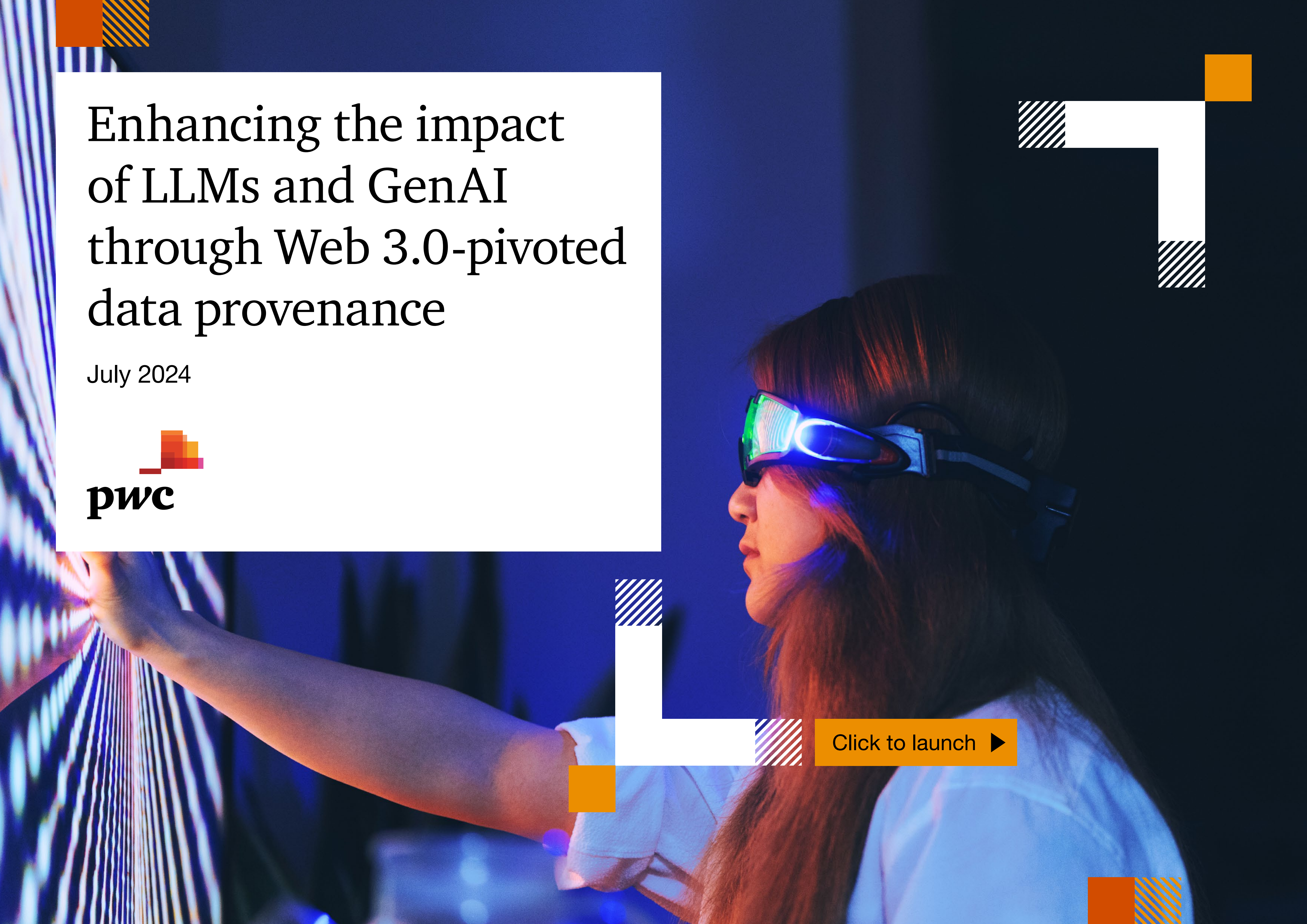
Enhancing the impact of LLMs and GenAI through Web 3.0-pivoted data provenance

July 2024



Click to launch ▶

Generative AI (GenAI): Growth and rapid adoption

GenAI and large language models (LLMs) have redefined innovation. The technology has rapidly and efficiently captured the attention of both individuals as well as businesses.

