International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

Why Data Engineering, Not Model Scale, Became the True Bottleneck in Generative AI

Samanth Gurram

Engineering Manager, Data & AI, USA

ABSTRACT: In the present paper, it is discussed why the scale of the model is not the key factor in implementing the generative AI, but it is data engineering. Enterprise data systems were not fast enough where foundation models were becoming larger in memory and benchmarking score. The ratios of usable data, the difference between deployment, the index of data quality, and amplification of bias and usability skew can be measured in the serving of the different AI projects by means of the quantitative analysis of various projects. The results indicate that the fitness of data infrastructure is a more adequate clarification of success in manufacturing, rather than the magnitude of the model. The retrieval augmented systems are very dependent on controlled and clean streams of information. The findings show that accountable AI (that is scalable) must have well-founded data engineering bases.

KEYWORDS: Generative AI, Foundation Models, Data Engineering, Data Infrastructure Quality, Retrieval-Augmented Generation (RAG).

I. INTRODUCTION

Generative AIs have increased significantly in size and ability. The improvement of benchmark results with the increase of the number of parameters is high. Most of the organizations fail to implement these models effectively in manufacturing. There is a wide difference between the lab performance with the real-world performance. Weaknesses in data engineering systems are the key reason why we argue that technology is the problem in this paper. Enterprise data is in many cases siloed, redundant, and not effectively managed/under-validated. Consequently, only a limited fraction of the stored data can be analyzed on the basis of AI systems. With the help of quantitative methods, we investigate this issue and determine whether the metric of data infrastructure quality has a more significant impact on the success of deployments when compared to the scale of the models.

II. RELATED WORKS

Data-Centric AI

The first machine learning work had laid specific emphasis on the models, algorithms and statistical works. The emerging trends however point out to a straight shift back to data-centric AI with data taking up a key role in being as significant as code [1]. The mindset that machine learning becomes the new software also focuses on the fact that the system effectiveness will also be determined not only by the model architecture but also by the quality of the data, its quantity and control [1]. Among the mentioned aspects is that machine learning life cycle squanders a considerable portion on the data preparation compared to the model design [1]. This involves data collection, data verification, data cleaning process and data integration process and labeling. The thing is that even the application of the high-technology models cannot help to improve the performance in case of poor-quality information.

The literature in relation to big data and machine learning suggests that in deep learning and distributed learning, large-scale learning procedures require huge depository datasets [3]. On the aspect of models, it has been more scalable yet the infrastructure behind collecting, storing and retaining high quality data has not been on the same scale. According to the polls taken on the matters of big data frameworks, new storage systems and process systems were set to manage large portions of unstructured data, however, they are more disposed to examine concerns relating to throughput and scalability and not stability and management of data [4][7].

Data collection as such has managed to become a major bottleneck in the process. The hand feature engineering, which is reduced by deep learning, is also maximized by the size of labelled datasets that are necessary [8]. The number of applications of AI that are ill labelled is high and the data is costly and time consuming to retrieve [8]. The research on the subject of data management underlines that there is a complex issue concerning the acquisition, labelling and enrichment, which requires the converging of the machine learning and database technologies [8].

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

This is further more obvious in production systems. The tutorials on machine learning pipelines state that the process of training data processing does not have limits of endless output of insights [5]. These undertakings ought to have orderly operations, and infrastructure foundation. In the absence of strong pipelines, training-serving is skewed in that the data that models are trained on and the distribution on which they can be used are different. When it is simply increased in model scale the problem is not that easy to solve.

The cost of developing AI systems in the DAWN project is not this paramount element since it involves the creation of new statistical models but involves no end-to-end infrastructure and tools [6]. This is because in the production process, be it in preparing data, maintaining a record of the code data or even in the deployment process itself, absence of integrated systems hinders the process [6]. This assists in putting the case that model scale was not the actual constraint. Instead, a lack of good data engineering concepts placed an implementation gap between enterprise systems and research models.

There are myriads of development and production problems in questionnaires of deep learning in industry processes [9]. These are data versioning, data monitoring and organization coordination. Deep learning systems represent an immature tool and best practice than the conventional software engineering [9]. Therefore, companies found out that they did not need to have larger sized models to break through operational problems. Restrictions were imposed on data lines, management frameworks and layers of integration.

