IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 4, Ser. 4 (Jul. – Aug. 2024), PP 12-19
www.iosrjournals.org

Navigating the Landscape of Text-based Gen-AI: Exploring Generic, Domain-Specific, and Indigenous LLMs

Roli Bansal[1], Richa Sharma[1#], Rahul Arora[1]
[1]*Department of Computer Science, Keshav Mahavidyalaya, University of Delhi, Delhi, India*
[#]*Corresponding Author*

Abstract:
Generative Artificial Intelligence (Gen-AI) has caught the attention of the world in a flash. It would be fair enough to say that such a steep escalation of Gen-AI has marked another era of technological revolution. Earlier, in the galaxy of technological innovations, there used to be a belief that whatever the invention is, it will take time for most people to understand and get used to it. But, Gen-AI has followed a different path, thanks to its convenient and user-friendly interface and functionalities. Gen-AI, especially LLMs (Large Language Models), have marked a significant transformation in the landscape of technology. LLMs have a multitude of applications in the fields of education, virtual assistance, software development and many more. These models have quickly evolved from early versions like GPT-1 to advanced iterations such as GPT-4. This paper aims to evaluate and compare the progress of LLMs across various dimensions like architecture capabilities, limitations, etc. We have tried to analyze the evolution of these models on similar grounds in order to have a fair and just comparison. Also, we have explored the advancements that the recent version upgrades have provided to the LLMs. In such fast-paced transformations, it is imperative to share our thoughts about future possibilities of the next-generation LLMs and their interdisciplinary applications as well. This study accentuates the significance of rapid innovations in state-of-the art LLMs, along with those designed in India, which are on the way to revolutionize fields from analysis to automation and beyond.

Keywords*: Large Language Model (LLM); Generative Pre-trained Transformer (GPT); Generative Artificial Intelligence (Gen-AI); Natural Language Processing (NLP)*

Date of Submission: 15-08-2024 Date of Acceptance: 31-08-2024

I. Introduction

Generative Artificial Intelligence (Gen-AI) is a class of algorithms that can generate new content, such as text, images, music or other data, based on the patterns that it learns from the training datasets. It has versatile applications from art and design to software development. Gen-AI models have progressively evolved in their complexity and scope. Large Language Models (LLMs) are specialized types of Gen-AI that can process, understand and generate human language based on training datasets. While understanding the definitions, the reader might question- "Are Gen-AI and LLM one and the same thing?" Essentially, LLM is a type of AI model specialized in language, whereas Gen-AI is a broader field that comprises LLMs and other models capable of creating different forms of content. So, it will be appropriate to say- "All LLMs are Gen-AI, but not all Gen-AI are LLM."

LLMs have played an important role in advancing the galaxy of Artificial Intelligence (AI), resulting in sophisticated language interpretation and interaction capabilities. LLMs like OpenAI's [1] trendsetter GPT series, Google's multimodal LLM – Gemini [2] and Meta AI's one of the recent innovations – LLaMA [3] have set higher levels of what AI can achieve, with each progression enhancing the models' linguistic capabilities. The objective of this paper is to analyze these advancements, comparing different versions of LLMs through the dimensions of architecture, capabilities, limitations, etc. Usually, when a trendy software tool comes into limelight, it seems to be a sudden discovery, but behind the scenes, it involves a lot of time and effort spent in understanding, developing and improving its fundamentals.

The roots of Gen-AI can be traced back to the early development of neural networks and rule-based systems, but ultimately it was the advent of deep learning and transformer architectures that enabled breakthroughs in the AI's creativity aspects. Generative Adversarial Networks (GANs) and Natural Language Processing (NLP) have set the stage for the modern LLMs, which are fueled by improved computational ability and access to vast amounts of training datasets. Before proceeding further, let's have a look at some key terminologies that will be implicitly used in the further comparative analysis.

