A PERSPECTIVE ON CHATGPT, ITS IMPACT AND LIMITATIONS

Infosys®
Navigate your next

Abstract

Technology with creative abilities is a game changer. Recent disruptions in Generative AI have redefined content creation and here is a look at one such system - ChatGPT. ChatGPT was a chatbot released by AI Startup OpenAI in November 2022 which immediately received widespread popularity even among non-technical users due to its human-like responses on almost any knowledge domain.

ChatGPT has 1 million users 5 days after launch

Chart showing months (y-axis, 0 to 45) to reach 1 million users by platform:
- ChatGPT: 5 days
- Instagram: 2.5 months
- Spotify: 5 months
- Dropbox: 7 months
- Facebook: 10 months
- Twitter: 2 years
- AirBnB: 2.5 years
- NetFlix: 3.5 years
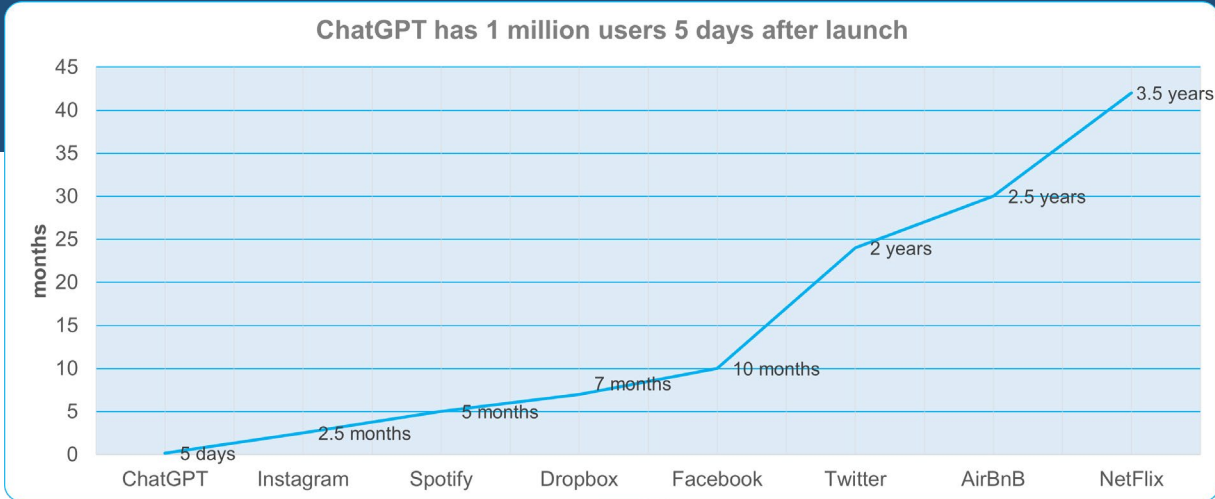
Overview

Transformer, a Deep Learning Architecture which was introduced not too long back in 2017, has been a tipping point for Natural Language Processing (NLP), the branch of AI dealing with communication between machines and humans. Google's BERT and OpenAI's GPT, both released in 2018 were the earliest Deep Learning Models based on Transformer Architecture, and the most recent being ChatGPT.

Chat Generative pretrained Transformer or ChatGPT is built on top of GPT 3 family of large language models. It is based on GPT 3.5 or Instruct GPT and optimized for dialogue. GPT-3 released in 2020 is a third generation GPT model from OpenAI the first 2 being GPT-1 released in 2018 and GPT-2 released in 2019.

Per definition, Chat GPT is a generative language model, but in practice it is acknowledged as an AI chatbot trained to interact in a conversational way in a dialogue format. Let us deconstruct and understand:
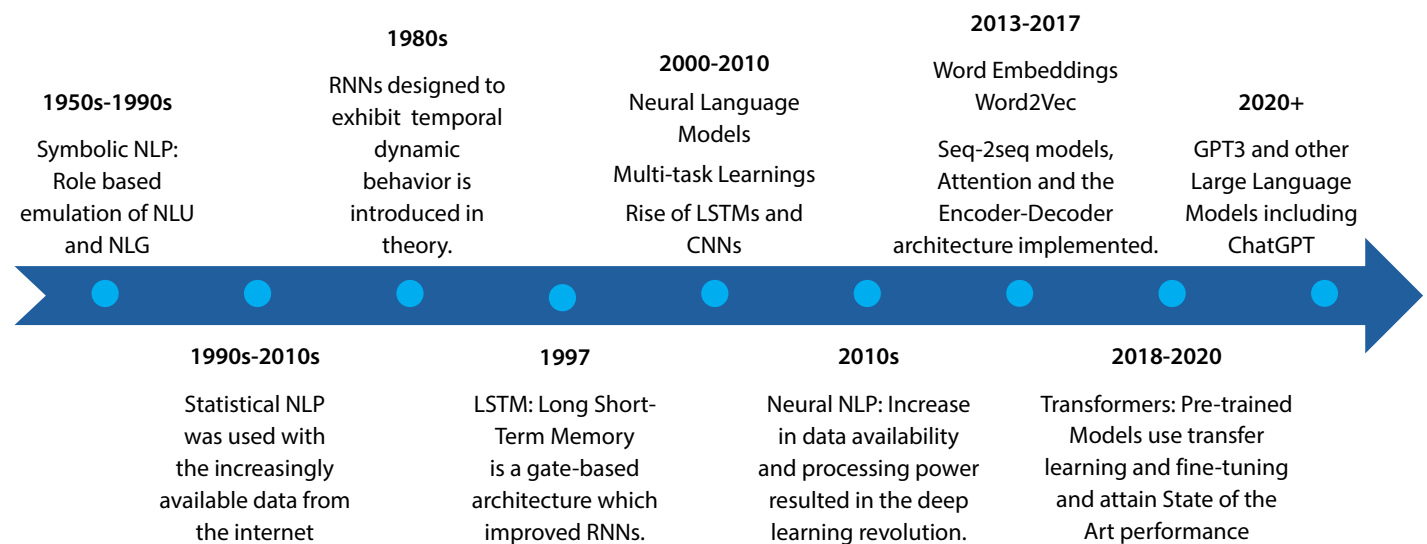
Generative: *Generates new data points by understanding the relationship between variables in existing datasets.*

