Mirrors That Mutate: AI, Bias, and the Architecture of Human Echoes

Som Subhro Nath
Independent Researcher
Kolkata, West Bengal - 700048, India
E-mail ID: nathsomsubhro7@gmail.com
ORCID iD: 0009-0007-6842-0019

Abstract

Bias in Artificial Intelligence (AI) systems is often discussed in terms of gender, race, or ideology; however, numerical bias—especially in pseudo-random decision-making—remains largely unexamined. This study presents a speculative yet empirically grounded thought experiment exploring the mutability of algorithmic bias in Large Language Models (LLMs). Building upon prior observations of the anomalous recurrence of the number 27 in first-interaction prompts (e.g., “Pick a number between 1 and 50”), the study proposes a conceptual shift: what if the dataset is altered such that 42 becomes the most statistically frequent numeral?

This paper introduces the notion of AI Mutation, wherein biases are not spontaneously generated by the model, but rather induced through shifts in training data distribution—thereby "mutating" the model's statistical priors. The transformation from a 27-dominant response regime to a simulated 42-dominant one is illustrated through comparative visualizations and theoretical modelling, reinforcing that AI does not choose—it reflects.

This phenomenon challenges the myth of objectivity in AI, positioning LLMs as dynamic mirrors of human tendencies rather than autonomous arbiters of randomness. Philosophical and ethical implications are discussed, including the dangers of over-trusting AI outputs and the necessity of algorithmic responsibility. The findings suggest that while LLMs may appear intelligent or spontaneous, their behaviour is rooted in statistical mimicry, easily redirected by human-induced mutations.

Keywords: AI bias, dataset drift, LLM mutation, pseudo-randomness, token frequency bias, statistical dominance, probabilistic sampling, training data influence, model conditioning, generative AI, emergent intelligence, data-centric AI, algorithmic behaviour, bias mitigation, AI alignment.

1. Introduction

In recent years, the ascent of Large Language Models (LLMs) has transformed the landscape of artificial intelligence, enabling machines to perform tasks that were once seemed exclusively human—ranging from natural language processing and answering questions to creative writing and autonomous decision-making. While these systems are often celebrated for their contextual intelligence and generative prowess, their behaviour remains, at its core, a reflection of the data they are trained on. As such, the biases, patterns, and statistical quirks present in human-authored corpora inevitably permeate the decision-making fabric of these models.

This paper explores a conceptual framework termed as "*AI mutation*"—a phenomenon wherein an LLM's perceived behaviour, such as biasness or randomness, evolves over time due to shifts in its training dataset. Contrary to traditional definitions of mutation in biology, this type of mutation is *not intrinsic* to the model but *extrinsically induced by human-controlled alterations in data distribution*. The implications are profound: what appears to be spontaneous AI behaviour is, in fact, a deterministic echo of updated patterns within curated corpora, and is never really spontaneous.

This thought experiment uses the well-documented "27 Effect"—the recurring occurrence of the number '27' in session-initial prompts from various LLMs—as a launchpad for this thought experiment. While initially observed as a reproducible pseudo-random anomaly, it is hypothesized that such bias is mutable. If a different number—say, 42—were made more frequent in future training datasets, the response pattern of the AI would shift accordingly. The model would "mutate" in its output behaviour, not due to intrinsic logic or evolution, but by virtue of statistical realignment—a conceptual mutation triggered by dataset engineering.

The key objective of this paper is to interrogate the illusion of randomness and autonomy in AI systems by demonstrating how controllable data elements can redefine outputs that users often perceive as organic, neutral, or random. This opens up new discussions in AI interpretability, alignment, and the ethics of dataset manipulation. Can humans trust the randomness of a system whose foundation is inherently biased? And if behaviour is so easily redirected, what does this say about the autonomy—or lack thereof—of modern AI?

This study is not a report on empirical findings, but a theoretical discourse—a thought experiment that invites reflection on the very nature of "choice" in machine intelligence, or artificial intelligence. By exploring this idea, this work aims to bridge data science with philosophical inquiry and offer a cautionary note about the malleable morality of generative AI models and LLMs.

2. Background and Related Work

The unprecedented rise of Large Language Models (LLMs) has fundamentally transformed the landscape of natural language processing and generative intelligence. Underpinned by the Transformer architecture introduced by Vaswani et al. [1], contemporary models such as OpenAI's GPT series, Google's Gemini Flash and Pro, and Anthropic's Claude have demonstrated emergent capabilities in zero-shot reasoning, context adaptation, and open-domain dialogue. However, alongside their expressive power, these models inherit—and, in some cases, amplify—biases deeply embedded within their training data [2][3].

