AMRUTVAHINI COLLEGE OF ENGINEERING, SANGAMNER

INTERNATIONAL CONFERENCE ON RECENT TRENDS AND ADVANCEMENTS IN COMPUTING TECHNOLOGIES,ICRTACT 2024

The Use of GPT-3 and Similar Models in Conversational AI: A Comparative Analysis

Saurabh Sharma, Satyendra Pratap Singh , Ragini Kumari

SCSE Galgotias University Greater Noida, India
saurabhsha2003@gmail.com, singhsatya.1999@gmail.com, ragini@galgotiasuniversity.e du.in

Abstract—*The fast growing conventional AI has given rapid rise to advance language models capable of generating human-like texts. These models have the capacity to create human-like content based on a huge collection of information, making them a vital instrument for different applications such as dialect interpretation, content summarization, and question-answering frameworks. As of late, two models, ChatGPT and GPT-3, have taken the AI community by storm with their progressed dialect capacities and real-world applications. Concurring to OpenAI, GPT-3 has been prepared on a gigantic corpus of 45 terabytes of content information, making it the biggest dialect show to date. On the other hand, ChatGPT may be a littler demonstrate prepared on a lesser corpus of information, however it still brags of impressive dialect capacities. This paper points to supply a comprehensive consider of these two dialect models, highlighting their specialized contrasts, advancements, limitations, and future suggestions. Based on broad investigate and investigation, we are going dig into the key zones of center for these models, counting pre-training, consideration components, fine-tuning procedures, show estimate and computing control, human-like responsiveness, dialect modeling, and the benefits of zero-shot exchange learning.*

Keywords – ChatGPT,NLP, BARD, Artifical Inteligence ,Deep Learning, BERT, T5, XLNet,

I. Introduction

In recent years, the field of Conversational AI has experienced a profound transformation, primarily driven by the advent of advanced language models like GPT-3 and its peers. These models, rooted in the realms of deep learning and natural language processing, have ushered in a new era of human-machine interaction. The integration of GPT-3 and similar models into Conversational AI has become a central point of interest, spurring extensive research and development. This comparative analysis embarks on a comprehensive exploration of the ever- evolving landscape of Conversational AI, delving into how GPT-3 and its contemporaries are harnessed, scrutinizing their inherent strengths and limitations, and delving into the far-reaching impact they have on a diverse array of industries, from revolutionizing customer service to enhancing healthcare practices.

II. The Evolution of Conversational AI:

Conversational AI, the branch of artificial intelligence focused on enabling machines to engage in natural language conversations, has witnessed a remarkable evolution over the years. Traditional rule-based chat bots, which often left users frustrated with their limited capabilities, have given way to more sophisticated models that utilize machine learning and natural language understanding. However, it was the introduction of transformer-based models, like GPT-3, that truly marked a turning point in this field.

(a) GPT-1

GPT-1 is the first version of the GOT language, released in 2018. It was based on the Transformer architecture, revolutionized natural language processing by enabling efficient parallelization of computations and capturing long-range dependencies through self-attention mechanisms. GPT-1 was pre-trained on large text data, which includes articles, books and webpages, using language modeling task. The model can predict next word in a sequence of text, given the previous words in the sequence. This pre-training process allowed GPT-1 to learn the patterns and relationship between words in large amount of text data. Following the initial pre-training phase, GPT-1 had the capability to undergo fine-tuning for particular tasks downstream, including activities like language translation, sentiment analysis, and text classification.

(b) GPT-2

GPT-2 was remarkable improvement over GPT-2, with 1.5 billion parameters, making it one of the largest language models at the time of its release. Like GPT-1, the model is trained on massive text data, which included webpages, books and other written content, using a language modeling task. The model was able to predict the next word in a sequence of text, given the previous words in the sequence. However, the difference was that the GPT-2 was able to generate longer and more clear sequences of text, and it demonstrated a greater ability to work on new tasks and domain. After the pre-training phase, GPT-2 had the flexibility to undergo fine-tuning for various downstream tasks, including activities like text classification, sentiment analysis, and question-answering. The model demonstrated cutting-edge performance on many of these tasks, excelling particularly in generating high-quality natural language text. One remarkable aspect of GPT-2 was its capacity to produce authentic and cohesive text, posing challenges in distinguishing it from human-authored content. This raised concerns about potential misuse, such as generating misleading information or propaganda. Consequently, OpenAI initially opted to release a scaled-down version of the model, withholding the full version due to these apprehensions.

