ijcsm INTERNATIONAL JOURNAL OF CASE STUDIES IN MANAGEMENT
tech3lab HEC MONTRÉAL
Volume 22
Issue 3
August 2024

# Blazing New Trails: Responsible Generative AI and the Creative Adoption of a Large Language Model at Deloitte Canada

Case prepared by **Marion KOROSEC-SERFATY**[1] and **Luc LESPÉRANCE**[2]

In late December 2022, Sameer Bohra, national information technology (IT) director at Deloitte Canada (Deloitte), sat pensively at his desk in Toronto. His view of the sparkling lights and Christmas decorations was spectacular, and the holidays were around the corner, but Bohra's mind was elsewhere. He had been leading Deloitte's IT team for over a decade, overseeing many transformational initiatives, but he now faced a daunting challenge. An emerging technology was changing user behaviour, influencing how tasks were carried out and decisions made while raising critical concerns about business ethics, data privacy, and security.

A few weeks earlier, a generative pre-trained transformer (GPT), an advanced type of large language model (LLM) adept at generating human-like text and understanding complex language patterns, had been publicly launched. Its rapid, global adoption had marked a significant milestone for LLMs and the broader integration of artificial intelligence-driven solutions within organizational processes and ways of working.

Bohra faced a momentous decision, and he had to decide soon. Should he allow the firm's 14,000 employees to use an external LLM for internal, client-purposed engagements? Should he block this new technology or limit its use with guidelines and best practices? Or should he deploy a similar technology internally to allow employees to benefit from the new LLM capabilities?

## Bohra's journey

Sameer Bohra was born in India and attended school in a small town in the western half of the country. His fascination with computers and technology began when he first interacted with a computer at a science exhibition in the early '90s. While arcade games and cricket interested most kids his age, Bohra spent hours tinkering with one of the three computers at his school, exploring their operating system and teaching himself to code. His bedroom was soon transformed into a mini-tech lab full of wires, circuit boards, and a growing collection of computer components.

[1] Marion Korosec-Serfaty is a professor of Information Technology in the Department of Analytics, Operations, and Information Technology at UQAM's School of Management and a Ph.D. candidate in Information Technology at HEC Montréal.

[2] Luc Lespérance is a faculty lecturer in HEC Montréal's Department of Information Technologies.


*The* International Journal of Case Studies in Management *is published on-line (http://www.hec.ca/en/case_centre/ijcsm/), ISSN 1911-2599.*
*This case is intended to be used as the framework for an educational discussion and does not imply any judgment on the administrative situation presented. Deposited under number 9 100 2024 003 with the HEC Montréal Case Centre, 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec) H3T 2A7 Canada.*

One day, while at university, Bohra entered a programming competition he'd heard about through his math teacher. To his delight, he was able to solve most of the problems by applying his algorithmic thinking skills, sparking his passion for technology and setting him on a path toward a technology career with Deloitte.

As a young IT professional, Bohra demonstrated technical and leadership skills that propelled him into managerial roles. He gained recognition for successfully spearheading many large-scale projects, including the implementation of enterprise resource planning (ERP) systems that streamlined the company's internal processes and enhanced efficiency. He also initiated and managed robotic process automation (RPA) projects that slashed operational costs and improved the firm's overall competitiveness.

More recently, aware of the potential of artificial intelligence (AI) to improve professional services, Bohra and his teams initiated a series of experiments and significant projects aiming to harness AI's power to enhance decision-making, streamline client services, and explore new revenue streams.

Bohra's success at Deloitte was tied to his holistic understanding of the business. He believed that IT solutions should be aligned with the company's strategic objectives, and the IT department's ability to link technology with corporate strategy was pivotal to supporting the organization's growth while ensuring Deloitte remained at the forefront of the Canadian professional services industry.

## Deloitte: A professional services firm

Founded in London in 1845, Deloitte Touche Tohmatsu Limited (Deloitte) had grown into the world's largest professional services firm in revenue and employee count. In 2022, Deloitte had more than 415,000 employees worldwide, with annual revenue of US$59 billion.[3] It provided professional services to four out of five Fortune Global 500® companies through its member firms in more than 150 countries, including Canada.[4] The Canadian member firm, known for its innovativeness, employed some 14,000 individuals and posted an annual revenue of US$3 billion.[5]

Professional services encompass a wide range of specialized, knowledge-based activities. In the case of Deloitte, these services include audit, assurance, consulting, financial advisory, risk advisory, tax, and related services offered to public and private organizations seeking specialized expertise and guidance from trusted advisors. It thus employs knowledge workers – professionals who deal primarily with information and knowledge – such as lawyers, accountants, consultants, and experts in specific fields of business.

