nasscom

Deloitte.

Large language Models (LLMs)- A Backgrounder

September 2023

Large Language Models (LLMs) – A Backgrounder

1. LLMs – The Story So Far

To begin with, Generative AI as a whole is powered by large and complex deep learning models pre-trained on vast amounts of data, commonly referred to as Foundation Models (FMs). LLMs are a variant of these Foundation Models which have been specifically trained on massive amounts of text data – including but not limited to books, articles, websites, code etc. LLMs use complex and sophisticated statistical models to analyze vast datasets, identify patterns and connections in the data between words and phrases, and leverage these to eventually generate completely new text. Their applications span from conversational AI that furnishes contextually relevant responses, to condensing intricate documents into concise synopses, effecting smooth language translations, and even auto-generating intricate code snippets.

Evolution of LLMs

The evolution of LLMs has been a constant journey of breakthroughs and innovations ever since it began with early neural network experiments in the 1950s. The pioneering step forward, however, was the creation of Eliza, a primitive chatbot, by Joseph Weizenbaum, an MIT researcher. Eliza highlighted the potential of natural language processing (NLP) and eventually began the journey towards LLMs as we know them today [1].

Three decades later, complex neural networks could be envisaged due to the advent of Long Short-Term (LSTM) networks in 1997. Moving ahead, in 2010, Stanford introduced the CoreNLP suite which enabled sentiment analysis and named entity recognition, further pushing practical applications of NLP.

In 2011, with Google's advent of word embeddings, a new level of contextual understanding was brought to NLP. This set the stage for the rise of transformer models in 2017, characterized by the development of GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers). Transformer models are advanced neural networks that learn context and consequently meaning by tracking relationships in sequential data. Transformers are considered as a breakthrough in this field because of their ability to leverage the concept of self-attention, unlike typical networks based on convolutions or recurrences.

Soon after, GPT-3 launched in 2020 with a massive 175 Bn parameter model, setting the standard for LLM capabilities. The introduction of ChatGPT in late 2022 brought LLMs into significant public attention which has been escalating ever since and generated exponentially more practical use cases than anyone could visualize. LLM evolution continues with GPT-4, which is a 1 trillion parameter model representing a quantum leap in size and capabilities of predecessors. Innovation in LLMs continues every moment and newer, faster, and more accurate models are being developed and deployed at all scales across the spectrum.

Key Aspects of an LLM

To quantify an LLM in terms of its size, model complexity and training resource intensity, it is prudent to look at Parameters and Tokens. These are two important metrics used to measure the size and complexity of an LLM.

- Parameters are the number of variables in the LLM's neural network. These variables represent the weights and biases that are used to learn the relationships between the input and output data. The more parameters an LLM has, the more complex it is and the better it can learn to generate text that is similar to the text it was trained on.
- Tokens are the basic units of text that the LLM uses to process and generate language. Tokens can be characters, words, or sub-words, depending on the chosen tokenization method. The more tokens an LLM has, the more expressive it can be in its output.

Metric	Description	Equivalence
Parameters	Number of variables in the LLM's neural network	More Parameters = More Complex LLM
Tokens	Basic units of text that the LLM uses to process and generate language	More Tokens = More Expressive LLM

As the complexity of LLMs increases, they tend to become more content-rich due to the greater number of parameters they possess. An LLM with a higher parameter count can ingest and process a larger volume of information from the training data. This expanded capacity enables the model to discern and differentiate finer details, better gauge relationships, and identify very subtle contextual cues within the text it encounters.

Challenges of LLMs

While LLMs represent a significant scientific breakthrough and can potentially provide significant benefits to humans in various fields, there is also a need to understand and be aware of concerns and challenges in using LLMs.

Accuracy

One of the more publicized issues of popular LLMs is a concept called 'Hallucination'. This refers to situations where LLMs tend to generate completely new information which is factually incorrect. Hallucination occurs because LLMs are trained on vast data sets which may not be well curated and complete, therefore, the model may associate certain phrases with concepts which are not relevant to it. It is therefore important that users of LLMs understand the hallucination aspect of the models and remain critical of the information provided by them in any field.

