
ITIHAASA POINT OF VIEW

MARCH 2023

UNDERSTANDING GENERATIVE AI *- HOW CAN INDIA LEVERAGE TOOLS LIKE CHATGPT?*

Kris Gopalakrishnan, Krishnan Narayanan, N Dayasindhu

OpenAI's ChatGPT has taken the world by storm. It is the fastest-growing consumer application in the history of internet applications, with over 100 million users in 2 just months.¹ Over 4 times faster than TikTok, and 15 times faster than Instagram to cross that milestone!² Many questions spring to mind. What is ChatGPT – is it a new technology, and how does it work? What explains its popularity? And besides chatbots, are there any other applications of this innovation? What are its limitations? What does it mean for India?

1. Looking under the ChatGPT hood

ChatGPT is not a new technology, but a new innovation, under the category of Generative AI, that has combined two existing technologies – chatbots and large language models (LLMs). We all have experienced customer service chatbots on many websites and apps, but these traditionally have been based on simple heuristics or a rules-based engine at the backend, and with limited conversational capabilities at the front-end. LLMs were introduced by Google in 2017 when it made its “transformer” AI architecture available as an open-source. Language models were developed for a simple objective – to predict the next word in a sequence of text / sentence, given the preceding set of words. We experience this in action in Gmail, where it prompts a set of words to complete a sentence, after one starts typing out a sentence. Other tools similar to ChatGPT include Chinchilla, LaMDA, Bloom and others.

What did OpenAI do to develop ChatGPT? The expansion of GPT – “Generative Pretrained Transformer” – gives a hint. OpenAI created a stack of these transformers and trained the language model on massive amounts of data gathered from the Internet and other sources (see box below for a comparison of the different versions of GPTs^{3,4} - ChatGPT is based on GPT3.5).

- GPT1 - 12 layers of transformers, 117 million parameters, trained on dataset of over 40GB of text from BooksCorpus
- GPT2 - 48 layers, 1.5 billion parameters, trained on 40 terabytes of text datasets from the internet sources including WebText (Reddit articles)
- GPT3 / GPT3.5 – 96 layers, 175 billion parameters, trained with 499 billion tokens with text from WebText, Wikipedia, Books1, Books2

Although language models typically only predicted the next word, but when these models were trained on sufficiently large and diverse corpora of text, they began to generate coherent, contextual, natural-sounding responses. Arithmetic capabilities (and fairly limited) in GPT-3 emerged from only a 30x increase in model size.⁵ Thus, OpenAI used pretrained transformers to create a generative AI chatbot, and hence its name, ChatGPT.

OpenAI paid particular attention to the user experience in creating their final product, and be very good at understanding user intent, maintaining context and remaining highly interactive throughout the conversation. It smartly used a technique called Reinforcement Learning from Human Feedback (RLHF) to produce that chat experience as if the user was interacting with another human (see box below for OpenAI's strategy⁶).

- First have human trainers write conversations – both as a user and as the bot. Use these conversations to train the GPT model. Thus, they created a model that could approximately answer like humans. Let's call this the "answering model".
- Next, they employed humans to rank the answers of the chatbot. They used this ranking data to train another model. Let's call this the "reward model".
- The final step was to use a reinforcement learning technique – similar to how we train a dog with a treat for a right task. The reward in this case was getting the highest grade for an answer. OpenAI pitted the answering model against the reward model a million times to fine-tune its algorithm, and achieve a human-like quality to its conversations.

2. Landscape and applications of Generative AIs and LLMs

Let us consider the landscape of Generative AIs, as suggested by Andreessen Horowitz (see figure below).⁷ The stack can be divided into three layers:

- Applications that integrate generative AI models into a user-facing product, either running their own model pipelines or relying on a third-party API
- Models that power AI products, made available either as proprietary APIs or as open-source checkpoints
- Infrastructure vendors (i.e., cloud platforms and hardware manufacturers) that run training and inference workloads for generative AI models

