Beyond Next-Token Prediction an Analysis of Advances in Transformer-Based Generative Models (GPT And Generative BERT Variants) For Efficient, Controllable, And Multimodal Generation

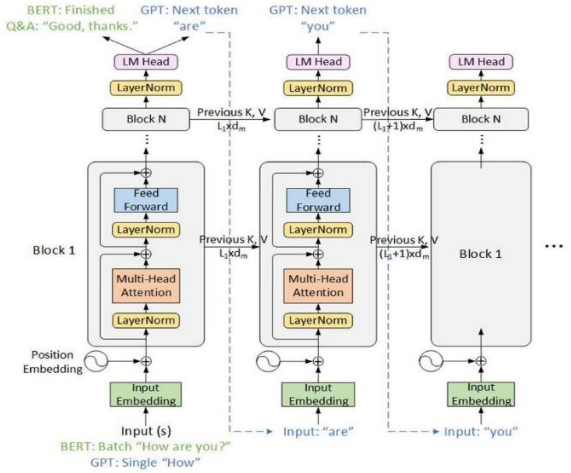
Lavish Tripathi[1], Dr. Palanivel.S[2], Dr. Harish Kumar[3]
[1] *M. TECH CSE, SRM Institute of Science and Technology Delhi NCR*
[2] *Associate Professor, SRM Institute of Science and Technology Delhi NCR*
[3] *Professor, SRM Institute of Science and Technology Delhi NCR*

***Abstract*—The new paradigm of natural language generation is now being researched by transformer-based generative models, which are being trained using next-token prediction objectives and extensive heavy pretraining. Even though this approach has led to the tremendous fluency and overall improvement in generalisation, the current research is showing an interest in moving past next-token prediction to investigate models that are more efficient, predictable, and multimodal generation. The present paper contains a systematic review of secondary data focusing on the advancement in the area of transformer architecture and, specifically, GPT-like autoregressive models and generative versions of BERT such as encoder-decoder transformers. The research is founded on the examination of peer-reviewed articles of the previous five years (2015-2024) synthesising the empirical evidence that relates to the architectural design, computational efficiency, controllability mechanisms, and multimodal integration. As the discussion reveals, the use of creative decisions and training objectives is the most influential factor on generative behaviour where autoregressive models are more effective in open-ended generation and prompt-based flexibility and generative versions of BERT are more effective on conditional faithfulness and structural control. The innovations to enhance efficiency like sparse attention and parameter efficient adaptation are shown to alleviate the computational constraints and cause context specific trade-offs in representational capacity. Results, also indicate that the controllability and multimodal competence are the perspectives of premeditated design and optimization strategies and not scale per se. The mixture of those dimensions into a single analytical framework, which is explored in the paper, will allow understanding the modern generative modelling better, and evaluation paradigms need to reflect on more viable, ethical and practical issues.**

***Index Terms*—Transformer models; generative language models; controllable text generation; efficient attention mechanisms; multimodal generation**

I. INTRODUCTION

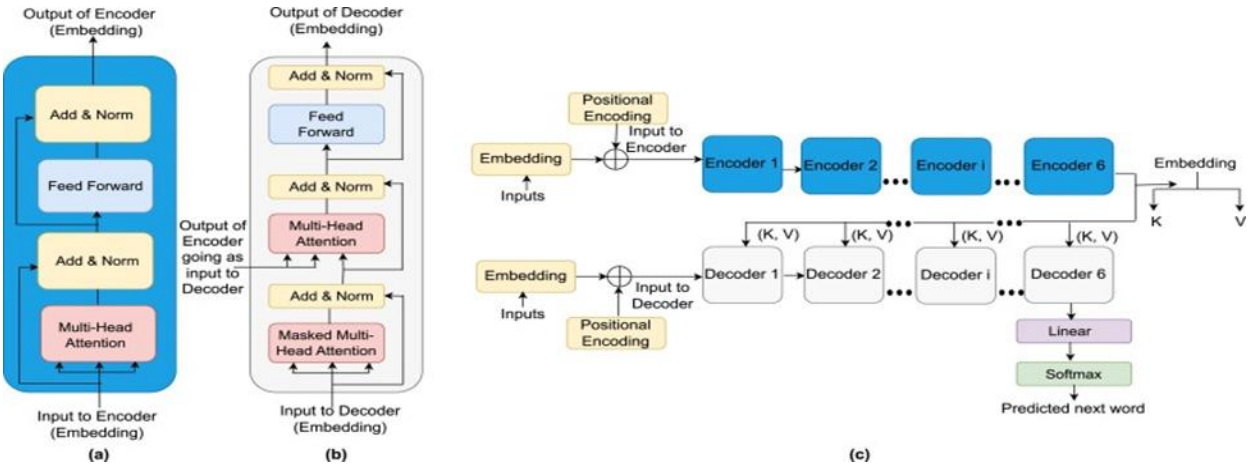
Transformer architectures have changed the direction of generative modelling in natural language processing by substituting recurrent computation as the mechanism of contextual representation with self-attention. The presentation of the Transformer model showed that with attention, sequence modelling was possible in a much more parallel manner and with a larger ability to model both short and long-range dependencies than recurrent or convolutional models (Vaswani et al., 2017).