(c) GPT-3

GPT-3 stands out as one of the most massive and potent language models ever developed, boasting
175 billion parameters, surpassing GPT-2 by several times. Trained on an extensive corpus of textual data, encompassing web pages, books, and diverse written materials, the model engaged in a language modeling task, predicting the next word in a text sequence based on preceding words. GPT-3 exhibited a remarkable capability to generate natural language text of superior quality, marked by coherence and realism. Notably, GPT-3 showcased versatility in performing various natural language processing tasks like text classification, sentiment analysis, and question-answering, all without the need for task-specific training data. This adaptability stems from the model's proficiency in learning diverse linguistic features and patterns during pre-training, enabling it to generalize across numerous tasks and domains. Introducing innovative features like multi-task learning, allowing simultaneous execution of multiple tasks, and few-shot learning, enabling the model to learn new tasks from minimal examples, further solidified GPT-3's reputation as a flexible and versatile language model. Its applications span real-world scenarios, including chatbots, language translation, content generation, and even code creation. GPT-3 has not only garnered substantial attention and enthusiasm in the artificial intelligence community but has also stimulated fresh research and development in the realm of natural language processing.

(d) ChatGPT

ChatGPT undergoes pre-training on an extensive collection of textual data, encompassing books, articles, and websites, where it engages in a language modeling task [88]. Through this pre-training process, ChatGPT acquires an understanding of patterns and associations among words and phrases in natural language. This proficiency enables ChatGPT to generate responses in conversations that are both coherent and realistic.

Table 1

Evolution of ChatGPT and their Characteristics.

Era	Key Characteristic
Rule-Based Systems (1960s-1990s)	• Predefined scripts and decision trees ● Limited flexibility in handling natural language ● Rigid, specific to programmed scenarios
Statistical and Machine Learning (1990s-2010s)	• Introduction of statistical models ● Improvement in dynamic responses ● Limited contextual understanding
Chatbots and NLP (2010s)	• Integration of NLP techniques ● Rise of chat bots with improved understanding ● Introduction of voice recognition (Siri, Google Assistant)
Deep Learning and Neural Networks (2010s-Present)	• Shift to deep learning and neural networks ● RNNs, Transformers (GPT) for natural language processing ● Significant improvement in language understanding
Pre-trained Language Models (2018-Present)	• GPT-3 and other pre-trained models ● Transfer learning for fine-tuning on specific tasks ● Reduced need for massive task-specific datasets
Multimodal Conversational AI (2020s-Present)	• Integration of text, voice, and visual inputs ● Enhanced context awareness and more natural interactions ● Advances in computer vision and speech recognition
Explainable AI and Ethical Considerations (2020s-Present)	● Emphasis on Explainable AI (XAI) ● Addressing ethical considerations such as bias and privacy
Conversational AI in Industry and Daily Life (2020s-Present)	● Applications in customer service, healthcare, education ● Integration with smart devices and home automation ● Personalized assistance in various domains
Ongoing Research and Future Trends	• Focus on context, sentiment, and intent understanding ● Research in more efficient and scalable models ● Exploration of new modalities (e.g., emotion recognition) ● Addressing ethical concerns in AI development and deployment

III. METHODOLOGY

- ChatGPT

The evolution of ChatGPT is intricately connected to the progression of OpenAI's GPT series, which initiated with the introduction of GPT-2 in 2019. GPT-2 garnered attention for its groundbreaking ability to generate coherent and contextually relevant text. Following this success, OpenAI unveiled GPT-3 in June 2020, a model that significantly surpassed its predecessor in both scale and capabilities, boasting an unprecedented 175 billion parameters. GPT-3 showcased unparalleled language understanding and demonstrated versatility across a spectrum of tasks, ranging from language translation to code generation.