Bias

Many contemporary LLMs are primarily built and trained using vast amounts of data parsed from the internet, and as such, any attributes of this data are automatically passed on to the model. The internet as a whole is comprised of an ocean of data with significant societal, geographical, and political biases encoded in it, to name a few. It is not feasible for humans to eliminate and train the model on a totally bias free dataset. The net result of this is that LLMs tend to exhibit a certain level of inherent bias, depending on the training dataset used, regardless of how well the developers attempt to eliminate such inherent differences.

Bias in LLMs is a significant detriment to its use in various fields and any users must be aware of this potential issue beforehand. For example, LLMs used in recruitment could potentially inject a pre-existing bias into shortlisting candidates for selection.

Perishability

Perishability, in the context of LLMs, is related to the extremely rapid pace of advancements being made in the field of AI, especially Generative AI. Developing a medium-high complexity LLM requires significant investment of resources like GPUs, servers, as well as a well-trained workforce.

Given the rapid advance and innovations across all the layers of the Generative AI tech stack, investment decisions may potentially be influenced. In the case of computing power advancements, models trained on current generation chips in a given amount of time may fall victim to updated silicon which enables significantly reduced training time and cost. The competitive landscape also has a significant impact as evident in the recent decline in ChatGPT active users – there are always newer LLMs coming up, each offering different solutions. Finally, the viability of maintaining an LLM over the long run is debatable – OpenAI reportedly spends US$ 700,000 daily to keep ChatGPT up and running [22] – while this is an example at the far end of the spectrum, any moderately complex LLM will require significant operational expenditure capability. Private investment in developing and deploying LLMs could therefore be deterred due to the perceived issues concerning perishability.

Different Types of LLMs

a. Publicly Available LLMs

Publicly available LLMs are typically massive in scale, with over 500 billion parameters, and are accessible to developers, researchers, and organizations through API calls. These models

are hosted on cloud platforms and can be used for various NLP tasks such as text generation, translation, summarization, and more. Some of the leading examples include OpenAI's GPT models, Google's PaLM (Path Language Model), Meta's LLaMA, and NVIDIA's NeMo ™.

These models have gained popularity due to their versatility and ease of integration into various applications. They have the advantage of being continuously updated and improved by the developers, which helps ensure they remain state-of-the-art in terms of performance. The use of these models allows for their continuous training.

b. Private LLMs

Private LLMs are models that are customized or tuned for specific tasks or industries, or organizations and are operated within a private cloud or an in-house infrastructure. These models could be either purpose-built from scratch or fine-tuned versions of publicly available models. The primary aim of these LLMs is to address specific business needs, confidentiality concerns, or compliance requirements.

Organizations may opt to develop and manage their LLMs to have more control over costs and data security. Examples of private LLM providers include NVIDIA, which offers smaller to medium-sized LLMs, and platforms like Hugging Face and MosaicML that facilitate fine-tuning and customization. For example, Bloomberg has created BloombergGPT, a 50-billion parameter private large language model, purpose-built from scratch for finance to better serve their customers [20].

c. Open Source LLMs

Open Source LLMs are models that are made available to the public with their source code and parameters, allowing developers to modify, enhance, and adapt them for various purposes. These models are often part of a hybrid strategy for enterprises looking to combine the benefits of both public and private LLMs. By using open source LLMs, organizations can avoid vendor lock-in and have more control over their AI initiatives. Falcon40B and its scaled-up version "Falcon 180B" by the UAE is a notable example of an open source LLM where recent announcements made it free of royalties for commercial and research use [2].