GenAI is considerably different from the conventional AI technology which focuses exclusively on analysing and optimising the information obtained via training datasets. On the other hand, GenAI leverages LLMs which can help generate new and unique content in the form of text, images and audios. LLMs use statistics to predict and generate the next word in a sentence of the output and use those patterns to generate content. LLMs can comprehend prompts provided by users and are able to respond to those by engaging in human-like conversations.

In the current day and age, LLMs provide the potential to innovate and architect a wide range of applications to benefit both individual users and businesses, such as:

- content creation that does not compromise on creativity, in terms of contextual responses and even mannerisms – to a degree.
- customer service experiences that lead to interactions with intelligent bots which facilitate human-like conversations.
- helping programmers to review, develop, debug and convert codes to different languages, facilitating transfer of projects to a new language from legacy, or empowering them to deal with unfamiliar syntax.

LLMs can be classified into public and private LLMs. A public LLM is trained on the data gathered from the internet – it leverages information based on web crawls and open databases on the web. Due to this unrestricted access to a massive amount of data, public LLMs can generate comprehensive answers for almost every query posed by the users.

On the other hand, private LLMs are trained on information sources which are ring-fenced, classified and not accessible by the public. Usually, such data belongs to a particular enterprise or is a part of private sources, databases or publications.

LLM hallucinations: Role of data provenance

Currently, most LLMs are trained using text, audio and video data. As these models depend upon vast amounts of data, whose provenance has not been validated, this poses a great challenge of data inconsistency or **'data hallucination'**. As a result, the outputs contain wrong facts and misleading information. As per **PwC's 27th Annual Global CEO Survey,** 'Trust' is a key ingredient in using technology to transform and grow a business. About 63% of the CEOs believe that GenAI is likely to increase the spread of misinformation in their company.[1]

Contrary to the belief that LLM hallucinations primarily occur due to statistical errors or incorrect algorithms, the reasons for these can be lack of data validation, data provenance and data democratisation.

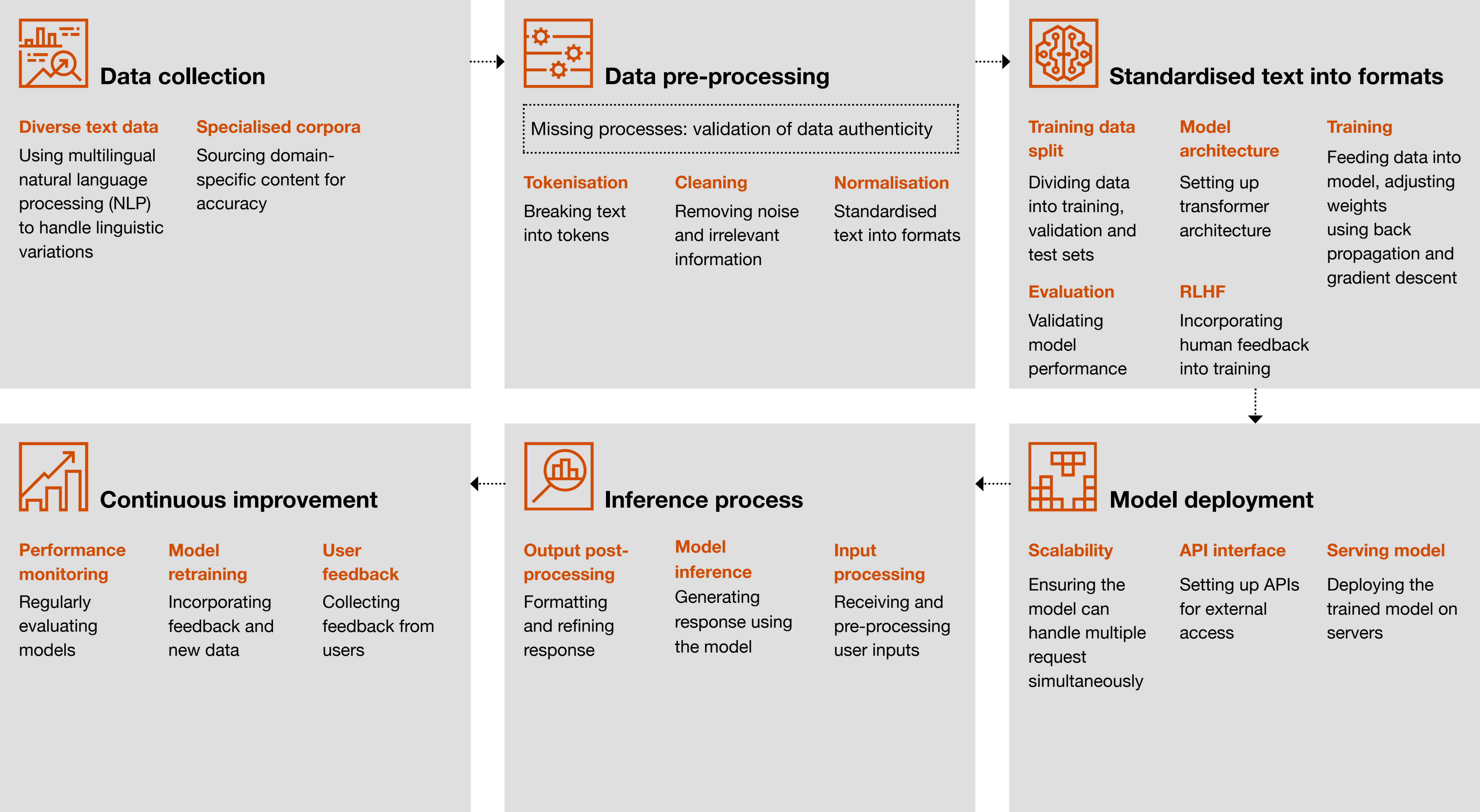
If **incorrect data** is used to train the model, it will lead to incorrect outputs and misinformation. This can cause existing biases, prejudices and faulty logic to creep into the training model and algorithm.

