models. Generally, hallucination refers to the AI's ability to generate content that deviates from conventional or expected outputs.

A particularly concerning hallucination is when the model generates an outcome that seems perfectly accurate but is not validated by verified facts. Even more concerning is the fact that users might accept the generated content as being accurate and inadvertently spread unverified or false information. The challenge is determining between probable, yet unsubstantiated content, and outputs that are grounded in validated facts.

> “
>
> Generative anthropomorphism – attributing human-like traits to nonhuman entities – becomes evident as AI systems learn and derive creativity from vast datasets.


Hallucination in AI
What is hallucination?
Confident response that is not justified by training data
Why is it important?
Trustworthiness and reliability
Implications in various fields
Examples of AI hallucinations
Incorrect predictions
Misinformation
Chat causes AI hallucinations?
Outdated or low-quality training data
Incorrectly classified or labeled data
Factual errors, inconsistence or biases
Lack of content
Struggle to infer intent

This imaginative capability leads to instances where AI-generated outputs blur the boundaries between the real and the imaginary:

- Content that deviates from conventional or expected outputs may sound valid but is not grounded in fact.
- Images depict mythical creatures, and texts may be crafted as narratives that defy conventional logic.

The causes of these hallucinations lie in the AI's tendency to generalize patterns from the data it has been trained on. The process of anthropomorphizing AI creativity often amplifies this tendency to generalize, giving rise to content that pushes the boundaries of reality.

GenAI hallucination can fuel innovation and unpredictability in creative processes, enriching the realm of AI-generated content. But it also presents risks, particularly when hallucinated outputs spread misinformation or perpetuate harmful stereotypes.

In addition, AI systems learn to make decisions based on their training data, which can include biased human decisions or reflect historical or social inequities, even if sensitive variables such as gender, race, or sexual orientation are not present.

Anthropomorphism helps us understand and manage the ethical and societal risks of hallucination, guiding the responsible usage and development of GenAI.

Types of GenAI hallucination

Hallucination in GenAI can take various forms, depending on the type of model and the specific task it is designed for. Common types of hallucination are described below.

Visual hallucination: In image-generation models, visual hallucination may involve the creation of images that depict objects, scenes or patterns that do not exist. These hallucinations can range from surreal and abstract art to entirely fabricated objects or creatures.

Textual hallucination: Language models may hallucinate text by generating sentences or paragraphs that contain fictional information or make false claims. Textual hallucinations can involve inventing events and "facts" that have no basis in truth.

Content-expansion hallucination: This phenomenon occurs when a model produces more information than what is present in the input data. For example, a model might add unnecessary details to an image or generate extensive narratives that go beyond the information provided.

Inference hallucination: In natural language processing tasks, inference hallucination can lead to incorrect assumptions or inferences. Large language models (LLM) may draw unwarranted conclusions from input data, leading to responses that misjudge or misrepresent the context.

Bias hallucination: Bias hallucination refers to the generation of content that mirrors or amplifies biases that are already present in the training data. This can result in outputs that display stereotypes and discrimination, or present unethical viewpoints.

Contextual hallucination: When language models generate text that seems contextually relevant but is factually incorrect or not representative of the actual context, this is known as contextual hallucination.

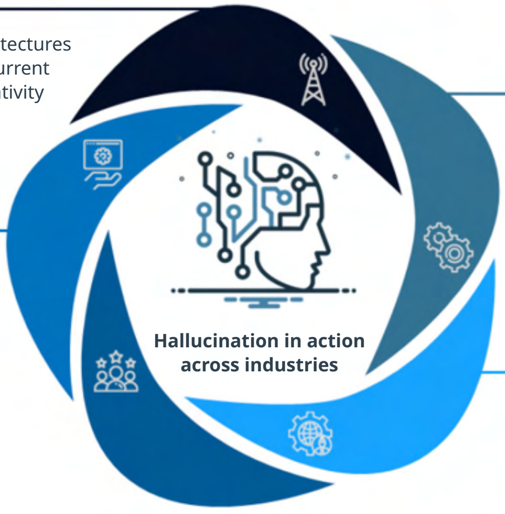
Both the type of hallucination and its effects vary among GenAI models. Mitigating and controlling these various forms of hallucination is a crucial factor in AI research and development to ensure that outputs are safe, reliable and aligned with the set objectives.