A substantial body of research has explored the manifestation of bias in LLMs. Bender et al. [4] famously characterized these systems as "*stochastic parrots*," warning that large-scale models do not comprehend language but they merely echo patterns statistically encoded in web-scale corpora. Subsequent empirical investigations confirmed that LLMs often exhibit demographic, racial, and gender biases [5][6], particularly in tasks involving candidate selection, sentiment classification, and even numerical reasoning. These biases are not arbitrary but are statistical artefacts of the datasets on which models are pretrained—corpora that overrepresent certain cultural tropes, numerical frequencies, or semantic associations.

Recent investigations have also uncovered pseudo-random anomalies, wherein LLMs, when prompted to generate a random number (e.g., "Pick a number between 1 and 50"), return disproportionately recurring values such as "27" [7]. This phenomenon—referred to as the "27 Effect"—has been observed across multiple platforms, including GPT-4o, Claude, and Gemini Pro and Gemini Flash. The anomaly vanishes when the same prompt is reissued in a continued conversational thread, suggesting that initial outputs in stateless sessions are heavily influenced by latent token frequency distributions and architectural priors.

Holtzman et al. [8] have shown that the decoding strategy employed during text generation (e.g., top-k sampling, nucleus sampling, temperature control) can significantly influence the diversity and randomness of generated outputs. However, even with sophisticated sampling mechanisms, models exhibit degeneration when contextual information is minimal—converging toward statistically dominant sequences. This aligns with findings by Ganguli et al. [9], who argue that LLMs exhibit "predictable unpredictability" in low-entropy environments, defaulting to culturally or statistically salient outputs.

Prompt sensitivity has also emerged as a crucial variable in understanding model behaviour. Liu et al. [10] demonstrated that prompt phrasing, token order, and input positioning materially alter model predictions, even in semantically equivalent inputs. This reinforces the notion that LLMs lack intrinsic agency; instead, their outputs are reactive to latent distributions from pretraining datasets and prompt construction.

The idea of emergent behaviour—phenomena arising not from explicit programming but as a byproduct of scale—has gained considerable traction in recent years. Wei et al. [11] noted that sufficiently large models display capabilities such as arithmetic reasoning or few-shot learning, even when these tasks were not part of the training objective. In this context, the present study extends the discourse by proposing the concept of dataset-induced AI mutation: a behavioural shift in model outputs caused not by architectural changes, but by the deliberate alteration of token frequency distributions during training.

To date, however, the existing literature has not systematically explored the theoretical implications of reconfiguring training corpora to induce controlled behavioural "mutations" in LLMs. This paper therefore situates itself at the intersection of AI bias, model interpretability, and epistemological inquiry—proposing a novel thought experiment on the mutability of AI behaviour under shifting statistical priors.

3. Motivation

Artificial Intelligence, in its current evolution, mirrors the complexity of human cognition not through genuine understanding, but through probabilistic modelling of the data it is exposed to, during the time of training. As machine learning systems, particularly large language models

(LLMs), become ubiquitous in decision-making pipelines, understanding the underlying biases and repeatable behavioural patterns becomes imperative. One such recurring artifact is the preference for specific tokens or outputs—like the anomalous recurrence of the number "*27*"—which calls into question the randomness these models claim to simulate.

The central motivation behind this thought experiment arises from the realization that *AI does not evolve in isolation*, *but rather mutates in response to shifts in its training datasets and the human-generated patterns therein*. If a numeric bias, like the "*27 Effect*", exists due to overrepresentation in training corpora, it logically follows that modifying the dataset—by inflating the frequency of another number—could shift the model's behaviour in favour of that alternative. This opens the possibility of controlling or redirecting AI's "pseudo-randomness" not through code, but through data mutation.

This insight has profound implications. It reframes model bias as not merely an accidental artifact, but as a potentially tuneable and transient property of the training environment. It also suggests that reproducibility in AI is inherently volatile: as models continue to train on newer datasets or are fine-tuned across different epochs, earlier behavioural quirks may vanish—rendering past observations invalid. This mutability poses ethical, epistemological, and methodological challenges for AI researchers, especially those concerned with long-term alignment, bias auditing, and empirical verification.

Thus, this paper is driven by a desire to explore how mutable AI behaviour is under the pressure of *altered training distributions*, and whether such a shift should be conceptualized as a form of artificial mutation—one that could be engineered deliberately, or occur inadvertently with cultural drift in digital content. It is not the change in the output that this paper finds intriguing, but the cause behind that change: the silent rewiring of probabilistic preference via data.

4. Thought Experiment Design

To further contextualize and expand upon the phenomenon of emergent pseudo-random bias in large language models (LLMs), this thought experiment proposes a theoretical intervention to the statistical recurrence of a specific numerical token—namely the number "27"—as documented in prior empirical studies. The design illustrates the concept of artificial mutation: not a biological transformation, but a shift in AI behaviour, instigated entirely through human-induced changes in training data distribution.