Deep Learning is a subset of machine learning that involves multi-layered neural networks which have the ability to model complex patterns in data. Neural Networks are inspired by the human brain, designed to recognize patterns and associations in data. Generative Adversarial Network (GAN) [4] is a type of machine learning framework that puts two neural networks against each other in a competitive scenario, in order to generate new data that is indistinguishable from real data. GANs usually have two components - Generator and Discriminator. Generator neural network creates new data instances and starts with random noise elements and then learns to produce more realistic outputs. Discriminator is a network that evaluates the data generated by Generator and determines its authenticity. This adversarial relationship helps both networks to improve over time, ultimately resulting in the generation of highly realistic data. Although GANs are not directly involved in the architecture of LLMs, still they play an important role in the development of LLMs. GANs help in generating high quality data, that can be used for training LLMs, consequentially leading to improved generalization ability of the models. The fundamentals of GANs continue to play an important role in the evolution of Gen-AI models. Natural Language Processing (NLP) is sub-field of AI that focuses on interaction between computers and human language. It interprets human language in various forms, then extracts information and meaning from the data and whenever required, also generates human-like text. NLP techniques set the fundamentals for LLMs to understand the nuances of human language, including syntax and semantics.

LLMs have quickly evolved from early generative pretrained transformer models like GPT-1 [1] GPT-2 [1], BERT [5] to comparatively recent models like GPT-3 [1] GPT-4 [1], LLaMA and Gemini. There are also other significant models such as T5 (Text-to-Text Transfer Transformer) [6], XLNet [7] , BLOOM [8] and many others. Early models like GPT-1, introduced in 2018, established the ground for transformer-based architectures. Then, GPT-2 followed, scaling up significantly to 1.5 billion parameters leading to enhanced accuracy. During the same time, BERT introduced a bidirectional approach which helped in better interpretation of context, subsequently becoming a benchmark for various NLP tasks. The introduction of GPT-3 brought major transformation, mainly due to its 175 billion parameters. It also worked well without the explicit task-specific training. GPT-4, one of the latest LLM models, launched in 2023, further improved this technology, including multimodal capabilities and also addressed a few glitches and biases which were the drawbacks of previous models. Models like T5 and XLNet unified task formats and improved dependency modeling. BLOOM, a multilingual model emphasized on the transparent and ethical usage of LLMs, while ensuring inclusivity in AI development. Many more models have contributed in different dimensions.

In the upcoming time, the use and integration of LLMs will keep achieving newer heights. There will be hardly any sector that will be untouched by its influence. Following are some of the potential domains where LLMs are proving to be effective.

- Healthcare: LLMs can analyze vast amounts of openly available medical data quickly. These models can also help medical professionals to create personalized treatment plans. By understanding complex scientific literature, LLMs can also help in predicting the implications of medicines.

- Environment: LLMs, by analyzing massive amount of climate data, can help in predicting pattern of environmental degradation, pollution levels and climate change. These models can also suggest optimal usage of natural resources.

- Education: Personalized learning patterns and guidance can be generated via LLMs. They can also generate educational materials like quizzes, mind-maps, etc. to cater to specific requirements of students.

- Finance: LLMs can analyze market trends and economic indicators to help in assessing investment opportunities for individuals and organizations. They can also help in predicting financial risks and developing strategies to avoid them.

Recent studies on LLMs have mainly focused on advancements in model architecture, such as evolution of transformer-based models. Although these studies have compared the capabilities of these models, but sometimes, they have overlooked the rapid evolution of new LLMs. Our paper aims to fill this gap by providing a comparative analysis of some of the latest generic LLMs, including Gemini and LLaMA. Also, we tried to evaluate the performance of these models by sample testing through various prompts. Subsequently, we also aim to explore the domain-specific and indigenous LLMs. Furthermore, our discussion on interdisciplinary applications and ethical considerations of LLMs adds a constructive dimension to ongoing explorations in this domain.

II. Comparative study

The most fascinating feature of this Gen-AI evolution is the progress that all tools have made together, while addressing each other's drawbacks and ensuring a competitive environment, which keeps on pushing the limits. This has ultimately led to such quick advancements and version upgrades. In this sequence, the first framework to discuss is the Generative Pretrained Transformer (GPT). GPT has the following features.