Hallucination in action in different sectors

Before exploring how hallucination influences various sectors, it's extremely important to understand the versatility of generative AI and its ability to visualize beyond conventional thinking. Just as human creativity knows no bounds, GenAI's hallucination presents a glimpse of imagination blending with technology.

Let's assume we have implemented GenAI in various sectors. The impact of the hallucination aspect is both appealing and transformative. From revolutionizing communication in the telecommunications sector to optimizing processes in business process outsourcing (BPO), hallucination metamorphizes uniquely in each sector, introducing innovative possibilities.

To effectively leverage hallucination in these sectors, we need to carefully evaluate hallucinated outputs for technical feasibility, adherence to industry standards and alignment with operational goals. Striking a balance between innovation and practicality is crucial to maximize the transformative impact of GenAI hallucination in specialized areas.



Telecommunications

GenAI may hallucinate innovative network architectures for communication protocols that are beyond current technological constrains. This hallucinatory creativity can spark disruptive ideas, but these demand rigorous evaluation before implementation.

Business process outsourcing (BPO)

GenAI in BPO could hallucinate the automation of processes and optimization of workflows. Evaluate the viability of these strategies in terms of resource efficiency, integration complexity and alignment with business goals. This will help determine which hallucinated processes can integrate effectively into BPO operations.

IT service management (ITSM)

GenAI hallucination could generate novel troubleshooting approaches or optimization strategies for IT processes. To ensure these solutions align with ITSM objectives, assess their practical implementation by considering scalability, compatibility with existing systems, and adherence to industry best practices.

Customer experience

GenAI may hallucinate unique, highly personalized customer engagement strategies based on complex behavioral models. To extract value from this hallucination process, evaluate these strategies for scalability, integration capabilities and alignment with organizational goals.

> "When GenAI is implemented in organizations and does not connect with data such as internal rules and work-related materials to generate content, it can lead to hallucinatory responses.

Transforming hallucination drawbacks into advantages

GenAI hallucinations, while often seen as a problem, can also be used positively. By understanding how and when hallucinations occur, we can develop techniques to use them for creative and innovative applications.

One use for GenAI hallucinations is to generate new ideas and concepts. For instance, by providing the model with an initial prompt and allowing it to hallucinate freely, we can create new product designs, advertisements or even movie plotlines.

Another way to leverage hallucinations is to use them to create synthetic data. Synthetic data is a form artificial data that is generated to mimic real-world data. It can be used for a variety of purposes, such as training machine learning models, testing new models and emulating or simulating complex systems. GenAI hallucination can be used to create synthetic data that is more realistic and varied than data derived from traditional techniques for generating synthetic data.

GenAI hallucinations can also be used to create new forms of art and entertainment. For example, a GenAI model could be used to create new music genres, films or video games. By allowing the model to hallucinate freely, we can create new and unique experiences that are not possible with traditional methods.

As GenAI models continue to learn and improve, we can expect to see even more innovative and creative applications of GenAI hallucination in the future.

However, because these hallucinations can be misleading and downright inaccurate, we need to use them with caution. Always verify the results of any GenAI model before using them in a real-world application.

> “Generative AI is impacting the automotive, aerospace, defense, medical, electronics and energy industries by composing entirely new materials targeting specific physical properties. […] Generative AI has already been used to design drugs for various uses within months, offering pharma significant opportunities to reduce both the costs and timeline of drug discovery.”[3]

[3] Jackie Wiles, Gartner (26 January 2023). Beyond ChatGPT: The Future of Generative AI for Enterprises. Accessed on 4 September 2024: https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises

Mitigating hallucination in GenAI

Hallucinations in GenAI, while fostering creativity, can create risks, especially when model outputs diverge too far from the intended domain. Mitigation strategies aim to ensure that the model's creative process remains aligned with human expectations and practical applications.

There are several techniques and strategies that can be used to mitigate hallucinations in GenAI models. Some of the most common include:

- **Prompting:** By providing the model with a clear and specific prompt, we can guide it to generate more accurate and realistic outputs.
- **Diverse dataset:** Training the model on a diverse and factual dataset of text and code can help it to learn a wider range of patterns and relationships. This makes it less likely to hallucinate and more likely to provide accurate data.
- **Improved GAN architecture:** Some model architectures, such as generative adversarial networks (GANs), are specifically designed to generate more realistic outputs.
- **Retrieval-augmented generation (RAG):** RAG provides the model with relevant context and information to improve the accuracy of its responses.
- **Output filtering:** The output of the model can be filtered to remove any hallucinations or inaccurate information.
- **Human in the loop:** A human evaluator can assess outputs of the model and provide critical feedback on whether the generated content aligns with real-world expectations and facts.