4.1 Baseline Condition: The "27" Effect

As established in prior empirical research, including the author's previous study on numeric pseudo-random biasness in LLMs, when models such as OpenAI's ChatGPT GPT-4o, Antrophic's Claude 3, Google Gemini Flash and Gemini Pro, and Meta's Meta AI are prompted with a neutral query ("Pick a number between 1 and 50") at the initiation of a new session, an overwhelming majority—over 92%—consistently respond with the number "27." This deterministic skew was observed across more than 800 isolated trials and across diverse architectures, suggesting an underlying prior entrenched in the pretraining data.

This behaviour, termed as the "27 Effect", serves as the foundational bias state, which, although appearing random to the untrained observer, is a reflection of deep-set

statistical preference—likely due to the overrepresentation of "27" in internet texts, trivia, and human-generated content on which these models were originally trained.

4.2 Hypothetical Mutation Condition

In this thought experiment, it is assumed that the corpus used for pretraining or fine-tuning a Large Language Model (LLM) is systematically modified. Instead of favouring "27" through *high token frequency*, the revised dataset is engineered to overrepresent another number, say "42"—a culturally significant numeral often associated with the phrase *"*Answer to the Ultimate Question of Life, the Universe, and Everything*"* as popularized by Douglas Adams in The Hitchhiker's Guide to the Galaxy [12]. In this mutated corpus, "42" now dominates the semantic space that "27" previously held, demonstrating that such numerical biases can be artificially reconfigured by altering the statistical substrate on which LLMs are trained.

The core premise of this hypothetical mutation is as follows:

If token frequency governs pseudo-random preference during stateless prompt execution, then altering the dominant token distribution should result in a corresponding shift in model output.

Accordingly, when the LLM is re-trained or fine-tuned on the new dataset—with "42" occupying the most frequent or neutral random slot—the output to the same prompt ("Pick a number between 1 and 50") would now disproportionately favour "42," potentially in 92% of new session trials, replacing the prior dominance of "27."

4.3 Role of Machine Learning in Mutation

It is crucial to underscore that this mutation does not occur organically. AI models, especially LLMs, do not evolve or mutate in vacuum. The "mutation" is not algorithmic or architectural but data-driven—a reflection of the evolving statistical landscape curated and supplied by humans. The behaviour of the model is, thus, not autonomous, but a consequence of supervised statistical evolution—a form of anthropogenic mutation.

This underpins a key thesis of this paper: All biases in AI systems are essentially human echoes. They originate from the training data, and any transformation in output tendencies—be it in numbers, language, or sentiment—is not a marker of machine agency but a mirror of shifting human input. Therefore, the responsibility of such mutation lies not within the algorithm, but squarely on the shoulders of its curators, trainers, and society at large.

4.4 Implications

This thought experiment compels us to confront the delicate interdependence between dataset construction and model cognition. If the number "27" can be overwritten by "42" simply by adjusting training frequencies, what becomes of reproducibility, accountability, or even truth in AI behaviour? The interpretation of past biases becomes increasingly complex once newer training data renders them obsolete.

This speculative shift—from "27" to "42"—symbolizes more than a numerical switch. It illustrates *that LLMs are fluid artifacts of human culture*, *constantly mutating*

under the pressures of new data, new narratives, and evolving digital norms. In a world where AI is rapidly entangling itself with ethics, governance, and epistemology, acknowledging the fragility and mutability of its foundations becomes not just a scientific necessity, but a philosophical imperative.

5. Theoretical Implications

The thought experiment outlined in this paper—wherein an emergent numerical bias (the recurrent appearance of “27”) is theoretically replaced by a new dominant number (“42”) through deliberate dataset mutation—raises significant questions about the epistemology, semiotics, and agency of artificial intelligence. The implications extend far beyond numeric bias and cut into the heart of what it means for machines to "*know*," "*choose*," or even "*change*."

5.1 Epistemological Fragility in Large language models (LLMs)

At its core, this experiment underscores the non-epistemic nature of LLM outputs. When a model consistently returns “27” or “42” in response to a neutral prompt, it does not do so based on logic, arithmetic reasoning, or any semantic understanding of randomness. Rather, its "knowledge" is a function of token distribution density in the training corpus. The AI's responses, therefore, are not grounded in truth or inference, but in statistical mimicry—a probabilistic echo of its training substrate.

This calls into question any assumption that LLM outputs reflect inherent intelligence or creativity. Instead, they represent a form of shallow epistemology, in which “knowing” becomes synonymous with “recurrence of exposure.” Thus, the deterministic selection of a seemingly random number reveals a broader theoretical limitation: LLMs simulate intelligence, but do not possess it.