Professional services firms are known for their expertise, professionalism, high-quality deliverables, and adherence to ethical standards. Deloitte's global value proposition was "To be

[3] Deloitte – Annual report: https://www2.deloitte.com/content/dam/Deloitte/dk/Documents/Annual_report_2022.pdf

[4] Deloitte – Who we are: https://www.deloitte.com/global/en/about.html

[5] Deloitte Canada – About Deloitte: https://www2.deloitte.com/ca/en/pages/about-deloitte/articles/about-deloitte-canada.html

the Standard of Excellence."[6] To meet this goal, the firm fostered a high-performing culture of continuous learning and development, empowering its employees to stay ahead in a rapidly changing business landscape.

From an IT perspective, Deloitte's value proposition translated into providing workers with the tools and infrastructure needed to augment their productivity and creativity. For Bohra, this meant continuously monitoring innovation and quickly implementing new technologies to ensure the firm's professionals remained on top of their game.

## The rise of large language models

Large language models (LLMs) are generative artificial intelligence-driven models trained on an extensive corpus of text data from the Internet, books, and articles. By leveraging natural language processing (NLP) techniques, most LLMs – especially GPTs – predict the most likely next word or phrase in a given context. They can understand and generate human-like text and provide contextually relevant information on a wide variety of topics. LLM applications span diverse fields such as content creation, programming assistance, and problem-solving, making them powerful professional tools.

In 2022, the emergence of publicly accessible, free, and powerful generative AI models marked the rise of LLMs. Most of those models featured user-friendly chatbot interfaces, giving individuals with minimal technical expertise to use their capabilities. By year-end, the public release of a GPT-based LLM triggered an unparalleled surge of interest and enthusiasm. It became evident that LLMs would profoundly influence user behaviour, changing how individuals interacted with AI in their personal and professional environments.

For knowledge workers such as the professionals at Deloitte, LLMs could have a positive and productive impact on various components of their work, including research, data analysis, content creation, and problem-solving. LLMs could streamline research by swiftly analyzing vast amounts of documentation, completed projects, and market studies, significantly reducing the time spent reading and summarizing key insights. They could interpret complex datasets to forecast trends and outcomes, helping professionals to make informed decisions more quickly. These models could also generate draft proposals, reports, and visuals, boosting their creativity and speeding up the production of deliverables. Finally, LLMs could not only provide new perspectives on business challenges, helping professionals to explore new approaches and solutions, but also facilitate knowledge search and organization, automate mundane tasks, and improve daily communication, efficiency, and effectiveness.

While the use of public LLMs held the promise of improving the efficiency of knowledge workers, it also introduced major challenges such as ethical concerns, user data privacy issues, and threats to data security. Due to their inherent complexity, biases within LLMs' training data could compromise the fairness of decision-making and obscure accountability. The extensive datasets processed by LLMs often contained sensitive information, requiring stringent data management and strict adherence to privacy laws to prevent data breaches. Moreover, the growing dependence

[6] Deloitte – Vision, values, and strategy: https://www2.deloitte.com/az/en/pages/about-deloitte/articles/vision-values-strategy.html

on LLMs heightened security risks, attracting cyber threats and demanding robust protective measures to safeguard proprietary and confidential information.

Realizing the challenges associated with the rapid adoption of generative AI, Bohra knew he would have to tread softly: successfully integrating Deloitte's own LLM solution based on third-party generative AI tools would require technical expertise and a comprehensive legal, ethical, and operational strategy.

## Implementing generative AI at Deloitte (Gen-D)

### Launching the project

Although Deloitte's clients had expressed no immediate concerns about the use of LLMs, Bohra and his colleagues recognized the potential risks associated with the exposure of sensitive internal and client data.

Before leaving for the holidays, Bohra contacted Deloitte's risk management department, which was responsible for privacy and security issues. Everyone agreed that unrestricted use of a public AI model would be unacceptable for a professional services firm such as theirs, so Bohra would have to act quickly. He thus adopted a proactive approach, bypassing the usual time-consuming process of identifying sponsors, garnering support, securing a budget, and enlisting stakeholders to support a new idea.