- Generative: Model's ability to generate text based on the input

- Pretrained: Model is trained on large datasets
- Transformer: Self attention mechanism to process input data

A transformer is a deep learning model architecture introduced by Vaswani et al. [9] in the paper "*Attention is All You Need*". It is based on the self-attention mechanism, which gives the model the capacity to evaluate the relative weight of various words in a sentence. This method has significantly increased the accuracy of understanding the context of the input data and users' requests. Transformers can handle massive amounts of data and they can learn complex patterns easily. This makes them highly scalable and hence, they have been integrated into models like GPT, BERT, T5, etc. which can handle large amounts of data. According to the basic mechanism of a transformer, as illustrated in Figure no. 1, the encoder processes the input data to generate its context-rich version. The encoder consists of multiple layers of self-attention and feed-forward networks. The encoder produces a sequence of vectors, which capture the meaning and relationships between the words in the input. On the other hand, decoder produces the output sequence. The decoder also comprises a self-attention mechanism that processes the generated tokens and a cross-attention mechanism that integrates information from the encoder's output. The final output sequence is typically generated, one token at a time.

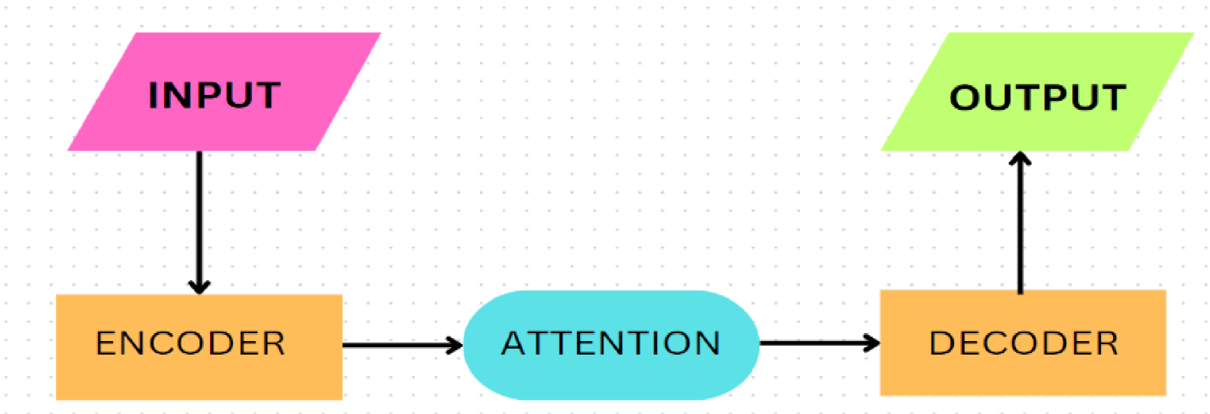

Figure no. 1: Block diagram of basic mechanism of a transformer

Training of LLMs:

As stated above LLMs have evolved from the pretrained transformer models. Thus, robust LLMs are the results of efficient training process. The basic procedure of training an LLM is illustrated in Figure no.2. The first step in training LLMs is gathering diverse datasets. The aim is to accumulate a corpus that helps the model to have a broad understanding of language patterns, topics, styles, contexts, etc. Then, the collected data needs to go under preprocessing to make it suitable and refined for model's training. Preprocessing includes several steps like removing noise, normalizing text and tokenizing sentences. Ensuring efficient preprocessing leads to consistent data for further training of the model. In the main training phase of the model, its parameters are adjusted to reduce the contrast between its predictions and the actual data. After training, the model undergoes evaluation on a validation dataset. These insights help in assessing model's accuracy and coherence. This also helps in identifying the focus areas, wherever improvement still needs to be done. After the evaluation, fine-tuning adjusts the trained model on a more specialized dataset to improve its performance on task-specific requirements.