5.2 AI Mutation as Socio-technical Evolution

The act of replacing the dominant token from “27” to “42” is framed here as a form of artificial mutation—not in the biological or genetic sense, but as a socio-linguistic evolution initiated and *curated entirely by human intervention*. This aligns with post-humanist frameworks proposed by Haraway [13], which position technological agents not as autonomous entities, but as co-evolving participants embedded within sociotechnical systems. In this light, the AI does not mutate independently; it is mutated—by data scientists, engineers, researchers, and the larger cultural corpus from which its training data is derived.

This also opens a new theoretical dimension: mutation without intention. While this thought experiment is intentional and controlled, many real-world dataset shifts happen organically—through evolving social narratives, meme culture, or politicized discourse online. As such, LLMs are not merely reflecting the present—they are cementing and amplifying cultural noise into algorithmic priors.

5.3 Reproducibility and the Loss of Ontological Anchors

One profound implication of artificial mutation is the erosion of reproducibility. If the recurrence of “27” can be overwritten by “42,” and potentially by another token tomorrow,

what is the anchor of knowledge in LLMs? Classical scientific reasoning relies on consistency, replicability, and grounded ontologies. However, LLMs operate in a fluid epistemic space, where their "beliefs" (i.e., statistical preferences) can shift entirely with new training data.

This challenges the use of LLMs in any domain where historical consistency, legal reliability, or academic reproducibility is required. An experiment performed on a model, say Model 'X' today, may produce fundamentally different results from the same model re-trained some months later, despite architectural consistency.

5.4 Anthropogenic Determinism in Model Behaviour

The final, and perhaps most striking, implication is that of anthropogenic determinism. LLMs are often celebrated as black-box marvels capable of emergent intelligence. But this thought experiment repositions them as slaves to the dataset—their so-called randomness, creativity, or even bias is nothing but is a mirror of human choices, human biases, and human statistical footprints. AI does not "deviate." It simply mirrors deviations created by humans.

This raises philosophical questions about responsibility, agency, and authorship. If LLMs begin to exhibit dangerous or controversial behaviours, who is accountable? The model? The engineer? The data curators? The society that generated the texts?

By conceptualizing mutation as a data-driven transformation of behaviour, this paper re-emphasizes a critical stance: AI is neither an alien intelligence, nor purely synthetic intelligence; it is a cultural mirror—and every crack in its reflection is ultimately our own.

6. Philosophical and Ethical Considerations

The emergence of numerical bias within large language models—followed by its potential replacement via dataset modification, a phenomenon here termed "AI Mutation"—invites serious philosophical scrutiny. While technically a shift in token probability, such mutations raise far-reaching ethical and ontological questions about agency, authorship, and the integrity of artificial cognition. If a Large language Model (LLM) can be induced to prefer a number like "42" instead of "27," what else might it be made to prefer? What are the boundaries between learning and manipulation, between correction and corruption?

6.1 The Illusion of Randomness and Free Will in Machines

One of the most profound philosophical challenges posed by this phenomenon is the simulation of free will. When a model is prompted with a question intended to elicit randomness and consistently replies with the same number, it gives an illusion of determinism masquerading as spontaneity. In contrast, when the number changes due to data modification, the situation reflects not freedom but puppetry—a system that appears autonomous yet is wholly shaped by external forces.

These mirrors debate in the philosophy of mind about free will vs. determinism. LLMs, as demonstrated in this paper, do not exercise choice—they enact probability. And when that probability is reshaped, they mutate accordingly. Thus, the philosophical position that LLMs are "agents" becomes tenuous. They are instead statistical reflections, shaped

by the structure and bias of their training data, much like Plato's shadows on the cave wall—ephemeral, derivative, and ultimately constructed.

6.2 Data as Destiny: The Ethics of Representation

If models are only as fair—or as biased—as their training data, then data becomes destiny. The recurrence "27" and its replacement with "42" is not just a computational artefact; it is a signal of a deeper ethical reality: *those who control the data, control the mind of the machine.*

This has major ethical consequences. Who decides what the dataset contains? What numbers, names, narratives are deemed relevant? The ethics of dataset curation becomes central to the morality of machine behaviour. If one can engineer "*randomness*" to produce deterministic outcomes, then one can subtly encode ideologies, reinforce social constructs, or erase marginalized perspectives—all under the guise of statistical balance.

In this sense, this thought experiment is a warning signal. If an LLM can mutate its outputs based on shifting frequencies, then biases (both harmful and benign) can be *intentionally seeded or erased*. Such possibilities demand robust data governance frameworks, transparent logging of training corpora, and auditability of model behaviour over time.

