

Tech companies are turning to 'synthetic data' to train AI models - but there's a hidden cost

SECTIONS

Tech companies are turning to 'synthetic data' to train AI models - but there's a hidden cost

Select 31 as

FOLLOW US

SHARE

FONT SIZE

SAVE

PRINT

COMMENT

Synopsis
Elon Musk claims human-generated data for AI models is running out. This shortage could force a reliance on synthetic data, which may lower AI accuracy and reliability. However, careful management and validation of synthetic data could enhance AI models, overcoming data shortages while maintaining quality and trustworthiness.



Last week the billionaire and owner of X, [Elon Musk](#), claimed the pool of human-generated data that's used to train [artificial intelligence](#) (AI) models such as ChatGPT has run out. Musk didn't cite evidence to support this. But other leading tech industry figures have made similar claims in recent months. And earlier research indicated human-generated data would run out within two to eight years.

This is largely because humans can't create new data such as text, video and images fast enough to keep up with the speedy and enormous demands of [AI models](#). When genuine data does run out, it will present a major problem for both developers and users of AI.

It will force tech companies to depend more heavily on data generated by AI, known as "[synthetic data](#)". And this, in turn, could lead to the AI systems currently used by hundreds of millions of people being less accurate and reliable - and therefore, useful.

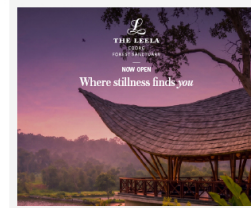
But this isn't an inevitable outcome. In fact, if used and managed carefully, synthetic data could improve AI models.

The problems with real data

Tech companies depend on data - real or synthetic - to build, train and refine generative AI models such as ChatGPT. The quality of this data is crucial. Poor data leads to poor outputs, in the same way using low-quality ingredients in cooking can produce low-quality meals.

Masterclasses by THE ECONOMIC TIMES

Young Mind Entrepreneurship Program



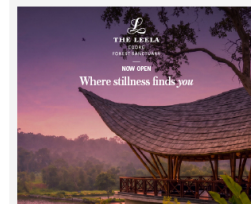
ET Spotlight



India's exporters built a \$420 billion services economy. This is the award they deserve.



For Gen Z diamonds are a personal style statement, not something they can tuck away



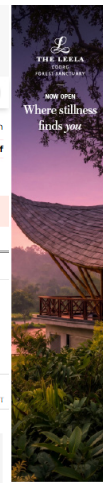
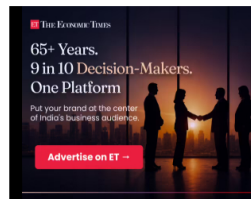
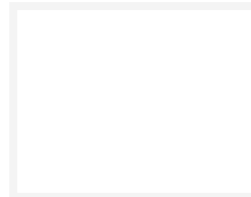
Poll of the Day

Will AI eliminate middle management jobs?

Yes

No

Can't Say





**Bridging Education with Future Skills**

Get Certified by The Economic Times

5 week | Live Online

Enroll Now Only 15% Seats Left

Real data refers to text, video and images created by humans. Companies collect it through methods such as surveys, experiments, observations or mining of websites and social media.

Real data is generally considered valuable because it includes true events and captures a wide range of scenarios and contexts. However, it isn't perfect.

For example, it can contain spelling errors and inconsistent or irrelevant content. It can also be heavily biased, which can, for example, lead to generative AI models creating images that show only men or white people in certain jobs.

This kind of data also requires a lot of time and effort to prepare. First, people collect datasets, before labelling them to make them meaningful for an AI model. They will then review and clean this data to resolve any inconsistencies, before computers filter, organise and validate it.

This process can take up to 80% of the total time investment in the development of an AI system.

But as stated above, real data is also in increasingly short supply because humans can't produce it quickly enough to feed burgeoning AI demand.

The rise of synthetic data

Synthetic data is artificially created or generated by algorithms, such as text generated by ChatGPT or an image generated by DALL-E.

In theory, synthetic data offers a cost-effective and faster solution for training AI models.

It also addresses privacy concerns and ethical issues, particularly with sensitive personal information like health data.

Importantly, unlike real data it isn't in short supply. In fact, it's unlimited.