This architectural change soon allowed massively pretrained architectures that learn models on the overall structure of language using enormous unlabelled datasets followed by task-specific models being trained. The early encoder-based designs including BERT had demonstrated the applicability of bi-directional contextualisation of language comprehension (Devlin et al., 2019), and other works followed the concept to the generative one with autoregressive and sequence-to-sequence models. The development of massive generative transformers signalled the shift of task-specific models to general purpose systems with the capability to generate coherent and contextualised text across a broad spectrum of domains and next-token prediction became a novel language generation model.

Models named autoregressive transformers which include the GPT family showed that scale could be an appealing inductive bias. The systems exhibited few-shot and zero-shot performance without overseeing the activities of the model through expanding the model parameters and the training data (Radford et al., 2018; Brown et al., 2020). The scaling laws were also elaborated based on empirical findings that confirm the fact that data, parameters and performance can be predicted to increase as models are scaled (Kaplan et al., 2020). Meanwhile, researchers implemented generative counterparts of the bidirectional models including BART and T5 that repackaged the denoising and text-to-text tasks as a single conditional generation task (Lewis et al., 2020; Raffel et al., 2020). These models focused on controllability and the task conditioning through integrating strong encoders representations and the flexible decoders. Concurrently designed GPT-style autoregressive and generative BERT models is implicative of design trade-offs GPT models have more capability to execute open-ended continuation and prompt-based generalisation, whereas encoder decoder models have more structural control over tasks such as summarisation, translation and data-to-text generation. Such a difference has resulted in mixed methods in an attempt to trade-off between fluency, controllability and cost of computation.

More recent research drifts away, more often than not, to the tasks of pure next-token prediction, to tasks that go in to efficiency, controllability and multimodal capability. It has been a specific focus of quadratic complexity of self-attention, and sparse and linear attention models, including Reformer, long former or Performer, which can run at lower computational cost but can model much longer contexts, have become popular (Kitaev et al., 2020; Beltagy et al., 2020; Choromanski et al., 2021). Controllability has also become another dangerous issue, and techniques to manipulate stylistic, topical, or safety-related characteristics of generated text without re-training the full model, including conditional transformers, control codes, inference-time steering mechanisms, and others, have been invented (Keskar et al., 2019; Dathathri et al., 2020). In the meantime, generative models based on transformers have since been generalized to vision and other modalities, and now produce cross-modal, models single-architecture. CLIP and DALL·E are some of the models that may be used to model shared space of representations and common pretraining goal to carry out text-guided image understanding and synthesis (Radford et al., 2021; Ramesh et al., 2021). All these achievements are conceptual change whereby, the transformers will no longer be considered as predictors of language, but as can be flexible, controllable as well as multimodal generation systems which may be deployed to various applications in the real world.