From Figure 1, we can see how an LLM is developed, and it is evident that data collection and processing of data goes directly into the model training pipeline. However, the crucial step of validating the authenticity of data is conspicuously missing.

In this approach, the effectiveness of LLM is evaluated right at the end of the development process where outputs are evaluated and corrected using reinforcement learning from human feedback (RLHF) through methods like parameter restricting and weightage. However, RLHF efforts are primarily on the context of the output rather the correctness of the content.

The mistakes that LLMs make in terms of generating incorrect information are not dovetailed into a single source of truth (SSoT) that is accessible to users or across the enterprise in a democratic manner.

[1] PwC's 27th Annual Global CEO Survey



Data collection

Diverse text data

Using multilingual natural language processing (NLP) to handle linguistic variations

Specialised corpora

Sourcing domain-specific content for accuracy

Data pre-processing

Missing processes: validation of data authenticity

Tokenisation

Breaking text into tokens

Cleaning

Removing noise and irrelevant information

Normalisation

Standardised text into formats

Standardised text into formats

Training data split

Dividing data into training, validation and test sets

Model architecture

Setting up transformer architecture

Training

Feeding data into model, adjusting weights using back propagation and gradient descent

Evaluation

Validating model performance

RLHF

Incorporating human feedback into training

Continuous improvement

Performance monitoring

Regularly evaluating models

Model retraining

Incorporating feedback and new data

User feedback

Collecting feedback from users

Inference process

Output post-processing

Formatting and refining response

Model inference

Generating response using the model

Input processing

Receiving and pre-processing user inputs

Model deployment

Scalability

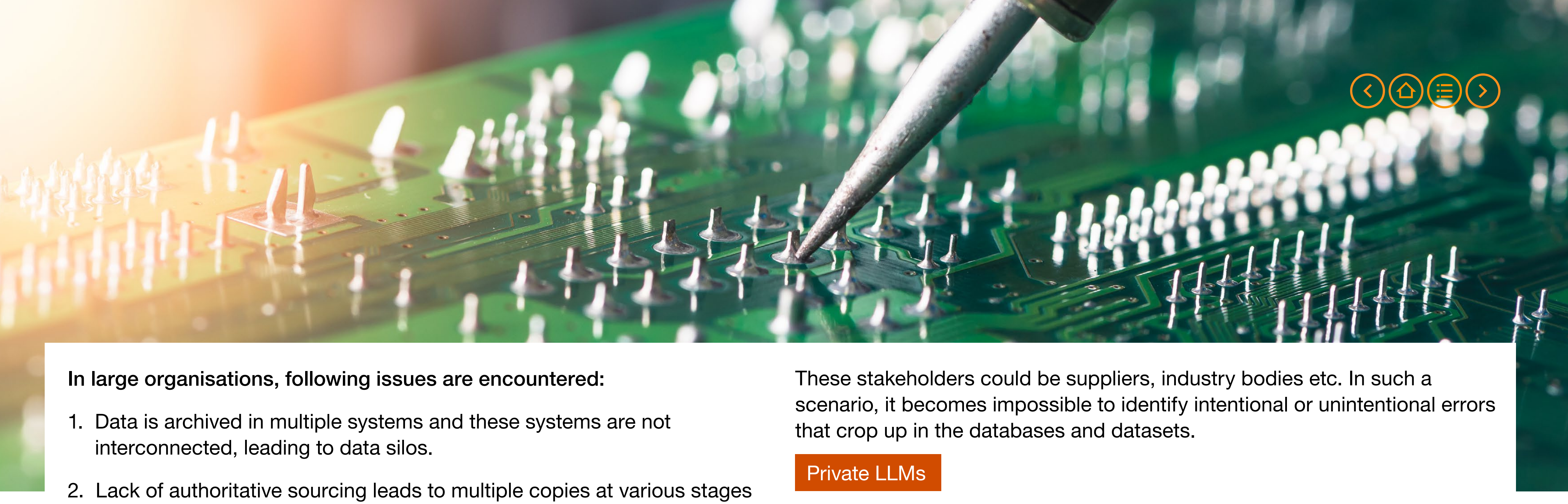
Ensuring the model can handle multiple request simultaneously

API interface

Setting up APIs for external access

Serving model

Deploying the trained model on servers



In large organisations, following issues are encountered:

1. Data is archived in multiple systems and these systems are not interconnected, leading to data silos.