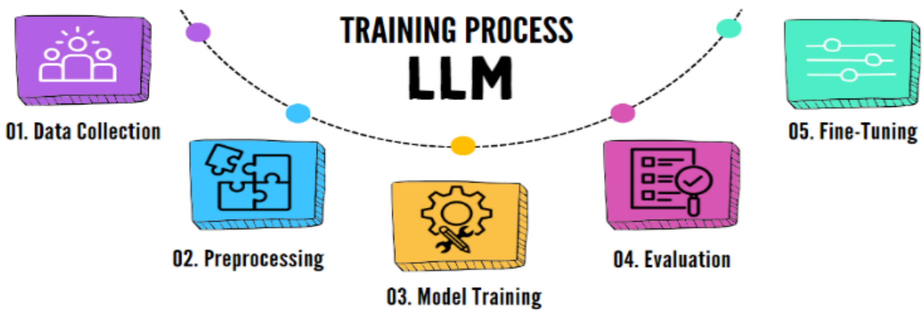

Figure no. 2: Training of LLMs

Now, let's delve into the comparative study of different LLMs and their available versions. For a fair, just and easily interpretable analysis, we first compare the GPT series models and then we move towards the comparative analysis of GPT with other popularly used models. Firstly, Table no. 1 presents the comparison of GPT-1, 2, 3 and 4. The data of the table has been compiled from various sources [19][20][21]:

Table no. 1: Comparison between GPT-1, 2, 3 and 4

Feature	GPT-1	GPT-2	GPT-3	GPT-4
Year of release	2018	2019	2020	2023
Architecture*	Transformer, 12 Layer Decoder	Transformer, 12-48 Layers Decoder	Transformer, 96 Layer Decoder	Transformer, with multimodal support
Parameters*	117 million	1.5 billion	175 billion	More than 1 trillion
Training data*	BookCorpus, English Wikipedia	8 million web pages	Multiple gigabytes of diverse datasets including web pages, books and Wikipedia	Even broader datasets compared to GPT-3
Capabilities	Basic text generation, translation and summarization	Better coherence and longer text sequences	Versatile task performance, comparatively higher fluency	Advanced reasoning, multi modal capabilities
Limitations	Limited coherence, lack of longer text sequences	Limited contextual understanding	Involvement of biases	Very high computational resources requirements
* Estimated values				

Due to the highly competitive environment, many companies do not disclose the exact architecture, training data set, number of layers used in their models. Hence, the above-mentioned numeric data-based comparisons are based on the estimated values.

All the 4 versions of GPT series are based on transformer architecture. The primary difference is reflected by the layers in their respective decoders, number of parameters and their training data. Deeper decoders can capture long-range dependencies between tokens in a sequence, consequentially leading to better contextual understanding. In the context of LLMs, parameters are the numerical values learnt by the model during training. These numerical values reflect the weights and biases that represent the model's behaviour. Usually, there is a positive correlation between the number of parameters and the performance of the model, helping the model to generate more relevant outputs.

The evolution of the GPT series marked a massive advancement in the domain of LLMs, each version building on the strengths of the predecessors while addressing their limitations. GPT-1 laid the foundation with the transformer architecture, enabling basic text generation, but it struggled with contextual understanding during long passages due to comparatively lesser number of layers in the decoder. GPT-2 extended the model's potential by increasing the parameter count to 1.5 billion, consequentially leading to better understanding of the context and improved coherence. Still, it had concerns about frequent generation of misleading content. The release of GPT-3 marked a substantial leap forward, with a massive count of 175 billion parameters, allowing the model to perform many tasks with minimal task specific training. However, it still couldn't get rid of biases, due to which it suffered from inaccuracy in few sophisticated tasks. GPT-4, the latest version available to users, further pushed the potential of LLMs. It included multi-modal capabilities and refined reasoning process, incorporated robust safeguards to reduce harmful outputs. However, all this necessitated the sacrifice of comparatively higher number of computational resources. After comparing the GPT series, we now present a comparison of the highest performing version of the GPT series, i.e., GPT-4 with Gemini and LLaMA in Table no. 2. The data of the table has been compiled from various sources [22][23][24].