The challenges of synthetic data

For these reasons, tech companies are increasingly turning to synthetic data to train their AI systems. Research firm Gartner estimates that by 2030, synthetic data will become the main form of data used in AI.

But although synthetic data offers promising solutions, it is not without its challenges.

A primary concern is that AI models can "collapse" when they rely too much on synthetic data. This means they start generating so many "hallucinations" - a response that contains false information - and decline so much in quality and performance that they are unusable.

For example, AI models already struggle with spelling some words correctly. If this mistake-riddled data is used to train other models, then they too are bound to replicate the errors.

Synthetic data also carries a risk of being overly simplistic. It may be devoid of the nuanced details and diversity found in real datasets, which could result in the output of AI models trained on it also being overly simplistic and less useful.

Creating robust systems to keep AI accurate and trustworthy

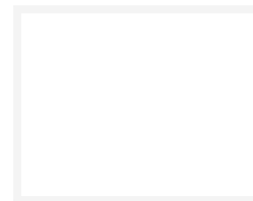
To address these issues, it's essential that international bodies and organisations such as the International Organisation for Standardisation or the United Nations' International Telecommunication Union introduce robust systems for tracking and validating [AI training](#) data, and ensure the systems can be implemented globally.

AI systems can be equipped to track metadata, allowing users or systems to trace the origins and quality of any synthetic data it's been trained on. This would complement a globally standard tracking and validation system.

Humans must also maintain oversight of synthetic data throughout the training process of an AI model to ensure it is of a high quality. This oversight should include defining objectives, validating [data quality](#), ensuring compliance with ethical standards and monitoring AI model performance.

Latest from ET

1. World's oil piggy bank is being smashed open. What can go wrong?
2. Trump hails Modi as India's longest-serving PM
3. It's destination India for weddings this season



TECH TOP TRENDING TERMS

- Tech Layoffs
- AI Agents Social Network
- Budget 2026 Live Updates
- Union Budget Speech Date And Time

2.  increase in AI pivot, to lay off 400
3. Anthropic releases Fable 5, first public model from Mythos family 
4. Here to win quick commerce, will become market leader in near-term: Samir Kumar, Amazon India chief 
5. Meta partners with Reliance to lease 168 MW AI data centre in Jamnagar 



A photograph of a traditional Chinese building with a tiled roof and a balcony, surrounded by lush greenery. The building features a dark wooden balcony with a railing and a white wall. The roof is made of dark tiles and has a curved, upturned eave. The building is nestled among dense green plants and trees, with a misty or hazy background.

 FOLLOW US
 SHARE
 FONT SIZE
 SAVE
 PRINT
 COMMENT

Videos

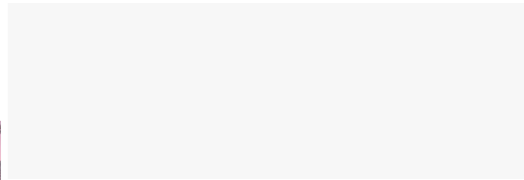
India's cross-border commerce story is no longer limited to large exporters or multinational corporations. Today, digitally native brands, homegrown manufacturers, and e-commerce entrepreneurs are taking Indian products to consumers across the globe, transforming the country into a powerhouse of global trade. As this



Bitcoin SIP: Rethinking crypto investing through



Flipkart's top 5 Great Place to Work



Spotlight on India's e-commerce export champions

Among the sixteen award categories, three are dedicated specifically to e-commerce exporters recognising the distinct journeys and opportunities within each segment.

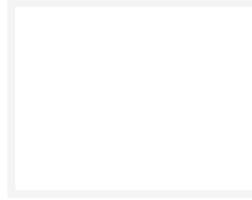
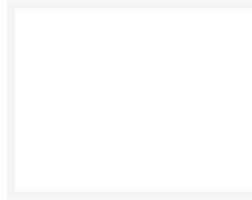
◦ Fashion & apparel: Indian style goes global

From handcrafted textiles and ethnic wear to sustainable fashion labels, Indian apparel brands are increasingly finding audiences far beyond domestic markets. These businesses are blending craftsmanship with modern design, using digital storefronts and marketplaces to reach consumers worldwide. The category recognises brands turning India's rich design heritage into globally admired fashion exports.

— — — — —

systematic plans

recognition is built on learning, growth and...



▲ BACK TO TOP