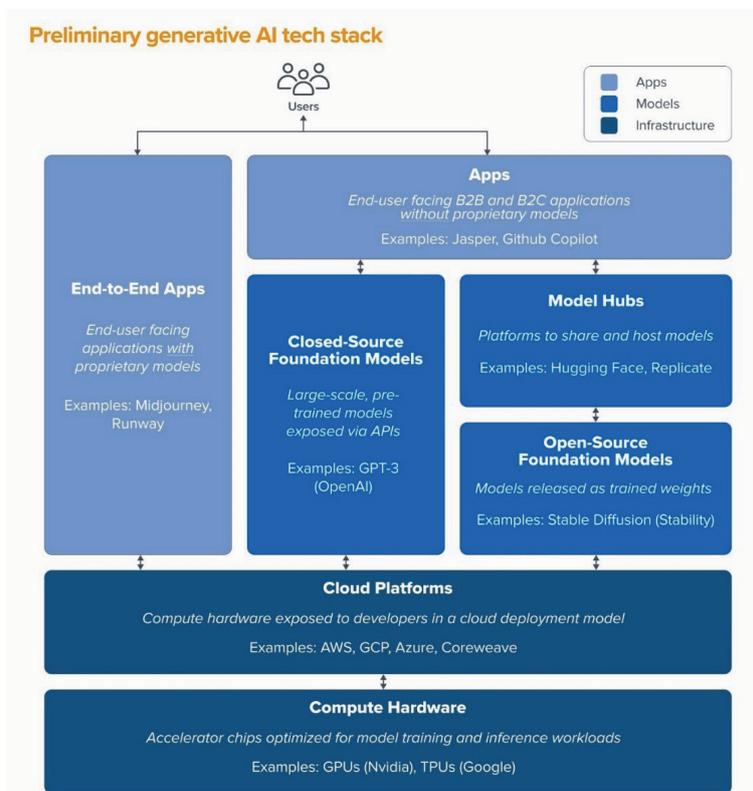


Figure: Preliminary generative AI tech stack; Source: Andreessen Horowitz

In this stack, let us consider examples, end-to-end, around OpenAI & GPT.

- Compute hardware – OpenAI had access to a 10,000 Nvidia V100 supercomputer when training GPT-3.⁸
- Cloud Platforms – ChatGPT and GPT 3.5 were trained on a Microsoft Azure AI supercomputing infrastructure.⁹
- Closed-source Foundation Models – GPT3 is an example of a foundation model, which Stanford has described as any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.
- Apps – Github Copilot, that emerged from a Microsoft – OpenAI collaboration, is an example of an app built on top of GPT foundation models. The GitHub Copilot feature, a new AI pair programmer, helps developers write better code. It uses the OpenAI Codex engine to suggest code and entire functions in real-time, right from the developer’s editor.¹⁰

Where can LLMs and foundation models like GPT be used?

For the vast majority of companies, it is difficult to train models from scratch – the infrastructure, quantity of data required and costs involved are too prohibitive. However, it is so much easier for them to pick a pre-trained LLM such as GPT-3, which is already trained on huge amounts of text data, and fine-tune it for a downstream task with company and industry-specific data.

The application areas are almost limitless, and we provide a few examples below.

<p>Search engines^{11,12}</p>	<p>Microsoft has just announced a ChatGPT powered Bing + Edge that provides a new search experience – Beyond generating a list of relevant links, the new Bing consolidates reliable sources across the web to give a single, summarized answer; users may also can chat naturally and ask follow-up questions to their initial search; and search may be used as a creative tool, write stories or share ideas for a project. Microsoft developed a proprietary way of working with the OpenAI model, called Prometheus model, that gives more relevant, timely and targeted results, with improved safety.</p> <p>Google augmented their search capability with the announcement of Bard, an experimental conversational AI service, powered by its foundation model, Language Model for Dialogue Applications (LaMDA).</p>
<p>Coding¹³</p>	<p>Developers can write software and teach robots physical tasks with large language models – The power of GPT3 has been brought to Microsoft’s Power Platform, the next generation business process automation and productivity suite. It is a low-code app development platform that helps everyone, from experienced developers to people with little or no coding experience, a.k.a “citizen developers” to build applications. Using GPT3, citizen developers or business users can describe what they want to do in natural language, and it will generate a list of the most relevant Power FX formulas for them to choose from.</p>

Healthcare¹⁴	<p>A number of opportunities to help healthcare providers, patients and in biomedicine</p> <ul style="list-style-type: none"> • Create efficient and accurate interface into EHRs (clinical notes, lab value histories and imaging files), helping healthcare providers create summaries of patient visitation, retrieving relevant cases and literature, and suggesting lab tests, diagnosis, treatments and discharges. • Provide an interface to patients, with relevant information about clinical appointments, answering patient questions related to preventive care, along with relevant medical explanatory information (e.g., text and graphics that explain conditions). • Researchers have shown that a transformer model trained on natural language could be adapted for other sequence-based tasks such as protein fold prediction, which is critical in biomedicine. • Given a set of drugs and a patient genome, foundation models may help predict which drug is likeliest to treat the patient with minimal side effects.
Customer service	<p>Improved customer service experience for any service provider (retailers, automotive OEMs and dealers, insurance providers, banks, telcos, etc.) – intelligent virtual assistants / chatbots that answer customer queries, help with comparisons and so on.</p>