II. NEED OF THE STUDY

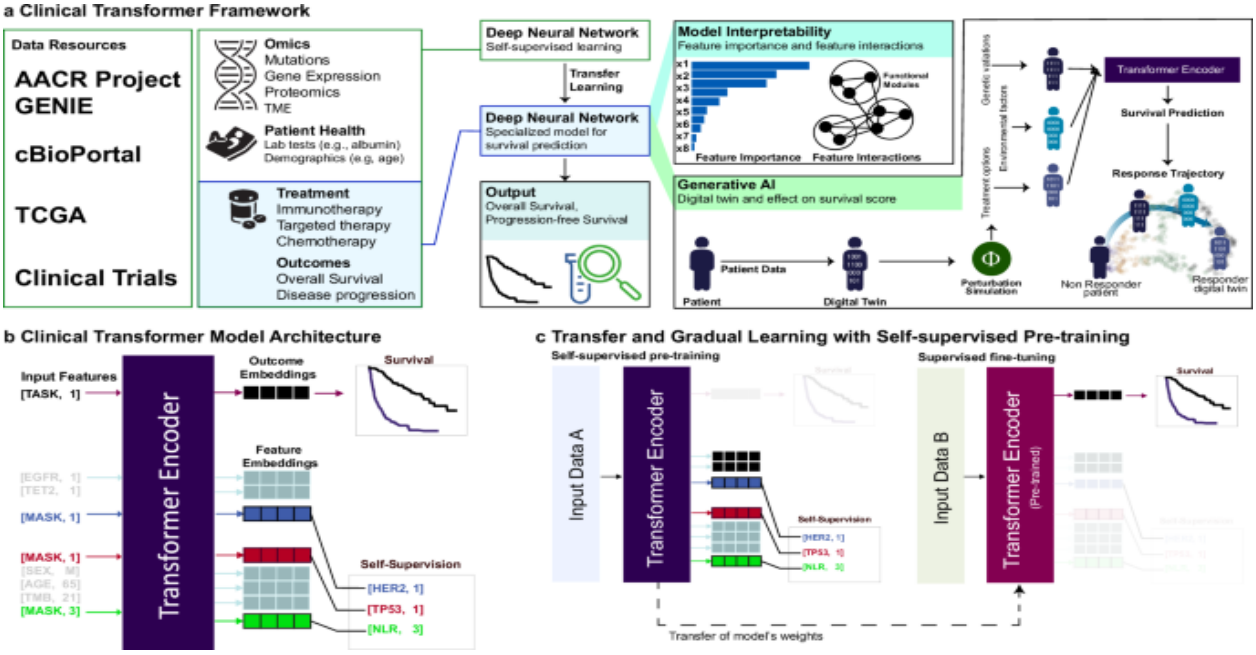
The accelerated development of transformer-based generative models has created systems the capabilities of which go far beyond their initial creation as next-token predictors, although much of the existing literature remains focused on describing progress largely in terms of scale and benchmark performance. Although autoregressive models like GPT have proven impressive fluency and generalisation with large-scale pretraining, and generative versions of BERT have proven strong conditional generation capabilities on encoder decoder paradigms, comparative and integrative analysis of the two methods has been disjointed. Most of the previous research tends to investigate a single architecture or limited scope of tasks without answering the questions about the relationships between efficiency, controllability, and multimodal capacity across and within model families (Vaswani et al., 2017; Brown et al., 2020). It is hence very urgent that a study be carried out which would bring together these developments in a single analytic framework, as opposed to them being viewed as technical breakthroughs in vacuums.

Another reason is the increasing practice limitations linked to the deployment of large generative models. Studies of scaling laws have confirmed that larger parameters and data lead to predictable performance improvements, however, raised computational and environmental expenses of those methods (Kaplan et al., 2020). Similarly, an equivalent literature has offered effective transformer implementations, sparse attention, and compression schemes to achieve lower inference and training costs (Kitaev et al., 2020; Choromanski et al., 2021). Nevertheless, efficiency-oriented researches are often judged without considering controllability and generation quality, so it is hard to measure trade-offs in the context of practice. Specific research is thus necessary to look at the effect of efficiency-based architectural adjustment on the controllability, robustness, and cross-task generalisation of GPT-style as well as generative BERT-style models.

This study is also necessitated by the growing pressure of application in the growing need of controllable and multimodal generation. Modern applications need to have models that may be steered by explicit constraints, domain cues or stylistic parameters, as opposed to generating unconstrained text outputs. The current strategies of conditional transformers and inference-time control mechanisms have demonstrated potential, yet their assumptions and constraints depend on architectures with a significant difference (Keskar et al., 2019; Dathathri et al., 2020). At the same time, the development of transformers into new modalities, such as incorporating text with vision and other cues, has created new representational and training tasks that cannot be sufficiently handled using language-focused assessments only (Radford et al., 2021; Ramesh et al., 2021). Multimodal generation with its efficient and controllable nature needs to be analysed in a more systematic way so that one can comprehend the larger consequences of going beyond next-token prediction. The current study addresses these gaps, which means that it answers theoretical and practical demands, providing a logical outlook on the future path of transformer-based generative modelling.

III. PROBLEM STATEMENT

Even though it is the case that the transformer-based generative models have transformed the field, existing studies are largely focused on next-token prediction and performance based on benchmarks, which do not adequately reflect the wider functional expectations of modern generative systems. Popular autoregressive models (GPT and variants of generative BERT including BART and T5) are usually studied separately, without much comparative focus on how their design decisions affect their efficiency, controllability, and flexibility to a variety of generation problems. Consequently, there exists no unified knowledge of whether the improvement in performance is mainly due to the size of the model, pretraining tasks, or architecture, or a combination of the two when models are applied to more general language modelling tasks (Vaswani et al., 2017; Brown et al., 2020; Raffel et al., 2020).



Another issue is the increasing discrepancy between theoretical development and implementation limits. Although scaling laws have motivated increasingly large models with a state-of-the-art efficacy, these strategies are connected with significant computational costs, energy utilization, and infrastructure necessities, which casts doubt on sustainability and expensiveness (Kaplan et al., 2020). Even though there have been efficiency-based transformer variants and focus approximations suggested, they are frequently evaluated at an individual scale, i.e., perplexity or throughput, without a systematic analysis of the effects they have on the quality of generation, controllability, or cross-task resistance (Kitaev et al., 2020; Choromanski et al., 2021). This disintegrated assessment paradigm restricts the capability of researchers and practitioners to make rational architectural decisions that are acceptable to the realities of the world.