As a specialized variant designed for conversational applications, ChatGPT emerged as a sibling model to GPT-3. Initially introduced as a research preview, ChatGPT allowed users to engage with the model and offer valuable feedback on its performance, strengths, and limitations. OpenAI iteratively improved ChatGPT based on user interactions and ongoing research in natural language processing, leading to refined architectures and enhanced training methods.

A pivotal moment in ChatGPT's timeline occurred in February 2023 with the commercial launch of the ChatGPT API by OpenAI. This made the model accessible to developers and businesses, enabling the integration of ChatGPT's language capabilities into external applications. The release of the API marked a significant step towards broader utilization of ChatGPT in various contexts.

Throughout its evolution, ChatGPT has been subject to research papers and comparative analyses, both by OpenAI and the user community. These endeavors aimed to better understand the model's behavior, uncover potential use cases, and address any inherent limitations. It's essential to recognize that the landscape of language models and conversational AI is dynamic, and subsequent developments may have occurred since my last update in January 2023. Researchers and developers continue to push the boundaries of what these models can achieve, contributing to the ongoing evolution of conversational AI.

- *BARD*

Bard, a conversational generative AI chatbot, was created by Google in response to the emergence of OpenAI's ChatGPT. Initially based on the LaMDA family of large language models (LLMs) and later PaLM, Bard was introduced in a limited capacity in March 2023, receiving a lukewarm response before expanding to other countries in May.

While LaMDA was announced in 2021 but not made public, the unexpected popularity of OpenAI's ChatGPT after its launch in November 2022 prompted a significant and rapid reaction from Google. The company, taken aback by ChatGPT's success, mobilized its workforce and hastened the development of Bard, which was officially launched in February 2023. Bard took center stage during the Google I/O keynote in May 2023, marking Google's comprehensive response to the evolving landscape of conversational AI.

- *BERT*

BERT, or Bidirectional Encoder Representations from Transformers, is a pivotal natural language processing (NLP) model that has significantly influenced the field of machine learning and AI. Developed by Google in 2018, BERT represents a groundbreaking shift in the approach to language understanding tasks.

Unlike traditional language models that read text input sequentially, BERT introduces a bidirectional context by considering the entire input sentence simultaneously. This bidirectional attention mechanism allows BERT to capture contextual information and relationships between words more effectively. BERT is based on the Transformer architecture, which has become a standard for various NLP applications.

Pre-training is a crucial aspect of BERT's effectiveness. The model is pre-trained on massive amounts of unlabeled text data, learning the intricacies of language structure, grammar, and semantics. This pre-training phase enables BERT to create contextualized word embedding, enhancing its ability to comprehend the nuances of language.

One of BERT's distinctive features is its versatility in handling a wide range of NLP tasks. Rather than being designed for specific applications, BERT can be fine-tuned for various tasks such as text classification, named entity recognition, question answering, and more. This adaptability has contributed to BERT's widespread adoption across different domains.

BERT's impact on search engines and information retrieval is noteworthy. Search engines, including Google, have incorporated BERT to better understand user queries and provide more relevant search results. The model's ability to comprehend the context of words within a sentence has led to more accurate and context-aware language processing.

Despite its success, BERT is not without challenges. The computational resources required for training and fine-tuning BERT are substantial,
limiting its accessibility for smaller projects. Additionally, BERT may struggle with out-of-domain or domain-specific language for which it was not explicitly trained.

In summary, BERT stands as a landmark in natural language processing, offering a powerful framework for understanding language context bi-directionally. Its impact extends across a spectrum of applications, influencing the way AI systems comprehend and generate human-like language.

- *T5*

T5, or Text-to-Text Transfer Transformer, is a cutting-edge natural language processing (NLP) model developed by Google Research. Introduced in a research paper in 2019, T5 is part of the Transformer architecture family, which has revolutionized the field of machine learning, particularly in the realm of language understanding and generation.