2. Lack of authoritative sourcing leads to multiple copies at various stages of transformation with unclear lineage, which results in compromising the concept of SSoT.
3. There are no mechanisms to create an SSoT by consolidating data from all systems into one, or to maintain distributed data capture across the enterprise. In such cases, any changes in data in one system does not get reflected immediately across the enterprise.
4. Sometimes private LLMs need to have access to a chunk of sensitive and private information in order to train the model. This information cannot be accessed openly by all and needs to be fully masked in such a way that required inputs are provided only to the model and cannot be read or interpreted by others.

Public LLMs

Public LLMs are not preferred by enterprises due to:

1. open access of LLM and data
2. increased time taken for data validation and verification as this data is obtained from multiple data sources. These stakeholders could be suppliers, industry bodies etc. In such a scenario, it becomes impossible to identify intentional or unintentional errors that crop up in the databases and datasets.

Private LLMs

From an enterprise perspective, it may be necessary to develop customised models for specific needs; private data and algorithms will be more useful in these cases.

In an enterprise, in order to use organisational data, various governance measures must be implemented – auditing, consent management and data protection. If these frameworks are not established to enable the same, compliance issues can arise.

Lack of enterprise governance and democratisation of data also contributes significantly to LLM hallucinations.

Web 3.0-pivoted data provenance and data democratisation may have the solutions to address hallucinations in both public and private LLMs.

Web 1.0 was referred to as the internet of connection, Web 2.0 as the 'internet of information' and Web 3.0 can be defined as the 'internet of value'. Here, we are referring to Web 3.0 in the context of decentralised architectures to enable data lineage and data provenance across the World Wide Web. Basically, the ownership of data belongs to the people who generate and use it. Everyone has equal rights and access to the data, thus fostering data democratisation.