Comparison of Contemporary LLMs

Numerous Large Language Models (LLMs) have been released, including prominent ones like OpenAI's GPT-4, Google's PaLM2, and Meta's LlaMA. These models differ based on their fundamental architecture, the data they were trained on, applied model parameters, and any specialized tuning for particular uses. Additionally, development costs and budget also play a role in shaping the performance of generative AI systems when responding to users' queries

Notable Large Language Models [3] [7]:

Company	Model	Launch Year	# Parameters in billions	#GPUs & Training Time
Open AI	GPT 3.5	2022	175	10k V100 GPUs/ 3500 A100 running for 240 Hours
	GPT 4	2023	1700	30K A100 GPUs, 34 days
Google and Deep Mind	Gopher	2021	280	-
	Chinchilla	2022	70	-
Google	PaLM	2022	540	6144 v4 TPUs/ 10,000 A100 GPUs for 1200hrs
	LaMDA	2022	137	-
Meta	OPT-175B	2022	175	1024 Nvidia A100 80 GB/2918 A100 40GB GPUs for 792 hrs.
	LlaMA	2023	65	2048 Nvidia A100 GPUs, 80GB for 500 Hours
Nvidia	NeMo ™	2021	530	-
Baidu	ERNIE 3.0	2021	260	-
BAAI [4]	Wu Dao	2022	1750	-

Company	Model	Launch Year	# Parameters in billions	#GPUs & Training Time
UAE	Falcon	2023	40	384 A100 GPUs for 1440 Hours
Amazon	AlexaTM	2022	20	-
Hugging Face	BLOOM	2022	176	-
Yandex	YaLM	2022	100	-
Anthropic	Claude	2023	175	-
Eleuther AI	Pythia	2023	1-12	32 A100 GPUs for 16 Hours to 64 A100 GPUs for 222 Hours

Source: nasscom and OECD Digital Economy Papers

3. Need for Sovereign LLM

Local Culture Aspect

Preserve and promote linguistic diversity and enhance cultural representation:

An LLM specifically for local languages would help preserve and promote linguistic heritage and represent the cultural nuances and sensitivities of local languages.

- **Improve accessibility:** A sovereign LLM would enable better access to information and services for speakers of local languages, reducing the digital divide and fostering greater inclusivity.
- **Support local economies:** An LLM tailored to local languages could bolster local economies by supporting the development of language-specific applications and services that cater to unique regional needs.
- **National security:** Developing a homegrown LLM allows for better control over data and technology, addressing national security concerns.

Accessibility & Affordability

The accessibility and affordability of running LLMs pose significant challenges for all the stakeholders involved. Although cloud computing might provide a more cost-efficient option, the overall expenses remain substantial, particularly when considering the extensive computational requirements of training models like GPT-3. As such, the financial barriers limit the ability of smaller companies and research groups with limited budgets to access and benefit from these advanced language models, potentially impeding innovation and democratization of AI technologies, making it imperative for the sovereign government to step in.

The significant training costs of LLMs, illustrated by the US$ 12 million price tag for training GPT-3 [26], display the roadblocks involved in further innovating in the field of Generative AI. GPT-3's training cost is 200 times greater than GPT-2, which accentuates the heavy scale-up required as the models become large and complex. As LLMs continue to evolve and expand in scale, the accessibility gap widens, making it challenging for startups, educational institutions, and researchers with limited funding to engage in meaningful AI research or develop innovative applications. In addressing these challenges, it is therefore crucial to ensure the availability and continuance of resources, both financial and operational, to create an indigenous LLM for the benefit of any sovereign nation.

Security Concerns

With the rapid rise and spread of ChatGPT and other publicly accessible LLMs, there has also been a simultaneous uptick in discussions and concerns related to data privacy and security when using these models. Illustrating an example in view of ChatGPT, any queries asked, information shared, or data generated can potentially be used by OpenAI to train, re-train, and update the foundation models in the background (unless the user opts out of sharing data).

Major corporations including JPMorgan Chase, Amazon, Verizon, and Samsung have opted to prohibit their employees from using ChatGPT due to apprehensions surrounding the potential exposure of sensitive data on OpenAI's servers [25]. The decision to implement such bans reflects the heightened sensitivity surrounding data security and privacy in today's digital age. These companies have raised concerns about the risk of confidential information inadvertently finding its way onto external servers, prompting them to prioritize safeguarding their proprietary and confidential data.