What sets T5 apart is its innovative "text-to-text" framework. Unlike traditional models designed for specific NLP tasks, T5 approaches all tasks as converting input text into output text. This unified framework simplifies the training and fine-tuning processes, making T5 highly versatile across a wide range of natural language processing tasks.

T5's pre-training involves exposing the model to diverse and extensive datasets, enabling it to learn general language patterns, syntax, and semantics. This pre-training phase is essential for T5 to understand the nuances of language and context, preparing it for various downstream applications.

The flexibility of T5 lies in its ability to be fine-tuned for specific tasks without task-specific model architectures. Whether it's text classification, summarization, translation, or question answering, T5 can be adapted with minimal modifications to its architecture. This adaptability has led to T5 being widely used in both research and practical applications.

The "text-to-text" approach promotes a unified way of framing different NLP tasks, fostering simplicity and consistency in model design and usage. T5 has demonstrated state-of-the-art performance in various benchmarks and competitions, showcasing its effectiveness in understanding and generating human-like text.

In conclusion, T5 stands as a significant milestone in the evolution of NLP models, offering a versatile and unified framework for addressing a multitude of language-related tasks. Its impact reverberates across academia and industry, shaping the landscape of natural language processing and pushing the boundaries of what AI systems can achieve in understanding and generating textual information.

TABLE-2

COMPARATIVE TECHNOGIES

Feature/Aspect	**Bard**	**ChatGPT**	**T5**	**BERT**
Developer/Origin	Google	OpenAI	Google	Google
Model Architecture	PaLM (formerly LaMDA)	Transformer	Transformer	Transformer
Release Date	March 2023 (limited), May 2023 (expanded)	June 2020	2019	2018
Primary Model Size	Not specified	175 billion parameters	Not specified	340 million parameters (BERT base)
Training Data	Not specified	45 terabytes of diverse text	Not specified	BookCorpus, English Wikipedia
Training Approach	Fine-tuning on PaLM/LaMDA	Pre-training and fine-tuning	Pre-training and fine-tuning	Pre-training and fine-tuning
Key Strengths	Conversational capabilities, Response to ChatGPT	Versatility, large-scale language comprehension	Text-to-text framework, Versatility	Contextual understanding, State-of-the-art performance
Applications	Conversational AI, Various industries	Conversational AI, Various industries	Various NLP tasks	Various NLP tasks
API Accessibility	Yes, limited capacity initially, expanded later	Yes, OpenAI API	Yes, accessible via Hugging Face Transformers library	No official API, but accessible through Hugging Face Transformers library
Fine-Tuning Flexibility	Fine-tuned on specific tasks	Fine-tuned on specific tasks	Fine-tuned on specific tasks	Fine-tuned on specific tasks
Release Impact	Lukewarm response initially, Expanded availability later	Significant impact on Conversational AI, API widely used	Significant impact on NLP research and applications	Pioneering impact on NLP research, widely used in various applications
Drawbacks/Challenges	Initial lukewarm response	Potential for biased outputs, Lack of true understanding	Computational demands, Ethical considerations	Sensitivity to input phrasing, Limited contextual understanding

Current Status/Updates	Available, continuous development	Available, continuous development	Available, continuous development	Widely used, continuous research

IV. Model Architecture

- *LaMDA*

LaMDA, short for Language Model for Dialogue Applications, represents a group of conversational large language models created by Google. Originally named Meena in 2020, the initial version of LaMDA was unveiled at the 2021 Google I/O keynote, with a second generation introduced the following year.

Language Model for Dialogue Applications (LaMDA) functions by processing and generating human-like text in response to conversational inputs. It leverages large language models developed by Google, aiming to engage in natural and contextually relevant conversations. LaMDA employs advanced natural language processing techniques, allowing it to understand the nuances of language and generate coherent responses. Through pre- training on extensive text datasets, LaMDA learns the patterns and relationships within natural language, enabling it to provide realistic and contextually appropriate answers in dialogue applications. The specifics of how LaMDA processes and generates text involve intricate algorithms and neural network architectures within the model, contributing to its ability to comprehend and respond to diverse conversational inputs.