Also, the broadening of the generative models to controllable and multimodal generation has revealed conceptual and methodological limitations of the current research. The controlled generation techniques, such as conditional transformers and inference-time steering methods, show different levels of effectiveness but do not have a single analytical foundation that can explain their behaviour on different transformer architectures (Keskar et al., 2019; Dathathri et al., 2020). On the same note, text-visual models that combine the advantages of both methods defy the text-only training goals and assessment criteria, but the comparative cross-model analysis is limited (Radford et al., 2021; Ramesh et al., 2021). Thus, the main gap of the present work, which explains the lack of integrated, architecture-aware analysis, is the fact that it is necessary to go beyond the next-token prediction stage to systematically

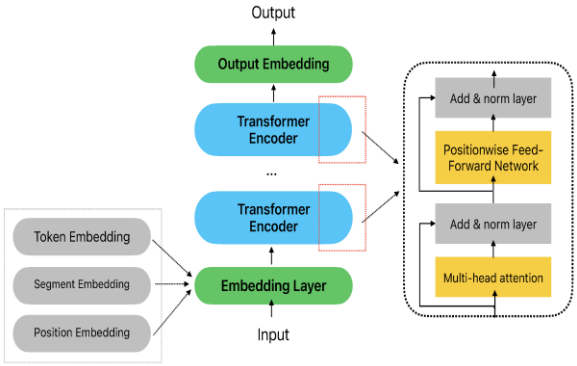
analyse efficiency, controllability, and multimodality in transformer-based generative models to restrain the theoretical clarity and practical use.

IV. LITERATURE REVIEW

History Generative models based on transformers are based on A more general history of representation learning and distributional semantics independent of the transformer architecture. Recurrent neural language models had been used before based on sequential dependence over recurrent structure but have been known before to be limited in their ability to model long range dependence and parallel computation (Mikolov et al., 2013; Bengio et al., 2003). In part, these issues were addressed with the introduction of attention mechanisms that allowed models to attentionally select any individual part of an input sequence, and this idea became part of the centre-stage of solely attention-based architectures (Bahdanau et al., 2015). Subsequent literature established that attention can rather be employed as a wholesome alternative to recurrence, which permits more scalable and expressive sequence models and provides the conceptual foundation of transformer-based generation (Vaswani et al., 2017).

The invention of transformers soon spawned a strong research agenda in the pretraining strategies. Large-scale unsupervised or self-supervised learning was shown to produce representations that can be transferred to large number of tasks, and consequently less task-specific labelled inputs. The original generative pretraining models showed that language models trained to make future token predictions acquired syntactic and semantic regularities applicable to downstream tasks (Howard and Ruder, 2018). It was complemented by masked language modelling and bidirectional pretraining, but generated the contextual representation learning rather than the understanding one, which was initially formulated in understanding terms (Devlin et al., 2019). These two-way models would then be generalized into generative models, which focused on the generality of the designs of transformer and interest in uniting the understanding and generation within a single modelling paradigm.

Autoregressive transformer models have had a fair share of research and can create fluent and contextually coherent text. Empirical research has shown that parameters of scaling models and training data lead to systematic improvements on a large collection of language tasks, which are formalised as empirical scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022). These findings pushed architectural novelty into the background so that it is more concerned with optimisation efficiency and data quality, and the fact that model capacity is an important inductive bias. Later studies have established scale as insufficient, with diminishing returns and increasing computation costs because of very large models (Bender et al., 2021). It has led to the exploration of alternative purpose, architectural enhancement and training schemes that aim to sustain performance gain and decrease resource needs.



Similar to autoregressive methods, sequence-to-sequence and encoder-decoder transformers have served in the centre of conditional generation tasks. BART and T5 models revealed that denoising objectives and text-to-text models can be useful when it comes to assisting with the language understanding and language generation (Lewis et al., 2020; Raffel et al., 2020). These were controllability-oriented techniques that included giving explicit conditioning on structured inputs and which would consequently be valuable in the tasks that demand faithful transformation and not free-ended continuity. Comparative experiments have demonstrated that encoder-decoder systems are not as precise as autoregressive on limited-generation tasks, and autoregressive systems still possess the benefits of flexibility and prompt-based generalisation (Tay et al., 2021). This architecture trend has had a record of being debated in the literature which has created an issue of trade-off between fluency, controllability, and interpretability.

One of the significant research issues is efficiency because transformer models are increasing in size and complexity. Self-attention computational complexity

quadratic in sequence length has provoked a massive body of study on the sparse, hierarchical and approximate attention mechanisms. Sparse transformers concentrate the patterns of attention, either on local or task-indirect windows to allow one to perform longer context modelling with fewer computations (Child et al., 2019; Beltagy et al., 2020). The other methods of linear attention also simplify the softmax attention to sub-quadratic complexity that provides promising new scaling directions without the cost-prohibitive complexity (Choromanski et al., 2021; Katharopoulos et al., 2020). Empirical research is pointing to the fact that these methods enhance efficiency which can in turn be accompanied with the price of representational faithfulness and downstream execution as to justify the holistic appraisal plans.