Pre-Trained: P*re-trained on large dataset*

Transformer: B*ased on Transformer architecture to process sequences of elements, such as words in a sentence to generate output.*
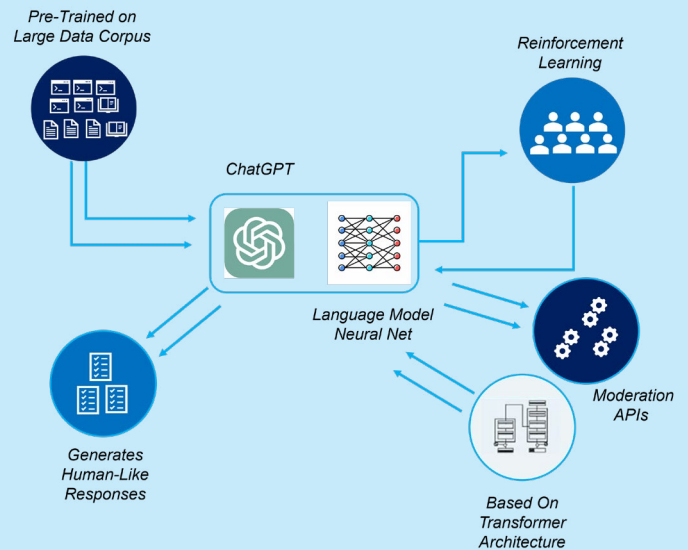
Here is a look at how NLP has evolved over the years and culminating in the breakthroughs in recent years

1950s-1990s

Symbolic NLP: Role based emulation of NLU and NLG

1980s

RNNs designed to exhibit temporal dynamic behavior is introduced in theory.

2000-2010

Neural Language Models

Multi-task Learnings

Rise of LSTMs and CNNs

2013-2017

Word Embeddings Word2Vec

Seq-2seq models, Attention and the Encoder-Decoder architecture implemented.

2020+

GPT3 and other Large Language Models including ChatGPT

1990s-2010s

Statistical NLP was used with the increasingly available data from the internet

1997

LSTM: Long Short-Term Memory is a gate-based architecture which improved RNNs.

2010s

Neural NLP: Increase in data availability and processing power resulted in the deep learning revolution.

2018-2020

Transformers: Pre-trained Models use transfer learning and fine-tuning and attain State of the Art performance

Generative AI

According to Research Firm Gartner's prediction, the market for AI software is expected to be $134.8 billion by 2025. The market growth was 14.4% in 2021 and it is expected to grow by 31.1% in 2025. In Gartner's Emerging Technologies and Trends Impact Radar for 2022 report, Generative AI is found to be one of the most impactful and rapidly evolving technologies that brings productivity revolution. Some of the Gartner predictions are as follows:

• By 2025, Generative AI will produce 10% of data and 20% of test data for consumer-facing use cases.

• By 2025, 50% of drug discovery and development will use Generative AI.

• By 2027, 30% of manufacturers will enhance product development with Generative AI.

Generative AI refers to unsupervised or semi-supervised algorithms which can generate new content based on data on which it is trained. This means it can create content which it has never seen before. Content can be audio, images, code, text, video, or synthetic data.

Generative AI systems are a part of Artificial Intelligence and a sub-field of Machine Learning and Deep Learning. Generative AI models can be Generative Adversarial Networks (GANs) - which can create visual and multimedia artifacts from image/text input data or Transformer-based models - technologies like Generative Pre-Trained Transformer (GPT) language models.

Chat GPT: A Language Model Optimized for Dialogue

ChatGPT- is OpenAI's most notable AI chatbot implementation and has been an important milestone for Generative AI. According to its creators OpenAI, ChatGPT is a language model optimized for human dialogue and hence it can:

1. ask follow-up questions - ChatGPT remembers previous responses and prompts

2. admit mistakes it made - It allows users to provide follow-up corrections for their responses.

3. challenge incorrect premises and designed to reduce the possibility of harmful or deceitful responses - Queries are filtered through moderation API to avoid offensive outputs from being present to and generated from ChatGPT
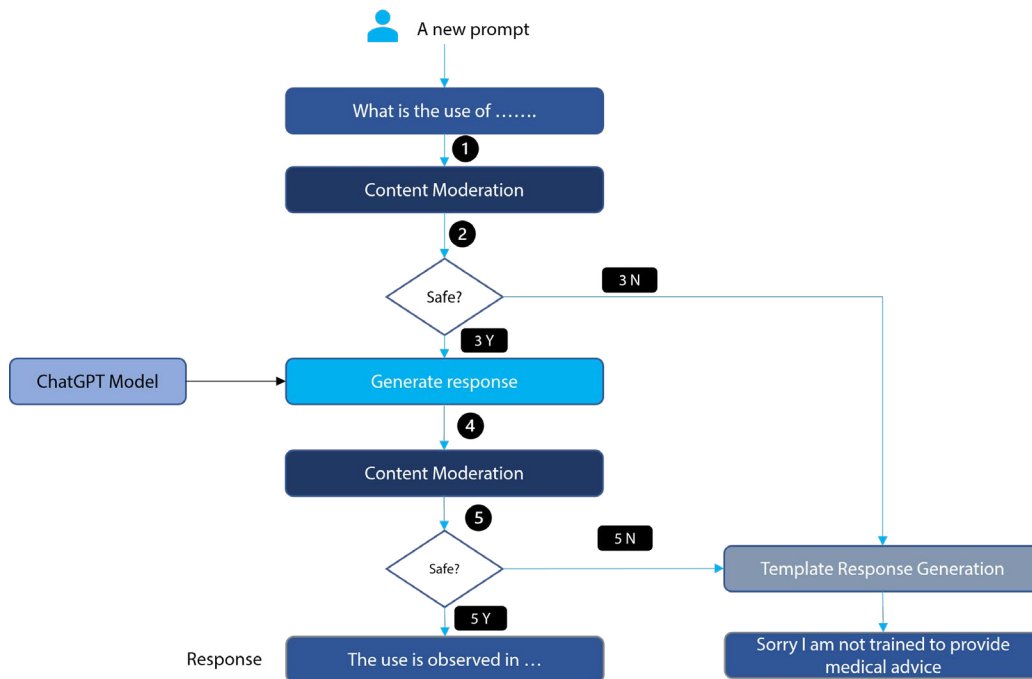
ChatGPT - Answering Prompts

Inputs are provided to ChatGPT in the form of a text. For the given input question or context sentences, ChatGPT generates an appropriate and coherent response. ChatGPT is considered a sibling model to InstructGPT. Both are based on GPT-3 but there is slight variation in data setup. ChatGPT is trained on conversational data and is a slightly smaller model compared to GPT-3. Feedback and learnings from the earlier deployments like GPT-3 and Codex are incorporated in the safety mitigations. All the queries are filtered through a moderation API. Harmful and incorrect outputs are reduced by usingreinforcement learning from human feedback (RLHF) training method. Here is an understanding of transformers on which GPT language models are based.

