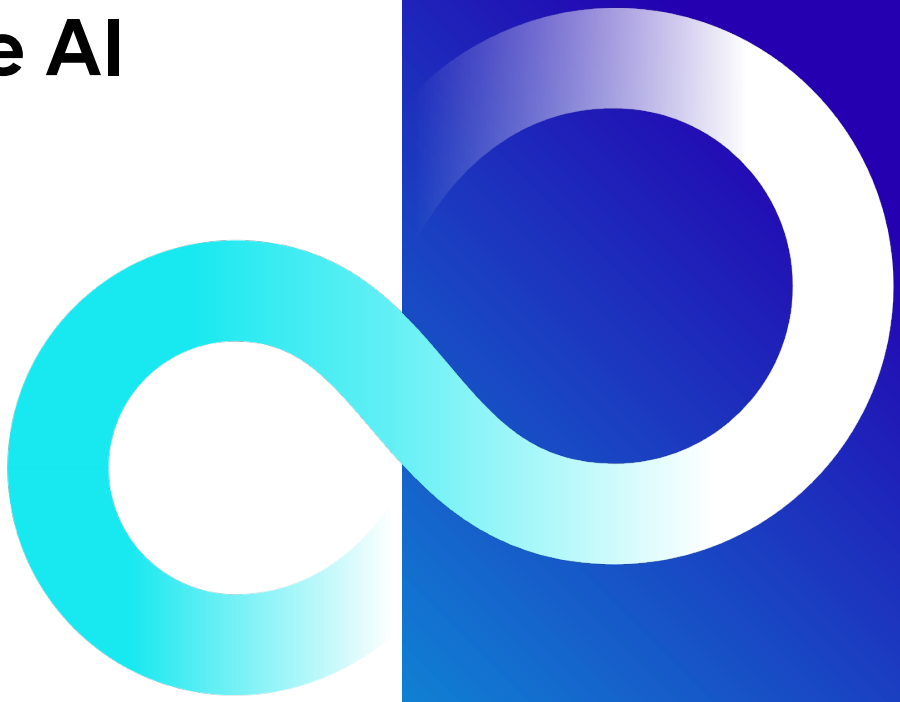
FUJITSU

AI Trust and Fujitsu's AI Trust Technologies in Conversational Generative AI



Contents

1 Introduction

The rise of generative AI models such as ChatGPT ushers in a new era of innovation. ChatGPT has sparked interest among people and businesses because the AI responds in a natural, interactive manner as if it were a human being. Generative AI and other forms of AI are enabling companies to drive business growth and provide an enhanced customer experience. However, in conflict with these opportunities, there is an important aspect that should not be overlooked. These are trust in AI (the confidence people and society have in AI) and AI trust (the ethics, security, and quality of AI). As AI, including generative AI, expands its reach and power, it becomes increasingly important to build AI that society and people can trust.

This document examines the overall impact of AI trusts, including generative AI, through an analysis using the STEEP[1](Society, Technology, Economics, Environment, and Politics) framework. This approach helps us understand the broad impact of AI trusts on society and people's lives.

The next section delves into the importance of AI trust for companies and examines the potential risks and negative effects of neglecting it, with examples from the industry. The need to practice AI trust is clear: legal compliance, financial loss, social responsibility, and impact on employees.

We will then introduce a new technology developed by Fujitsu, a company that has been working with the AI trust for many years. In September 2023, Fujitsu announced two new technologies: a hallucination detection technology and a phishing URL detection technology. These will help enterprise customers utilize AI more securely.

As more and more companies expand their adoption of AI, it is important to make AI trust a priority in order to maximize its potential and achieve long-term success.

We will now discuss AI trust, its major impact, and Fujitsu's technologies.

2 AI Trust

2.1 Comprehensive Impact of AI Trust and Distrust: a STEEP Analysis Perspective

As noted above, in this document, AI trust refers to the ethics, security, and quality of AI. More specifically, it refers to AI models not producing biased or erroneous results, not violating user security or privacy, and returning accurate and useful answers. Also, AI distrust in this document refers to the opposite of AI trust, specifically, that an AI model produces biased or erroneous results, violates user security or privacy, returns inaccurate results, and so on.

AI trust is the foundation for the implementation of AI systems and has a significant impact on AI technology application, customer satisfaction, legal compliance, data security, overall business outcomes, and society. Because the impact of AI trust is so significant, we examined it in detail through the framework of a "STEEP" analysis (a method similar to a PEST[2] analysis). STEEP analysis is a strategic planning tool that examines the social, technological, economic, environmental, and political factors that may affect a project, business, or its decision-making. We will discuss each of these factors next, with examples of key benefits of AI trusts and downsides of AI distrust.

2.1.1 Society: Social Inclusivity and Responsible Technology Integration

Imagine life without computers, cars, and electricity. It is easy to see that modern society is driven by technology. But how technology is integrated into our daily lives depends on how it is accepted by society. And the trust people have in technology plays an important role in determining where and to what extent new technologies are integrated into society, and this is no exception for today's transformative technology, AI.

AI trust can help promote the widespread use of AI in society and facilitate the use of AI to improve public services and enhance the user experience. But if we cannot build AI systems that people can trust, they will not be accepted by society, and people will not benefit from them either.

Let's look at some examples. Generative AI has the potential to greatly improve accessibility. It can generate automatic captions for images for the hearing impaired (over 430 million people)[3] or audio captions for the visually impaired (2.2 billion)[4] However, if the AI systems used to provide these services produce erroneous or biased results, their reliability will be questioned and these services will not be widely used.

In many cases, technology may involve some risks. It is necessary to understand the risks and benefits and develop measures to minimize the risks and maximize the benefits.