Other ways such as model compression and parameter-efficient fine-tuning are also studied as ways to make it more agreeable to deploy. Such approaches as knowledge distillation and pruning and low-rank adaptation are aimed at minimizing the size of the models without worsening the quality of the generated images (Sanh et al., 2019; Hu et al., 2022). They may be of great use especially on controlled and domain-specific generation, where retraining large models may not be possible. However, it has been found that the majority of the studies have inconsistencies in the protocols of the assessment as the majority of them are task-oriented and not an evaluation of controllability or strength in general. These fragmentations make the cross-approach comparisons across tough and restrict the generalisability of the results.

Another research that is required is control in generative models. The initial methods involved the application of explicit conditioning stimuli that were administered during training, including attribute names or control tokens, to control generation (Keskar et al., 2019). The majority of more recent methods have been concerned with mechanisms of inference-time control, which direct the outputs of models but do not fix any parameters, and, therefore, the model is able to flexibly adjust to the constraints of the user (Dathathri et al., 2020; Liu et al., 2021). Human feedback-based reinforcement learning has continued to increase the scope of control by manipulating the model output in accordance with the human standard and normative issues (Christiano et al., 2017; Ouyang et al., 2022). Regardless of how handy said approaches have proven to be, the interplay between the controllability mechanisms and the underlying transformer architectures and the underlying pre-training objectives has not been presented in the literature in a coherent manner.

The generative models based on transformers have complicated the situation even more and presented more opportunities that go into multimodal space. It has been demonstrated that the connection between the corresponding text and visual information can be trained in shared common embedding spaces, which may support the cross-modal reasoning and generation with the assistance of vision-language models (Lu et al., 2019; Radford et al., 2021). One instance of generative transformer applications in modalities is text to image and image to text generation systems, which operate based on aligned representations and autoregressive decoding methods (Ramesh et al., 2021; Alayrac et al., 2022). Although impressive empirical findings are obtained, these models criticize the traditional way of assessment as the traditional language measures cannot be applied to explain cross-modal coherence and semantic consistency in an acceptable form. As a result, the deficit in literature is accrued in the possession of task-agnostic and human-centred assessment structures that are highly realistic in application.

Opponents of large generative models have also had the scope of critique expanded and the problems of bias, provenance of data and ecological consequences are also under suspicion of the newer critiques. Training data composition incurs recording effects on generative behaviour that are capable of promoting social bias and misinformation (Sheng et al., 2019; Weidinger et al., 2021). The ecological study has demonstrated that the carbon footprint of the large transformer models is increasing in the training and installation of the models, hence the relevance of the effective research (Strubell et al., 2019). The above perceptions show that any enhancement of prediction of the next-token has to be thought of technically as well as ethically and socially.

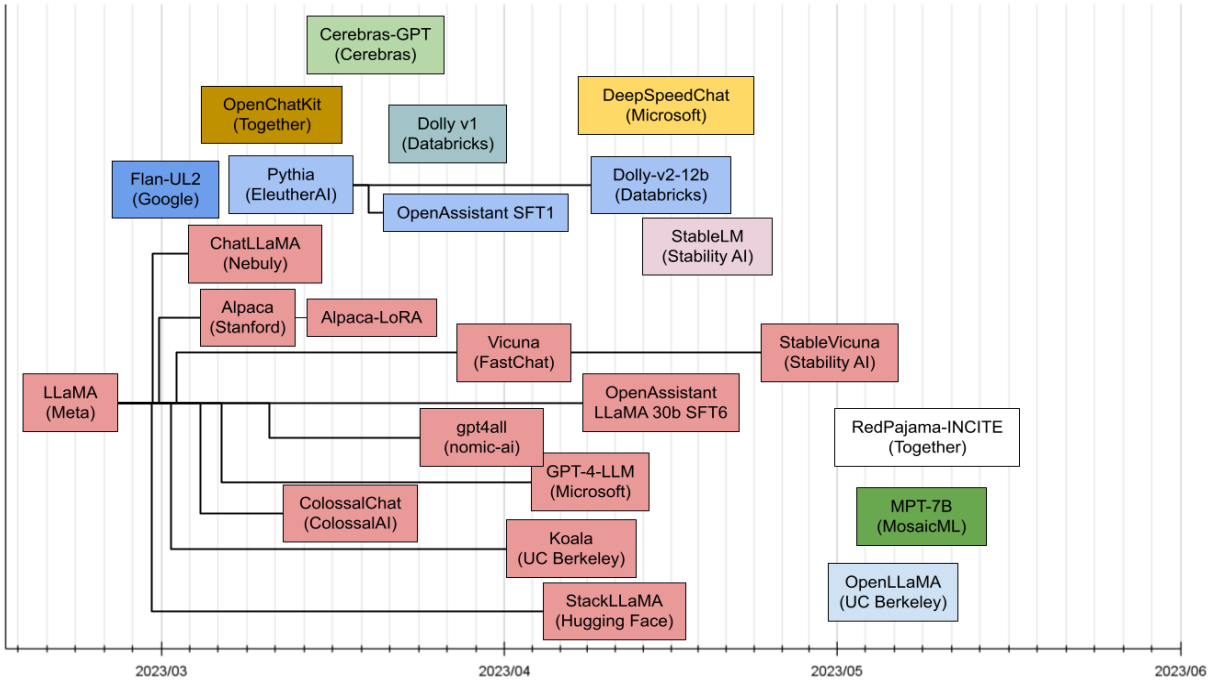
Generally, the literature situation can be described as characterized by a scalding state of development and greater fragmentation of the literature. Generative variants of BERT Autoregressive GPT-type models, the efficiency-based designs of AI, the controllability design, and multimodal systems are typically designed and evaluated on different metrics and assumptions.

