UNIVERSITY OF CALCUTTA

TOPIC: CHATGPT

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

EXAMINATION: M.Sc. Computer Science Semester-III

PAPER CODE: CSMP306

NAME: Rahul Biswas

REGISTRATION NUMBER: 057-1112-0181-20

UNIVERSITY ROLL NUMBER: C91/CSC/231019

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| Serial Number | Contents | Page Number |
| 1 | Introduction   * What is ChatGPT? | 2 |
| 2 | Transformer architecture   * What is Transformer architecture? * How Transformer works? * What is attention mechanism? | 4 |
| 3 | Generative Pre-trained Transformer (GPT)   * What is GPT? * How has GPT evolved over the years? * What is the working principle of GPT? | 9 |
| 4 | Applications of ChatGPT   * What is the impact of ChatGPT on different fields? | 13 |
| 5 | Conclusion | 17 |
| 6 | References | 18 |

**Introduction**

ChatGPT, founded on the Generative Pre-trained Transformer (GPT) model created by OpenAI,[3] exhibits outstanding efficacy in Natural Language Processing (NLP). This has markedly expedited our endeavors to develop a machine capable of communicating with people as efficiently as feasible. GPT is based on transformer architecture, which solely depends on the attention mechanism, thereby eliminating convolution and repetition totally. It demonstrates superior performance relative to Recurrent Neural Networks due to its parallel computation capabilities.

A language is only a set of rules or symbols used to convey meaning through speech or writing. Natural Language Processing (NLP) is especially helpful for users who are not proficient in programming or formal machine languages, as it allows them to interact with systems more easily without the need or desire to learn a new technical language. [1].With the advent of NLP, more natural interactions between computers and humans have been possible, completely changing the face of communication. One of the main reasons natural language processing has advanced so quickly is the explosion of textual content on the internet. Innovative approaches to these problems have recently become possible as a result of scientific advances. An important step forward in NLP has been the development of GPT. [2]. Launched by OpenAI, a research group committed to the advancement of AI technology, ChatGPT catapulted GPT to fame. Machine translation, text categorization, sentiment analysis, language modeling, and language production are just a few of the many possible uses for GPT, a deep learning model that has been pre-trained on large text corpora. [4].

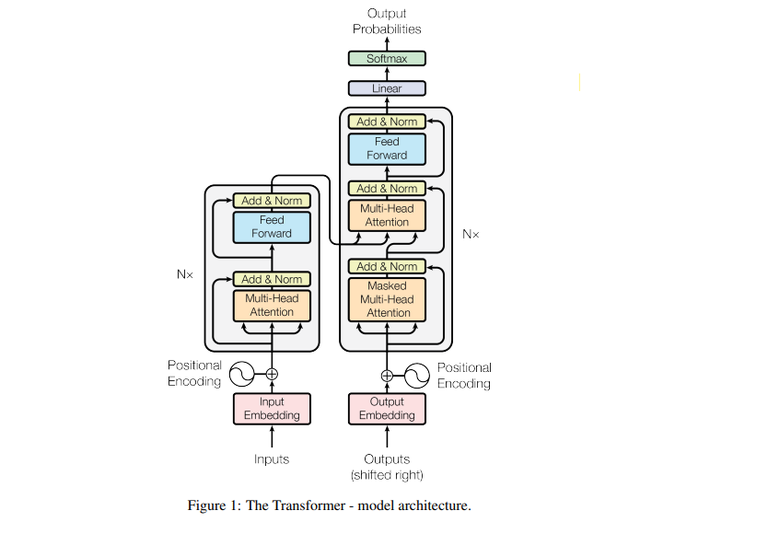
Previous methods of natural language processing, such as RNN and CNN, were significantly inferior to GPT's transformer design. It improves the model's ability to understand and produce language by using a self-attention mechanism that allows it to consider the entire phrase's context while generating the next word. In order to produce the output text, the decoder makes use of the input representation. Reducing sequential processing is the foundational idea behind ByteNet, ConvS2S, and the Extended Neural GPU. At their heart, these systems are convolutional neural networks, which process hidden representations at all input and output points simultaneously. The operational needs of signal correlation between any two input or output locations in these models increase as the distance between them increases. While ByteNet displays a logarithmic pattern, ConvS2S displays a linear one. Because of this, it is more difficult for distant sites to become interdependent. [5]