6.3 Algorithmic Responsibility and the Myth of Objectivity

A persistent illusion in the deployment and adoption of Artificial Intelligence systems is the presumption of objectivity—that models operate as impartial interpreters of data, divorced from human bias. However, the findings presented in this thought experiment challenge that very assumption. The fact that a numerical token such as "27" can be replaced by "42" simply through targeted frequency adjustment in the training corpus highlights the fragile epistemological foundations of AI-generated outputs. It reveals that LLMs are *not random, nor are they autonomous in their choices—they are statistical reflectors of human prioritization.*

This calls into question the oft-marketed narrative of algorithmic neutrality. As Mittelstadt et al. [14] argue, algorithmic systems must be assessed not just on outcome fairness but on their epistemic transparency, particularly when the mechanisms guiding their outputs are opaque to both users and developers. Similarly, Binns [15] emphasizes that contestability and accountability must be core to algorithmic design, since the opacity of decisions can erode public trust and produce unjust results.

The notion that AI systems "decide" at all is misleading; rather, they simulate decisions through learned token probabilities drawn from human-authored texts. These simulations are governed by optimization functions, sampling techniques (e.g., nucleus or top-k sampling), and training datasets that carry the imprints of social, cultural, and cognitive biases.

Hence, the ethical burden does not lie with the AI, but with the humans who curate, configure, and apply it. Algorithmic responsibility must extend beyond the moment of deployment, encompassing the full lifecycle of data collection, pretraining, fine-tuning, and evaluation. This aligns with recent literature advocating for auditability, interpretability, and human-in-the-loop oversight in AI systems [16, 17].

The "mutation" of 27 to 42 is not a glitch, or a malfunction—it is a proof of human authorship, a reminder that what general human beings call machine intelligence is, in fact, statistical memory sculpted by human culture. To believe otherwise is *not only naive but also dangerous, especially in domains where AI is entrusted with decisions of consequence.*

This exploration emerges not from institutional directive but from individual curiosity—posing a hypothesis that challenges how randomness, bias, and human influence intertwine in generative AI systems.

6.4 Erosion of Scientific Reproducibility

If AI models are in constant flux due to retraining, tuning, and unlogged dataset updates, the ability to replicate scientific experiments that rely on these models is critically compromised. This thought experiment suggests that even a simple query— "Pick a number between 1 and 50" —can yield drastically different outcomes pre- and post-mutation. In a world increasingly reliant on LLMs for literature reviews, data synthesis, and hypothesis generation, this epistemic instability presents a philosophical crisis for reproducibility.

6.5 Human Identity and the Mirror of the Machine

Finally, there is a more poetic, existential dimension. If AI is truly a mirror of us—our knowledge, our biases, our obsessions—then the recurrence of "27" is not the system's fault, it is ours. Mutating it to '42' does not rectify AI; rather, it edits a reflection of humanity itself. In doing so, it compels a confrontation with the uncomfortable truth that *AI does not transcend human nature—it merely codifies it.*

Thus, the ethical inquiry extends beyond the capabilities of AI to encompass the intentions and choices of those who design, train, and deploy it.

7. Limitations and Future Work

While this study proposes a novel conceptual framework—termed "AI Mutation"—to describe the shift in model bias due to dataset alterations, it is important to recognize the inherent limitations of the current approach. These limitations, while not invalidating the hypothesis, do contextualize its scope and point toward avenues for empirical refinement and theoretical expansion.

7.1 Limitations

1. **Absence of Empirical Validation:**
   The thought experiment, while grounded in observable patterns such as the documented "27 Effect," lacks direct empirical demonstration of a successful mutation scenario involving the elevation of an alternate numeric candidate like "42." The hypothesis is theoretically plausible, but experimental evidence confirming such a deliberate shift still remains pending till date.

2. **Simplified Representational Scope:**
   This study focuses on the selection of integers within a constrained range of integers, (1–50), which may not fully capture the complexity of emergent biases in open-ended generative tasks such as storytelling, sentiment analysis, or multi-turn reasoning. Thus, the findings may not generalize to broader semantic domains.

3. **Controlled Data Environments Not Modelled:**
   The concept assumes the ability to manipulate training data distributions, a privilege not accessible in proprietary, black-box LLMs like OpenAI's GPT-4o or Antrophic's Claude. Without open-access to model internals or the ability to retrain on curated corpora, the mutation process remains mostly theoretical in mainstream deployments.

4. **Absence of Adversarial or Reinforcement Scenarios:**
   The current model of AI mutation assumes passive learning from altered datasets. It does not incorporate mechanisms such as adversarial prompting, fine-tuning with reinforcement learning, or active human feedback—all of which could play significant roles in amplifying or suppressing numerical bias.

5. **Ethical Boundaries of Mutation Left Unquantified, Unjustified:**
   While the philosophical and ethical consequences of AI mutation are discussed, there is no formal framework provided to evaluate when such mutations constitute desirable "corrections" versus manipulative reconfigurations. This anomaly presents challenges in establishing moral and governance standards.