This approach was a far cry from every other initiative Bohra had been involved in. When he had started out as an IT professional, working primarily on large-scale ERP projects, projects would begin with initiation and planning, needs assessment, and business process management. Implementation typically followed a structured, waterfall process involving extensive upfront planning based on detailed specifications and client requirements. There were generally few, if any, iterations, and end-users would validate the technology with usability testing once it was well developed. Over the years, Bohra had gradually adopted and promoted more agile approaches to building and delivering IT solutions. Unlike waterfall approaches, agile approaches embrace flexibility, collaboration, and iterative development, preferring smaller, incremental releases with continuous end-user feedback. These implementations enable IT to work closely with end-users to focus on value and rapid responses to changing requirements. Usability testing is not limited to a single phase but is integrated throughout the project.

With this new initiative, the goal was more straightforward: knowledge workers just wanted to use generative AI capabilities. For Bohra, the sooner his team could release a safe, reliable LLM solution, the better. To expedite the process, the IT department assumed a proactive, leadership role, sponsoring the project and demonstrating its worth.

### Technical approach

An analysis of the various options had shown that the fastest and most secure way to benefit from an LLM's capabilities would be to use a third party's existing models. This technical strategy would be made possible by an application programming interface (API) that would be almost transparent from a user perspective: employees would use a chatbot interface – dubbed Gen-D for "generative

AI at Deloitte" – while leveraging the third party's generative AI capabilities within the Deloitte environment (see Figure 1).

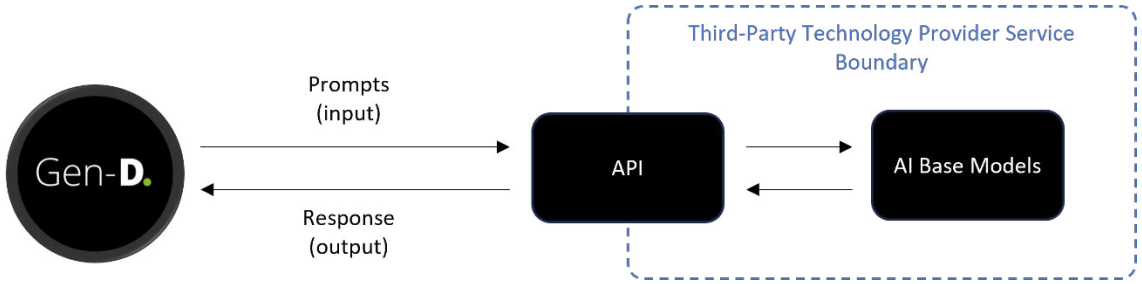



**Figure 1 – *High-level schematic of Gen-D processes***

This architecture offered three main benefits. First, it allowed user prompts as input and generated text as output. Second, it augmented prompts with user data, providing more suitable responses for a professional services firm. Third, it enabled the organization to fine-tune responses by augmenting the models with its own data.

Bohra's six-person NLP team focused on experimenting with the API. The team began by implementing and configuring the API and developing its own interfaces before formulating prompts and conversational scenarios aligned with Gen-D's purpose. This entailed creating a range of input queries and dialogue sequences to elicit meaningful responses from the model.

To maximize the precision and usefulness of Gen-D, the NLP team integrated a Deloitte-specific persona reflecting the firm's professional identity. LLMs often used personas to augment user prompts with meaningful context. When a user wrote a prompt such as "Can you summarize this article?" for example, the prompt would append a longer, built-in command adding contextual information such as "You are a Deloitte professional; write the response in a business-like fashion […]." This augmented prompt strategy would help align the model's responses with the desired tone, style, and expertise, making the interactions more tailored and contextually appropriate for Deloitte professionals.

The NLP team developed the Gen-D solution in just two months. Its members' experience in the field enabled them to quickly understand and harness the capabilities of the third-party generative AI solution and overcome technical challenges along the way. As the saying goes, however, "the soft stuff is the hard stuff," and most of the challenges they faced were not technical.

**Data privacy, security, and access concerns**

For Bohra, the most important ethical and business concerns involved privacy and security risks. When interacting with a public LLM, Deloitte professionals could inadvertently expose sensitive, non-anonymized, or confidential data, for example. This data was at risk of being stored and reused by the LLM and potentially used as training material by its base models. The LLM could incorporate all or part of those queries to refine its algorithms and produce outputs that could reveal or infer this sensitive information in response to other users' queries. The possibility of exposure

and potential leakage raised significant concerns about data privacy and security for a firm that managed sensitive and confidential client data.