Table no. 2: Comparison between GPT-4, Gemini, LLaMA

Feature	GPT-4	Gemini	LLaMA
Developer	OpenAI	Google DeepMind	Meta
Year of release	2023	2023	2024
Parameters	Estimated > 1 trillion	Estimated > 1 trillion	70 - 405 billion
Architecture	Transformer	Transformer	Transformer
Open source	No	No	Yes
Versions	GPT-4o, GPT-4o Mini, GPT-4 Turbo	Gemini- Ultra, Pro, Flash, Nano	LLaMA 3.1, 8B, 70B, 405B

OpenAI's GPT-4 and Google's Gemini, both involve over 1 trillion parameters, extensive training datasets and significant computational resource requirements. They are designed for large scale applications with advanced fine tuning and multi-modal capabilities. Meta's LLaMA is specifically more focused on efficiency and accessibility, optimized for deployment in diverse environments, including those with limited resources. While exploring the capabilities, we tried to observe the outputs of the above-mentioned Gen-AI models on several input prompts. Although the responses were overall similar, but there were differences in the structure of output, possibly due to the variations in contextual understanding capabilities or some internal algorithms. The prompts and the respective outputs generated by GPT-4o (based on GPT-4), Gemini and LLaMA 3.1, respectively, are presented in Figure no. 3 (a-c).

- Prompt 1: "*What came first, egg or hen?*"

  As per the outputs given in Figure no. 3(a), GPT-4o considered the scientific and philosophical perspectives to answer the question. Gemini considered Biological and Semantic perspectives to answer the question. It gave contrasting answers in these two perspectives. LLaMA considered the fundamentals of evolutionary biology and also considered the views of scientists and experts. It also tried to provide the reason for the answer in a chronological way.

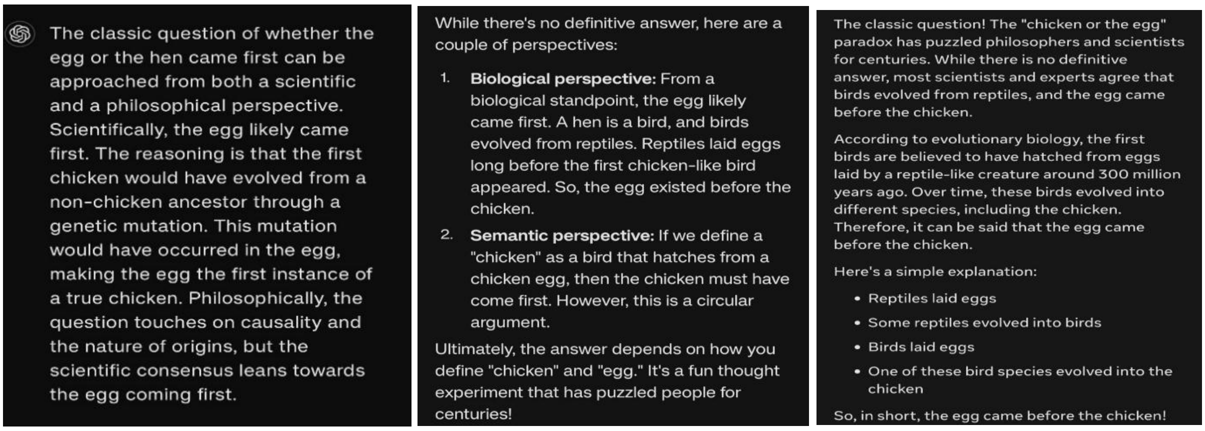

Figure no. 3(a): A sample output generated by GPT 4-o, Gemini and LLaMA 3.1's respectively for prompt 1

- Prompt 2: "*Who is the captain of Indian team?*"

  As per the outputs given in Figure no. 3(b), all the three models considered 'Men's Cricket' as the context automatically, even when the sport's name wasn't mentioned in the prompt. This reflects that there are still loopholes in contextual understanding. GPT 4-o provided a partially incorrect answer, which reflected its loopholes in accessing accurate real-time, latest information from web sources. On the other side, both Gemini and LLaMA were able to answer the question correctly, which shows their potential of accessing and utilizing real-time, latest information to generate output.