- *PaLM*

PaLM 2, the latest language model, represents a significant advancement over its predecessor, PaLM, boasting enhanced multilingual proficiency and reasoning capabilities with greater computational efficiency. Built on the Transformer architecture, PaLM 2 underwent training using a diverse range of objectives. Rigorous evaluations across English and multilingual tasks, as well as reasoning challenges, demonstrate a considerable improvement in task quality across various model sizes. Importantly, PaLM 2 showcases accelerated and more efficient inference compared to PaLM, facilitating broader deployment and enabling quicker responses for a more natural interaction pace. The model excels in robust reasoning, exhibiting substantial enhancements over PaLM in tasks like BIG- Bench. PaLM 2 maintains stable performance in responsible AI evaluations, providing inference-time control over toxicity without compromising other capabilities. Overall, PaLM 2 achieves state-of-the-art performance across diverse tasks and capabilities.

It's essential to note that when discussing the PaLM 2 family, distinctions should be made between pre-trained models, fine-tuned variants, and user-facing products, which often involve additional processing steps. User-facing products may also see changes in the underlying models over time, so exact performance matching with the reported results may vary.

- *Transformer Architecture*

The Transformer is a type of neural network architecture designed for processing sequential data, encompassing text, audio, video, and images. Unlike traditional architectures, it forgoes recurrent or convolution layers, relying instead on a fundamental layer known as Attention. Additionally, it incorporates essential layers like fully-connected layers, normalization layer, embedding layer, and positional encoding layer. The subsequent sections will delve into the specific functions of each of these layers.

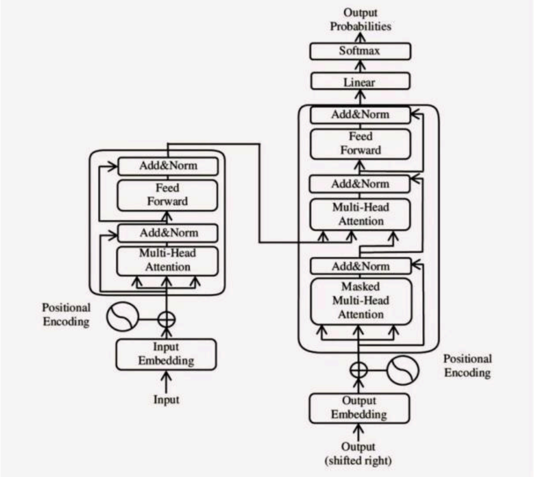

Figure: An illustration of main components of the transformer model

As mentioned earlier, the transformer was originally designed for machine translation, a task involving the processing of two sequences the encoder, decoder, and additional layers will be elaborated below.

1. Encoder

The encoder constitutes a crucial component within the transformer architecture, positioned right at the input of the input sequence. Its primary function is to convert the input sequence into a condensed representation. In the original transformer design, the encoder was iteratively replicated six times, although this can be adjusted based on the overall size of the architecture. Each encoder block encompasses three key layers: multi-head attention (MHA), layer normalization, and multi-layer perceptrons (MLPs) or feedforward layers, as per the paper.

The transformer paper terms multi-head attention and MLPs as sub-layers. Layer normalization and dropout are situated between these sub- layers, with residual connections also present to maintain the correct flow, as depicted in the diagram.

As mentioned earlier, the number of encoder layers was set to 6. Increasing the number of encoder layers corresponds to a larger model, enhancing the model's capacity to capture the global context of input sequences and, consequently, improving task generalization.

2. Decoder

The decoder closely resembles the encoder, with the addition of an extra multi-head attention mechanism that operates on the output from the encoder. The primary objective of the decoder is to merge the encoder output with the target sequence and generate predictions, essentially predicting the next token. In the decoder, the attention directed towards the target sequence is masked to prevent the current token being processed from attending to subsequent tokens in the target sequence. This masking is crucial to avoid the scenario where the decoder has full access to the target sequence, preventing the model from generalizing beyond the training data.

Similar to the encoder, the decoder is typically repeated the same number of times. In the original transformer design, there were also 6 blocks of decoder layers.