Research advancements and best practices

There is a growing body of research on how to mitigate hallucinations in GenAI models. Some of the most recent advancements include:

- **Hallucination detection:** Researchers have developed new algorithms to detect hallucinations in GenAI outputs and remove them before they are used in a real-world application.
- **Hallucination correction:** There are new algorithms to correct hallucinations in GenAI outputs, thereby improving the accuracy and realism of GenAI outputs.

There is no single solution that can eliminate hallucinations in GenAI. However, by using a combination of techniques and strategies, we can significantly reduce the risk of hallucinations and improve the accuracy and realism of GenAI outputs.

OpenAI's potential new strategy for fighting the fabrications is to train AI models to reward themselves for every correct step of reasoning they take in order to arrive at an answer, instead of just rewarding a correct conclusion. This approach is called "process supervision," as opposed to "outcome supervision," and could lead to better explainable AI.[4]

3 best practices for mitigating hallucinations in GenAI

1. **Use a variety of mitigation techniques**
There is no single technique that can eliminate hallucinations in GenAI. The best approach is to use a variety of techniques in combination.

2. **Monitor the model's output**
It is important to monitor the model's output regularly for signs of hallucination. This can be done manually or by using an automated hallucination-detection algorithm.

3. **Be aware of the limitations of GenAI**
GenAI models are still under development and are not perfect. It is important to be aware of the limitations of these models and to use them with caution.

[4] Marvie Basilan, International Business Times (6 January 2023). OpenAI Announces New Approach To Fight AI 'Hallucinations' After Legal Fabrications Accessed on 4 September 2024:

Ethical and societal implications of hallucination in AI

Although hallucination in AI models has a great deal of creative potential, it also requires a thoughtful examination of the ethical considerations and societal impacts that accompany this technology.

Ethical considerations

Truth and falsity

The creation of content that may not align with reality raises ethical concerns. AI-generated hallucinations can blur the lines between fact and fiction, potentially leading to the dissemination of false information, which in turn has consequences for informed decision-making and trust.

Bias and discrimination

Hallucinations produced by AI models may amplify biases present in training data. This raises concerns about perpetuating stereotypes, reinforcing societal prejudices, and exacerbating issues related to bias and discrimination.

Privacy and consent

Generating content, especially in personal contexts, can infringe upon individuals' privacy and consent. Issues relating to consent, data usage and potential harm to individuals featured in hallucinated content must be addressed.