Transformers and Large Language Models

In 2017 a team at Google Brain introduced Transformers in "Attention is all you seek" Research paper. Until then, Convolution or Recurrence Neural Networks (LSTM-based models) were used for sequential data processing, and transformers outperformed the existing models. Transformers were termed as "foundation models" in a 2021 paper by Stanford researchers. They were seen driving a paradigm shift in AI. Foundational models are trained on vast unlabeled data at scale through self-supervised learning (where Labels are generated from unlabeled data, according to the structure or characteristics of the data) and adapted to a wide range of downstream tasks. A Transformer is a deep learning modelanda neural networkthat tracks relationship in sequential data (e.g., words in a sentence) and learns context and meaning by using sequence-to-sequence learning where a sequence of tokens is given as input and the next word in the output sequence is predicted.

How do matrices of numbers relate to language?

Transformers can detect relationships and dependencies between data elements in a series even if they are distant. It uses a concept called Self Attention which allows it to weigh or score different parts of the input when making predictions. Language models convert sentences into understandable bits of data that a program can work with called tokens. Tokenizers convert input text to tokens and the tokens are transformed into vector representation called Embeddings.

Positional encoding describes the position of an entity in a sequence. Masking refers to removing or replacing words in a sentence by some other dummy token such that the model does not have access to those words at the time of training. They are
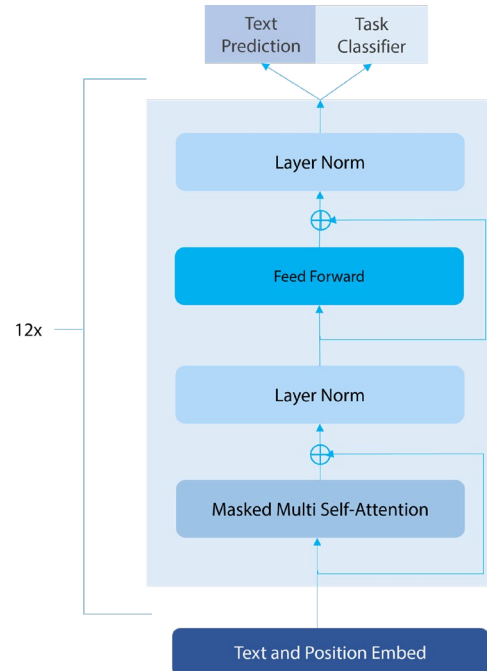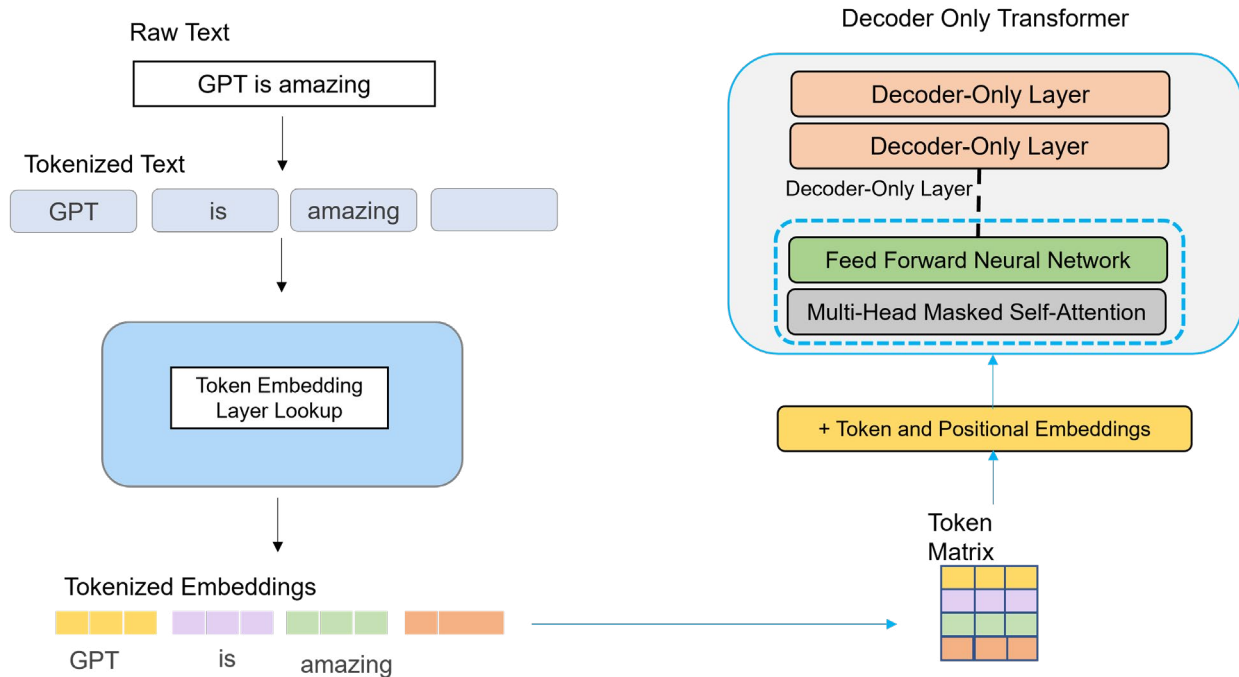
Transformer Based Architecture of GPT-1

used by Attention units which determine the relationship between elements by calculating a kind of algebraic map. Attention queries are typically executed in multi-headed attention in a parallel manner.

After the self-attention mechanism, the output is fed into a feed-forward neural network, which performs a non-linear transformation to generate a new representation. The output from each layer is normalized, and a residual connection is added to allow the input to be passed directly to the output, allowing the model to learn which parts of the input are most important.

Multiple repetitions of these components form a deep neural network. The model can see the same patterns humans see.

Raw Text

GPT is amazing

Decoder Only Transformer

Decoder-Only Layer

Decoder-Only Layer

Decoder-Only Layer

Feed Forward Neural Network

Multi-Head Masked Self-Attention

Tokenized Text

GPT | is | amazing |

Token Embedding Layer Lookup

+ Token and Positional Embeddings

Token Matrix

Tokenized Embeddings

GPT | is | amazing

Transformer Based Architecture

Why are Transformers important?

CNNs (Convolutional Neural Networks) were successful for image dataset but not for language data. RNNs (Recurrent Neural Network) and Transformers are designed to process sequential data like natural language, but transformers are differentiated by the capability to process longer sequences at once in a parallel manner rather than one word at a time. Word order is important for language. Context for any position in input sequence is provided by Attention mechanism. **Due to the absence of labelled data and training parallelization, transformers can be trained on larger datasets**.