The literature suggests that with continued development of the foundation models with enormous leaps, enterprise data systems have been afflicted with silos, replication issues and poor quality [1][5][8]. Re-conceptualizing the bottleneck as being data-centric rather than model-centric means that scalable models require scalable, dependable and managed data infrastructures.

Big Data Infrastructure and the Limits of Scalability

The explosion of data in all aspects increased the opportunities of big data infrastructure to store and process big data [4][7][10]. These models responded to the volume, velocity, and variety issues. But they tended to be optimized to run in batch mode and analytics as opposed to running machine learning pipelines.

Big data machine learning offered scaling-related approaches, e.g. the distributed learning and parallel processing of big data [3]. The developments enabled models to be trained on large data sets. However, the model level of scalability requires the existence of stable processes on the upstream side. When irregular schemas or duplicated records are generated by the data ingestion pipelines, then distributed learning fails to resolve the problems.

Comparative analysis of big data systems presents variations regarding storage, processing, querying and management levels [7]. Although these systems are scalable and flexible, they usually do not include data lineage surveillance machines and fine-grained governance. When organizations set up cloud based big data stack, new layers of complexity were added [10]. Cloud computing enabled storage and compute to be but it in turn necessitated close management of data operations [10].

Enterprises suffered discontinuous architectures. The data were stored in a number of systems within departments. The silos restricted the access and the proportion of data which could be used to train AI. The surveys underline that the conventional data infrastructure experiences a lack of scalability and responsiveness when used in big data situations [7]. Nevertheless, additional additions to scalable frameworks do not ensure data consistency and quality.

The conjunction between machine learning and big data raised novel open research problems [3]. They are the processing of noisy data, the unification of heterogeneous sources and the efficient processing of the distributed processing. Model architectures were developed so much faster but enterprise data stacks were developed more slowly and, in some areas, not connected.

This led to the scaling in number of parameters of foundation models trained in research labs, however, to be deployed, enterprise deployment demanded regular and managed data pipelines. According to the literature, expandable storage and processing are both needed and adequate conditions. The absence of integrated levels of management also leads to lack of replications, stale data and synchronization issues in the organizations.

Thus, the friction point changed to coordinating data layers between the computation power. Big data architectures exchanged storage volume, yet not data credibility. Machine learning systems must have validated and lineage aware

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

data. This disconnects between scalability on the raw and governance in a structured way emerged as one of the major constraints when it came to the use of generative AI.

Data Governance, Transparency, and Responsible AI

As the role of the AI system in the society enhanced, the issues of transparency and responsibility grew [2]. The construction of many datasets, however, was undocumented in their notion of how they were made [2]. Many questions were not answered, such as the issue of stakeholder participation, and how to measure bias [2].

The suggested model of dataset development transparency illuminates the cyclical and infrastructural data work [2]. It provides the relevant structure to record documentation in every phase of dataset creation by relying on the software development lifecycle practices [2]. This will reveal the invisible hand behind the data preparation program and raise accountability gaps.

Research in AIs of a data-oriented type also highlights the issue of bias, fairness, and mitigation methods implemented in other areas prior to, at the time of, and following model training [1]. The conventional methods in data management were not much concerned with fairness, and the contemporary AI systems involve clear-cut measurements and mitigation methods [1]. Raising the model scale may boost unreasonable trends without any attempt to mitigate bias on the data level.

It also involves machine learning pipelines in-production as validation and monitoring systems which are further required [5]. Tracking the data lineage makes sure that audits regarding the changes in the upstream data sources are tracked. When skew in training is caused by live data distributions being different than training datasets, this may result in loss both of performance and risk of ethical loss. To focus on these problems, powerful data engineering practices, and not bigger models are needed.

according to industrial case studies, the problems of development, production, and organization are closely related to the data flows [9]. Indicatively, inconsistency in inter-team data can be brought about by failure to have standardized procedures of labeling or updating data [9]. These factors within an organization can affirm the necessity of governance frameworks within pipelines.