Limitations of LLMs like GPT

While ChatGPT is an exciting innovation, has caught the attention of the consumers worldwide, and has immense potential for all businesses, it is not an AI silver bullet. ChatGPT hallucinates i.e., it imagines things which are not factual and shares it in a very convincing way. CNET, which began publishing articles written by ChatGPT as-is, started reviewing the accuracy of all its AI-written articles after several major errors were discovered.¹⁵

ChatGPT may have systemic and unwanted biases, based on the underlying datasets it was trained on. It has been modelled for ethical / appropriate behaviour. For instance, it is trained not to give answers that suggest violence. But ChatGPT's guardrails aren't borne of some kind of conceptual understanding of what "violence" means, but of something far more superficial, just the word. Thus, with devious prompting, ChatGPT can be easily tricked.¹⁶

Companies that are looking to adopt ChatGPT based applications in their work should keep in mind two big challenges – i) the individual company norms and values aren't reflected in ChatGPT, and it will have to be specifically trained on company culture and values; and ii) intellectual property issues are not clearly established – who can lay claim to new product ideas / designs that such generative AI apps recommend?¹⁷

We posed this very question to ChatGPT itself. See the table below for its responses.

Prompt: What are the potential challenges or risks of using LLMs for business applications?

While large language models have many potential benefits and applications for businesses, there are also a number of challenges and risks to be aware of. Here are a few:

- **Bias and discrimination:** LLMs are trained on large datasets, and if these datasets are biased in any way, the models can perpetuate and amplify these biases in their outputs. This can lead to discriminatory or unfair outcomes, particularly in sensitive areas like credit scoring or criminal justice.
- **Privacy concerns:** The large amounts of data processed by LLMs can raise concerns about the privacy of individuals and their personal information. This is particularly relevant for businesses handling sensitive information such as financial or health records.
- **Explainability and accountability:** Because LLMs are trained on vast amounts of data and generate outputs based on complex algorithms, it can be difficult to understand why they make certain decisions. This can make it challenging for businesses to ensure that their use of LLMs is fair, transparent, and in compliance with regulations and ethical principles.
- **Data quality and reliability:** The quality and reliability of the data used to train LLMs is crucial to the accuracy and utility of their outputs. Poor quality data or data that is not representative of the target population can lead to incorrect or misleading outputs.
- **Technical challenges:** Building and deploying large language models can be a technically complex process, requiring specialized expertise and significant computational resources. This can be a challenge for businesses, especially small or medium-sized enterprises, that may not have the necessary technical capabilities in-house.

3. How can India leverage tools like ChatGPT?

All the application areas that we explored in the previous section is relevant to Indian businesses as well. We will focus on two application areas which may be particularly important to India – i) language translation, and ii) coding.

Language translation-based applications

LLMs are particularly suitable to translate between English and regional languages, making information and services more accessible to a wider audience. Thus ChatGPT-based apps or chatbots can provide real-time customer support for e-commerce platforms, banks, and other businesses in India, especially in local languages. Similarly, applications may be developed to provide information on health conditions, advisories on precision agriculture, personalized investment advice, all in regional languages.

One exciting new pilot is by Bhashini, a team at the Ministry of Electronics and IT (MeitY), Government of India, which is building a WhatsApp-based chatbot that leverages content generated by ChatGPT.¹⁸ Satya Nadella described this pilot project: "A demo I saw was a rural Indian farmer trying to access some government programme. He just expressed a complex thought in speech in one of the local languages that got translated and interpreted by a bot, and a response came back saying 'go to a portal and here is how you will access the programme'. He said, 'I'm not going to go to the portal, I want you to do this for me.' The bot completed it, and the reason why it was able to complete it was because a developer building it had taken GPT and trained it over all of the Government of India's documents and then scaffolded it with the speech recognition software."

This is an example of combining two AI models – one is the ChatGPT model trained on Indian government policy / scheme documents and data, and the other is the language model trained on large datasets of the various local languages spoken in India. Bhasha Daan, is an ambitious project that aims to crowdsource voice datasets in multiple Indian languages. The pilot project supports 12 languages including English, Hindi, Tamil, Telugu, Marathi, Bengali, Kannada, Odia, and Assamese. Thus, if a user (farmer) were to ask a question i.e., send a voice note in any of these languages, the chatbot will successfully respond.

LLM-based apps to support coding

LLM-based apps can greatly enhance a coder's productivity.

