VIEW POINT

GENERATIVE AI IS ONLY AS GOOD AS THE DATA FOUNDATION IT STANDS ON

Abstract

Generative AI (Gen AI) is driving transformation across business functions and industries. To generate authentic content, build trust and make ethical data-driven decisions, high-quality data is needed and will be foundational to the success of Gen AI in business.

Infosys®
Navigate your next

Generative AI (Gen AI) applications can add trillions of dollars to the global economy. A McKinsey research found that Gen AI could potentially add $2.6 to $4.4 trillion per year to the global economy, increasing the impact of all AI from 15 to 40 per cent. The potential impact of Gen AI is anywhere between $200 billion to $660 billion a year, depending on the industry. Some of the sectors that are expected to have the highest impact include life sciences, banking and financial services (BFSI), retail and consumer goods. Generative AI applications can understand human language, making them suitable for knowledge-based work, and increasing the productivity of the workforce enormously. Workforce transformation is already underway, and an estimated 50 per cent of activities are expected to be automated between 2030 and 2050.

Gen AI applications are built on large language models (LLM) giving them the ability to create and innovate. Machines can now create text, speech, images, and videos, and even participate in designing candidate molecules for developing new drugs.

Use cases for Gen AI

Several industries such as defence, aerospace, automotive, and electronics will benefit from the adoption of Gen AI in material sciences, giving them the ability to actively identify required materials, rather than wait for years on end of passive research. The pharma industry is set to reap the rewards of Gen AI - drug design and discovery is expected to accelerate from three to six years, to merely months, also resulting in cost savings. Healthcare can leverage Gen AI to create synthetic data which protects patient privacy.

Businesses are deploying Gen AI to accelerate customer onboarding. Gen AI applications can be used to obtain customer information in natural language, get it validated at the backend and present a friendly interface to enhance customer experience. In industries such as banking, where customer onboarding is an intricate and time-intensive process, and requires due diligence checks by compliance officers, Gen AI can be deployed to automate Know Your Customer (KYC) and Anti-Money Laundering (AML) processes.

Gen AI finds several use cases in customer service experience too. For instance, agent assistance at real-time can provide live transcription, real-time suggestions for customer queries based on knowledge repositories, comprehensive summaries, and recommendations about agent behaviour and responses once the call is complete. Customer experience enhanced by AI is reshaping business models and operating structures of enterprises. With live language neutralisation customers and agents or bots can chat in different

External Document © 2024 Infosys Limited

languages, enabling smooth customer experience. Technologies such as Microsoft's Smart Assist are redefining how humans and machines interact with each other for effective data management, allowing human agents to leverage AI capabilities in real-time, and provide intelligent customer service. Good quality of master data lays the foundation for digital brain of an enterprise.

These rapid developments are exciting, however, there are a significant number of challenges to address. First amongst these is the quality of data, which is paramount to the success of Gen AI. In fact, good quality of master data lays the foundation for the digital brain of an enterprise.

Why data quality is pivotal to Gen AI

The initial data that is used to develop the machine learning model is called the training data. This data could be in the form of text, images, audio, video or from sensors. The model creates and refines the rules from this data, thereby impacting future performance. Data readiness is important to start the AI journey. If the data that is used for training the model is inaccurate, inconsistent, or incomplete, the results from the AI application will be the same, resulting in erroneous predictions and decisions. This can have serious implications. High-quality data is essential to the success of business transformation with Gen AI. Let's see why.

To improve language understanding: Chatbots are a common application of Gen AI but imagine a chatbot that gives irrelevant responses. However refined the model may be, high-quality training data is necessary to derive the benefits. For instance, a healthcare chatbot may ask for patient symptoms, and suggest a physician that the patient can consult. Accurate responses are crucial in such scenarios. When the training data consists of correct and consistent information, the model can generate precise responses that improve the overall performance of the application. High-quality training data also boosts the understanding of semantics, syntax and the context of natural language which leads to better output. The data should also accurately capture the relationship between business entities.

To lower bias and discrimination: Human biases can infiltrate AI systems. Data needs to be complete, valid, accurate and relevant to ensure that the final decision is not influenced by the unavailability of the right information. Race and gender stereotypes could in fact be reinforced and perpetrated when the training data has a poor sample selection or when the process of data collection itself is biased. For example, if an AI image generator is asked to produce a picture of a business leader, it may produce a white male, due to the historical bias that has crept into the training data. Choosing a representative sample, a fair data measurement process and human oversight are necessary to lower AI bias.

To generalise the model: The AI model needs to be able to adapt to previously unseen data. This is possible only when the training dataset is large, consistent, and diverse enough to enable generalisation. The model needs to be able to capture the context, relevance, and patterns from the training data, but not "memorise" it. For example, an AI-based application

External Document © 2024 Infosys Limited

that trains on data from heart arrhythmia patients should be able to suggest a treatment protocol for new patients requiring cardiac care. This is possible only if the training data consisting of electrocardiography images is labelled and annotated by clinicians and qualified medical assistants and is a sufficiently large and diverse dataset.

To promote ethical AI: High quality data is necessary to build gen AI applications that result in fair and transparent output. The data needs to represent the real world, and be free of gender and racial bias, especially when gen AI applications are used for decision making in areas such as law enforcement, healthcare, hiring, or financial transactions, where the impact can be significant. Harmful or inappropriate content needs to be filtered out. Without a deep consideration of data ethics, gen AI applications can have unintended, and even potentially dangerous consequences.

To drive innovation: Gen AI applications need to keep adapting to new data and situations. Consequently, a one-time high-quality training dataset is not enough. To scale AI in business, all data, both master and transactional, must be interoperable, consistent, and accurate across disparate sources. A robust data management platform is necessary. Any new business entities and relationships between them should also fall into the data governance framework.

How to get ahead of data quality issues

Businesses need to establish a strong data governance framework to ensure that the data meets the highest standards A data governance framework with well-defined roles and responsibilities for human & AI interventions is the key success factor for next-gen enterprises. The data governance framework must establish the policies and procedures for the collection of data, quality standards, and how it is stored and consumed. The framework should also specify data ownership, as well as the roles and responsibilities of the data stewards who ensure data integrity and high-quality data. All data consumed by the enterprise must be validated during input, to ensure that it is correct and complete. Data cleansing is necessary to remove redundancies, inconsistencies, and errors. Automated data validation and cleansing tools can be used, but human oversight may be necessary too. Data labels and annotations are required both for better understanding by the AI models, and to identify anomalies. Enterprises need to ensure data privacy and data security from the perspective of privacy laws and regulations, as well as building user trust and acceptance.

Gen AI offers organisations the ability to dramatically increase productivity and efficiency and reduce costs, all while innovating to explore new streams of revenue. To harness the true potential of Gen AI responsibly, a firm data governance framework is needed for sustained data quality that drives the organisation of the future.

For more information, contact infosysbpm@infosys.com

Infosys®
Navigate your next

© 2024 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.