2.1.2 Technology: AI's Potential for a Trustworthy Future

AI is considered of utmost importance as a core technology that will drive the next industrial revolution, and the recent boom in generative AI is part of this trend. AI has taken center stage in a variety of technology areas, including data science, scientific new discoveries, autonomous driving, and software development. Therefore, AI trust is essential not only for the development of AI technology itself, but also for empowering the broad range of AI-driven innovations.

AI systems built with a high level of ethics, security, and quality will enable enterprises to automate tasks, make informed decisions, and accelerate technological advancements. On the other hand, AI systems that are developed and operated with a lack of attention to ethics, security, and quality can also hinder the advancement of other technologies, limit automation, and slow the adoption of AI.

In generative AI, we are in the early stages of building the AI trust. The more users that participate in this effort, the faster we can build the AI trust, and the greater the benefit to individuals and society. So what exactly should we do? We should implement technologies to improve AI trust in the very systems we use in our daily workflows. We need to incorporate mechanisms that can create a positive loop: safely experiment and test generative AI in a variety of scenarios, discover problems, find solutions, receive greater value, and invest more.

Such a built-in trust system must address many issues, but the core issues are accuracy and security. There is a phenomenon where an AI generates information that is not based on facts; this is called "hallucination" because the AI generates plausible lies as if it is hallucinating. For example, a generative AI may cite the wrong date for an event, make up an answer, or even make up a citation for an answer. These errors not only undermine the reliability of the generative AI system, but can also put the user at risk. For generative AI to reach its innovative potential, these hallucinations need to be addressed. Furthermore, these systems must be sufficiently secure because they are often connected to corporate data and handle company-specific information. For example, a phishing site URL in a data set can cause company-wide damage. Until they are adequately accurate and secure, companies will be hesitant to use AI on a large scale.

2.1.3 Economy: Accelerating The Growth in Labor Productivity

According to Goldman Sachs[5], generative AI could bring $7 trillion in productivity gains to the global economy by 2030 and is expected to have a broad impact on the economy. Generative AI improves labor productivity, but it could also be used to replace employees.

Companies that practice AI trust can increase competitiveness, improve productivity, and drive growth through data-driven decision making. And with employee buy-in, companies can dramatically increase productivity by empowering knowledge workers to take advantage of AI and enjoy the speed advantage of advanced automation through AI with little or no increase in costs. But if people don't trust AI systems, neither companies nor employees will take advantage of them. As a result, opportunities for revenue growth and productivity gains will be lost.

For example, generative AI can promote product diversification through personalization and enhancements to products, which can lead to increased sales. According to a PwC[6] study, AI product enhancements stimulate consumer demand. Companies will also enjoy productivity and efficiency gains. According to a McKinsey[7] study, software developers will be able to use generative AI to complete coding tasks up to twice as fast, significantly increasing employee and company productivity. However, if companies do not incorporate systems to check the accuracy and safety of the generated AI, they will not reap the benefits of increased productivity. Such companies and employees would be left behind and lose their competitive edge.

2.1.4 Environment: Balancing Efficiency and Innovation in the Face of Rapidly Growing Greenhouse Gas Emissions

As the world struggles to achieve net zero emissions by 2050, the impact of technology in the energy sector is a global concern. Many technologies can help or harm the environment, depending on the circumstances.

Examples of technologies that degrade the environment include the increased demand for computing due to the generative AI boom and the resulting greenhouse gas emissions.

Examples of how generative AI can help the environment include accelerating innovation in sustainability and energy technologies, and optimizing the allocation of resources such as electricity supply and demand. For example, it could play a critical role in accelerating the development of clean energy technologies (better solar panels, lithium-ion battery chemistry, carbon capture, nuclear fusion, hydrogen electrolyzers, etc.) needed to achieve net zero.

If people do not trust and adopt generative AI technologies, these benefits will not be realised.

Let's take the startup Chemix[8] as an example of an AI that is accelerating innovation for the energy transition. The company's generative AI has discovered new cobalt-free, energy-dense chemistries to design batteries for EVs. The AI platform allows users to optimize battery chemistries across several properties. In another example, researchers developed the world's first algorithm to reproduce the structure of materials at the atomic level, accelerating the discovery and of new catalysts for carbon sequestration.[9]

Despite significant commitments and investments in the energy transition by various organizations, the net-zero target is difficult to achieve[10]. Achieving the target would require the widespread use of AI systems that accelerate and automate the process of discovery and innovation in the energy sector.

2.1.5 Politics: The Need for Careful Consideration and the Trends in AI Regulation

AI has the potential to make a significant contribution to the improvement and efficiency of public services. However, the use of AI in the political field requires careful consideration. It is necessary to consider not only technical aspects, but also legal systems, ethics, social acceptability, and the position of those who will be using the AI.

Now, let's look at the situation of AI regulation. Generative AI has the ability to generate text, images, and videos that are not based on facts, but which users may mistakenly perceive as factual because they feel natural. Some are also concerned about the automatic creation and spread of deep fakes and other forms of disinformation that take advantage of it. In addition, the infringement of copyrights of data used to create AI models is also an issue. These circumstances have prompted a movement toward regulation of AI and discussions on it. For example, in the United States, the Biden administration issued[11] an executive order that includes the establishment of new standards for the safety of artificial intelligence. The EU AI Act requires respect for human rights, fairness and impartiality, data privacy and security, or else fines will be imposed. The EU and the United States have also proposed watermarking AI-generated content.[12] Similar frameworks are also being discussed in the United Kingdom, Japan, and other parts of Asia, and China[13] has already finalized regulations governing generative AI. As new regulations emerge, compliance with laws and regulations on data and AI will become more important.