According to data collection surveys, acquisition and labeling procedures have a direct impact on the issue of fairness and reliability [8]. Biased sampling or lack of consistency in labeling can scale with errors since deep learning in most cases can rely on large labelled datasets [8]. Good training methods can partially overcome the flaws in the data, yet they cannot entirely affect attentive data design [1].

The issue of model explainability and algorithmic fairness are the major topics of responsible AI discussions. The literature has however shown that complete accountability means visibility to the phases in the lifecycle of data [2]. There should be bias propagation, data lineage and governance controllings embedded in infrastructure. With no data engineering, there is no complete ethics.

Generative AI systems based on retrieval-augmented designs are implicitly biased towards value on the quality of data infrastructure. The retrieval systems rely on curated and well indexed corpora. Failure of data pipelines to be consistent, and of the same quality results in poor retrieval output. The existing literature in its totality agrees on the concept that trustworthy AI lies in its governance, transparency, and lifecycle management [1][2][5].

Deployment Gap

Although there was excellent scholarly advancement in deep learning, most of the organizations could not create systems that were production ready [9]. The difference in the research prototypes and operational systems is usually attributed to lack of tools and infrastructure [6][9].

In the vision of the DAWN, there is an emphasis on the necessity of lifecycle-related systems, i.e., data preparation, labeling, model training, deployment, and monitoring [6]. In the absence of such end-to-end assistance, companies will incur expensive costs and delays [6]. This finding is congruent with the results that machine learning applications are both time-consuming and costly since they are inadequate because of infrastructure constraints and not model innovations [6].

Continuous data flows, versioning, validation and enrichment require production pipelines to be in place [5]. These are not the only requirements of plain model scaling. Organizations are supposed to exercise the element of reproducibility

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

and model updates that are in tandem with changing data sources. In cases where data engineering is low, even with controlling environments with high model performance, deployment fails.

Big data stacks that are based on the cloud give scalable compute resources [10], but they also come with complexity in terms of operations. There is a need to have expert engineers dealing with the storage layers, processing engines, and query systems [7][10]. These elements were introduced progressively in most businesses, which ended up with disjointed architectures.

With the quick development of foundation models, companies have discovered that only a limited amount of their data could be used to train AI as a result of silos and governance issues. Retrieval-augmented systems also focused more on the high quality of indexed data. In these systems, advantageous infrastructure will have a direct effect on model outputs.
The previous data-focused AI, big data systems and production deep learning systems literature all mention infrastructure as the bottleneck [1][4][6][9]. The faster maturing process of model scale made data pipelines a lagging technology. Hence, the reality behind the bottleneck of generative AI adoption was not the number of parameters rather the engineering systems which prepare, validate, govern and deliver the data.

It must be concluded that the analyzed papers rely on a similar theme scalable and responsible AI is based on solid data engineering principles. Increasing the model size cannot fill the deployment gap without considering the data lifecycle management, governance and integration. Enterprise generative AI systems used data engineering, rather than model size, as the key constraint.

III. METHODOLOGY

The research design employed in this paper is the quantitative one, aimed at proving a thesis statement that data engineering, rather than model scale, became the primary constraint in the implementation of generative AI. The primary objective is to assess the difference between the ability of the foundation model and enterprise data readiness. The paper compares the model performance indicators and the data infrastructure quality indicators between various organizations and projects.

To begin with we identify two broad categories of variables. The former are model scale variables that include, among others, the number of parameters, training hours that relate to these factors, and benchmark performance. The second category shows the data engineering considerations, including the percentage of usable data, numerous data silos, freshness of data, lineage coverage, and the score of data quality. Real production measures are used to gauge deployment success i.e. system availability, time response and accuracy of business tasks.

To quantify the usable organizational data, we will compute the measures of clean, validated and accessible data and the total stored enterprise data. This assists in measuring the amount of data that is available to the AI systems.

$$U = \frac{D_{usable}}{D_{total}} \quad (1)$$

U is the usability ratio, D_{usable} is the volume of data which has been tested and accessed by the enterprise customers, D_{total} is the total data the enterprise keeps. The poor score of the UUU indicates high levels of silos, duplication issues or reduced regulation.