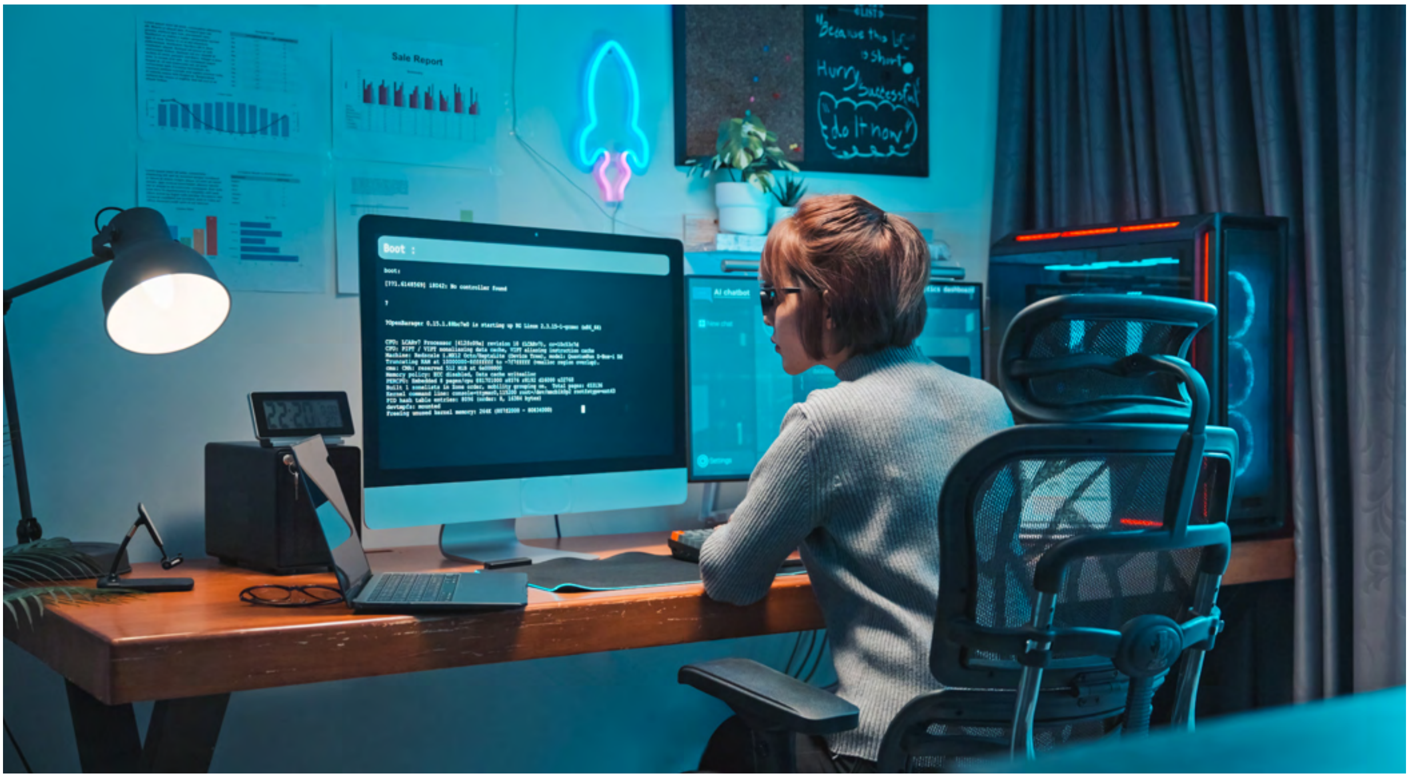
Potential impacts on society

Misinformation and/or disinformation

Hallucinations can become vehicles for the spread of misinformation and disinformation, impacting public discourse and trust in information sources. It may require active measures to combat the dissemination of false or misleading content.

Security concerns

AI-generated hallucinations can be exploited for malicious purposes, including deepfakes or deceptive content used in cybersecurity attacks or fraud.



Guidelines for the responsible use of GenAI

Transparency and accountability
Developers and users of AI systems must prioritize transparency in the generation process and be accountable for the content produced. Clearly indicating that content is AI-generated can help mitigate the risk of misinformation.

Bias mitigation
Implementing bias-reduction techniques and actively working to debias training data are primary steps in mitigating the amplification of biases in hallucinated content.

Ethical guidelines
Framing and adhering to ethical guidelines for AI model development and usage can set standards for responsible AI deployment and content generation.

As we continue to explore and understand AI hallucination, it is crucial to remain focused on these ethical and societal considerations. Responsible development, usage and guidelines are essential to maximizing the benefits of this technology while minimizing its negative effects.

How NTT DATA is charting the future with GenAI

Developments in GenAI and the phenomenon of hallucination bring us to the crossroads of immense potential and profound responsibility.

The creative capacities of GenAI, exemplified by hallucination, open doors to unprecedented possibilities in numerous sectors. From telecommunications to customer service, this technology can reshape how we perceive and interact with each other and the digital world.

However, the generation of content that blurs the lines between fact and fiction necessitates an ethical framework that values transparency, accountability and responsible use, and helps to mitigate bias, combat misinformation and safeguard individual privacy.

Responsible development and adherence to ethical guidelines are central to harnessing the power of AI’s creativity while mitigating its potential pitfalls. IT industry leaders such as NTT DATA have a great responsibility to chart the way forward, guided by innovative practices, transparency, collaboration and a commitment to responsible AI use.

LITRON

NTT DATA’s LITRON® Generative Assistant is an AI service that generates responses by securely connecting various data types like internal rules, work-related materials and external data with generative AI.

Coding by NTT DATA

Coding by NTT DATA is a cutting-edge platform that transforms the way custom code is created and modernizes legacy applications.

Let's get started

See what NTT DATA can do for you.

- Deep industry expertise and market-leading technologies
- Tailored capabilities with your objectives in mind
- Partnerships to help you build and realize your vision.

Visit nttdata.com to learn more

NTT DATA is a trusted global innovator of business and technology services, helping clients innovate, optimize and transform for success. As a Global Top Employer, we have diverse experts in more than 50 countries and a robust partner ecosystem. NTT DATA is part of NTT Group.

NTT DATA