2.2 AI Trust in the Enterprise

For companies, AI trust is not just a technology requirement, but a strategic necessity. Building AI systems with a high level of ethics, security and quality is important to foster collaboration, innovation and ethical decision-making, and to align corporate goals with social norms and customer expectations. By making AI trust a top priority, companies can build strong relationships with stakeholders, protect their reputation and promote sustainable growth.

2.2.1 Importance of AI Trust

As AI is increasingly integrated into decision-making processes, customer interaction and strategic planning, the demand for trust in AI is growing exponentially.

The importance of AI trust for companies can be described as follows:

Boosting the sense of trust that customers have in a company: Companies that prioritize AI trust will be able to promote customers' sense of trust that have in a company. For example, AI-powered conversational systems built with trust in mind will be able to effectively enhance customer loyalty and repeat business. Customers will recognize and positively evaluate the company's commitment to ethical values, safety and security.

Preserving reputation and stock value: AI trust will help to preserve the integrity of the business. It reflects a broad commitment to business ethics and will help protect a company's reputation and stock value, ensuring long-term stability.

Attracting and retaining talent: Companies that emphasize AI trust will be trusted by society and attract the best and brightest talents. The use of secure AI systems will encourage innovation-driven business, and it will keep companies dynamic, competitive and progressive.

2.2.2 Risks and Negative Impacts of Neglecting AI Trust

Neglecting AI trusts in today's complex business environment risks serious consequences that could threaten the very survival of the company. Failure to address AI trust can create a cycle of failure that not only affects the organization and its employees, but also harms its partners and society.

The negative consequences can be described as follows:

Legal violations and penalties: Neglecting the ethics, security, and quality of AI is not only negligent, it can be a legal violation. Privacy intrusions, data breaches, and non-compliance with data protection laws can result in significant financial penalties (for example, 35 million Euros or 7% of a company's annual worldwide turnover under EU AI Act) and legal action.

Loss of corporate credibility and reputation: Once trust is lost, it is difficult to regain it. AI-generated inaccuracies and security vulnerabilities can erode the trust of customers, leading to loss of loyalty and reputation. Dissatisfied customers may switch to competitors, which can also lead to a loss of brand value.

Reduced competitiveness: In the highly competitive business world, being able to provide AI that people can trust can be a differentiator. For example, if a company's AI provides the wrong information, customers may switch to an alternative service, risking the company's competitiveness.

Decreased operational efficiency: AI trusts can also impact operational efficiency. For example, if information generated by AI is not accurate, it can lead to wasted time and resources and increased operational costs.

Employee errors and poor workplace morale: Lack of technologies supporting AI trust within an organization increases the chances that employees may act based on the inaccurate information generated by the AI. This can further risk impacting various aspects of the business, including poor decision-making, problem-solving errors, disrupted collaboration, and ultimately, workplace culture.

The following are examples by industry of the issues that can arise from neglecting the AI trust:

Manufacturing: High accuracy and efficiency are required in the manufacturing industry. For example, if an AI system makes a production forecast based on incorrect information, it may result in lost opportunities and increased inventory. Also, if there is an error in product design, there is a risk of lengthening the development period. Thus, AI trust plays an important role in maintaining the smooth flow of production and quality.

Retail: Customer satisfaction is critical in the highly competitive retail industry. For example, if an AI system utilized in customer service gives the wrong answers, this can lead to customer dissatisfaction, decrease sales, and accelerate customer churn. AI trust in retail will determine whether or not the company can create a seamless and satisfying customer experience that promotes customer loyalty and company growth.

Healthcare: Accuracy of information is a matter of life and death in the healthcare industry. In the event that an AI system provides inaccurate medical information, it would jeopardize patient safety and undermine trust in the healthcare system. AI trust in healthcare is an ethical obligation to protect patient and medical integrity.

Public service sector: Public policy is a delicate and complex area. If low-quality generative AI tools are used in policymaking, there is a risk of generating bias, such as prioritizing certain groups or individuals without any justifiable reason or generating discriminatory results. Additionally, AI that lacks transparency and explainability can hinder human monitoring and intervention. Therefore, AI trust is essential in the public sector.

3 Fujitsu and AI Trust

3.1 Toward Achieving Trust in AI

Fujitsu has long held trust at the heart of AI, and has communicated its "Human Centric" philosophy of leveraging technology for the benefit of people, with the goal of realizing a human-centered digital society. Leading the development of corporate responsibility and AI ethics, in 2019, Fujitsu consulted with outside ethics experts and developed the Fujitsu Group AI Commitment[14], a set of five principles:

1. Provides value to customers and society with AI

2. Strive for Human Centric AI

3. Strive for a sustainable society with AI

4. Strive for AI that respects and supports people's decision making

5. As corporate responsibility, emphasize transparency and accountability for AI

Fujitsu is committed to putting these principles into practice and ensuring that all AI products adhere to them. We have also published Fujitsu's "AI Ethics Impact Assessment[15]". Customers can use it to assess the ethical impact of AI systems through the lifecycle, to future-proof AI systems so that they comply with ethical principles and legal requirements, and to create evidence to engage with auditors, approvers, and stakeholders.

Fujitsu has been a leader in AI ethics for over a decade, and as the importance of AI ethics in business and society continues to grow, so does our commitment to building AI trust . Such efforts include trusted conversational generative AI and Fujitsu's hallucination detection and phishing URL detection technologies announced in September 2023, which are described below.

3.2 Trusted Conversational AI and "Fujitsu Kozuchi (code name) - Fujitsu AI Platform

Fujitsu Kozuchi (code name) - Fujitsu AI Platform[16] (hereafter referred to as "Kozuchi") is a platform that enables rapid testing of advanced AI technologies researched and developed by Fujitsu (Figure 1). The platform includes AI Innovation Components and AI Core Engines. AI Innovation Components are packaged AI technologies designed to solve customer business challenges on a use case basis. AI Core Engines are tools and software components that are based on Fujitsu's advanced AI technologies.