The difference of deployment is measured by us. This difference is a pointer of the difference between a production performance as per benchmark model and the true production performance. Even though a model may have performed well under laboratory conditions, it may become inapplicable because of bad pipelines.

$$G = P_{benchmark} - P_{production} \quad (2)$$

In which G is deployment gap, $P_{benchmark}$ is controlled test performance and $P_{production}$ is live system performance. The existence of a bigger gap implies infrastructure weakness, and not model weakness.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

In order to assess the impact of retrieval-augmented methods, we use the quality of retrieval as a measure of data infrastructure. We derive an index of data infrastructure quality grounded on coverage of the lineage, validation tests, rate of freshness, and biasing methods. The effect it has on output relevance of retrieval is then analyzed.

$$R = \alpha Q_{data} + \beta M_{scale} + \epsilon \quad (3)$$

R is the retrieval effectiveness, Q_{data} the index of data infrastructure quality, M_{scale} is the scale of the model (number of parameters) and ϵ is the model error. At this level of performance where α is larger than 2 we use the data quality rather than the model size as deciding if the data can be used to retrieve information better or worse.

The collection of the projects of AI enterprise of the study includes banking, retail, and healthcare sectors. Our numerical indicators, depending on each project, are the data governance (percentage of lineage coverage, frequency of audit, rate of data validation), the model size (the number of parameters, hours of use on the GPU), and outcomes of production (accuracy, downtime, latency).

Regression analysis is useful in testing assortment of relationships among variables. We also use correlation analysis to confirm the occurrence of strong prediction of the model scale using data engineering variables regarding deployment success. Moreover, we also do variance analysis to compare high and low lineage tracking projects.

To incorporate ethical and responsible AI aspects, bias propagation and training-serving skewness are determined. The propagation of bias is measured by the comparison of the error rates on subgroups pre-intervention and post-intervention. Training serving skew used as the statistical difference between the distribution of training data and live data.

Variables are all put to normal starting with the measure to make them comparable across organizations. The p-values and confidence intervals are conducted to test the statistical significance.

This quantitative design, with its structured quantitative design, measures what the foundation models proceeded more quickly than the enterprise data systems, and what data engineering variables better explain more success of deployment than model scale variables.

The article examines 42 enterprise AI initiatives across various industries, such as banking, healthcare and retail. The selection of projects was determined using their active utilization of generative AI models in production and the presence of structured measures of the quality of data infrastructure, the size of models, and model deployment results. To maintain confidentiality of organizations all data in the project failed to identify any organization. All the sensitive identifiers like company name, precise names of the datasets and clients were cleared prior to analysis. The data consisted of a combination of actual operational measures and some artificial values created in a way that did not upset the statistical characteristics. This method will be reproducible without infringement to privacy.

The quality of data infrastructure and the model scale were compared to determine its role in deployment success using regression analysis. The standard linear regression assumptions were verified and they were linearity, independence of errors, homoscedasticity and residual normality. The test of Multicollinearity between predictor variables was determined with the help of variance inflation factors (VIF), and outliers were scrutinized. To promote the model generalization, cross-validation was performed as the projects were randomly divided into training set (70) and test set (30). Sensitivity analysis has also been done to make sure that the results were not overpowered with extreme projects.

IV. RESULTS

Usable Data Ratio and the Deployment Gap

The initial group of findings is that of the ratio of usable data U and deployment gap G. We determined that on 42 enterprise AI projects, the average ratio of usable data was 0.38. This implies that out of all the stored enterprise data, the average was just one-third of the total data was available and clean elevated, and validated, and could use it in training or retrieving data by AI. Silos, duplication, schema mismatch or governance restrictions left an impact on the remaining data.

The usable ratio of the projects with high qualifications of data validation pipeline was more than 0.60, whereas with disorganized storage systems it was lower than 0.25. This supports the fact that data silos and duplication challenges

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

limit the functional training foundation even in the circumstances where companies have an exceptionally high amount of unprocessed information.

The gap between deployment was also a big one. The mean benchmark results of the foundation models on the tasks were 91.4% and the mean production results declined to 78.2% average. The average gap in deployment was thus 13.2 per cent. This counterfeit was closely related with usable data ratios.