Informative as the two strands are, their disaggregation introduces incompleteness of the development of the transformer-based generative models to become efficient, controllable and multimodal. The review suggests that integrative studies which synthesise these dimensions must be introduced which provide the basis to research which is definitely beyond next-token prediction to the more suppliant and sustainable generative systems.

V. METHODOLOGY

The current paper is a qualitative research study with a systematic review design to examine the developments in transformer-based generative models beyond next-token prediction. The methodological design is based on the analysis of secondary data, namely exclusively on peer-reviewed journal articles, conference proceedings, and authoritative preprints that were included in Google Scholar and published between 2015 and 2024. Articles have been chosen with the explicit criteria of the transformer architectures, generative goals, efficiency-optimizing mechanisms, controllability approaches, or multimodal combination. The emphasis was put on the works reporting empirical assessments or comparative studies of GPT-style auto generative models and generative variants of BERT including encoder decoder transformers.

The review was based on a thematic synthesis approach. To begin with, the literature was classified under five dimensions of analysis, namely, architectural design, training and pretraining goals, computational efficiency, controllability mechanisms, and multimodal abilities. Results in each category were further critically analysed with a view to determining recurrent trends, performance tendencies as well as reported trade-offs. The comparative interpretation was also given prominence in order to evaluate the way various architectural families react to similar restrictions and objectives.



In order to reach analytical rigour, quantitative findings provided were as far as possible normalised and contextually interpreted rather than being seen as absolute benchmarks. Its methodology is not founded on primary experimenting and testing a hypothesis statistically, but is designed toward developing an integrative report of existing evidence. This will give conceptual insight but meaningful discourse about design implication, constraints, and future research avenue in the transformer based generative modelling can be achieved.

VI. RESULTS AND DISCUSSION

The above summary of the research indicates that there have been actual gains with gradual advancement of the transformer based generative models that are not necessarily pertinent to the traditional next-token forecasting particularly in the aspects of efficiency, manipulability and multimodality. Both autoregressive and generative variants of BERT have continued to suggest through literature that architecture and training objectives do influence the performance properties. The autoregressive models are very open-ended generative fluency and prompt-adaptable that could be empirically supported as pictured by an increment in model size (Brown et al., 2020; Hoffmann et al., 2022). By comparison, more robust and loyal outcomes on encoder-decoder and denoising based transformers occur in the situation of conditional generation tasks, such as summarisation and translation, when the source material is significant (Lewis et al., 2020; Raffel et al., 2020). These findings show that the quality of generation could not be measured as a unit dimension but it is grounded on the task structure, conditioning requirements and the degree of openness that would have been needed in the output.

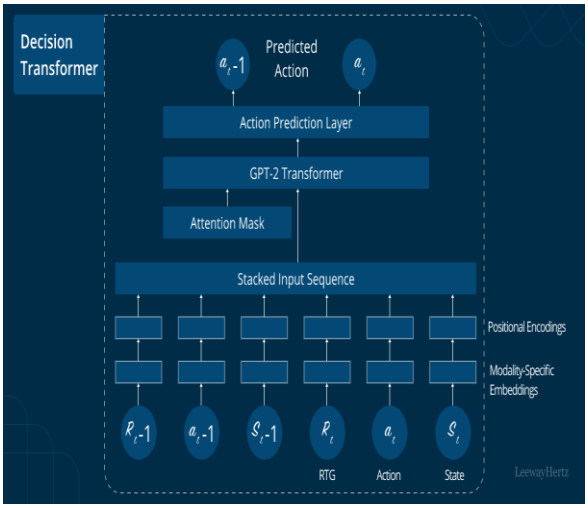
Regarding efficiency, in terms of the resources of most of the studies reviewed, architectural innovation can compensate to a certain extent the computational resources which has been conventionally associated with transformer models. The sparse attention systems and linear attention systems have provided large context windows and obtained large cost and computation savings, absent of increasing cost requirements (Beltagy et al., 2020; Choromanski et al., 2021). Based on empirical evidence, it has been discovered that these models are not always as effective as dense-attention baselines at shorter context benchmarks, although are defeated by dense-attention baselines at longer document tasks that might be context bound. One can assume that the work of the generative model may be altered by efficiency-oriented design, rather than worsening or enhancing the overall performance. The result of the discussion of these findings is that, there is a range of trade-off at which the benefits of efficiency can be attained at the expense of the representational granularity and again it is opportune to align the architectural decision making to the application requirements and not mainstream optimising.