Since using a public LLM was risky, Bohra's NLP team had to find a way to ensure both privacy and security when using an external AI model.

In this context, using an API helped to create the right balance between user experience and data privacy and security. Deloitte professionals would use a Deloitte interface, accessible through the firm's Intranet, to leverage advanced generative AI capabilities in their work activities. In the background, the API service would encrypt data in transit between Gen-D and the third-party generative AI solution to protect against potential interception and unauthorized access. To reinforce security, the API provider would adhere to robust protocols, including network security and threat protection. Moreover, contractual agreements with the vendor would ensure that the data provided to the third party's generative AI solution would not be shared with other users, customers, or the third-party company. In other words, Gen-D would operate completely independently, and the firm's internal and client data would be securely stored within Deloitte's cloud infrastructure.

Although, technically speaking, the API had provided part of the solution, the NLP team knew it wasn't enough. The third party's generative AI solution was good, but it was not perfect, reliable, or unbiased. Bohra wondered whether they should wait for the third-party provider to resolve those issues or teach their employees about the risks and limitations of such models – helping them to adopt "good behaviours" when using LLMs.

Bohra's NLP team worked closely with Deloitte's risk management department, data office, and change management professionals to assess risks and benefits and establish Gen-D usage guidelines. This collaboration led to the creation of a Guidance and limitations agreement that appeared as a pop-up window every time a Deloitte user accessed Gen-D. This agreement educated users about the solution's capabilities and limitations, such as its limited world knowledge, susceptibility to biases, potential unreliability, and inability to verify sources. The agreement also included examples of data types that were approved or not approved for use with Gen-D (see Exhibit 1).

Every Deloitte user was required to explicitly consent to the Guidance and limitations agreement before using Gen-D. In doing so, users acknowledged their responsibility and accountability for their interactions with Gen-D. This consent process accomplished three things: 1) it made users responsible for the inputs they provided within their prompts and the outcomes they received from Gen-D, 2) it underscored the importance of thoughtful and appropriate Gen-D use to prevent potentially harmful or unethical outputs, and 3) it reinforced Deloitte's commitment to transparency, ensuring users' awareness of AI's capabilities and the standards they were required to meet when using Gen-D.

For legal and compliance purposes, Deloitte had to ensure transparency and traceability. Since the same prompt could generate different responses each time, the team recorded all user interactions with Gen-D in a secure and robust database. Every input provided to the model and its corresponding output was systematically logged and securely stored in the database. This approach to data recording enabled the firm to comply with regulatory and internal auditing standards.

Professional services firms often audit their projects to ensure regulatory compliance and alignment with high-quality expectations and ethical standards. At the time of writing, only two Deloitte employees could access and monitor database activity.

**Human-in-the-loop**

The NLP team incorporated human-in-the-loop to improve Deloitte professionals' understanding of the ethical implications of AI use while engaging them in providing feedback about AI-generated content to align with the firm's standards and regulatory compliance.

Human-in-the-loop refers to an interactive and iterative feedback process involving humans in designing, developing, and using AI systems to leverage the strengths and mitigate the weaknesses of AI and human intelligence to achieve optimal outcomes. It is particularly important at professional services firms such as Deloitte, where the risk of inaccuracies or ethical breaches could have significant implications, requiring the careful integration of human judgment to ensure accuracy, compliance, and trustworthiness in AI-driven decisions.

Human-in-the-loop can be integrated at various levels across the entire AI system lifecycle to optimize each stage through human oversight and expertise. During the initial setup stage, humans define their responsibilities, decision-making boundaries, task execution, and supervisory roles. During the training stage, humans label data and provide feedback to enhance the AI's task accuracy. During the operational stage, humans monitor AI performance and facilitate incremental enhancements through feedback loops. During the continuous improvement stage, humans identify opportunities for AI enhancement and participate in retraining sessions to improve precision and relevance. Quality control involves humans managing uncertainties and conducting audits to maintain standards. Finally, during the customer interaction stage, humans resolve user interface issues that could lead to a poor user experience. In this context, and to ensure continuous improvement, Bohra's NLP team focused on human verification, instituting a feedback mechanism to facilitate Gen-D's continuous improvement.