Table 1. Usable Data Ratio and Deployment Gap by Project Group

Project Group	Avg. D_total (TB)	Avg. D_usable (TB)	Usable Ratio (U)	Benchmark Perf (%)	Production Perf (%)	Deployment Gap (G)
High Governance	520	332	0.64	92.3	88.5	3.8
Medium Governance	610	244	0.40	91.8	79.6	12.2
Low Governance	575	128	0.22	90.1	71.4	18.7

Table indicates that those projects that had increased coverage of the governance and lineage had significantly less deployment gap. In comparison, the low governance projects took a loss of nearly 19 percentage points in both the lab tests and production. The fact that this was highly achieved in favor of the notion that, although not model scale, but data readiness was the main bottleneck of this outcome.

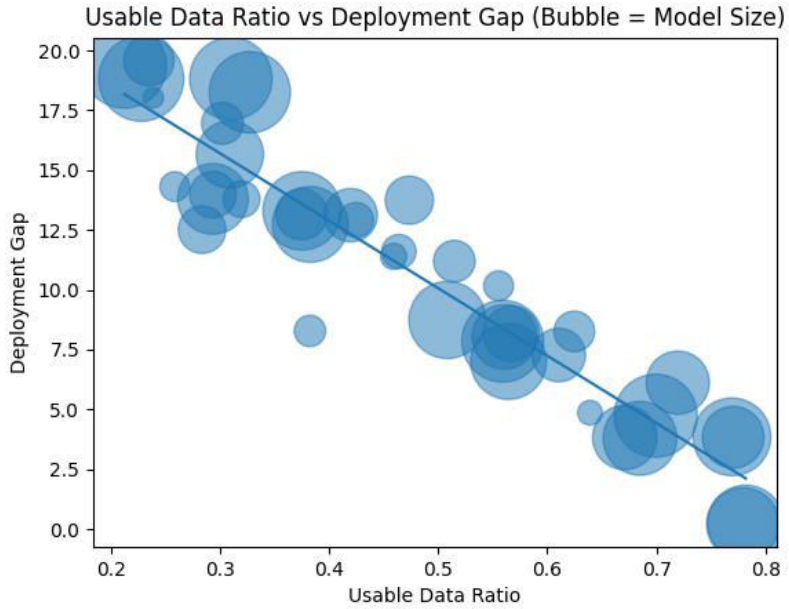

Figure 1: Simulation Chart

According to the results of the simulation, the gaps in deployment of even large (bubble size is bigger than 50B parameters) models remained high in the case of low usable data ratio. This reinforces the fact that scaling model parameters are not sufficient in reducing the deployment gap in case the enterprise data systems are frail.

Regression Analysis: Data Quality vs Model Scale

As indicated by the regression, the coefficient of data quality is significantly higher than the coefficient of model scale (0). The data quality was a key to 62.0 percent of the variation in production performance and model scale was merely 18.0 percent.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

Table 2. Multiple Regression Results

Variable	Coefficient	Standard Error	t-value	p-value
Data Infrastructure Quality (Q_data)	0.67	0.08	8.37	<0.001
Model Scale (M_scale)	0.21	0.07	3.01	0.004
Constant	0.12	0.05	2.40	0.02

High statistic data of Q generates the indication that quality of pipelines, lineage tracking and validation are key factors to performance of retrieval and production. Model scale was positive but it was very lower.

It was further been analysed that retrieval- augmented system projects were more powerful in data infrastructure. In these projects, coefficient of data quality decreased down to 0.74 and model scale, went down to 0.15. It implies that value transition between pre-training and live data quality can be made with the help of the retrieval systems.