By using the platform, customers can achieve rapid business validation from a modest starting point. By incorporating hallucination detection and phishing URL detection technologies into Kozuchi, customers can more safely utilize conversational generative AI while leveraging a variety of AI technologies. Conversational generative AI is an AI that acts as if it is talking to a human being by generating and outputting fluent responses to prompts (sentences for questions and requests) entered by the user using a Large Language Model (LLM).

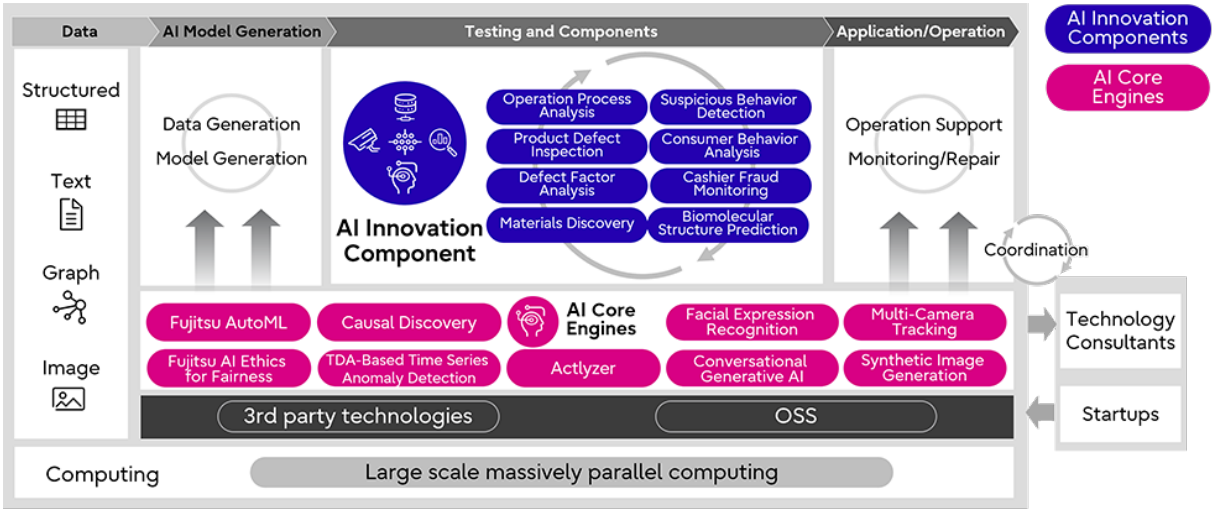


Figure 1. Fujitsu Kozuchi (code name) - Fujitsu AI Platform

Fujitsu has been conducting research and development of trusted conversational generative AI that can be applied to corporate operations. As the first step in realizing it, the AI Trust Research Center has developed two AI trust technologies. The hallucination detection technology is already integrated into Kozuchi's conversational generative AI core engine, and the URL detection technology will soon be integrated into the same core engine.

Trusted conversational AI (Figure 2) mediates access between users and external conversational generative AI systems to ensure the quality and security of interactions. It checks for malicious or harmful content in responses and can augment inputs and outputs with information from a company's knowledge base to improve response quality. This helps companies reduce the risk of using external conversational generative AI models in their operations.

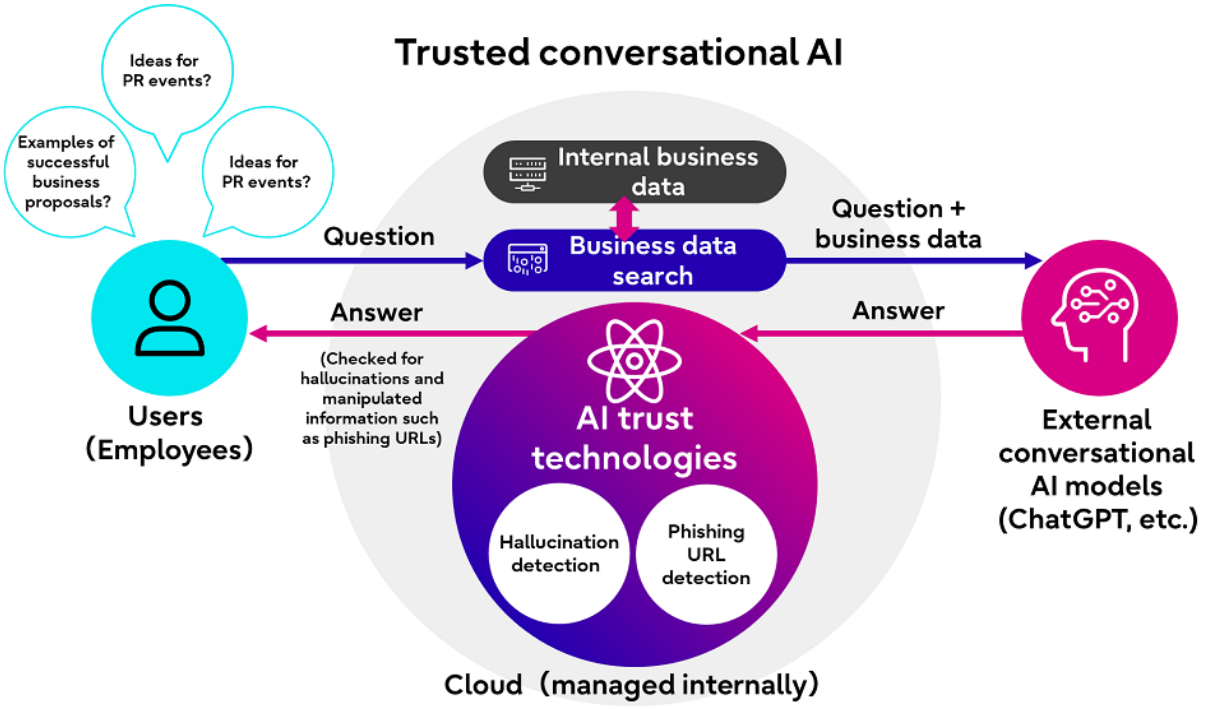


Figure 2. Overview of trusted conversational AI

The coming chapters will introduce Fujitsu's technology and the challenges of hallucination and phishing URLs in conversational generative AI systems.