It's plausible therefore that organizations with sophisticated technology teams are likely to proceed with Private LLMs and will aim to put appropriate guardrails in place, such as:

- Training the LLMs in secure in-house data centres or single tenant private cloud servers
- Leveraging the internal organizational workforce for training the LLM instead of external parties
- Establish privacy and security controls within the organization to ensure compliance with relevant local regulations.
- Conduct thorough testing of the model and mitigate any potential issues of bias, hallucination, or inaccuracies, and certify usage accordingly.
- Establish governance programs, such as LLMOps, to prevent model drift, establish trust, and promote safe and responsible use of output generated through the model.

Customizability & Flexibility

In addition to the aforementioned reasons, nations need their LLMs to be able to build specific toolsets to address gaps in critical areas. The ability to customize the model will ensure that the country can tailor the customer-facing applications to different sectors as needed. To take a few examples, applications in the Healthcare sector would require the model to be trained on medical jargon and gather contextual information on terminologies, treatments, and results. On the other hand, the Education sector would require a different approach with more focused training on a diverse set of topics and in helping the educator by generating training material. Therefore, having a domestic LLM is paramount to ensure a high degree of customizability and flexibility in further applications of Generative AI.

Global Examples of Sovereign LLMs

Japan

Several LLMs are being developed in Japan based on localized datasets [8] [9] [10] [11]. Some examples are as below –

- SoftBank is developing an LLM based on Japanese datasets that will serve as the basis of its generative AI. The number of parameters is currently about 1 Bn and is expected to go up to 60 Bn. Unlike existing models based on English and Chinese data, SoftBank's approach aims to capture local nuances.

  "The significance of a domestically produced generative AI is that it will be developed on a Japanese language data set making it more suited to Japanese business practice and culture. We believe that it will be able to handle unique expressions in Japanese, such as those used in public and medical services." – As per SoftBank CEO and President Junichi Miyakawa during an earnings presentation [9] [29]

- CyberAgent released an LLM in May that specialises in Japanese language and culture.
- Rinna Co., a startup headquartered in Bit Valley, has released a large-scale LLM specializing in Japanese.
- Fujitsu Ltd. and the Tokyo Institute of Technology have begun developing LLMs specializing in Japanese with the help of the supercomputer Fugaku

Israel

Both government and private sectors are investing in the development of multi-lingual and customizable LLMs [12] [13]:

- A critical goal of Israel's Innovation Authority and AI experts under the National Program for AI Infrastructure is the development of a state-of-the-art large-scale national language model that will function in Hebrew and Arabic
- In March 2023, AI21 Labs, an Israel-based AI research lab that offers its NLP capabilities as a service, released an upgraded version of its LLM called Jurassic-2 – which is claimed to be the world's most customizable large language model for organizations that want to build advanced, chat-based AI applications at large scale.

UAE

In 2019, UAE unveiled a roadmap to position UAE as an AI global leader by 2031. As part of this, the government set up the Mohamed bin Zayed University of Artificial Intelligence, an institution solely focused on driving AI education in the country. They also released an open-source large language model called "Falcon 40B" shortly

followed by its scaled-up version "Falcon 180B", available for research and commercial use.
While Falcon 40B is a model with 40 Bn parameters and trained on one trillion tokens developed by the Technology Innovation Institute (TII), the latest version is a model with 180 billion parameters and was trained on 3.5 trillion tokens using Amazon SageMaker for a total of ~7,000,000 GPU hours. Ranked at the top of the Hugging Face Open LLM leaderboard, it outperforms competitors like LLaMA, StableLM, RedPajama, and MPT. It is trained primarily in English, German, Spanish, and French, with limited capabilities in Italian, Portuguese, Polish, Dutch, Romanian, Czech, and Swedish [14] [16][31].