One of GPT’s key strengths lies in its capability to interpret natural language (NLU), enabling it to grasp meaning, detect entities, and identify relationships within sentences. It also performs remarkably well in natural language generation (NLG), where it produces original text—useful for tasks like content writing or crafting informative responses. Additionally, GPT serves as a code generator, capable of writing code in various programming languages such as Python and JavaScript. In the domain of question answering, it can either generate narrative-style answers from source material or produce concise factual summaries. GPT is also proficient in translating text between different languages and summarizing lengthy content, such as news articles or academic reports, into shorter, digestible formats. [2]

**Transformer Architecture**

Modern neural sequence transduction systems often rely on encoder-decoder frameworks for high performance. In this setup, the encoder transforms the input sequence of symbols x = (x1,…,xn).into a continuous vector representation z = (z1,…zn) Using this encoded representation, the decoder then constructs the output sequence y = (y1,…,yn), generating one symbol at a time. This generation follows an auto-regressive process, meaning each new output is influenced by previously generated outputs.

As shown in Figure 1, the Transformer architecture is designed with multiple layers of self-attention and fully connected feed-forward networks in both the encoder and decoder components.



Encoder: The encoder component includes six repeated layers (N=6N = 6N=6), and each layer contains two essential modules: one for multi-head self-attention and another that’s a basic feed-forward neural network. To stabilize training and preserve information flow, each sub-layer adds a shortcut connection (residual) before applying layer normalization. For an input x, the output is computed as **LayerNorm(x + Sublayer(x))**, where **Sublayer(x)** indicates the operation performed by the sub-layer. These residual connections are consistently applied across the entire model, including the embedding stage, and the output dimensions are fixed at 512.

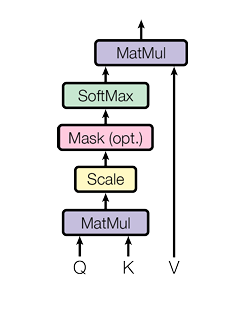
Decoder: The multi-head self-attention layer in the decoder differs slightly from that in the encoder. It conceals all tokens to the right of the token for which the representation is being calculated, so ensuring that the decoder can only focus on tokens before the token it aims to forecast. This is illustrated in Figure 1 as "masked multi-head attention." The Decoder incorporated an additional sublayer, which is a multi-head attention layer applied to all outputs of the Encoder.

Attention: An attention function is defined as a mapping between a set of key-value pairs to an output using vector representations for the query, keys, values, and output. The result is calculated as a weighted sum of the values; the weight of each value is determined by a compatibility function between the query and the corresponding key.

A. Scaled Dot-Product Attention: This particular method of focus that we have developed is called "Scaled Dot-Product Attention". Dimensional queries, keys(), and values() make up the input. Using the attention layers that are running in parallel, we determine their dot products.

To get the data weights, run a query with all keys, divide them all by, and then use the softmax method. on a set of queries structured as a matrix Q, we calculate the attention function all at once. The keys and values are consolidated into matrices K and V. We calculate

There are primarily two ways to pay attention: additive and dot-product (multiplicative). Our approach is identical to dot-product attention with the exception of the scaling factor of . The compatibility function of a feed-forward network with just one hidden layer can be calculated by applying additive attention. Even if dot-product attention and enhanced matrix multiplication both have comparable theoretical complexity, the latter is far more efficient in terms of space and time.

Figure 2: Scaled Dot-Product Attention

For smaller values of ​, both approaches show comparable performance. However, as ​ increases, additive attention tends to scale more efficiently than dot-product attention. When becomes sufficiently large, it is believed that the dot products grow significantly, causing the **softmax** function to operate in areas where gradients are minimal. To counteract this issue, the dot products are scaled by a factor of

B. Multi Head Attention: Instead of applying a single attention mechanism using queries, keys, and values of dimension, it proved more effective to first project them h times into lower-dimensional spaces via independently learned linear transformations to dimensions ​ and ​, respectively. These projected versions—queries, keys K(), and values V()—are then simultaneously fed into the attention mechanism. This produces output vectors with a dimension of [5].