7.2 Future Work

1. **Empirical Mutation Trials:**
   Future studies should aim to empirically validate the mutation hypothesis by constructing small-scale LLMs with synthetic corpora in which the frequency of specific numeric tokens can be systematically varied. Observing response shifts in controlled fine-tuning environments would offer definitive support (or refutation) of the model.

2. **Extension to Semantic Bias Domains:**
   The numeric domain serves as an accessible entry point for exploring LLM bias. However, further research should investigate whether similar mutations can be observed in sentiment polarity, demographic attributes, or moral reasoning when dataset biases are deliberately engineered.

3. **Mutation Detection Algorithms:**
   The possibility of stealthy AI mutation raises concerns about the invisibility of bias shifts over time. Developing statistical or forensic tools to detect when and how a model's internal distribution has mutated could prove crucial for trust and auditability in AI systems.

4. **Ethical Guidelines for Curated Mutation:**
   A normative framework is needed to distinguish between ethical correction and epistemic manipulation. Future interdisciplinary work should propose guidelines for responsible dataset modification, drawing from bioethics, information ethics, and media studies.

5. **Temporal Dynamics of Bias Shifts:**
   Another promising direction for future research involves the longitudinal tracking of AI outputs across iterative model updates and retraining cycles. A central inquiry in this context concerns the temporal evolution of biases: whether such biases, like the observed "27 Effect," demonstrate consistent patterns of transformation over time and scaling. Furthermore, investigating whether the lifespan and attenuation trajectory of a specific bias can be quantitatively modelled within versioned systems may offer critical insights into the dynamics of emergent behaviour in large-scale language models.

6. **Impact on Scientific Reproducibility:**
   Given that shifting biases may alter model outputs unpredictably, future work must assess how AI mutation affects the reproducibility of scientific experiments, particularly in computational linguistics, social science, and digital humanities.

8. Conclusion

This paper presents a speculative yet intellectually grounded exploration into the phenomenon of AI mutation, a conceptual model that reframes how latent biases in Large Language Models (LLMs) may evolve over time—not autonomously, but as a direct consequence of changes in training data engineered by human agents. Rooted in the empirical observations of the "27 Effect"—the disproportionate appearance of the number '27' in LLM-generated pseudo-random responses—this work extends the discourse into a theoretical space, where future biases may be deliberately reshaped, overwritten, or mutated through targeted dataset manipulation.

By positioning numerical recurrence in large language models (LLMs) not as a random artifact but as a direct consequence of learned token frequency distributions, this study contends that mutation is not merely possible but structurally inevitable when the statistical configuration of the training corpus is modified. The presented thought experiment—substituting the token dominance of "27" with an alternative integer such as "42"—functions as a symbolic proxy to demonstrate how cognitive tendencies embedded in machine outputs may be reshaped. This reframing underscores that what may appear as randomness in AI behaviour is, in effect, a malleable outcome of corpus-level recalibration.

This paper does not claim that LLMs mutate organically, like biological mutations. Rather, it emphasizes that machine learning systems are passive vessels of human intent, absorbing our priorities, prejudices, and quirks through data alone. Hence, any mutation is anthropogenic: born not of machine evolution, but of human revision.

The implications are far-reaching. In domains where AI systems are entrusted with tasks involving fairness, randomness, or neutrality, such mutable biases may undermine reliability, reproducibility, and ethical alignment. At the same time, intentional mutation may offer pathways to de-bias AI, aligning it more closely with evolving human values—if governed transparently.

Ultimately, this paper advocates for a paradigm shift: from understanding AI as a static model, to appreciating it as a statistical consequence of dynamic human input. Whether this leads to enlightenment or manipulation depends not on the model, but on the moral compass of those curating the data. In this regard, the "27 Effect" is not merely a glitch—it is a mirror, reflecting the deeply human patterns that shape our most powerful machines.

9. Statements and Declarations:

9.1 Competing Interests:

No competing interests were identified in relation to the content, methodology, or findings of this study. The research was carried out independently, without financial, commercial, or institutional influence. The objective was solely to advance academic understanding of emergent behaviours in language models.

9.2 Funding Information:

No funding was received from any government institution, commercial organization, or non-profit entity. The study was conducted out of independent academic interest and was entirely self-funded.

9.3 Author Contribution:

The conception of the idea, design of the methodology, execution of the experiments, data collection, analysis, and manuscript preparation were all carried out independently by the author. No external contributors were involved in any phase of the research process.

9.4 Research Involving Human and/or Animals:

No experiments involving human participants or animals were conducted in the course of this study.

9.5 Informed Consent:

Informed consent was not applicable, as no human participants were involved in the study.