TII in collaboration with LightOn is also developing a large-scale Arabic NLP model called 'NOOR' with 10Bn parameters and features applications in automated summarization, chatbots, and personalized marketing. NOOR was trained on a High-Performance Computing resource with 128 A100 GPUs, allowing for the distribution of computations and ensuring efficient use of available hardware resources.[32][33]

Europe (France, Spain and Sweden)
Europe has also seen several examples of advanced language-specific models across Spain and Nordic regions. Some examples are quoted below -

- In 2022, Hugging Face released its large open-access multilingual language model (BLOOM) in cooperation with French governmental agencies. BLOOM includes 176 billion parameters and is developed on France's Jean Zay public supercomputer [28]
- Madrid-based AI startup Clibrain has released a **Spanish-instruction tuned LLM called Lince Zero;** focus on language will enable this model to understand Spanish nuances better than an average LLM [17]
- **The GPT-SW3 initiative** in Sweden aims to develop a large GPT-model for **Nordic languages, primarily Swedish**. The aim is to create a foundational resource for Swedish NLP that represents the entire Swedish language, and the entire Swedish population [18] [19]

4. Key Considerations for India

Reviewing all the key aspects and arguments, it is imperative that while the necessary compute infrastructure is being set up, India should also pursue strategic efforts to initiate the process of developing a sovereign LLM, starting with major languages which have significant data availability across modalities (text, images, audio, video, code). To influence the direction of AI development and global priorities, India must focus on nurturing its capacities and enhancing its global competitiveness. While India has made multiple strides in AI skilling and strengthening its data ecosystem, concerted efforts are required to strengthen the overall AI ecosystem through key pillars:

Compute Infrastructure: Most of the world's AI compute infrastructure is based on NVIDIA's GPUs [23]. China and the United States are leaders in the AI compute infrastructure while other nations lag.

Large Language Models: The majority of current LLMs originate from the US and China, lacking representation of diverse languages and cultures. Therefore, echoing the initiatives undertaken by countries such as Japan and UAE which have stepped up efforts to develop local LLMs, India needs to invest in creating substantial and culturally relevant AI models that are trained in myriad local languages to cater to non-English speakers.

Access to High Quality Training Data: Current datasets, especially high-value datasets (HVD) suffer from several challenges impacting the usability of these datasets and limiting the scope of data-driven research and innovation ecosystem growing to its full potential in India. HVDs are typically 80% unstructured data and 20% from structured datasets.

Open-source initiatives like Bhasha Daan under Bhashini are positive initiatives in this regard and play a pivotal role in advancing this endeavor. They contribute to the accumulation of language-specific datasets, allowing researchers, developers, and the community to collaborate in collecting, curating, and sharing resources [37]. By collectively contributing to such efforts, stakeholders can ensure the quality and coverage, making it possible to train LLMs that better understand and cater to India's linguistic diversity.

The full value of India's data can be unlocked only if the datasets are relevant, complete, timely (updated), consistent and reliable. [34]

Access to AI Finance: There is a scarcity of patient capital for cost-intensive foundational model work due to a dearth of Indian deep-tech funds. Global generative AI startups have raised a cumulative funding of USD 19Bn+ during 2021-2023 of which only USD 475Mn has been raised by Indian Generative AI startups (70% of this funding has come in 2022 alone). There are 550+ global Gen AI startups worldwide (the majority of which are in the United States), compared to only 60+ Gen AI startups in India [27]. Hence, the need of the hour is to provide Indian AI startups access to risk capital at various stages of development to cover the exorbitant cost of building and maintaining sovereign compute infra and native datasets.

Take into account the Total Cost of Ownership

One of the key considerations as India sets up its compute infrastructure is to keep the Total Cost of Ownership (TCO) in mind. Total Cost of Ownership can be broken down into capital expenditure in the form of building the LLM and operational expenses accrued in training and inference of the Generative AI models developed on top of the base infrastructure.

Establishing, training, maintaining, and updating a local Large Language Model (LLM), entails significant operational expenses. Due consideration also needs to be given to the high rate of technological obsolescence and thereby continuous investments are required given the dynamic Gen AI ecosystem which requires constant innovation and adaptation.

Given the scale of costs, a more practical and feasible way to encourage developers and investors to build India's models would be to offer incentives that can defray the large on-going expenditure.