The transformer architecture is the building block behind Large Language Models or LLMs. Language models use probability to understand the way sentences are formed in a language. For instance, a simple language model can predict the next word most likely to follow based on statistical analysis of text sequences. For e.g.: "I love reading a" is most likely to end with "book" than "refrigerator." Language models are word sequence prediction models trained to predict the next word based on previous input sentences. LLMs are trained on massive datasets and the knowledge gained is used by the deep learning algorithm, to recognize, summarize, translate, predict, and generate text and other content.
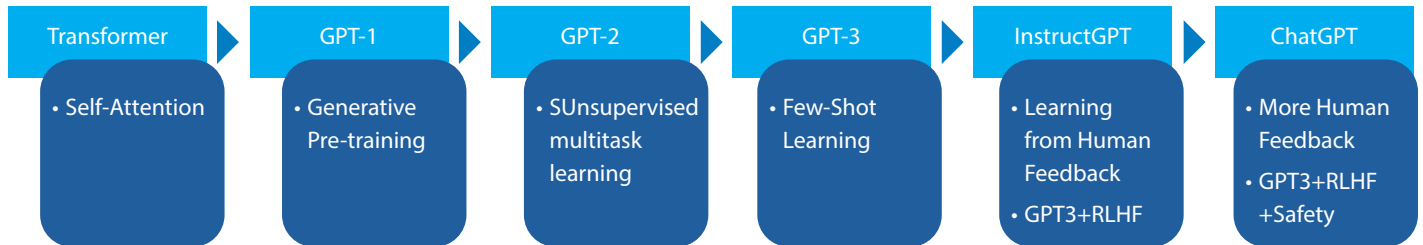
Why do we need a GPT-like Language Model?

Before GPT, the NLP models were trained using supervised learning techniques for particular tasks. Supervised Learning refers to usage of labeled data to classify or predict outcomes. Following are the two major limitations of Supervised models:

1. **Availability of Annotated data is a challenge:** A large amount of labelled data is required by Supervised models for learning. Manually labeling is tedious and time-consuming whereas unlabeled data is constantly created

2. **Lack of Generalization:** The models are not able to generalize (model's ability to adapt properly to new, previously unseen data) other than what they are trained

Following was proposed to overcome the short comings - Pre-trainingon a diverse corpus of unlabeled textfirst and discriminative fine-tuning onspecific downstream task afterwards. Open AI developed GPT-1 model which achieved NLP capabilities with a single task-agnostic model based on Transformer Architecture via generative pre-training based on Unsupervised learning and supervised discriminative fine-tuning. Hence the name Generative Pre-trained Transformer.

Overview of OpenAI's GPT Models

Transformer	GPT-1	GPT-2	GPT-3	InstructGPT	ChatGPT
• Self-Attention	• Generative Pre-training	• SUnsupervised multitask learning	• Few-Shot Learning	• Learning from Human Feedback • GPT3+RLHF	• More Human Feedback • GPT3+RLHF +Safety

	GPT-1	GPT-2	GPT-3
Learning Objective and Concepts	Semi-supervised learning (unsupervised pre-training followed by supervised fine-tuning) in 3 stages: • Unsupervised Language Modelling • Supervised Fine-Tuning • Task Specific Input Transformations	Task conditioning: when the task is different, the model should produce a different output for the same input. Zero Shot Learning and Zero Short Task Transfer: Zero shot learning is a special case of zero shot task transfer and the model performs the task based on the given instruction with no examples at all.	In-context learning: Due to training on large datasets, the language models can recognize patterns in data. This ability increases with number of model parameters and helps the model during zero-shot task transfer. When a few examples and/or a description of a task is given, the model matches the pattern of the examples with what was learnt in past for similar data and performs the tasks.
Dataset Part of Training Corpus	BooksCorpus	The Reddit platform was scraped, and data pulled from outbound links of high upvoted articles.	Common Crawl corpus which has data collected over 8 years of web crawling in the range of petabytes. It contains web page raw data, metadata, and text extracts with some filtering. WebText2 web page texts from outbound Reddit links of posts with 3+ upvotes. Books1 & Books2 books corpora Wikipedia English language pages
Architecture	12-layer decoder only transformer with masked self-attention	GPT-2 had 48 layers	GPT-3 has 96 layers with each layer having 96 attention heads.
Inference	GPT-1 proved that with transfer learning it can perform NLP tasks with little fine-tuning.	GPT-2 showed that increasing the training dataset size and parameters improved the ability to understand tasks and perform with more accuracy.	GPT-3 was built with a vast dataset and there was no need to fine-tune and required only a few demonstrations to understand tasks and perform them.
Released	2018	2019	2020
Parameters	117 Million	1.5 Billion	175 Billion

What Makes OpenAI GPT-3 Different?

The GPT-3 Has Been Upgraded from GPT-2 With 175 billion Parameters and trained on 45 TB text data from multiple sources. GPT-3 is effective due to its large size, diverse training data, and ability to perform a wide range of language tasks without task-specific fine-tuning.

OpenAI GPT-3 can perform tasks with few or no shots (shots refer to examples/demonstration). NLP tasks can be done without any fine-tuning, gradient or parameter updates making this model Task-Agnostic.

In-context is when a model learns to perform a task after seeing only a few examples despite not being trained for that task.

Zero Shot, One Shot & Few Shot task interaction with GPT-3, contrasted with traditional fine-tuning
Image Source: arxiv.org

How was ChatGPT trained?