3. Attention

Attention serves as a fundamental component within the transformer architecture. At its core, attention is a mechanism enabling the neural network to focus more on the portion of input data containing significant information while reducing attention to the remaining input.

The utilization of attention predates the introduction of the transformer architecture and was a key element in various tasks. Initially employed in neural machine translation (NMT), the attention concept aimed to identify positions in the input sentence where the most relevant information is concentrated. The breakthrough came as attention-based NMT allowed for joint or simultaneous alignment and translation, outperforming earlier methods. The visual representation below illustrates the network's ability to correctly order words in a translated sentence, a notable achievement that prior neural machine translation approaches struggled to attain.

(a) Self-Attention

The self-attention mechanism is a fundamental aspect of the transformer architecture, revolutionizing the processing of sequential data. At its core, self-attention empowers the model to dynamically weigh the significance of different elements within the same sequence, enabling nuanced and context-aware information processing. In the transformer's self-attention mechanism, each word in the input sequence is associated with Query (Q), Key (K), and Value (V) vectors, derived from the input embedding. By calculating attention scores through the dot product of Query and Key vectors, the model can emphasize important words and generate weighted sums based on their contributions, fostering a rich contextual understanding.

This self-attention mechanism proves pivotal in capturing intricate relationships and long-range dependencies between words, making transformers highly effective in natural language processing tasks. The ability to consider the contextual relevance of each word to others within the sequence enhances the model's capacity for tasks like language modeling, machine translation, and text generation. The transformer's self-attention mechanism stands as a testament to its adaptability and proficiency in handling diverse sequential data with a focus on context-aware information processing.

(b) Multi-Head Attention

Multi-Head Attention is a pivotal mechanism in the transformer architecture, enhancing its capacity to capture intricate patterns in sequential data. Instead of relying on a single attention mechanism, multiple attention heads operate in parallel. Each head independently processes the input sequence, allowing the model to attend to diverse aspects simultaneously. This parallel processing enriches the model's ability to understand both local and global dependencies in the data, contributing significantly to its success in various natural language processing tasks.

The main difference between self-attention and multi-head attention lies in their scope of processing. Self-attention, also known as intra-attention, involves attending to different positions in the same sequence. It allows the model to weigh the significance of each element in the sequence concerning the others within that sequence. On the other hand, multi-head attention extends this concept by employing multiple attention heads that operate in parallel. Each head focuses on a distinct subspace of the input, facilitating the model in capturing diverse relationships and patterns simultaneously across the entire sequence. While self-attention is more localized within a sequence, multi-head attention broadens the model's understanding by considering various aspects in a parallel manner.

4. *Feedforward Neural Network*

After the self-attention mechanism, the transformer uses feedforward neural networks for further processing. In the transformer architecture, feedforward neural networks (FFNN) serve as a crucial component, contributing to the model's ability to capture complex patterns and relationships within the input sequences. Positioned after the self- attention mechanism in both the encoder and decoder blocks, the FFNN provides a non-linear transformation to the attended representations. Typically consisting of two linear transformations separated by a rectified linear unit (ReLU) activation function, the FFNN allows the model to learn intricate mappings and higher-level features from the input data. This transformation enhances the model's capacity to understand and represent abstract patterns, contributing to its effectiveness in various natural language processing tasks. The inclusion of feedforward neural networks in the transformer architecture contributes to the model's ability to handle intricate sequential data and perform tasks such as language translation, text summarization, and sentiment analysis.

5. *Layer Normalization and Residual Connections*

In the transformer architecture, layer normalization and residual connections play vital roles in enhancing the training stability and gradient flow within the model. Layer normalization is applied before each sub-layer, such as self-attention and feedforward neural networks, normalizing the activations and reducing internal covariate shift. This normalization ensures that the model remains robust during training by maintaining consistent input distributions across layers. Residual connections, inspired by the concept of skip connections, involve adding the input of a sub- layer to its output, creating a shortcut connection. This allows the gradients to flow more easily during backpropagation, mitigating the vanishing gradient problem and promoting effective learning. Together, layer normalization and residual connections contribute to the overall stability, convergence, and training efficiency of the transformer architecture, facilitating its successful application in various natural language processing tasks.