Trained on billions of lines of code, GitHub Copilot turns natural language prompts into coding suggestions across dozens of languages. It shares recommendations based on the project's context and style conventions. Developers can quickly cycle through lines of code or complete function suggestions, and decide which to accept, reject, or edit. A recent analysis found that on average more than 46% of code is being built using GitHub Copilot across all programming languages, and 61% among developers using Java. Since its release, GitHub Copilot has transformed developer productivity for more than one million people, helping developers code up to 55% faster.¹⁹

IBM CodeNet is a large dataset aimed at teaching AI to code – a project where computers can program computers. It consists of about 14M code samples and around 500M lines of code in more than 55 different programming languages (modern ones like C++, Java, and Python, to legacy languages like COBOL, Pascal, and FORTRAN).²⁰ Project Wisdom, built on CodeNet, enables generation of code for developers on Red Hat Ansible (an IT infrastructure automation platform) and generates automations for the hybrid cloud through a natural-language interface. While Wisdom had only 350 million parameters in its foundation models (as compared to 12 billion for CoPilot), it was 2.7 times more accurate than the other solution.²¹

Given the significant base of IT services in India, the question of how such LLM-based apps would affect coding becomes important. We believe that such apps will act as productivity enhancers for the developers, but not replace them entirely. See the table below for coding scenarios where such LLM-based apps can assist.²² Companies will start adopting offerings like GitHub Copilot for Business or CodeNet. At the same time, care and caution should be exercised before placing confidential customer code in publicly available LLM-based apps for coding suggestions. We are also likely to see the emergence, from Indian IT companies, of end-to-end, developer-facing, generative AI apps that are built on proprietary models.

- **Explaining code** – Prompt: “What is this function doing?” – during requirements gathering phase, use LLM-based app to understand the existing code
- **Improve existing code** – Prompt: “I want to raise an exception if the authentication is incorrect” – the developer prompts the LLM-based app with descriptions of the requirements, and the app provides instructions on how to improve code or in some cases even provide some modified code
- **Rewrite code** – Prompt: “Rewrite this function following Python's naming conventions” – when refactoring code written by a developer not native in a particular language, the LLM-based app can rewrite code using the correct styles, and also explain the reasons for change

- **Simplifying code** – Prompt: “Simplify the following code” – The LLM-based app can provide a more compact version of the original code
- **Writing test cases** – Prompt: “Write a few test cases to ensure the new X function is correct” – The LLM-based app can help write test cases in order to test certain functions
- **Translating code** – Prompt: “Translate this function from JavaScript to Python” – The LLM-based app can help port the code from one language to another
- **Writing documentation** – Prompt: “Write documentation for this function” – The LLM-based app can provide documentation for the program, and would include usage examples too
- **Tracking bugs** – Prompt: “Find the bug in the following function that finds whether a supplied value X exists in an array” – The LLM-based app can quickly find a bug in the code
- **Scaffolding** – Prompt: “Write a function that reads a text file from disk and uploads it to an S3 given as a parameter” – the LLM-based app kicks-off the structure of any new code to be written

References

- 1 <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- 2 <https://www.msn.com/en-in/money/other/chatgpt-reaches-100-m-active-users-in-record-time-breaks-facebook-tiktok-s-record/ar-AA171wBP>
- 3 <https://iq.opengenus.org/gpt-3-5-model/>
- 4 <https://olivermolander.medium.com/chatgpt-llms-and-foundation-models-a-closer-look-into-the-hype-and-implications-for-startups-b2f1d82f4d46>
- 5 <https://crfm.stanford.edu/commentary/2021/10/18/steinhardt.html>
- 6 <https://twitter.com/AlphaSignalAI/status/1600209950860451841>
- 7 <https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform/>
- 8 <https://olivermolander.medium.com/chatgpt-llms-and-foundation-models-a-closer-look-into-the-hype-and-implications-for-startups-b2f1d82f4d46>
- 9 <https://openai.com/blog/chatgpt/>
- 10 <https://github.com/features/copilot>
- 11 https://www.bing.com/new?icid=mscom_marcom_H1a_BingAI
- 12 <https://blog.google/technology/ai/bard-google-ai-search-updates/>
- 13 https://itihaasa.com/public/pdf/Becoming_a_Co_Creative_Living_Enterprise_in_the_X-verse_dec_2022_final.pdf
- 14 <https://arxiv.org/abs/2108.07258> - "On the Opportunities and Risks of Foundation Models"
- 15 <https://gizmodo.com/cnet-ai-chatgpt-news-robot-1849996151>
- 16 <https://garymarcus.substack.com/p/inside-the-heart-of-chatgpts-darkness>
- 17 <https://www.mckinsey.com/capabilities/quantumblack/our-insights/generative-ai-is-here-how-tools-like-chatgpt-could-change-your-business>
- 18 <https://indianexpress.com/article/explained/explained-sci-tech/chatgpt-on-whatsapp-bhashini-welfare-schemes-8442622/>
- 19 <https://github.blog/2023-02-14-github-copilot-for-business-is-now-available/>
- 20 <https://research.ibm.com/blog/codenet-ai-for-code>
- 21 <https://research.ibm.com/blog/ai-for-code-project-wisdom-red-hat>
- 22 <https://twitter.com/svpino/status/1624392828154712064>