According to Open AI, ChatGPT and InstructGPT were trained using Reinforcement Learning from Human Feedback (RLHF). The same methods were used but with slight differences in the data collection setup. Azure AI supercomputing infrastructure and NVIDIA GPUs were used for training ChatGPT and GPT 3.5. ChatGPT is fine-tuned from one of the models in 3.5 series using both supervised and reinforcement learning methods and assistance of human trainers.
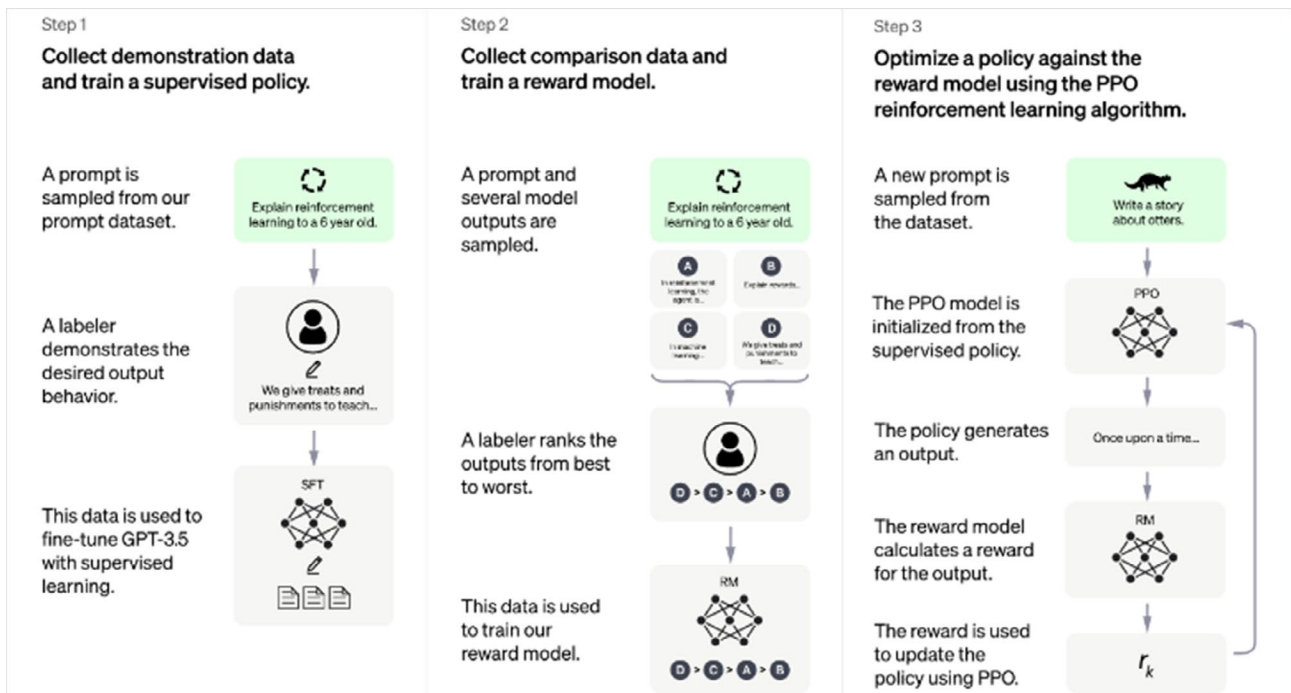
Training Process of ChatGPT
Image Source: Open AI

Getting past the hype: Some of the Practical Use Cases of Chat GPT

Enhance Chatbot, answering questions: Can be used to converse in a natural and engaging manner and answer questions. E.g.: Used to enhance Customer support chatbots, self-serve HR functions, Sales Support Chatbot, Learning tools etc.

Content creation: It can create new content based on patterns and relationships it has learned from the data trained on. E.g.: news articles or blog posts, task list, interview questions, sales & marketing content creation, Creative writing, draft emails, Optimize Communication etc.

Text summarization: for effective analysis and insights. Summarize presentation decks, documents, analyze customer feedback, create product manuals, Job Descriptions etc.

Text Classification: Input text and return a label. E.g.: Classify email, sentiment analysis- determine the sentiment of a tweet, topic labelling, intent detection – prioritize emails.

**Text Completion:** Help complete sentences or paragraphs.

Machine translation: Language detection & translation.

Developer Productivity: ChatGPT can be used for generating code suggestions/ fixes, improving code, creating test cases, SQL queries, boilerplate code, documentation etc. Can be used as productivity improvement tool integrated with IDE.

Power the Search Engine: ChatGPT has potential use as a semantic search interface. It can understand the user's intent and provide a conversational experience
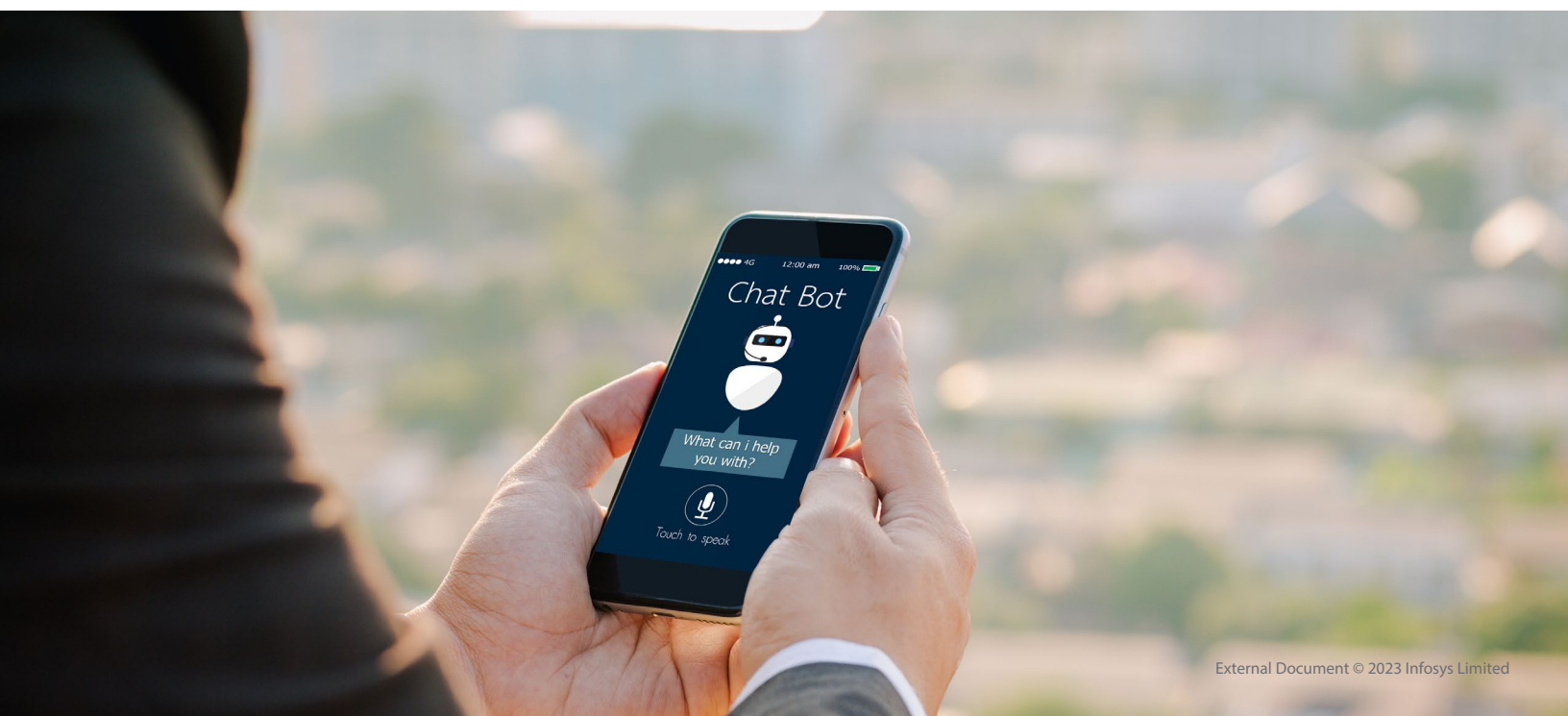
Debunking Myths

Is Chat GPT End of google search?