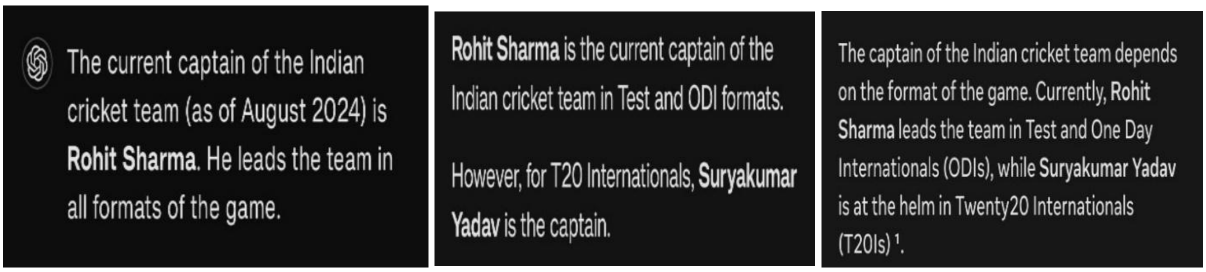

Figure no. 3(b): A sample output generated by GPT 4-o, Gemini and LLaMA 3.1's respectively for prompt 2

- Prompt 3: "*Describe the world in 5 lines.*"

  GPT 4-o generated the answer in exactly five lines. It generated a balanced answer, including the characteristics of the world along with the challenges faced by the world. Gemini generated a comparatively brief answer. It also didn't mention about the concerns, issues faced by the world. On the other hand, LLaMA utilized its entire capacity of 5 lines (as per input), but its answer was not as balanced as that of GPT 4-o. (Note: Assessing the comprehensiveness of the output is subjective)

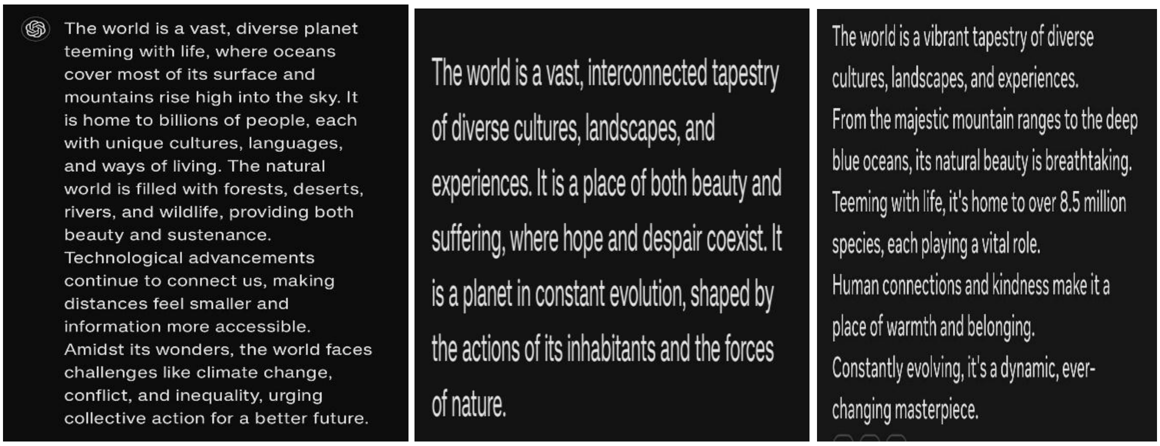

Figure no. 3(c): A sample output generated by GPT 4-o, Gemini and LLaMA 3.1's respectively for prompt 3

Although a few prompts cannot be used to arrive at a conclusion, still, the experiment adhered to the comparative analysis that we have seen earlier. GPT 4-o (a version of GPT-4) tried to generate comprehensive answers, but it faces drawbacks during generation of real time accurate information from the different web sources. Gemini's responses were also quite similar to that of GPT, but not as much comprehensive as GPT-4. since it didn't receive sufficient context in order to generate a specific response. So, it can be observed that Gemini is trying to focus on the accuracy of the output and waits for sufficient contextual details from the user. LLaMA 3.1 generated comparatively optimized responses by combining the related content in the response. Still, it could have been more comprehensive. Both, Gemini and LLaMA have been able to generate real-time accurate information, usually.