Decentralisation of data storage, equal rights to monetise data by its owners, tracking and tracing the data from source to origin, establishing SSoT are the key principles of Web 3.0-based architecture. Moreover, blockchain-based technology further helps accomplish these key principles and features of Web 3.0.

When we refer to the word data provenance, which is also called data lineage in the industry, it refers to a documented trail that accounts for the origin of a piece of data and where it has moved from to where it is presently.[2] Data provenance ensures that there is a traceable history of the data being used from its source of origin to where it is currently deployed.

With blockchain bringing in decentralisation, data can be distributed across multiple locations, thus ensuring that there is no concentration of data over a single entity. This reduces the risk of data manipulation. Blockchain brings in the element of SSoT. This immutability offered by blockchains fosters transparency and ensures that the data is accurate and unaltered and allows the tracking of data.

To facilitate data provenance for datasets that work as inputs for public LLM training models, large-scale interventions across the World Wide Web are required. These cannot be accomplished by the action of few operators alone and should instead include a systematic overhaul in terms of the implementation of:

i. self-sovereign identities (SSIs)-based standardisation across the world wide web

ii. a strong code of practice

iii. exhaustive community support like developers, forums, online communities etc.

However, such interventions may take several years to get implemented. In case of private LLMs, the same can be accomplished with concerted efforts in few months.

Hence, for an enterprise looking to improve the effectiveness of a private LLM, we would recommend it to implement decentralised data architectures, and set up SSoT designed on Web 3.0 principles within the enterprise and the ecosystems owned by them.

This will allow private LLMs to address hallucinations and prevent incorrect outputs and falsification, as detailed in the rest of the paper.

[2] https://www.nnlm.gov/guides/data-glossary/data-provenance#:~:text=The%20term%20%E2%80%9Cdata%20provenance%E2%80%9D%2C,to%20where%20it%20is%20presently

High-level overview of the proposed solution

The primordial step in the development of any LLM is creating the datasets and providing them as inputs to training LLMs. The current practice, as illustrated in the previous diagram, whether it is public or private LLM, is the provision of data without any validations. This needs to be corrected.

The first intervention for building an LLM includes addressing data validation and correction even before it is furnished to the LLM's training model. We recommend the below-mentioned steps for LLM development:

1. Data integration
2. Data verification and data governance
3. Data democratisation

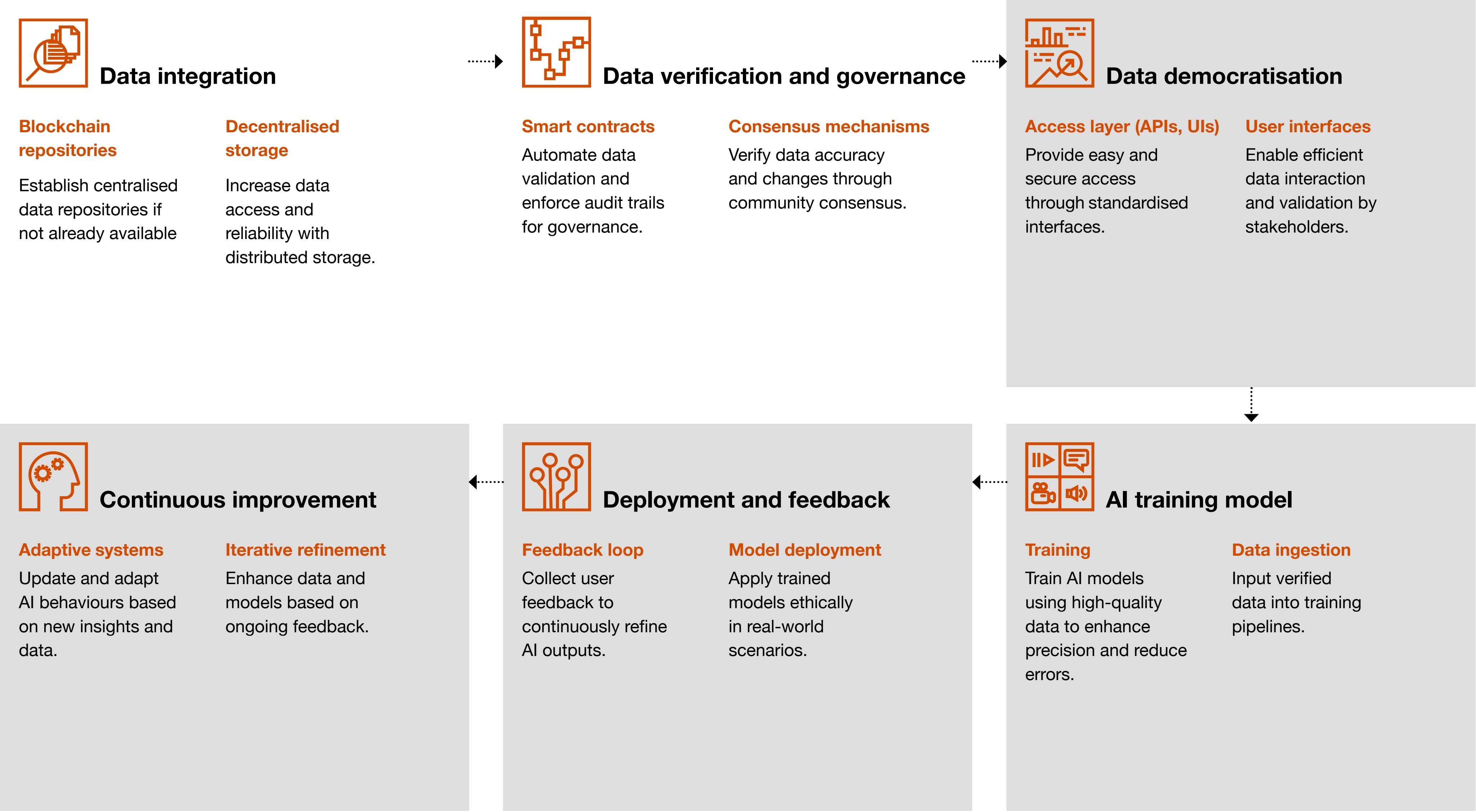
An enterprise is required to facilitate these frameworks and deploy them even before the efforts are made to develop the LLM training pipeline and data assets.

Thereafter, an SSoT should be created within the enterprise LLM to ensure:

- Data integration: Data originates from various decentralised sources like blockchain repositories within the enterprise. If blockchain repositories are not available, enterprises should create those first. To create such repositories, data needs to be extracted and consolidated from all underlying systems after data validation, in order to create an SSoT. Once all the data has been aggregated and an SSoT has been created, data verification and governance mechanisms can be set up.

- Data verification and governance: Data passes through verification mechanisms using smart contracts, blockchain-based hashing validation and consensus methods to ensure its authenticity and integrity.
- Audit trail implementation and validation for any change in upstream or downstream data within the organisation or its ecosystem will help enterprises achieve data governance and maintain the integrity throughout the data trail.
- Data democratisation: Verified data is made accessible across the enterprise through application programming interfaces (APIs) and user interfaces, allowing broader access and contribution. This fosters transparency and collaborative data curation.

Figure 2. High-level solution using Web 3.0 architecture for LLMs



Data integration

Blockchain repositories

Establish centralised data repositories if not already available

Decentralised storage

Increase data access and reliability with distributed storage.

Data verification and governance

Smart contracts

Automate data validation and enforce audit trails for governance.

Consensus mechanisms

Verify data accuracy and changes through community consensus.

Data democratisation

Access layer (APIs, UIs)

Provide easy and secure access through standardised interfaces.

User interfaces

Enable efficient data interaction and validation by stakeholders.

AI training model

Training

Train AI models using high-quality data to enhance precision and reduce errors.

Data ingestion

Input verified data into training pipelines.

Deployment and feedback

Feedback loop

Collect user feedback to continuously refine AI outputs.

Model deployment

Apply trained models ethically in real-world scenarios.

Continuous improvement

Adaptive systems

Update and adapt AI behaviours based on new insights and data.

Iterative refinement

Enhance data and models based on ongoing feedback.

Once the above are accomplished, AI model training can begin.