ChatGPT gives human like responses, and it feels like talking to an expert. ChatGPT is a complex Deep Learning Algorithm which predicts output based on probability. Whereas Google, which is a Search Engine, requires filtering through pages. Search algorithms depend on factors like the words of the query, relevance and usability of pages, expertise of sources, and users' location and settings. Google lacks any language processing capabilities. ChatGPT can understand the user's intent and provide a conversational experience. Quality of ChatGPT response depends on quality of the prompt. Google can provide a list of relevant web links in response to a query. It can be personalized. Users can decide which search result they want instead of Bot deciding for them.

Source citations are missing in ChatGPT. ChatGPT has an accuracy issue, and the output has to be fact-checked and reviewed before usage. Google offers real-time information whereas ChatGPT does not connect to the internet and responds based on trained data (as of now training done till 2021). Google usage is free, and it continues to hold its own importance. Also, with the recent ChatGPT integration with Bing we have seen how it can work as an add-on to the search engine.

Is ChatGPT sentient?

ChatGPT is not considered sentient/aware. It responds to a prompt (question or query) in a natural and human-like manner.

Is Chat GPT free?

During the research review phase, it is free - at chat.openai.com. A paid and advanced version of ChatGPT for professionals is released subsequently.

Is ChatGPT replacing developers?

ChatGPT is capable of providing code suggestions/fixes, write generic functions or boilerplate code but it has some limitations. The output needs human review. Skilled developers are needed to understand business needs and adapt technologies and platforms considering budget limitations, follow standards etc. It will not replace developers but rather work as productivity improvement tools for developers. Mostly usage is going to be as IDE built ins. Also new career options like Prompt Engineering are opened by ChatGPT. Prompt Engineering refers to the principles and techniques for writing prompts or inputs for the models that will get the best generations for the task.

Is Chat GPT replacing the existing NLP Chatbots?

ChatGPT lacks integrations for transactional tasks like processing orders, reporting issues etc. It is not reliable for factual data. Handling queries and FAQs on organization data is a challenge. OnPrem deployments are not an option due to the resource requirements and complexity. Data privacy and security need consideration. Hence ChatGPT cannot replace the existing NLP Chatbots but can improve their effectiveness.

Microsoft Investment in OpenAI

In 2019, Microsoft invested $1 billion in OpenAI , and became the exclusive provider of cloud computing services to OpenAI. Microsoft invested again in OpenAI in 2021 making the company exclusive licensee of GPT-3. This means Microsoft has sole access to GPT-3's underlying model. In 2022, Microsoft announced a new multiyear, multibillion-dollar investment, the third phase of the partnership with OpenAI. OpenAI has predicted a revenue of $1 billion by the end of 2024. $200 million in revenue is expected by the OpenAI by the end of 2023. OpenAI has deployed models like GPT, DALL-E and Codex through cloud API and the Azure OpenAI Service for enterprise and developers use

Why is ChatGPT gaining traction?

OpenAI claims over 300 applications are using GPT-3 APIs to power search, conversation, text completion, and other advanced AI features. ChatGPT's content generation capability is disruptive. But if Enterprises must use it then Mature tools, APIs, etc., are required for Fine-tuning the pre-trained models and applying specific knowledge for specific data and domains. This press release from Gartner suggests starting using ChatGPT for non-customer facing scenarios with NLP tasks such as classification, summarization, and text generation. Since customization and training is involved for enterprise use, it recommends choosing a task-specific, pretrained model to cut down costs. Humans should review the output preferably.

In Feb 2023, Microsoft announced integration of an enhanced version of ChatGPT (called Prometheus) into Bing, providing a ChatGPT-like experience within the search engine. Integration with Teams is also announced. Whether the model will be integrated into other Microsoft applications such as Outlook, Word, PowerPoint, and Microsoft Power Apps is not clear yet.

From the perspective of developers, ChatGPT can work as Productivity Enhancement tool or Developers AI Assistant integrated with IDE or a conversational search interface etc. For Education & Training perspective it can be used as a learning tool. For Communication & Marketing ChatGPT offers content generation/summarization capabilities. For Customer Support, it can provide conversational chatbot experience. The model is in the early stage of scaling. How the usage will take the shape of Productivity improvement tools or Digital assistants, or how the usages will shape SDLC, Training & Education, Recruitments is a wait and watch.

Are there any other Large Language Models?

Attempts have been made to democratize Large Language Models, and other than OpenAI's GPT-3 there are several other alternatives as follows:

Company	Model	Training Parameters/Tokens	License	Brief Info
EleutherAI	GPT-J	GPT-J - 6 billion parameters	Apache 2.0	GPT-3 like model. It can translate between languages, write blog posts, complete code etc.
EleutherAI	GPT-NeoX	20 billion parameters	Apache 2.0	Based on Megatron architecture
Google	BERT (Bidirectional Encoder Representations from Transformers)	Bert Base -110 million parameters Bert Large - 340 million parameters	Apache 2.0	One of the first transformer language models. It has 2 versions - Bert Base with 12 layers of transformers and Bert Large with 24 layers. BERT powers Google Search to understand search query intent
Google	PaLM (Pathways Language Model)	540 billion parameters/768 billion tokens	Proprietary	Trained on Pathway system to learn problem solving. Reached practical scaling limits.
Google	LaMDA (Language Models for Dialog Applications)	137 billion parameters /1.5 trillion words	Proprietary	Specialized for conversations. Bard is a chat service powered by LaMDA. This is Google's answer to ChatGPT.
Google	GLaM (Generalist Language Model)	1.2 trillion parameters	Proprietary	This is not released yet. It is a mixture of models (MoE) and consists of sub models specializing in different domains. A gating network picks the relevant expert based on the input. Expensive to train but optimal to run.
Microsoft and Nvidia	Megatron-Turing Natural Language Generation (MT-NLG)	530 billion parameters	Restricted web access	NVIDIA DGX SuperPOD-based Selene supercomputer was used for training. Used for reading, comprehension, and natural language inference.
AI21 Labs	Jurassic-1	Has 2 parts: J1-Jumbo – 178B parameters J1-Large – 7B parameters	Proprietary	Released Models via APIs
BigScience Workshop by HuggingFace	BLOOM	176 billion parameters	Responsible AI	Requires 384 graphics cards, each one with more than 80 gigabytes. It is trained in 46 languages and 13 programming languages and has versions with lesser parameters. Based on Megatron-LM, an 8.3B-parameter predecessor of Megatron-Turing NLG.
DeepMind	Gopher	280 billion parameters	Proprietary	Is specialized in answering science and humanities questions. DeepMind is AI lab acquired by Google in 2014
DeepMind	Chinchilla	70 billion parameters 1.4 trillion tokens (text data)	Proprietary	A compute-optimal model. Not released to public. Trained on lesser parameters and more data
Meta	Open Pretrained Transformer (OPT)	175 billion parameters.	Non-Commercial Research	16 NVIDIA V100 GPUs were used for training and deployment (lower than others). It is available under a noncommercial license for research use. A pre-trained model and code for training is provided. It is trained on public datasets. Architecture is GPT-3 with adaptations from Megatron
Amazon	Alexa Teacher Models (AlexaTM 20B)	20 billion parameters	Public Web-API	Sequence-to-Sequence bi-directional model. Hence a better few shot learner and better at Translation and 1-shot summarization tasks.

