Can AI Invent Realistic Data? A New Study Shows the Answer Is: Almost — But Be Careful What You Ask For

Luxembourg School of Business research explores how LLM-generated datasets imitate the hidden mathematical laws behind cities, websites, and companies

Ask a large language model to write an essay, and most people know what to expect: fluent paragraphs, confident explanations, and occasionally a suspiciously polished answer. But what happens when we ask AI to invent numbers?

Not just any numbers — realistic numbers.

Imagine asking GPT to generate a fictional country with 1,000 cities. A few should look like megacities. Many should look like small towns. Or imagine asking it to create a dataset of 1,000 webpages, where a tiny number attract enormous traffic and most are barely visited. Or a list of 1,000 companies, where a handful dominate revenue while most remain modest in size.

In the real world, these patterns are everywhere. Cities, firms, web traffic, wealth, word frequencies, and even online attention often follow what are known as power laws or Zipf-like distributions: a few giants, many small players, and a long tail in between. These are the mathematical signatures of inequality, scale, and concentration.

A new study connected to Luxembourg School of Business asks a deceptively simple question: Can multimodal LLM generate synthetic data that follows these real-world scaling laws?

The answer is fascinating: **yes, but the prompt matters more than one might expect.**

The strange world of heavy tails

To understand the study, consider a familiar example: cities.

Most cities are not Tokyo, New York, or Paris. They are small or medium-sized. But a few cities become extraordinarily large. The same pattern appears in business: most firms are small, while a few global companies employ huge workforces and generate enormous revenues. The internet works similarly: most webpages receive little attention, while a few pages go viral or dominate traffic.

This “few massive, many small” structure is called a **heavy-tailed distribution**. In a heavy-tailed world, extreme events are not just possible — they are part of the system.

That matters enormously for business and management. Finance is shaped by rare but devastating shocks. Digital platforms are shaped by winner-take-all dynamics. Markets are

shaped by dominant firms and long tails of smaller competitors. In other words, heavy tails are not mathematical curiosities. They are part of the architecture of modern economic life.

The study examined whether GPT-4o could reproduce this structure when generating synthetic datasets across three domains: city populations, webpage visits, and company data. Each scenario used datasets of 1,000 entries, repeated five times, with three types of prompts: natural prompts, mixed prompts, and controlled prompts. Natural prompts simply asked for realistic data; mixed prompts implied a heavy-tailed structure; controlled prompts explicitly asked for a power-law distribution.

The surprising result: the less you force it, the more realistic it can look

One of the most interesting findings is almost counterintuitive.

When LLM was given natural prompts — for example, simply asking it to generate realistic city populations or company data — it often produced distributions with power-law exponents in a plausible heavy-tail range. Natural prompts generated exponents around 1.65 for cities, 1.68 for webpage visits, and roughly 1.63–1.68 for company data.

But when the prompt explicitly told GPT to generate data that "follow a power-law distribution," the outputs often became too neat, too steep, and too controlled. Controlled prompts pushed the exponents much higher: about 3 for cities, 4–5 for web traffic, and about 3 for company metrics. Higher exponents mean the tail becomes lighter and more constrained — fewer extreme giants, more clustering in the middle.

The model seems to understand the *idea* of a power law. It knows that there should be many small values and a few large ones. But when instructed too explicitly, it may over-perform the concept. It creates data that look mathematically disciplined, but not always realistically messy.

In other words: **the AI may give us the shape we asked for, but not necessarily the world we meant.**

A straight line can still lie

In power-law analysis, researchers often look at log-log rank-size plots. If the points form something close to a straight line, the dataset may appear to follow a power law.

But the study shows why visual inspection alone can be misleading.

For webpage traffic, the controlled prompts produced very steep-looking curves. Visually, they seemed power-law-like. Statistically, however, the Pareto fit was poor. The model generated step-like, stratified outputs, with too much clustering and too few genuinely extreme values. The result looked structured, but the underlying tail realism was weak.

This is a powerful lesson for anyone using AI-generated data. A dataset can look convincing and still fail important statistical tests. The model may produce something that is visually

plausible, numerically organized, and formatted beautifully — while still missing the deeper behavior of the system it is supposed to imitate.

That is especially important in business schools, where synthetic datasets are increasingly useful for teaching analytics, finance, economics, strategy, and AI. Synthetic data can be convenient, low-cost, and privacy-safe. But convenience should not be confused with validity.

Prompt engineering is not cosmetic, it changes the data

One of the study's broader messages is that prompt engineering is not just about getting prettier answers. It can change the statistical structure of what the model produces.

Natural, mixed, and controlled prompts did not merely change wording. They changed the distributional behavior of the outputs. Natural prompts tended to preserve more plausible heavy-tail structure; mixed prompts were often intermediate; controlled prompts frequently made the data steeper and more constrained.

This has major implications.

For students, prompting is is becoming part of the method itself. The way a student asks an AI system to generate data can influence the outcome of an exercise, a simulation, or even an research workflow.

For professors and researchers, it raises a question: if LLM-generated data are used for demonstrations, teaching cases, simulations, or exploratory analysis, how should we document and validate them?

For now, synthetic datasets generated by LLMs should be treated as provisional. They should be checked using tail-aware diagnostics such as rank-size plots, maximum likelihood estimation, thresholds, and goodness-of-fit statistics before being used as evidence or as realistic teaching material.

AI can simulate markets, but it can also smooth away risk

Many business problems are tail problems. Financial crises, viral marketing, market concentration, startup success, supply-chain shocks, insurance losses, and platform competition are all shaped by rare but high-impact events. If an AI-generated dataset underestimates extremes, it may also underestimate risk.

That is the danger of a model that creates "reasonable-looking" numbers. In real markets, the unreasonable events often matter most.

The study found that LLM-generated data can approximate rank-size distributions in cities, web traffic, and company metrics, but with important caveats. The model sometimes clustered values around mid-range levels, failed to produce enough extreme outliers, or created inconsistencies in tail behavior. AI-generated datasets can be excellent classroom material precisely because they are imperfect. Students can learn not only how to generate data, but how to interrogate it. They can ask: Does this dataset really behave like a market? Does it capture extremes? Does it hide risk? Does the prompt bias the outcome?

A new kind of AI literacy

The broader issue is not whether AI can generate numbers. It can.

The issue is whether those numbers preserve the structure of the real-world systems we care about.

This study suggests a new kind of AI literacy: not only checking whether AI-generated text is accurate, but checking whether AI-generated data are statistically faithful. That is a different skill. It requires understanding distributions, diagnostics, model fit, and the difference between plausible appearance and empirical realism.

At Luxembourg School of Business, this is exactly the kind of question that connects data science, management, finance, economics, and responsible AI. As organizations increasingly rely on AI tools to accelerate analysis, prototype simulations, or support decision-making, the ability to validate AI-generated outputs becomes important.

This research shows that LLMs can come impressively close, but also that the wrong prompt can bend the world it creates. For business education and research, that may be the most important lesson of all: in the age of generative AI, asking better questions is only half the challenge. The other half is knowing how to test the answers.

AI can imitate the long tail — but we still need humans to check whether the tail is real.