- AI model training: High-quality data helps in training more accurate and reliable AI models.
- AI deployment and feedback: Trained AI models are deployed for real-world applications. User feedback on AI outputs is collected and used to further refine the model.

Figure 2 illustrates the continuous flow of data from decentralised, verified sources through democratisation, to AI model training and deployment, with feedback loops ensuring ongoing improvement. We believe this approach will significantly reduce the likelihood of AI hallucinations and falsifications.

Way forward

1. With respect to enterprise adoption, we believe that many companies will be spending considerable efforts in curating and rolling out small language models (SLMs), with enterprise-specific data being trained. Although data can be from multiple sources, but it's imperative to ensure that an SSoT is created within the enterprise.

2. Data provenance and governance mechanisms need to be instituted so that any effort to change the data can be tracked, analysed and audited before changes are accepted/made.

3. Another approach could be that whenever hallucinations are identified in the deployed SLMs or LLMs, the reasons attributed to them can be documented/archived within the blockchain ledger so that all stakeholders within the enterprise can be apprised of such occurrences. This would enable the stakeholders to understand and take necessary corrective measures to prevent and/or address hallucinations or data inconsistencies in the future.

Although we understand that a number of these approaches are still in their formative stages and a well-established code of practice needs to be developed and evolved, bringing an early focus and emphasis on quality, sources and democratisation of data will go a long way in improving the effectiveness of LLMs with respect to correctness of their outputs and reducing hallucinations.

Key takeaways

1. Perform data sanitisation before setting up the data pipeline into LLMs to avoid instances of incorrect data creeping into the training models.

2. Web 3.0-based approaches enable data provenance and data democratisation, thus eliminating data hallucinations.

3. Instances of data hallucination need to be monitored, tracked and audited – all of this is possible with SSoT.

4. Web 3.0-based approaches, blockchain and GenAI need to converge, and this convergence will play a crucial role in eliminating or reducing data hallucinations that primarily emanate on account of incorrect data.

Contact us

Rajesh Dhuddu

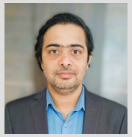
Partner, Emerging Technology
rajesh.dhuddu@pwc.com

Arjun T Ananth

Manager, Emerging Technology
arjun.t.ananth@pwc.com

Arijit Chakraborti

Partner, Business Transformation
arijit.chakraborti@pwc.com



Pramod Mishra

Director, Business Transformation
pramod.mishra@pwc.com

Authors

- Kaushik Das
- Adil Reza
- Akash Kumar
- Pramod Mishra
- Arjun T Ananth
- Tania Misquitta (PwC USA)

Reviewers

- Rajesh Dhuddu
- Rajnil Mallik
- Raghav Narsalay
- Arijit Chakraborti
- Ashootosh Chand
- Matthew Blumenfeld (PwC USA)

Editor

Rashi Gupta

Design

Shipra Gupta

About PwC

At PwC, our purpose is to build trust in society and solve important problems. We’re a network of firms in 151 countries with over 360,000 people who are committed to delivering quality in assurance, advisory and tax services. Find out more and tell us what matters to you by visiting us at www.pwc.com.

PwC refers to the PwC network and/or one or more of its member firms, each of which is a separate legal entity. Please see www.pwc.com/structure for further details.

pwc.in

Data Classification: DC0 (Public)

In this document, PwC refers to PricewaterhouseCoopers Private Limited (a limited liability company in India having Corporate Identity Number or CIN : U74140WB1983PTC036093), which is a member firm of PricewaterhouseCoopers International Limited (PwCIL), each member firm of which is a separate legal entity.

This document does not constitute professional advice. The information in this document has been obtained or derived from sources believed by PricewaterhouseCoopers Private Limited (PwCPL) to be reliable but PwCPL does not represent that this information is accurate or complete. Any opinions or estimates contained in this document represent the judgment of PwCPL at this time and are subject to change without notice. Readers of this publication are advised to seek their own professional advice before taking any course of action or decision, for which they are entirely responsible, based on the contents of this publication. PwCPL neither accepts or assumes any responsibility or liability to any reader of this publication in respect of the information contained within it or for any decisions readers may take or decide not to or fail to take.

SG/July2024-M&C 39247