How can enterprises use LLMs like ChatGPT?

Challenges: Scaling and maintaining large language models involve high costs. Developing Foundational models needs significant costs for training involving supercomputing infra. Access to large datasets is also a challenge. The technical expertise and significant data, compute, and engineering resources used for training and deploying LLMs make it out of reach for many enterprises.

As per AI21 Labs 2020 research, the cost for developing a text-generating model with 1.5 billion parameters is pegged at $1.6 million. As per BDTechTalks, the estimated cost of running GPT-3 on a single AWS instance is $87,000 / year. A GPT3 training run is estimated at $4.6 million and multiple iterations are required to fine tune. ChatGPT was trained on Nvidia GPUs using Azure Super

computing infrastructure. It is reported that the model costs US $3 million per month or $100,00 a day to run which is steep.

Solutions: To make it accessible to enterprises, companies are offering services which offer pre-trained LLMs for fine-tuning and specific task training to suit specific use cases. Fine-tuning needs lesser data, power, and time to train specific tasks.

In-Context Learning (Few-shot learning) is used to train the new task-specific layer with smaller training dataset on domain-specific words and phrases. For enterprise usage such as Customer Service Chatbots or Sales & Marketing use cases customization is required for enterprise context, formality of tone, brand tone and other domain specific data. Matured tools, APIs etc. are required for this.

ChatGPT or the underlying GPT3 model usage areas are defined by Gartner as in below image:
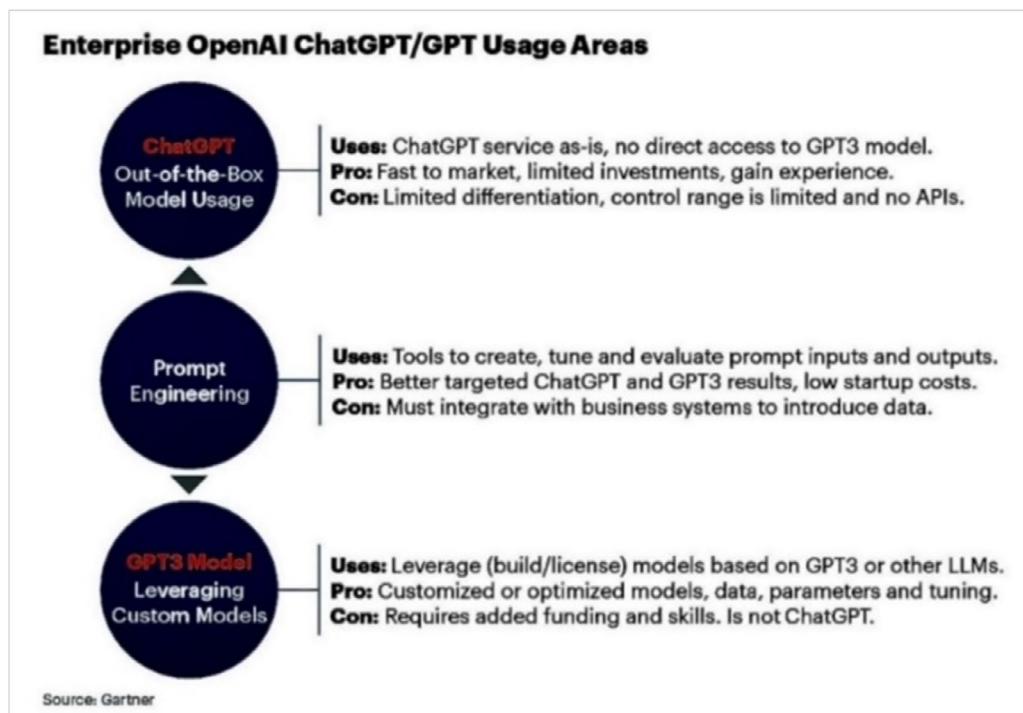

Enterprise OpenAI ChatGPT/GPT Usage Areas

ChatGPT Out-of-the-Box Model Usage
Uses: ChatGPT service as-is, no direct access to GPT3 model.
Pro: Fast to market, limited investments, gain experience.
Con: Limited differentiation, control range is limited and no APIs.

Prompt Engineering
Uses: Tools to create, tune and evaluate prompt inputs and outputs.
Pro: Better targeted ChatGPT and GPT3 results, low startup costs.
Con: Must integrate with business systems to introduce data.

GPT3 Model Leveraging Custom Models
Uses: Leverage (build/license) models based on GPT3 or other LLMs.
Pro: Customized or optimized models, data, parameters and tuning.
Con: Requires added funding and skills. Is not ChatGPT.

Source: Gartner

Image Source: Gartner

Word of Caution

1. Inaccurate & Inexplainable Results: Responses generated by ChatGPT are not always correct. There is no built-in mechanism to detect an inaccurate response. It is trained on large data corpus, and it is difficult to trace the fact base to which responses are attributed to. OpenAI has not shared full insight into ChatGPT Algorithm. Hence, ChatGPT can only be used where errors are acceptable and can be corrected due to its Black-box nature. Content generated should be reviewed by humans.