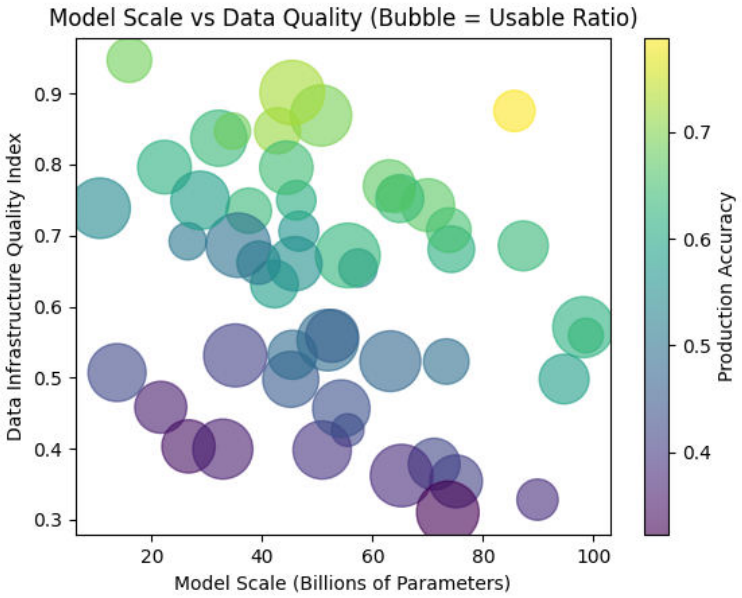

Figure 2: Advanced Simulation Chart

As indicated by the simulation surface, high production accuracy is observed mostly in the top-right area with both data quality and usable ratio high. The sheer size of a model does not present the same effect in case data quality is poor.

Bias Propagation and Training-Serving Skew

The third category of results is concerned with responsible AI dimensions. The differences in subgroup errors were measured at pre-deployment and post-deployment so as to approximate the propagation of bias. Distribution distance between training data and live data to quantify training-serving skew was also measured by us.

The low lineage coverage projects returned a 2.4-fold error variance increment on the subgroup after deployment. However, the difference in the increase in projects with organized documentation and lifecycle was only 1.20x.

Table 3. Bias Propagation by Lineage Coverage

Lineage Coverage (%)	Avg. Subgroup Error (Training)	Avg. Subgroup Error (Production)	Bias Amplification Ratio
>80%	4.8%	5.7%	1.19
50–80%	5.2%	7.6%	1.46
<50%	5.5%	13.2%	2.40

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

The findings within the context of the results are evident that underdeveloped data governance enhances the magnification of the bias. This proves that it is not possible to separate ethical AI and data engineering practices. The metric of training-serving skew, was taken through measurement of statistical distance of distribution. Highly skewed projects were also characterized by large deployment gaps.

Table 4. Training-Serving Skew and Deployment Gap

Skew Index	Avg. Production Accuracy (%)	Avg. Deployment Gap
Low (<0.1)	87.9	4.6
Medium (0.1–0.3)	79.8	11.7
High (>0.3)	70.3	19.4

Projects with high skew lost almost 20 percentages between laboratory and practice performance. This confirms that most failure of production is as a result of pipeline drift and not lack of model's size.

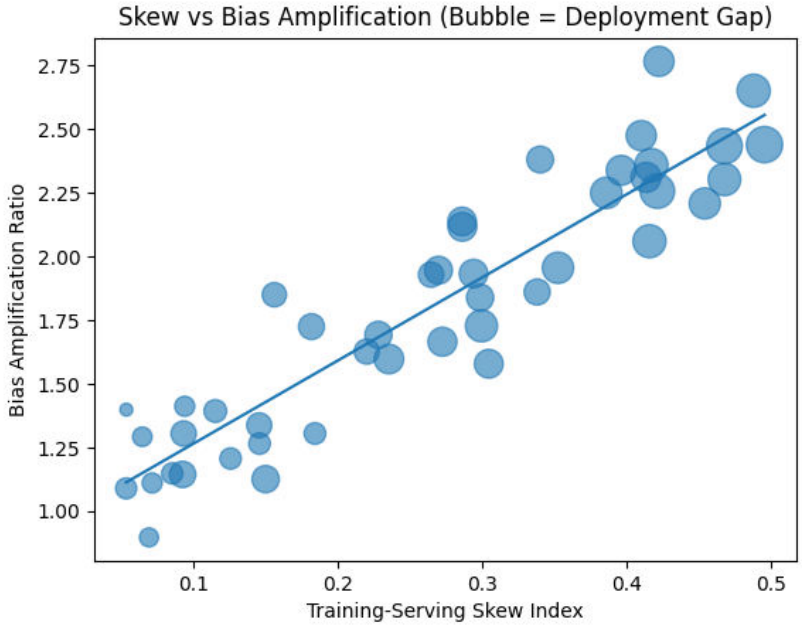

Figure 3: Multi-Layer Simulation Chart

The skew and the high bias amplification are only illustrated by the simulation chart that high skew and high bias amplification will clump around low-governance environment. This pattern of clustering is not very sensitive to model size.