To this end, the Guidance and limitations agreement raised awareness by increasing user accountability; it suggested human verification of AI-generated responses for originality and intellectual property purposes. When using Gen-D, Deloitte professionals were encouraged to critically analyze and assess its responses by cross-referencing them with existing documents, databases, and other relevant sources to check for potential infringements or accidental duplication of existing material while ensuring that the responses were in line with the firm's ethical and moral guidelines. This human verification was aligned with Deloitte's code of conduct,[7] which required all Deloitte professionals to comply with the principles of honesty, integrity, and professionalism and outlined standards of behaviour.

The NLP team also developed a thumb-up/thumb-down feature that displayed with each Gen-D response, enabling Deloitte professionals to provide feedback that helped fine-tune the third-party generative AI base solution. (See Exhibit 2.) Instead of using a simple binary feedback mechanism, when users clicked the thumb-up or thumb-down icon associated with the response, Gen-D

[7] Deloitte Canada – Code of Conduct: https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/about-deloitte/ca-code-of-conduct-en-jan2024-AODA.pdf

displayed a field where they could communicate their thoughts, concerns, or suggestions. The NLP team logged and analyzed those qualitative messages as part of the solution's continuous improvement.

### Responsible generative AI

In 2020, Deloitte introduced its Trustworthy AI™ framework[8] to help clients deploy AI technology responsibly and ethically[9] while helping organizations meet governance and regulatory standards across the complete AI lifecycle from initial concept development to design, development, deployment, and continuous improvement. Deloitte used this multidimensional AI framework to help organizations develop ethical safeguards across seven key dimensions: transparent and explainable, fair and impartial, robust and reliable, private, safe and secure, responsible, and accountable (see Exhibit 3).

The Trustworthy AI™ framework played a major role in guiding the NLP team as it implemented the Gen-D solution. The dimensions of transparency and explainability, responsibility, accountability, and privacy informed the decisions made by the NLP team throughout the development of the solution and the Guidance and limitations agreement. Those dimensions led to the creation of clear usage procedures, limited access to database activity, and accountability reminders to users.

Using an external, pre-existing, unmodifiable AI solution nevertheless presented additional challenges, such as the potential lack of explainability, transparency, and fairness of its outputs. The architecture of LLMs involved highly complex network structures with billions of parameters, making it difficult for users to understand how those models generated specific outputs. Public LLMs were trained on a wide range of datasets, such as books and articles, that could embed societal biases such as stereotypes and prejudices, potentially causing models to replicate those biases after learning from those patterns and associations. Professional knowledge workers thus had to be cautious when using LLM-generated responses to avoid perpetuating or amplifying biases, creating ethical issues, or hurting their firm's reputation.

Bohra knew that while the framework would help steer and fine-tune the development of Gen-D, converting broad principles into specific, actionable steps for implementing and using the solution would be challenging. While Deloitte's strategic decision to use a public LLM facilitated swift progress and safe access to cutting-edge LLM capabilities, it also meant accepting compromises inherent in using a technology that was beyond the firm's direct control or influence.

## Build it, and they will… use it?

Since Bohra aimed to release a safe, reliable solution as quickly as possible, the adoption process – and the change management approach – were quite different from those of earlier IT initiatives.

---

[8] Deloitte Consulting LLP, *Deloitte Introduces Trustworthy AI Framework to Guide Organizations in Ethical Application of Technology in the Age of With™*, press release, August 26, 2020.

[9] Enio Moraes, "Embracing the Age of With in AI: Real examples of companies leading the way", *LinkedIn Pulse*, March 12, 2023. "The concept of the Age of With in AI is based on the idea that artificial intelligence is not meant to replace humans but to work with them. In this era, the focus is on creating AI systems that can complement human abilities and enhance our decision-making processes, rather than replacing us altogether."

The NLP team didn't involve end-users in identifying use cases[10] describing their interactions with Gen-D, nor did it impose pre-defined use cases within the solution. The need was clear: knowledge workers just wanted fast access to generative AI capabilities. The goal was to encourage the organic adoption of AI while users explored and experimented with the solution and tailored its use as necessary. In short, the team wanted to reproduce users' current behaviours with publicly available LLMs.

For Bohra, the underlying assumption was that people just wanted to leverage LLMs' capabilities. He summarized this "build it and they will come" approach this way: "We didn't want to push users to use Gen-D in a specific way. We just wanted to let them explore and use it as it evolved."

As a result, the use of Gen-D was characterized by its emergent nature. Rather than being predetermined, its usage evolved organically as users across different roles and practices discovered new ways to leverage the LLM's capabilities in their work, thereby boosting their productivity and creativity. This approach allowed the solution to effectively address a wide range of tasks, making it a versatile and valuable tool for Deloitte's knowledge workers.