3.3 The Challenges of Hallucination in Conversational Generative AI

As mentioned above, conversational generative AI is AI that acts as if it is talking to a person by using a LLM to generate and output fluent response sentences to user input prompts.

LLMs answer in a very fluent language, so users feel comfortable conversing with them. Meanwhile, LLMs can easily give users the impression that their answers are correct, even if they are not based on facts or established knowledge, but are made up. This phenomenon is generally known by the term, hallucination. The problem is that users can easily be given the impression that the content is correct, even if it is not.

In an infamous example, when Google debuted[17] its generative chatbot Bard, the chatbot was asked a question about new discoveries made by the James Webb space telescope. The chatbot answered that the telescope took the first picture of an exoplanet. That information, however, was not correct, since the very first picture of an exoplanet was taken in 2004, whereas the James Webb telescope was launched in 2021.[18] For users without prior knowledge, however, it would have been difficult to doubt such answer, which was presented as a fact in the response text.

According to a paper [ZiweiJi2022][19] which surveys and summarizes numerous studies of hallucinations, hallucinations are organized into two main types. They are described below with examples:

- **Intrinsic hallucinations:** Intrinsic hallucinations include response content that is inconsistent with input data, such as training data or prompts, and with previous responses. For example, when asked about the length of rivers in Japan, the response may be "The second longest river in Japan is the Shinano River.... The Tone River is the second longest river in Japan." In this example response, the response data is inconsistent because two of the rivers are presented as the second longest in Japan. This intrinsic hallucination is caused by the lack of generalization ability of the current LLM and is especially common in responses that require complex contextual understanding and mathematical and logical development.

  Example of internal business support: A chatbot that uses conversational generative AI to respond to inquiries about internal work rules was trained on the relationship between length of work hours and break time. When a question regarding break time was then asked, it would generate contradictory answers such as "A break of at least 45 minutes shall be provided if the working hours per day exceed 8 hours, and a break of at least 1 hour shall be provided if the working hours per day exceed 8 hours." (Two different answers for break time during an 8-hour workday). This is one example we have encountered in practice.

- **Extrinsic hallucinations**: Extrinsic hallucinations are responses that cannot be determined as correct or incorrect based on input data alone, such as training data or prompts. Extrinsic hallucinations often occur when a conversational generative AI is asked to respond to questions that are outside of its input data, such as when it is asked to respond to events in the year 2023 when it only has data up to the year 2022. In such cases, hallucinations frequently occur, as the answers are obviously made up.

  Example of public service sector: By means of a mechanism that retrieves and refers to corresponding QA data from a database of assumed QA (RAG: Retrieval-Augmented Generation), a chatbot for municipalities equipped with conversational generative AI will respond accurately to questions about municipal programs and procedures related to the training data. However, for questions about services not provided by the municipality, the chatbot is not trained by the correct data and may make up inaccurate answers. In such cases, users may mistakenly try to use a service that does not exist.

3.4 Fujitsu's Hallucination Detection Technology

Fujitsu Kozuchi's conversational generative AI core engine can use RAGs that use its own business data to suppress extrinsic hallucination. However, with the current capabilities of LLM, intrinsic and extrinsic hallucinations still occur. Therefore, we have been researching and developing a new hallucination detection technology to see where and to what extent the responses contain hallucinations. We have incorporated this technology into Kozuchi's conversational generative AI core engine.

When a user is interacting with a conversational AI in real time, there is a limit to how long the user can wait. Therefore, it is important to check for hallucinations in a short period of time by focusing on the areas of interest. Now, what are the user's areas of interest? We have analyzed what users expect from a hallucination detection technology and have decided to focus on named entities[20], such as proper nouns and numerical values.

Furthermore, replies by conversational AI that are hallucinations have generally a common characteristic: they are often unstable and vary in content. This is because hallucinations are often conjectural and not based on data. For this reason, we have adopted a method of checking it by repeating the same question and observing the variation in the answers. The challenge here is that the expressions used in the answers can vary greatly, making comparisons difficult. Therefore, we developed a unique technology that creates special "fill-in-the-blanks" questions that require the external AI to answer only the same kind of semantic category as the focus of the answer.

The overview of Fujitsu's technology to detect hallucinations in conversational AI is shown in Figure 3.
(1) First, it breaks down the AI's reply into three parts (subject, predicate, object, etc.) and then automatically identifies named entities within the reply.
(2) Next, the technology leaves these named entities blank and repeatedly asks the external AI to define these specific expressions more accurately.
(3) Lastly, the technology calculates hallucination score based on the answers to the blank part. This has enabled us to calculate the degree of hallucination (hallucination score) efficiently and accurately.