6. *Pretraining and Fine-Tuning*

Transformers are commonly pretrained on extensive text datasets through self-supervised tasks, such as predicting masked words within sentences. Following this pretraining, these models undergo fine-tuning for specific applications to adapt their acquired knowledge to the target tasks.

In the context of ChatGPT, the transformer architecture is customized for the task of generating text in a conversational style. Input prompts are processed through the encoder-decoder transformer, which incorporates self-attention mechanisms, position encodings, and other components to produce coherent and contextually fitting responses.

The transformer architecture's proficiency in capturing intricate dependencies and word relationships has positioned it as a fundamental component in contemporary natural language processing, supporting applications like machine translation, text generation, sentiment analysis, and more.

V. Market Survey and analysis

Due to its capacity to comprehend context and offer pertinent information, ChatGPT holds significant potential as a valuable tool for gathering, analyzing, and comprehending market trends. The tolol offers several advantages, including the streamlining of existing processes, the facilitation of qualitative data collection through conversational surveys, the analysis and extraction of information from extensive unstructured data, the provision of valuable market intelligence, and the considerable time and effort savings for researchers.

In the survey, 115 senior market researchers worldwide were polled regarding ChatGPT and its implications in the market research industry. The results indicated that 45% of respondents believed ChatGPT would profoundly positively impact the industry, while 36% were in the process of exploring and monitoring ChatGPT, anticipating an increased utilization in the years to come. The survey findings emphasized that ChatGPT is poised to influence market research, although certain considerations need to be incorporated to enhance the effectiveness of traditional market research methods. It is anticipated that an escalating number of market research firms will delve into exploring ChatGPT and its potential positive effects on their market research procedures.

In May, OpenAI's ChatGPT experienced a decline in global visits, stabilizing at around 1.8 billion, still surpassing Microsoft Bing, which received approximately 1.25 billion global visits. Notably, Bing, operational for 14 years, had fewer visits compared to ChatGPT, which was introduced at the end of November 2022.

These traffic figures encompass repeated visits by the same users, and while unique visitor data for May is pending, April witnessed 206.7 million unique ChatGPT visitors resulting in
1.76 billion visits. The total user count since ChatGPT's launch in November is likely higher, although there isn't an official estimate from Similarweb regarding the overall user count.

Concurrently, Google's Bard chatbot observed a surge in visitors, reaching 142.6 million in May, a notable increase from 49.7 million in April according to preliminary data. Both Bard and the Bing chatbot, featuring OpenAI's latest GPT-4 model (accessible only with paid ChatGPT accounts), are free alternatives leveraging search engines, providing access to more current information compared to ChatGPT.

1. ChatGPT, hosted at chat.openai.com, registered 1.8 billion global visits in May, reflecting a 2.8% increase from April.
2. Analyzing daily visits for May (31 days) and April (30 days) shows a marginal monthly increase.
3. Overall traffic to openai.com, mainly from the ChatGPT subdomain, rose by 2.2% in May, showcasing a substantial year-over-year surge of nearly 3,700%.
4. A year ago, OpenAI was a relatively unknown research lab, primarily recognized by AI researchers.
5. Character.ai, the second-largest AI chat platform, experienced a significant 62.5% month-over-month increase, recording 281.4 million visits in May.
6. Founded by former Google engineers dissatisfied with the company's approach to AI chat products, Character.AI has become a prominent player.
7. Google's Bard, launched in beta at bard.google.com in February, saw a remarkable spike with 142.6 million visits in May, indicating a 187.2% surge from April.
8. Bard's success contributes to Google's credibility in the generative AI domain, aligning with efforts to integrate these capabilities into core products, including the search engine.
9. Google officially announced Bard's general availability during the Google I/O conference in early May, eliminating the waitlist.