Further, this study also explored some domain specific LLMs developed around the globe and then indigenous LLMs developed to cater to multiple applications in the Indian context.

III. Domain Specific LLMs

Domain-specific LLMs are general models that are trained and designed to perform specific tasks which are related to the main objective of the organization, field or domain. In facilitating domain-specific tasks, usually they outperform generic LLMs because of requirement specific training and fine-tuning. For example, a word like "board" has different meanings in different contexts. A generic LLM might not be able to produce the required results here, but a domain specific LLM would be easily able to associate the context to the input. Usually, in domain specific LLMs, fine-tuning is used by tuning existing LLMs to a specific domain. During the process of fine-tuning, the model alters the parameters that were used in the training of the model, as per the requirements of the domain. Now, let's have a quick look at the features of some domain-specific LLMs:

- BioGPT [10]: It is designed for the domain of biomedical research. It has the potential to be used in medical literature summarization and clinical decision analysis and support.
- Codex [11]: It translates natural language to code. Its training data contains natural language and lot of lines of source codes from openly available sources.
- BloombergGPT [12]: Developed by Bloomberg, it is a 50 billion parameter LLM that is trained on financial data, including news articles, analyst research, financial reports, etc.
- ClimateBERT [13]: A model that is trained on climate data from various sources, ClimateBERT leads to lowering error rates for various climate-related tasks, including analyzing companies' climate-risk disclosures.

IV. LLMs Developed in India

India is also taking huge leaps in the development of LLMs. While some models are domain specific, some others are specifically developed keeping in mind the linguistic diversity present in India. Even, the LLMs developed in India have also created a diversity of their own. The unity exists by the shared belief of accessible and affordable technology for all. Let's have a look at the features of some LLMs developed/being developed in India:

- Krutrim [14]: Developed by Ola, Krutrim is a multilingual AI model. Its base model is trained on 2 trillion tokens (estimated) and the larger multi-modal version is designed for advanced problem solving and complex task fulfilment.
- Bhashini [15]: An initiative from Ministry of Electronics and Information Technology, Bhashini aims to translate and facilitate communication across diverse languages. It aims to provide the access to approximately 10-15 courses in 12 Indian languages to the students.
- OpenHathi [15]: Developed by Sarvam AI, its version Hi-v0.1 is the first Hindi LLM version in its OpenHathi series. It's built on Meta AI's LLaMa2-7B architecture and the developers believe that its performance is almost similar to GPT 3.5 for Indic languages.
- Indus Project [16]: The Indus Project, by Tech Mahindra aims to empower all Indic languages that have originated out of the Indus civilization. The company aims to build an Open-Source Large Language AI model that can serve the needs of 25 percent of world's population.
- Project Vaani [17]: Being collaboratively developed by Indian Institute of Sciences (IISc), AI and Robotics Technology Park (ARTPARK) and Google, Project Vaani's initiative is based on transcribing open-source anonymized speech data from almost all districts in India.
- Vaidya [18]: Developed by AI firm Fractal, Vaidya – India's first medical LLMs are fine-tuned open-source base models that involve 30-70 billion parameters and these base models are trained on more than 600k+ images and 200k+ text inputs. Essentially, it is a general-purpose assistant which can help users with diagnosis, treatment and medical advice. Ethically, it encourages the users to still take the advice of their trusted health professionals and use the model's advice just a guiding light and not the ultimate suggestion/remedy.