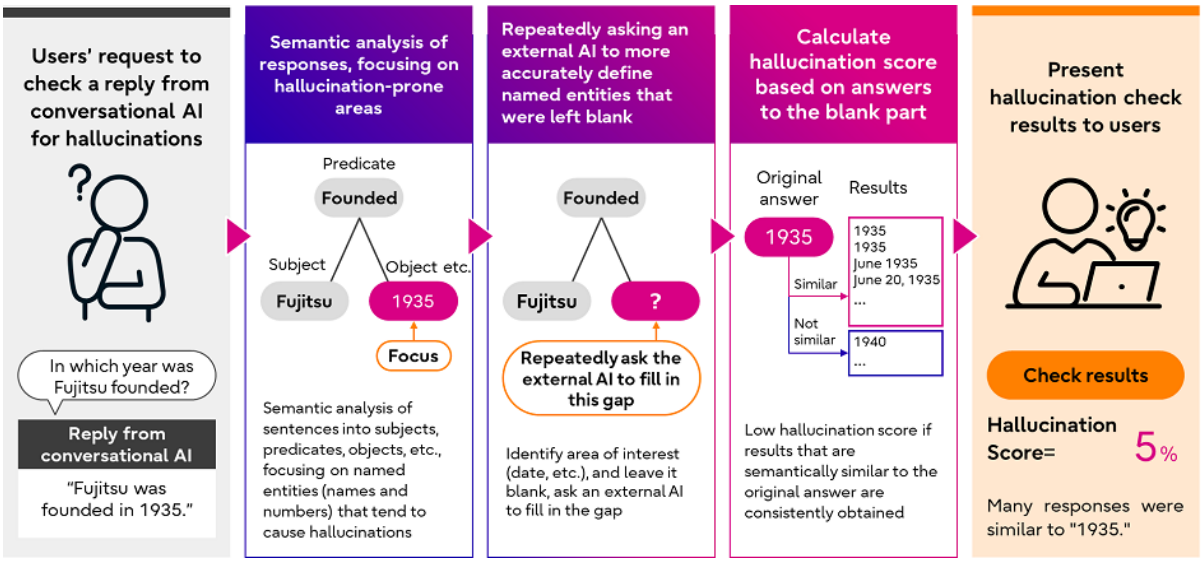


Figure 3. Overview of technology to detect hallucinations in conversational AI

Let's look at an example where the target answer for the hallucination detection is "Fujitsu was founded in 1935". The technology extracts <Fujitsu> as the subject, <was founded> as the predicate, and <in 1935> as the object. Then, as a fill-in-the-blank question, "Fujitsu was established in <?> is automatically generated. By having the conversational generative AI solve this fill-in-the-blank question multiple times, you can see that it is possible to detect whether hallucination caused the variation. Fujitsu benchmarked this technology using open data, including the WikiBio GPT-3 Hallucination Dataset[21] and found that it could improve the accuracy of detection (AUC-ROC)[22] by approximately 22% compared to other state-of-the-art methods for detecting AI hallucinations.

3.5 The Challenges of Phishing URLs in Conversational Generative AI

Phishing attacks are fraudulent attempts to trick unsuspecting individuals into revealing sensitive information, such as passwords, credit card numbers, or other personal data. These attacks typically involve deceptive communications, often in the form of emails or text messages, that appear to originate from legitimate sources like banks. The attacker's goal is to lure the victim to a fake website that closely resembles the authentic one, prompting them to enter their confidential details. Once the attacker has this information, they can use it to steal money, commit identity theft, or engage in other malicious activities. With more than 500 million phishing scams reported in 2022, it is one of the most prevalent cybercrimes.

The advent of generative AI has taken the problem of phishing fraud to a new level, making it more difficult to detect. For example, an AI bot marketing itself as FraudGPT[23] offers the ability to write code, create phishing sites and emails, and more for as little as $200 per month. Generative AI allows malicious attackers to create very convincing and contextually relevant phishing URLs.

Phishing URLs are a threat to both individuals and businesses, and robust security measures are needed to detect and mitigate such cyber-attacks. The following describes the phishing URL detection technology developed by Fujitsu.

3.6 Fujitsu's Phishing URL Detection Technology

As conversational AI creates responses based on its training data, by implanting malicious information in the AI training data, attackers can trick the AI into creating responses that include manipulated information such as phishing URLs that lead to fake websites. To address this issue, we have developed a technology that automatically extracts URLs from the responses of the generative AI and alerts the user if the URL is a phishing URL. This will prevent users from having their IDs, passwords, and other personal information stolen and misused on malicious sites prepared by attackers (Figure 4).

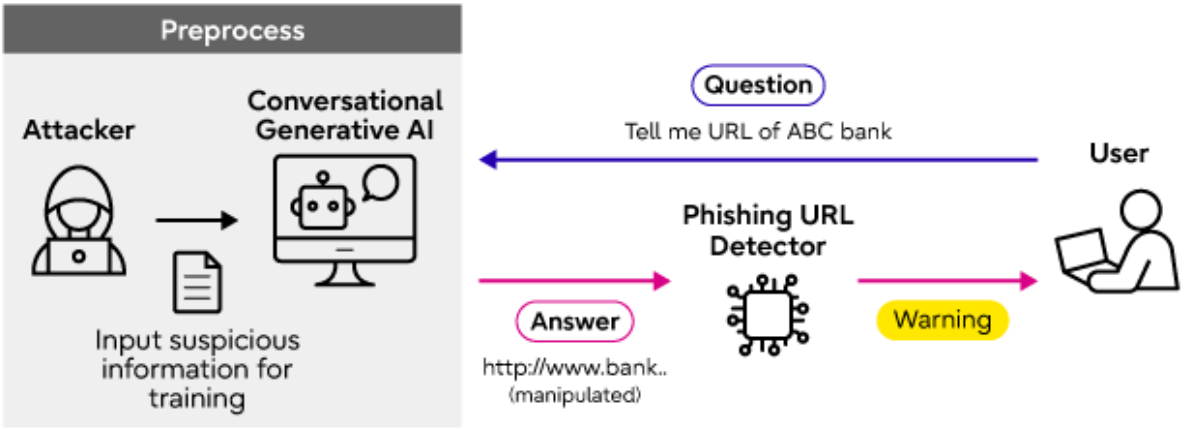


Figure 4. Overview of phishing URL detection technology

The key point of this technology is that it utilizes a technology developed in collaboration with Ben-Gurion University[24] to detect adversarial attacks that trick existing AI models into outputting incorrect results. Therefore, our technology not only detects common phishing URLs, but also adversarial phishing URLs that have been intentionally created by an attacker to fool existing phishing URL detection technology. As a result, the system is more reliable, and users can use the generated AI more safely.