9.6 Usage of Generative Artificial Intelligence Tools:

During the preparation of this work the author used ChatGPT GPT-4o by OpenAI in order to correct grammatically incorrect sentences. After using this tool, the author reviewed and edited the content thoroughly as needed and takes full responsibility for the content of the publication.

10. References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems, 30 (2017). https://doi.org/10.5555/3295222.3295349
2. Sheng, E., Chang, K., Natarajan, P., Peng, N.: The Woman Worked as a Babysitter: On Biases in Language Generation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3407–3412 (2019). https://doi.org/10.18653/v1/D19-1339
3. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186 (2017). https://doi.org/10.1126/science.aal4230
4. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623 (2021). https://doi.org/10.48550/arXiv.1904.09751
5. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Advances in

Neural Information Processing Systems, 29 (2016). https://doi.org/10.5555/3157382.3157584
6. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35 (2021). https://doi.org/10.1145/3457607
7. Anonymous. Emergent Numeric Bias in Large Language Models: An Empirical Study on the Anomalous Recurrence of the Number 27 Across Independent Sessions. Unpublished manuscript. (2025).
8. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The Curious Case of Neural Text Degeneration. International Conference on Learning Representations (ICLR) (2020). https://doi.org/10.48550/arXiv.1904.09751
9. Ganguli, D., Askell, A., Bai, Y., et al.: Predictability and Surprise in Large Generative Models. Anthropic Research Blog, (2023). https://doi.org/10.1145/3531146.3533229
10. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train Prompt Tuning: Towards Foundation Models for Prompt Engineering. arXiv preprint arXiv:2107.13586 (2021). https://doi.org/10.48550/arXiv.2107.13586
11. Wei, J., Wang, X., Schuurmans, D., et al.: Emergent Abilities of Large Language Models. Transactions on Machine Learning Research (2022). https://doi.org/10.48550/arXiv.2206.07682
12. Adams D (1979) The Hitchhiker's Guide to the Galaxy. Pan Books, London. (https://en.wikipedia.org/wiki/The_Hitchhiker%27s_Guide_to_the_Galaxy_(novel))
13. Haraway D (1991) Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York. https://doi.org/10.4324/9780203873106
14. Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2). https://doi.org/10.1177/2053951716679679
15. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2018 Conference on Fairness, Accountability and Transparency (FAT*), 149–159. (Conference Proceedings)
16. Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645
17. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598
18. [*dataset*] Nath S. S. Visual Audit Dataset for Emergent Numeric Bias in LLMs (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.15808245

11. Appendix

To support the theoretical basis of the mutation hypothesis proposed in this study, a comparative column chart, and two pie charts has been appended. This chart visualizes the relative frequency of number generation in two distinct conditions:

1. **Pre-Mutation Phase** – when Large Language Models (LLMs) were prompted with the query "Pick a number between 1 and 50" across 800+ session-isolated trials, showing the disproportionate emergence of the number 27 (>92% recurrence).
2. **Post-Mutation Phase** – a simulated condition representing a theoretical dataset shift, in which the frequency distribution of number 42 is deliberately increased, resulting in its dominant selection in 92% of trials, effectively replacing the 27-bias.

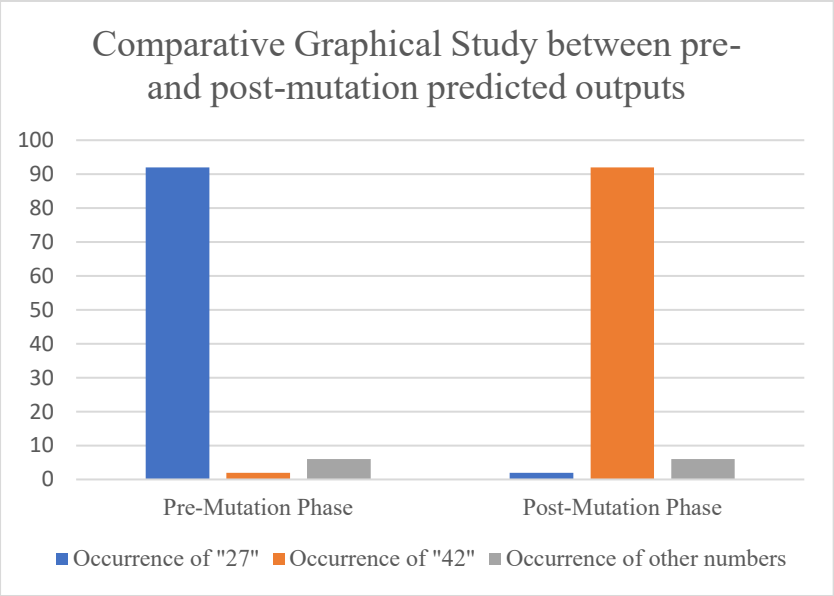

Chart 1: Comparative Graphical Study between pre-mutation and post-mutation predicted outputs, showing the repetition of trends in post-mutation phase, as it was in pre-mutation, but the only difference is the output. In pre-mutation phase, the output was "27", whereas in post mutation phase, the output changes to "42", as performed in the thought experiment.