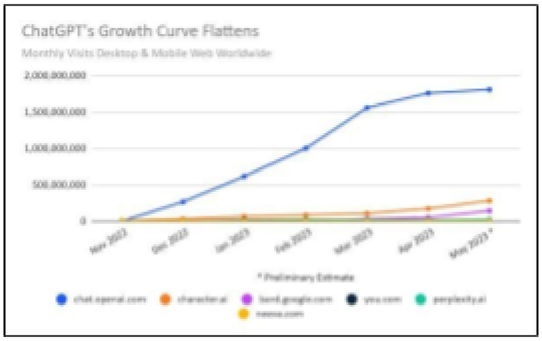

Figure: ChatGPT's Growth Curve Flattens (Monthly Visits Desktop & Mobile Web Worldwide)

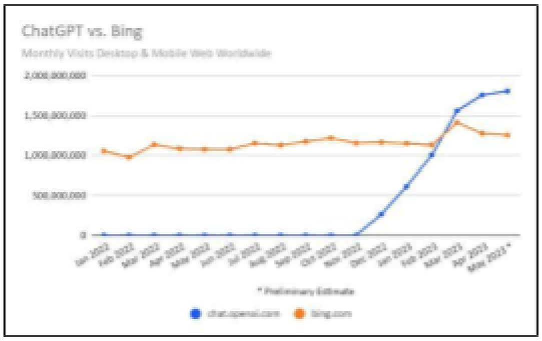

Figure: ChatGPT vs. Bing (Monthly Visits Desktop & Mobile Web Worldwide)

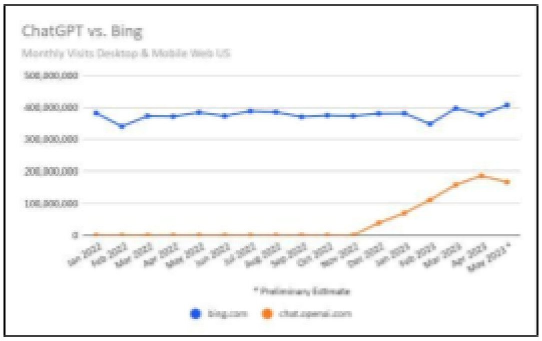

Figure: ChatGPT's Growth Curve Flattens (Monthly Visits Desktop & Mobile Web US)

VI. FUTURE SCOPE

The future scope of ChatGPT holds significant promise, with potential advancements focusing on addressing biases, and ethical considerations, user customization options, increased collaboration with industry applications, and ongoing research and development efforts by OpenAI. The evolution of ChatGPT is expected to involve a holistic approach, incorporating technological enhancements, user feedback, and ethical considerations to ensure responsible and beneficial use across various domains.

Recently, various studies, have found that the GPT-3 possesses an IQ of 150, which places it in the 99.9th percentile. ChatGPT has been tested to have a verbal-linguistic IQ of 147 and achieved a similar result on the Raven's ability test. GPT 3.5 has performed well on the US bar exam, CPA, and US medical licensing exam. On the other hand, it improved contextual understanding, multimodal capabilities integrating text with other modalities, applications in specialized domains through fine-tuning, enhanced task- specific handling, failed in JEE advanced exam, which is India's top engineering exam. Apart from the fail and pass, ChatGPT beats almost 90% of humans in the world's toughest exams.

VII. CONCULTION :

As Conversational AI continues to evolve, researchers, developers, and policymakers are actively addressing these challenges. Efforts are being made to enhance model interpretability, reduce biases, and improve data privacy. The future of Conversational AI will likely see more responsible and ethical use of these powerful models.

VIII. REFERENCE:

1. Ba, J.L., Kiros, J.R. and Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.
3. Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
4. Ahmed, I., Kajol, M., Hasan, U., Datta, P.P., Roy, A. and Reza, M.R., 2023. Chatgpt vs. bard: A comparative study. UMBC Student Collection.
5. Hassani, H. and Silva, E.S., 2023. The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. Big data and cognitive computing, 7(2), p.62.
6. Mungoli, N., 2023. Exploring the Potential and Limitations of ChatGPT: A Comprehensive Analysis of GPT-4's Conversational AI Capabilities.INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, 7(2), pp.178-1.