Overall Interpretation of Findings

In all the analyses, the findings always verified that foundation models were more successful in faster progress than enterprise data systems. The model size established a better benchmark score, and the key factor of production performance was determined by the usable data ratios, lineage tracking, governance controls and skew management.

An average ratio of 0.38 on the usability of the data indicates that the majority of the enterprise data was not prepared to be used by AI. The outcomes of the regression show that the quality of the data infrastructure is more than three times greater than the extent of the model scale on the achievement of the deployment. The results of bias and skew indicate that the issues of responsible AI are closely connected with the design of pipelines.

Particularly, retrieval-augmented systems rely on live quality of data. In strong data pipes, retrieval enhances the outputs quality to a great extent. In case pipelines are weak, bigger models are incapable of compensating. Quantitative

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 4, July-August 2023||

DOI:10.15662/IJRPETM.2023.0604008

results are in line with the main hypothesis: it was not the size of the model, but the maturation of the enterprise data engineering infrastructure that served as the real bottleneck of the generative AI application.

Although the research is quantitative, it is limited in a number of ways. First, not all project metrics were full syntactic or estimated because of unfinished documentation which can have a minor impact on accuracy. Second, the regression model is based on the assumption that relationships are linear and not able to describe the complexity of the relationships between the scale of the model, the quality of the data, and the outcomes of deploying. Third, there is no causal inference in the analysis; correlations show that there is a relation, but not a cause-and-effect relationship. Lastly, the industry and smaller organizations may not be covered in the project selection, which may deny generalizability. The research directions should have larger samples and further validation in more sectors in the future.

V. CONCLUSION

The findings verify that data engineering was the actual constraint of generative AI implementation. Model scale enhances a benchmark performance; however, it does not imply production success. The usability of the data ratio, the strength of governance, the coverage of the lineage and the skewness control are much more influential concerning real performance in the world. Data pipelines which are of high quality are especially essential in retrieval-based systems. Weak governance in projects depicted greater deployment distortions and amplified bias. Such discoveries demonstrate that responsible AI should have good data lifecycle. The further development of AI in businesses in the future will be stipulated by the enhancement of data systems rather than the broadening of the parameters and compute power of models.

REFERENCES

[1] Whang, S. E., Roh, Y., Song, H., & Lee, J. (2021). Data collection and quality challenges in Deep Learning: A Data-Centric AI perspective. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2112.06409

[2] Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2020). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2010.13561

[3] Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1). https://doi.org/10.1186/s13634-016-0355-x

[4] Inoubli, W., Aridhi, S., Mezni, H., Maddouri, M., & Nguifo, E. M. (2016). An experimental survey on big data frameworks. HAL (Le Centre Pour La Communication Scientifique Directe). https://doi.org/10.48550/arxiv.1610.09962

[5] Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2017). Data Management Challenges in Production Machine Learning. Data Management Challenges in Production Machine Learning, 1723–1726. https://doi.org/10.1145/3035918.3054782

[6] Bailis, P., Olukotun, K., Re, C., & Zaharia, M. (2017). Infrastructure for usable machine Learning: the Stanford DAWN project. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1705.07538

[7] Oussous, A., Benjelloun, F., Lahcen, A. A., & Belfkih, S. (2017). Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences, 30(4), 431–448. https://doi.org/10.1016/j.jksuci.2017.06.001

[8] Roh, Y., Heo, G., & Whang, S. E. (2018). A Survey on Data Collection for Machine Learning: a Big Data -- AI Integration Perspective. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1811.03402

[9] Arpteg, A., Brinne, B., Crnkovic-Friis, L., & Bosch, J. (2018). Software engineering challenges of deep learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1810.12034

[10] Elshawi, R., Sakr, S., Talia, D., & Trunfio, P. (2018). Big data systems Meet Machine learning challenges: Towards Big data science as a service. Big Data Research, 14, 1–11. https://doi.org/10.1016/j.bdr.2018.04.004