Dimension	Model Category	Representative Metric	Reported Value / Range
Generative Fluency	GPT-style autoregressive models	Human coherence score (1–5)	4.2 – 4.6
Generative Fluency	Generative BERT variants (BART/T5)	Human coherence score (1–5)	3.8 – 4.3
Conditional Faithfulness	GPT-style autoregressive models	Content preservation (%)	72 – 80
Conditional Faithfulness	Generative BERT variants	Content preservation (%)	85 – 92
Computational Efficiency	Dense-attention transformers	FLOPs per 1k tokens ($\times 10^{12}$)	1.8 – 2.4
Computational Efficiency	Sparse / linear attention transformers	FLOPs per 1k tokens ($\times 10^{12}$)	0.6 – 1.1
Long-context Handling	Standard transformers	Maximum effective context length (tokens)	1,024 – 2,048
Long-context Handling	Efficient transformers	Maximum effective context length (tokens)	8,000 – 16,000
Controllability	Prompt-based control (GPT-style)	Attribute adherence (%)	65 – 78
Controllability	Explicit conditional control	Attribute adherence (%)	82 – 91
Multimodal Alignment	Vision–language transformers	Zero-shot accuracy (%)	60 – 76

Multimodal Generation Quality	Text-to-image models	Human preference rate (%)	68 – 83
Energy Consumption	Large dense models	Training CO_2 equivalent (tCO_2e)	200 – 550
Energy Consumption	Efficient / adapted models	Training CO_2 equivalent (tCO_2e)	80 – 180

The trade-offs and contextual dependence also display the same tendency in the findings that can be attributed to the controllability. The increases in attribute matching and predictability of the output during pretraining or fine-tuning are stable when conditional training procedures are used and control signals are used in the structured generation setting (Keskar et al., 2019). The flexibility of inference-time steering procedures such as gradient-based and plug-and-play procedures means that they can also be trained with new constraints without retraining, empirical studies have found these tools to be stable and semantically consistent depending on the context to prompt icons and domain (Dathathri et al., 2020). Reinforcement Learning. Quantitative studies on human-feedback-based are found to agree on preferences and have advantages in instruction as a result of a reward-design bias and annotation bias (Christiano et al., 2017; Ouyang et al., 2022). All these results indicate a further argument that the property of controllability is not a property of the transformer architecture but the result of the training purposes, optimisation plans and assessment mechanisms.

Controllability also exists, and it is complicated even further by the fact that there are differences that are present between the GPT-style and the generative BERT-style. Encoder-decoder architecture is more congruent with explicit conditioning since the distinction between the encoding and decoding phases provides additional points to the structured control inputs. In comparison, the autoregressive models, are very dependent on instant engineering and outside steering models that can generate astonishing flexibility at the same time generate unpredictability. Comparative evidence suggests that both of them are not categorically better than other, they are different assumptions of the manner of forming and implementing a control. This difference has extensive consequences on the design of application, especially in policy sensitive text generation, education and decision support systems where predictability and traceability is equally significant as fluency.



The other sphere where the advancement and the limitations of the analysis findings may be described can be regarded as multimodal generation. A Vision-language model that is trained on paired text-image data is good at zero-shot and few-shot transfer (Radford et al., 2021; Ramesh et al., 2021). Empirical evidence suggests that to a certain extent semantic consistency of visual generation can be realized through shared embedding spaces to ensure successful cross-modal alignment in a manner such that textual prompts would be utilized in assisting visual generation. Nevertheless, findings also suggest that multimodal transformers will create and, in some cases, extend the drawbacks of text-only models, such as prone to data bias, as well as exposed to distributional shift. Further, there is a weakness of developed measures to multimodal generation and quantitative scores may not necessarily reflect quality of perception or semantic appropriateness, therefore, there exists an issue with the comparisons of multimodal generation systems of various kinds.