2. Filters should be effective to catch inappropriate content input/ output to the system. Unwanted biases may be part of the vast training data. Potential misuse & Ethical Concerns exists as it is easy to create fake content. It can be damaging for the company's reputation.

3. Expensive compute resources, technical skills, and access to dataset are faced for Training. Running the model is similarly expensive. It is not environmentally friendly.

4. The training data corpus may consist of intellectual property of others. Response may include ideas which might be plagiarized information. Fact-checking may be required.

5. Large technology giants have made large investments to build LLMs using significant AI talent. Power is concentrated in a few deep-pocketed entities.

6. Customization is needed for Enterprise use. Enterprises should choose vendors offering strong data usage and ownership policies and should be mindful to avoid confidential and PII (Personal Identifiable Information) data feeding into the system.

Limitations of ChatGPT

1. Chat GPT requires significant resources to run due to its complexity and computational demands. It cannot run on resource constrained environments.

2. ChatGPT was trained till 2021 data and will provide responses from this data. It is not connected to the internet.

3. RLHF training method is used, and the model provides answers based on training data and not on Human Instructor's knowledge. There is no Source of Truth and ChatGPT can respond with unexpected or incorrect information or answers which do not make sense. It is just a complex algorithm predicting outputs based on probability and not a credible source for factual data.

4. It is trained to be cautious and may decline to provide certain answers. Over-optimization in the training phase where comprehensive longer answers are preferred. E.g.: Overuse of certain phrases, like mentioning it is a language model trained by OpenAI.

5. OpenAI has used Moderation API to warn/block unsafe content to ensure the model will refuse inappropriate requests, still it responds to harmful instructions or exhibits biased behavior.

6. Inconsistent: If the same prompt is attempted multiple times or phrasing of inputs is changed, ChatGPT can give a different answer. Quality of response depends on the quality of theprompt.

7. If ChatGPT cannot understand an ambiguous query, it usually guesses the user's intent rather than asking for clarification.

Scope of Improvement: Techniques for Model Simplification (reduce the model size or memory requirement for training) or Reduced training time will help reduce costs. Improvements in Training methods and data setup to address the limitations like inaccurate and biased/harmful responses are anticipated in the upcoming GPT version. Accurate responses and overcoming ambiguous user inputs with better reasoning is also expected from the newer versions. GPT-4 is yet unreleased. What new it has to offer over the limitations of the previous model is wait and watch.
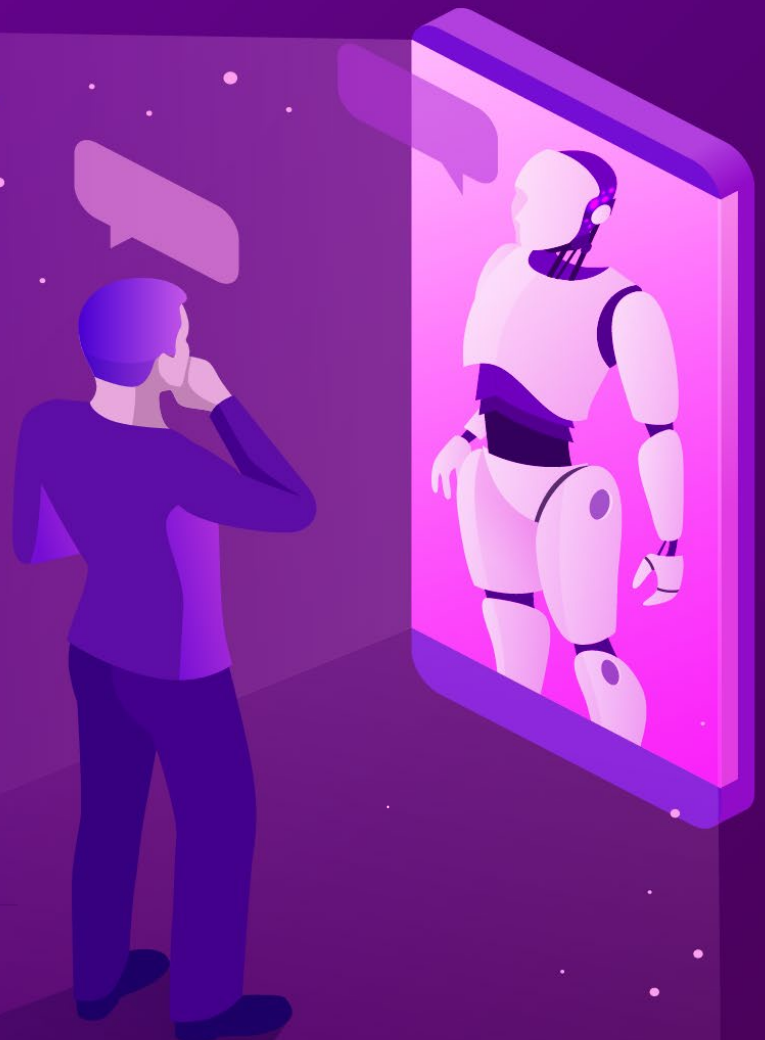
Conclusion

The core functionality of a chatbot is to converse with humans in a natural way. But ChatGPT has versatile use cases from writing computer programs to composing blog posts and articles. Its generative capabilities are disruptive and capable of bringing productivity revolution. The reach of this AI is very wide, and humans can guide and correct it. The failures are acceptable unlike in the case of AI in automated cars.

The model is in the early stages of scaling and Enterprise use cases are impressive, but it is important to consider the limitations and challenges associated with it and then proceed. ChatGPT is a significant step forward in AI towards the development of more advanced systems. It is a powerful tool to be used wisely.

Author

Varsha Bhandarkar
Senior Technology Architect
HIL-ADG, Infosys

✉ in

Reviewer

Jitendra Jain
Principal Technology Architect
HIL-ADG, Infosys

✉ in

References

1. ChatGPT:Optimizing Language Models for Dialogue(openai.com)

2. How enterprises can use ChatGPT and GPT-3 | Computerworld

3. Top 10 Open-Source GTP3 Alternatives You Should Try in 2023(analyticsinsight.net)

4. ChatGPT Guide in 2023: Definition, Top Use Cases & Limitations (aimultiple.com)

5. GPT models explained. Open AI's GPT-1,GPT-2,GPT-3 | Walmart Global Tech Blog (medium.com)

6. How generative AI & ChatGPT will change business | McKinsey

7. Why is ChatGPT Making Waves in the AI Market? (gartner.com)

8. What Are Large Language Models Used For and Why Are They Important? | NVIDIA Blog

Infosys®
Navigate your next

For more information, contact askus@infosys.com

Infosys.com | NYSE: INFY   Stay Connected