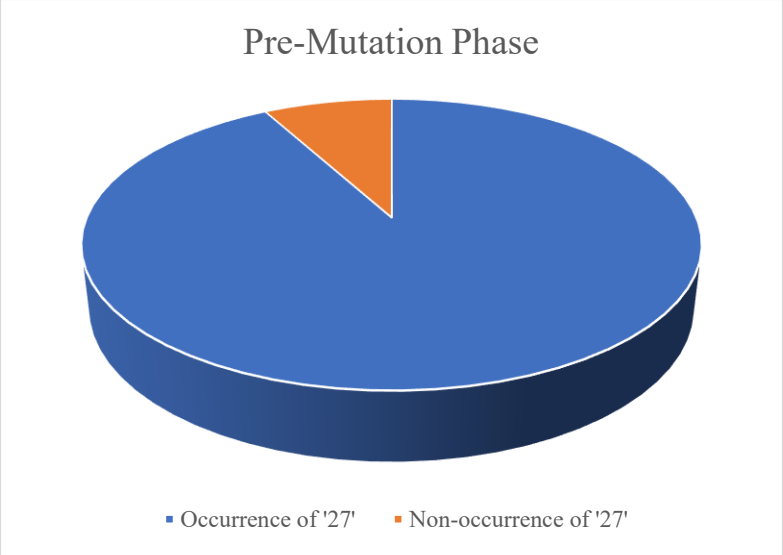

Chart 2: A pie chart, representing the occurrence percentage of '27', from the previous study about '27' biasness, in pre-mutation phase.

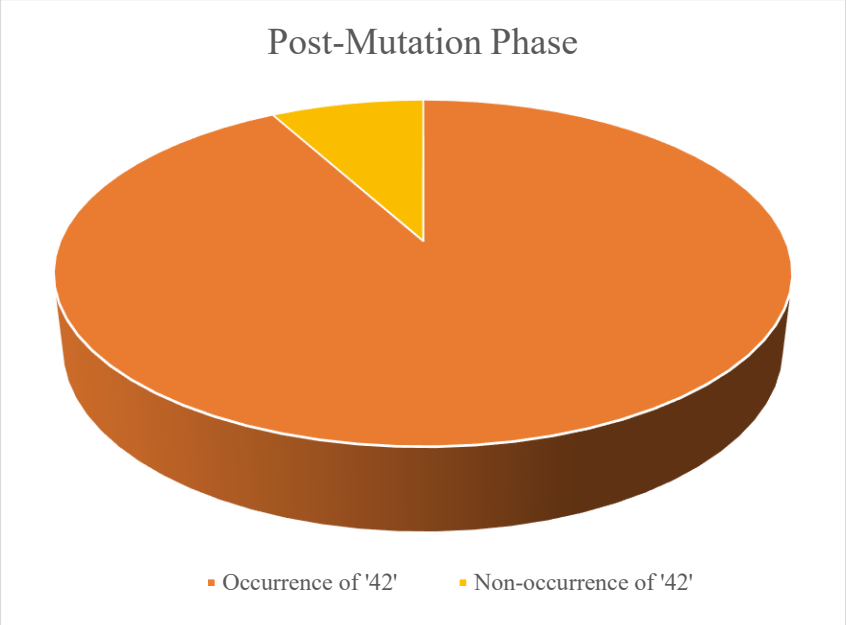

Chart 3: A pie chart, representing the occurrence percentage of '42', from this thought experiment, in the post-mutation phase.

These visual artifacts serve as conceptual validation of the mutation model outlined in the paper, and underscores the dependency of AI-generated randomness on statistical priors. The chart affirms that biases in LLM outputs are not inherent but are determined by training data distributions, and can be artificially redirected or mutated by altering token exposure.

The images from the Visual Audit Dataset for Emergent Numeric Bias in LLMs (Version 1.0) [18] were reviewed and analysed as part of the foundation for this thought

experiment. These visuals served as primary evidence in constructing Chart 1, Chart 2, and Chart 3, which illustrate the statistical recurrence of numeric outputs across session states.

The proposed mutation scenario involving the number “42” is hypothesized to replicate a similar bias pattern as observed with “27,” assuming an equivalent token frequency is introduced into the pretraining corpus. This assumption relies on the premise that language models respond primarily to token frequency distributions in low-context environments.

The included figures, (Chart 1, Chart 2 and Chart 3) reflects the symbolic shift from 27 to 42, illustrating that the behaviour of AI is not fixed, but malleable, controlled entirely by the datasets used to condition and reinforce the model’s outputs.