As illustrated in Figure 3, the outputs from each attention head are concatenated and passed through a final linear transformation to produce the result. [5]

The multi-head attention mechanism allows the model to attend to information from different representation subspaces at multiple positions in parallel. This capability is diminished when using only a single attention head, as averaging reduces its expressiveness.

Where the projection are parameter matrices ,

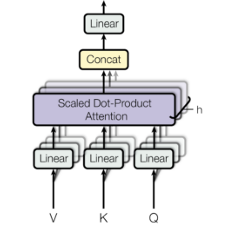
 here h is the number of parallel attention layers.

Figure 3: Multi Head Attention

Position-wise Feed Forward Network: Each layer of our encoder and decoder makes use of a fully linked feed-forward network, and both have attention sub-layers. Every location receives this network in a consistent and autonomous fashion. This contains two linear transformations and an activation function called ReLU.

The linear transformations employ distinct parameters for each layer, while remaining constant elsewhere.

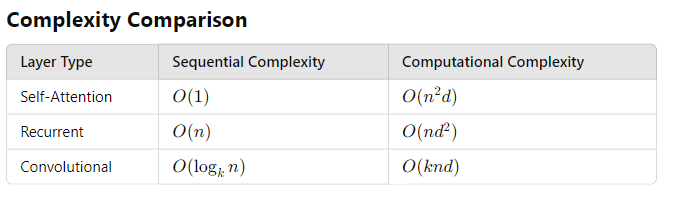
Self-attention: The core mechanism of Transformers is self-attention, enabling the model to efficiently capture relationships between words in a sequence regardless of their distance. Self-attention connects all words in constant time, in contrast to recurrent layers that rely on O(n) sequential processes, hence significantly enhancing parallelization and computational efficiency. Acquiring long-range dependencies—where remote words influence each other—poses a significant challenge in sequence modeling. In recurrent networks, information must traverse many time steps, complicating the models' ability to preserve context across prolonged sequences. Despite necessitating multiple stacked layers and thereby elevating processing complexity, convolutional layers can mitigate this issue. Self-attention minimizes path length and enhances dependency learning through direct connections among all input tokens. Self-attention additionally enhances interpretability. Attention heads, specializing in diverse syntactic and semantic patterns, can highlight important textual linkages. This enables models such as BERT and GPT to excel in tasks including machine translation, summarization, and question-answering. Although extended sequences are computationally expensive, self-attention can enhance efficiency by restricting focus to certain areas. Self-attention is crucial in modern NLP models as it facilitates accelerated training, enhanced long-range dependency acquisition, and improved interpretability.[4] 

Figure 4: Complexity Comparison of Self-attention, Recurrent and Convolutional models

For long-range dependencies, self-attention shortens the path length to O(1), therefore optimizing deep learning tasks.

**Generative Pre-trained Transformer (GPT)**

1. Definition of GPT: The GPT model generates a substantial volume of complex and relevant machine-generated text from minimal input. GPT models are classified as language models that emulate human text using deep learning methodologies. They operate as auto regressive models, wherein the present value is affected by preceding values.

a. “GPTs are language models pre-trained on vast quantities of textual data and can perform a wide range of language-related tasks.”

b. “A GPT is a language model relying on DL that can generate human-like texts based on a given text-based input.”

c. “GPT is a language model developed by OpenAI to help give systems intelligence and is used in such projects as ChatGPT.”

2. Evolution of GPT model: Generative artificial intelligence has arisen from significant advancements in natural language processing technology facilitated by GPT models. Before GPT, NLP models relied on extensive annotated data specific to various tasks. The primary limitations of this strategy were that models were constrained by their training datasets and acquiring sufficient labeled data was difficult. Their inability to generalize from training data renders them less adaptable to new challenges.

To address these challenges, OpenAI developed GPT-1, a generative language model trained on unlabeled data. Through the generation of suitable responses to incoming text, GPT-1 might be enhanced for many tasks, including sentiment analysis, categorization, and question-answering, thus exceeding established classifications. This transformed NLP, as GPT-1 demonstrated the ability of generative models to execute diverse linguistic tasks. Released in 2018, GPT-1 was a major advancement in enabling computers to generate and comprehend language more naturally. It employed a 12-layer transformer architecture featuring a self-attention mechanism, grounded in extensive literature. One of its most significant achievements was its ability to perform zero-shot tasks, indicating that generative pretraining can be effective with minimum fine-tuning.