V. Conclusion

One common aspect that we have observed in the above comparisons, is the need for massive amounts of training datasets. Today, quality data collection for training has become a costly task in itself. So, at this point, the ability to self-generate the training data is probably the next big thing in the universe of Gen-AI. Any frequent user of the trending LLMs and Gen-AI chatbots might have seen a similar disclaimer written on the interface- "*This model can make mistakes. Kindly verify the data.*" But, in the rapidly evolving technological world where even few milliseconds matter, the users want a typical "All-in-One" structured model which can even automatically assess the facts or content generated by it. Already, many developers have started working or researching on it. Enhanced reasoning capabilities would make LLMs more transparent and understandable, allowing the users to observe the actual procedure of decision-making by the model. This will also ensure accurate feedback from users. There should be increased focus on mitigating biases in the LLMs, in order to generate reliable content.

There are two sides of a coin, and LLM is no exception. While it has huge potential to benefit the mankind, but its misuse can have detrimental impacts of the same intensity. LLMs can be exploited to generate fake news and misleading content. They can also be utilized to generate sophisticated phishing mails and they can even automate the discovery of security vulnerabilities in software systems. The huge amount of training datasets also give rise to privacy violations. Such threats highlight the significance of ethical practices in the development, deployment and usage of such models. The motive of these technical innovations is to help the mankind, not to disrupt the society. The onus lies on the stakeholders, including the developers and users to ensure the ethical usage of LLMs for greater good of the society. Promoting responsible use of these LLMs and the related chatbots, should also be the focus for upcoming versions of Gen-AI.

References

[1] Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. In Future Internet (Vol. 15, Issue 6). MDPI. https://doi.org/10.3390/fi15060192

[2] Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: a review of emerging educational technology. In Smart Learning Environments (Vol. 11, Issue 1). Springer. https://doi.org/10.1186/s40561-024-00310-z

[3] Team, L., & Meta, A. @. (2024). The Llama 3 Herd of Models.

[4] Xu, Y., Li, Q., Zhao, X., Liu, D., & Shi, N. (2024). The Applications of Generative Adversarial Networks in Architecture Design: A Systematic Review. Journal of Design Service and Social Innovation, 1–15. https://doi.org/10.59528/ms.jdssi2024.0130a14

[5] Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (n.d.). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://github.com/tensorflow/tensor2tensor

[6] Fang, S. (2024). A Comprehensive Survey of Text Encoders for Text-to-Image Diffusion Models. EAI Endorsed Transactions on AI and Robotics, 3. https://doi.org/10.4108/airo.5566

[7] Xiao, L., Li, M., Feng, Y., Wang, M., Zhu, Z., & Chen, Z. (n.d.). Exploration of Attention Mechanism-Enhanced Deep Learning Models in the Mining of Medical Textual Data.

[8] https://medium.com/@marketing_novita.ai/bloom-an-in-depth-guide-for-the-multilingual-large-language-model-a9aaa7cd08a4

[9] Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (n.d.). Attention Is All You Need.

[10] https://huggingface.co/docs/transformers/en/model_doc/biogpt

[11] https://openai.com/index/openai-codex/

[12] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (n.d.). BloombergGPT: A Large Language Model for Finance.

[13] https://www.chatclimate.ai/climatebert

[14] https://www.olakrutrim.com/

[15] https://www.analyticsvidhya.com/blog/2023/12/llms-that-are-built-in-india/

[16] https://www.techmahindra.com/insights/press-releases/tech-mahindra-launches-project-indus-large-language-model-llm/

[17] https://www.moneycontrol.com/news/technology/indias-ai-uprising-can-a-challenger-emerge-in-the-llm-marathon-11931681.html

[18] https://economictimes.indiatimes.com/tech/artificial-intelligence/ai-firm-fractal-has-built-indias-first-medical-llm-vaidya-ai/articleshow/112517944.cms?from=mdr

[19] https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2

[20] https://hgs.cx/blog/from-gpt-1-to-gpt-4-a-look-at-the-evolution-of-generative-ai/

[21] https://360digitmg.com/blog/types-of-gpt-in-artificial-intelligence

[22] https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

[23] https://explodingtopics.com/blog/gpt-parameters

[24] https://llama.meta.com/