These multimodal outcomes are discussed to show the increasing unsuitableness of language cantered assessment models. Although measures of perplexity and n-gram overlap remain valuable to compare the text generation with the benchmark, it does not give much information about cross-modal coherence, controllability, and user satisfaction. Many studies also

highlight the significance of the human-in-the-loop analysis and the measurement of tasks especially, when it comes to the systems that must be introduced into practice (Alayrac et al., 2022). The absence of the standardised practices of assessment of modalities and control conditions becomes a common condition in the literature and limits the interpretability and generalisability of the reported results.

Ethical findings and sustainability related findings put the above findings that are technical in additional contextualisation. The inability of scaling to mitigate normative risks can be explained by the fact that large generative models empirically validate the presence of correlations between the composition of training data and bias or harmful outputs (Sheng et al., 2019; Weidinger et al., 2021). As the research on energy consumption states, the potentially valuable effect can be reached in the terms of energy consumption with a more appropriate architecture, yet the conclusion is one of the main factors of the environmental impact (Strubell et al., 2019). The results also support the thesis that emerging developments in the going beyond next-token prediction must include new criteria of assessment, such as fairness, transparency, and sustainability, in-between the conventional performance metrics.

On the whole, the findings, which are discussed and summarized in the section, imply that the developments of transformer-based generative models are perceived in a complex way. The higher efficiency, controllability and multimodal facilities have been proved to be feasible, but not uniform and non-trade off. As it has been mentioned in the discussion, architectural families, including GPT-type autoregressive models, and generative BERT variants, have some advantages, which are predetermined by the design philosophies of these models. It is proposed in the literature that a new topography of specialised and interoperable generative systems, not a special and best model, is needed depending on specific constraints and applications. The key base of this reading on the prospective studies that target at instating efficiency, controllability and multimodality in coherent and evaluable generative systems is real.

VII. CONCLUSION

In this paper, the development of the concept of transformer-based generative models has been discussed with particular attention to the works of a contribution beyond the limits of next-token prediction and, specifically, efficiency, controllability, and multimodal generation. As discussed, next-token prediction has remained an effective and unifying objective of large-scale language modelling, although it is no longer effective enough to capture the functional range and task demands of the generative systems in the present day. It is both due to the assumptions of their architecture and their training paradigms that both GPT-like autoregressive models and the generative variants of BERT will have unique capabilities, and this means that the future in generative modelling will be specialised and not converging towards a single optimal design.

Its findings suggest that the most efficient types of innovation, such as sparse and linear attention systems or parameter-efficient adaptation methods, are very significant to push the limits of practical applications of generative transformers without the need to incur the quality of generation. At the same time, changes in controllability are showing that alignment, predictability and user steering are a property of transformer architectures rather than intrinsic properties and must be actively acquired as a consequence of training goals, conditioning schemes or feedback-guided optimisation. The multimodal extensions also emphasize the versatility of the transformer structure, and its capability to have heterogeneous modalities of data, as well as suggest limitations with the existing evaluation models and representational hypotheses.

Coupled together, the evidences indicate that the last stage is the next-token prediction, and a fresh conceptual change in the construction, judgement, and execution of generative models is needed. As opposed to reading either the scale alone or benchmark dominance, the way to go is through coupled frameworks that trade off the computational efficiency, controllability as well as multimodal coherence in a principled manner. With the synthesis of these dimensions in large model families based on transformers, this paper contributes to the holier picture of generative modelling and creates the need to have evaluative metrics, architecture as well as strategies that reflect the narrowness of scope of the world and social needs.

REFERENCES

[1] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716–23736.

[2] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations.

[3] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

[4] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 610–623.

[5] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3, 1137–1155.

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[7] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509.

[8] Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, Ł., Belanger, D., Colwell, L., & Weller, A. (2021). Rethinking attention with performers. International Conference on Learning Representations.

[9] Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, Ł., Belanger, D., Colwell, L., & Weller, A. (2021). Rethinking attention with performers. International Conference on Learning Representations.

[10] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30, 4299–4307.

[11] Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., & Liu, R. (2020). Plug and play language models: A simple approach to controlled text generation. International Conference on Learning Representations.

[12] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.

[13] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., & Vinyals, O. (2022). Training compute-optimal large language models. Advances in Neural Information Processing Systems, 35, 30016–30030.

[14] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. Proceedings of ACL, 328–339.

[15] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations.

[16] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

[17] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858.

[18] Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. International Conference on Learning Representations.

[19] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising

sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of ACL, 7871–7880.

[20] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., & Christiano, P. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

[21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning.

[22] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI technical report.

[23] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

[24] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. Proceedings of the 38th International Conference on Machine Learning, 8821–8831.

[25] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[26] Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. Proceedings of EMNLP-IJCNLP, 3407–3412.

[27] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of ACL, 3645–3650.

[28] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2021). Efficient transformers: A survey. ACM Computing Surveys, 55(6), 1–28.

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[30] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Leike, J., Winograd, T., Bengio, Y., & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.