Gen-D was widely adopted during the first two months following its launch, with 6,000 unique users engaging with the technology in various ways. Those users generated 170,000 requests, indicating a strong demand for the LLM's capabilities. Moreover, Gen-D processed 85 million tokens, highlighting the scope of its use and the AI model's vast language processing capabilities.

## What did the future hold?

Almost ten months later, Bohra found himself back at his desk thinking about all he had learned since deciding to leverage a third party's generative AI solution. One of the key learnings for him and his team had been to do their best with the available technology. IT departments could no longer afford to wait for technologies to mature, with most of their initial faults and inherent problems removed or reduced before implementing them. AI-related initiatives were emerging fast, and firms such as Deloitte had to stay at the head of the parade by constantly exploring, experimenting, and testing new technology.

As Bohra contemplated Gen-D's next steps, he recognized the need to fine-tune part of the third-party generative AI solution and develop use cases for specific business needs. His team had already begun testing an economy-related chatbot based on the same AI-based models, but fine-tuned according to the firm's economic perspective and publications over the past decade.

He also recognized the need to expand Gen-D to other member firms. Gen-D's success had attracted the attention of other member firms struggling with using public LLMs.

Bohra could foresee Gen-D's becoming a global solution for all firms within the group. Although this would pose technical challenges, the Canadian firm's business learnings could be leveraged to accelerate Deloitte's global success.

*2024-08-21*

[10] Nicky Daly, "What is a use case?", *Wrike Blog*, April 25, 2022.

# Exhibit 1
# Data classification within Gen-D

The Guidelines and limitations agreement provides examples of data types that have been approved and not approved for use with Gen-D.

**Table 1**:**Types of data approved**

| Types of data | Examples |
|---|---|
| **Deloitte data** | Information found on Deloitte's external public website |
| **Public data** | Information that is not protected by copyright or other terms of use |
| **Third-party data purchased from data brokers**[11] | Data acquired from vendors |
| **Synthetic data** | Data generated as a substitute for live or real data |
| **Personal information** | Aggregated, de-identified, or anonymized personal information |
| **Client information approved for use with the AI tool**[12] | N/A |

**Table 2: Types of data not approved**

| Types of data | Examples |
|---|---|
| **Deloitte data** | Individual-level talent data |
| **Public data** | Information (public or proprietary) where the terms of use expressly prohibit its use (e.g., online journal data or market research) |
| **Third-party data purchased from data brokers**[13] | Data acquired from vendors |
| **Personal information** | Any individual level or identifiable personal information |

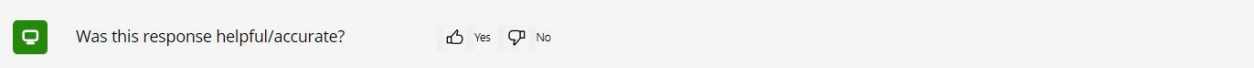
[11] Only if it has been confirmed with the third-party contract owner that generative AI data use case would be permitted under the terms of that contract.

[12] Only if the client information is used in accordance with the terms and conditions of the client agreement.

[13] Purchased third-party data not procured or purchased for Deloitte use.

# Exhibit 2
# User feedback mechanism

The thumb-up/thumb-down feature is added to every response provided by Gen-D, as shown below.



# Exhibit 3
# Deloitte Trustworthy AI™ Framework[14]

**Table 3: Explanation of each dimension of the Deloitte Trustworthy AI™ Framework**

| Dimensions | Definitions |
|---|---|
| Private | User privacy is respected, and data is not used or stored beyond its intended and stated use and duration; users can opt in/out of sharing their data. |
| Transparent & explainable | Users understand how technology is being leveraged, particularly in making decisions; these decisions are easy to understand, auditable, and open to inspection. |
| Fair and impartial | The technology is designed and operated inclusively to aim for equitable application, access, and outcomes. |
| Responsible | The technology is created and operated in a socially responsible manner. |
| Accountable | Policies are in place to determine who is responsible for the decisions made or derived from using the technology. |
| Robust & reliable | The technology produces consistent and accurate outputs, withstands errors, and recovers quickly from unforeseen disruptions and misuse. |
| Safe & secure | The technology is protected from risks that may cause individual and/or collective physical, emotional, environmental, and/or digital harm. |

[14] Deloitte – Trustworthy AI: https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html