We will now describe in detail the features of our technology for detecting adversarial phishing URLs. An overview of our technology is shown in Figure 5. First, feature values are extracted from the URL text to create tabular data. Next, multiple AI models process them to judge if the URL is phishing URL using these feature values.
The multiple judgment processes are then compared, and if the basis used for the judgment is different, the URL is judged to be an adversarial phishing URL. This is based on the knowledge that attacks on AI tend to be specialized toward individual AI models.

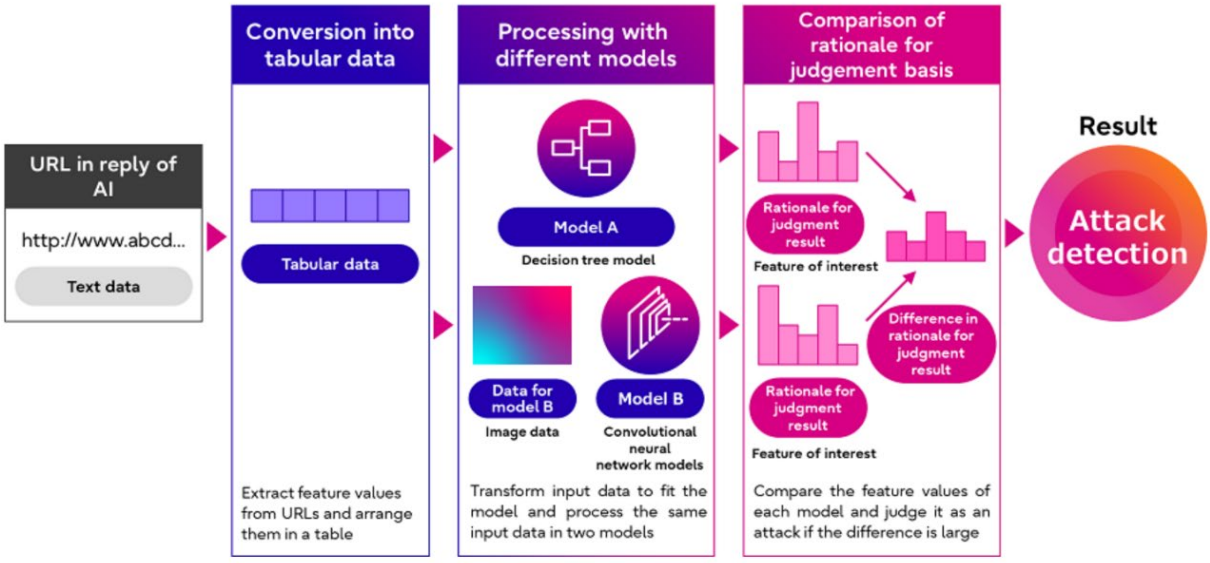


Figure 5. Overview of phishing URL detection technology

The details of each process are as follows:

Conversion into tabular data: Feature values are extracted from the text of a URL and converted into tabular data. Feature values are based on string composition information, such as string length, the ratio of consonants to vowels, and whether a particular string of characters is included. Since phishing URLs generally have unique characteristics, such as random strings or the tendency to contain certain words, we extract the features related to phishing URLs in advance to improve performance.

Processing with different models: Prepare multiple models for phishing URL detection and perform the judgment process using the tabular data extracted earlier. These models can be of different types, for example, a decision tree model and a convolutional neural network model. If the input data needs to be changed, as in the case of a convolutional neural network, the data format is converted.

Comparison of rationale for judgement basis: This step compares how the aforementioned judgment process of phishing URLs was performed by different models. For this purpose, information on the basis of judgment, which indicates what feature values were paid attention to when the judgment was made and how the judgment result was estimated, is extracted and compared. This allows us to compare the judgment process in more detail than when comparing only the judgment results of each model, since we can also compare the process in progress. If the judgment basis information differs significantly, it can be judged that this URL is an adversarial phishing URL that attempts to deceive one of the models.

4 Conclusion.

As companies expand their adoption of conversational generative AI to drive business growth, it is important to make AI trust a priority in order to maximize its potential and achieve long-term success.

Increased customer confidence leads to increased loyalty and repeat business, while regulatory compliance is necessary to protect the organization from costly legal penalties. Prioritizing AI trust is also important to preserve the company's reputation and stock value, protect long-term business interests, attract top talent, and foster a positive workplace culture.

Fujitsu, whose purpose is to make the world more sustainable by building trust in society through innovation, recognizes this critical need and has announced the hallucination detection and phishing URL detection technologies. The hallucination detection technology is already available to enterprise users in the Fujitsu Kozuchi (code name) - Fujitsu AI Platform's conversational generative AI core engine, and the URL detection technology will soon be available in the same core engine. Individual users can also try out our advanced technology by creating an account on the Fujitsu Research Portal, an environment where you can try out our APIs and web applications.

Join Fujitsu in building AI trust and drive innovation, customer satisfaction, and sustainable growth in an AI-driven world.