In further support of the national AI ecosystem and empowering Generative AI ambitions, India has recently witnessed the participation of the private sector in the development of LLMs:

- Reliance and Nvidia have announced a partnership to develop an indigenous foundational large language model trained on local languages and tailored for Generative AI applications in India [35]
- Tech Mahindra has announced the launch of Project Indus, an Indic-based foundational model for Indian languages, The project will be built on 7-billion parameters and would initially support 40 different Hindi dialects, with more languages and dialects to be added subsequently.[36]

To develop a domestic LLM, a set of strategic choices needs to be reviewed for implementation

- **Talent and Resource Procurement** - Efforts need to be directed towards the identification and acquisition of essential human capital and resources. This is imperative to assemble proficient teams and the necessary resources to ensure the success of the overall initiative.
- **Curating Vernacular Data Sets** – A wide assortment of diverse datasets needs to be curated, incorporating a wide array of sources such as news articles, websites, literary works, scripts, audio recordings, video and any other relevant sources. These datasets will serve as the bedrock for training the model with vernacular language capabilities.
- **Gradual Advancement via Incremental Model Development** - Adopt a phased approach, commencing with the creation of preliminary models in major languages. These models need to subsequently undergo an initial public beta testing phase, to collect and action on user feedback, creating a more refined model in the process.

Size Appropriate LLMs

One of the key strategic choices to be made in the process of building a domestic LLM is to agree upon the size and complexity required of the model. Depending on the envisaged usage of the LLM, the capabilities required for the potential use-cases, and the support available in terms of resources and long-term funding, India may potentially consider starting with a small to medium sized model.

Building a large language model similar to GPT-4 (>trillion parameters), or even GPT-3.5 (175Bn parameters), demands significant GPU compute capacity, a resource which is currently in high demand and short supply globally, notwithstanding the geopolitics of the supply chain

involved. Notably, ChatGPT required 10,000 Nvidia GPUs [21] for its training, a number which will present a considerable challenge in procurement in the short term. Therefore, to initiate its LLM journey, it will be more practical for India to consider crafting Small to Medium sized LLMs within the 5-50 billion parameters range. Focusing mainly on language/text and omitting images or other modalities could trim down resource demands, costs, and model complexity. This staged approach would enable India to gradually scale its language capabilities and foster innovation without facing challenging roadblocks.

Focused LLMs

LLMs can be built to generate content across modalities and subjects – however, building and maintaining a broad LLM is expensive and resource intensive. In contrast to developing broad-spectrum models like GPT3.5/4, a more strategic direction involves deploying focused LLMs with targeted expertise. For instance, a finance focused variant like BloombergGPT has been engineered to cater specifically to the nuances of the finance domain. This type of approach, which looks at a specific area, optimizes resource allocation, ensuring a higher degree of relevance and effectiveness in addressing specific areas of interest.

Incorporate into India Stack

In a similar vein to how the GPU compute should be part of the India Stack, the LLMs built on top of this infrastructure should also be hosted on the platform providing unrestricted access to Indian users. Integrating LLMs into India's Digital Public Infrastructure can democratize access to advanced AI capabilities, fostering innovation, education, and citizen engagement. By offering free access to LLMs across languages, this initiative could enable inclusive innovation, provide educational resources, enhance government services, and support local language communication. Pursuing this approach has the potential to position India at the forefront of AI-driven development while promoting collaboration, cultural relevance, and self-reliance.

Next Steps for India

- Collaborate with AI research institutions and experts to design language-specific architecture tailored to local language intricacies
- Review different options available for collaboration with developers and industry players including exploring collaborative models supported through incentives
- Establish partnerships with educational institutions, cultural organizations, and language experts to gather and curate diverse linguistic datasets
- Develop guidelines for data annotation incorporating linguistic nuances, idioms, and local context
- Leverage existing models and fine-tune them on the datasets to ensure accuracy and context relevance
- Engage with domain experts from sectors such as healthcare, agriculture, legal systems etc. to develop specialized LLMs catering to specific industries' terminology and requirements
- Collaborate with content creators, journalists, and educators to generate quality content in local languages
- Establish training programs and workshops for developers, researchers, and linguists interested in working on LLMs