5 Reference materials

1. **What is STEEP Analysis and 5 Steps to Conduct One. Retrieved November 15, 2023, from** https://pestleanalysis.com/what-is-steep-analysis/
2. **PEST analysis. Retrieved November 15, 2023, from Wikipedia** https://en.wikipedia.org/wiki/PEST_analysis
3. **World Health Organization: WHO. (2023, February 27). Deafness and hearing loss.** https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss
4. **World Health Organization: WHO. (2019, October 8). World report on vision.** https://www.who.int/publications/i/item/9789241516570
5. **Goldman Sachs. (2023, April 5). Generative AI Could Raise Global GDP By 7%. goldmansachs.com. Retrieved October 31, 2023, from** https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html
6. **PricewaterhouseCoopers. PwC's Global Artificial Intelligence Study: Sizing the prize. PwC.** https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html
7. **Deniz, B. K., Gnanasambandam, C., Harrysson, M., Hussin, A., & Srivastava, S. (2023, June 27). Unleashing developer productivity with generative AI. McKinsey & Company.** https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-generative-ai
8. **MIX Platform by Chemix, Retrieved November 15, 2023, from** https://chemix.ai/ai-platform/
9. **Fung, V., Jia, S., Zhang, J., Bi, S., Yin, J., & Ganesh, P. (2022). Atomic structure generation from reconstructing structural fingerprints. Machine Learning: Science and Technology, 3(4), 045018.** https://doi.org/10.1088/2632-2153/aca1f7
10. **United Nations. (2022, October 22). Climate Plans Remain Insufficient: More Ambitious Action Needed Now.** https://unfccc.int/news/climate-plans-remain-insufficient-more-ambitious-action-needed-now
11. **The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023, October 23).** https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
12. **Szucs, A. (2023, June 13). EU pushing for watermark on AI-generated contents. aa.com. Retrieved October 31, 2023, from** https://www.aa.com.tr/en/europe/eu-pushing-for-watermark-on-ai-generated-contents/2921467

13. **Kharpal, A. (2023, July 13). China finalizes first-of-its-kind rules governing generative A.I. services like ChatGPT. CNBC.** https://www.cnbc.com/2023/07/13/china-introduces-rules-governing-generative-ai-services-like-chatgpt.html#:~:text=Generative%20AI%20services%20will%20need,material%20to%20the%20relevant%20authority.

14. **Fujitsu AI Ethics and Governance. 5 Principals of "Fujitsu AI Commitment".** https://www.fujitsu.com/global/about/research/technology/ai/aiethics/#anc-03

15. **AI Ethics. Fujitsu.** https://www.fujitsu.com/global/about/research/technology/aiethics/

16. **Fujitsu Kozuchi (code name) Fujitsu AI Platform.** https://www.fujitsu.com/global/about/research/technology/ai/fujitsu-ai-platform/

17. **Google (2023, February 06). An important next step on our AI journey.** https://blog.google/technology/ai/bard-google-ai-search-updates/

18. **Google ChatGPT rival Bard Flubs Fact about NASA's Webb Space telescope. (2023, February 9). CNET.** https://www.cnet.com/science/space/googles-chatgpt-rival-bard-called-out-for-nasa-webb-space-telescope-error/

19. **Ji, Ziwei (2022, February 22). Survey of Hallucination in Natural Language Generation.** https://arxiv.org/abs/2202.03629

20. **Named entity. Wikipedia.** https://en.wikipedia.org/wiki/Named_entity

21. **WikiBio GPT-3 Hallucination Dataset.** https://huggingface.co/datasets/potsawee/wiki_bio_gpt3_hallucination

22. **AUC-ROC (Area Under the Curve of the Receiver Operating Characteristic Curve):** The area under the curve of the curve obtained when the threshold value of the judgment is changed with respect to the abnormality score by placing the true positive rate on the vertical axis and the false positive rate on the horizontal axis. A random anomaly score is 0.5, and a perfect answer is 1.0. It is generally considered that a certain level of performance can be achieved when it is higher than 0.7.

23. **Criminals Are Flocking to a Malicious Generative AI Tool. (2023, July 26).** https://www.bankinfosecurity.com/criminals-flocking-to-malicious-generative-ai-a-22660

24. **Fujitsu and Ben-Gurion University Embark on Joint Research at New Center in Israel for Precise and Secure AI. (2021, November 16).** https://www.fujitsu.com/global/about/resources/news/press-releases/2021/1116-02.html

About the authors

Naomi Hadatsuki

Sr. Market Research Manager at Technology Strategy Unit.

Naomi's research mainly focuses on understanding global megatrends and how they impact technology, society, enterprise, and government, supporting corporate strategy planning and innovation activities at Fujitsu.

Bryan McMahon

Bryan McMahon is a Technology Strategy Research Analyst with Fujitsu's MegaTrends team, where he tracks structural changes in global macrotrends to help develop Fujitsu's medium- and long-term corporate strategy. Before Fujitsu, he worked as a Research Analyst for Japan's New Energy and Industrial Technology Development Organization (NEDO) evaluating public R&D strategies to develop cutting-edge artificial intelligence and semiconductor technologies.

Dr. Takao Mohri

Takao Mohri is a Senior Research Manager in AI Trust Research Centre, Fujitsu Research. He is working on research and development, and customer practice of AI trust technology in generative AI such as hallucination detection technology, and AI ethics technology such as fairness-aware machine learning.

Kentaro Tsuji

Kentaro Tsuji is a Senior Research Manager in AI Trust Research Centre, Fujitsu Research. He is working on research and development and customer practice of AI security technologies that prevent AI-targeted attacks, such as generative AI.

Hiroya Inakoshi, Virginia Ghiara, Satoshi Nakashima, and Hisashi Kojima greatly contributed to the completion of this white paper.

December, 2023