OpenAI launched GPT-2 in 2019, expanding upon the achievements of GPT-1. This model, with 1.5 billion parameters—tenfold that of GPT-1—demonstrated significant scale improvements. It was designed to handle a wide range of NLP tasks, including translation and summarization, without the necessity for extensive training data. GPT-2 significantly enhanced language prediction by demonstrating an exceptional ability to comprehend long-range relationships inside text. Processing unrefined input text enables the generation of coherent and contextually pertinent text sections, so illustrating the expanding capabilities of generative models.

The subsequent major iteration, GPT-3, provided a groundbreaking advancement in artificial intelligence language processing. With an impressive 175 billion parameters—100 times greater than GPT-2—it is among the largest and most powerful language models ever created. GPT-3, trained on an extensive dataset of 500 billion words, was capable of generating language that nearly resembled human writing. It possessed exceptional abilities in essay composition, addressing complex inquiries, basic arithmetic, and coding. GPT-3 was primarily accessible via a cloud-based API due to its substantial scale and computational demands, enabling developers to integrate it into various projects. Despite its complexity raising concerns around artificial intelligence transparency, its ability to generate very realistic language led to widespread acceptance in both creative and commercial domains.

OpenAI introduced ChatGPT in 2022, building upon GPT-3.5. This version improved upon GPT-3 by refining its ability to generate human-like responses. GPT-3.5 was trained on extensive datasets comprising social media posts, Wikipedia entries, and news articles, utilizing a diverse amalgamation of text and code. In conversational applications, this enhanced its ability to comprehend context and generate more accurate responses. Released in March 2023, GPT-4 is the most sophisticated multimodal language model offered by OpenAI.[2] It features superior reasoning capabilities and an extended context window of up to 32,768 tokens. GPT-4 was trained using reinforcement learning on a combination of public data and licensed datasets to align with human expectations. Providing enhanced precision, greater fluency, and an expanded array of functionalities, it signifies the subsequent stage in the evolution of generative artificial intelligence, building upon prior iterations.

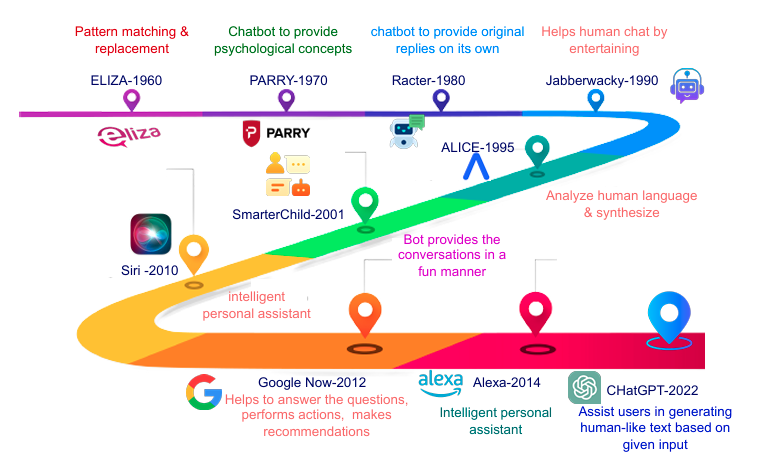


Figure 5: GPT road map

3. GPT and Its Architecture: What are known as pre-trained transformer models are neural networks that have been trained for tasks in natural language processing (NLP), including language modeling, text categorization, and generative pre-training (GPT). For text processing, GPT relies on the Transformer paradigm, a successful architecture in natural language processing. Because it uses self-attention methods, this model can evaluate input sequences of different lengths. While the conventional Transformer model incorporates both encoder and decoder blocks, GPT streamlines its architecture by exclusively using decoder blocks. By predicting subsequent words in a sequence based on their predecessors, this model was pre-trained on a big text dataset using unsupervised learning techniques. With this groundwork in place, the model may learn to represent natural language in a way that is unique to each task. [2]

To prepare the input words or subwords for processing by Transformer blocks, the input embedding layer converts them to continuous vector representations. To address the inherent disorder in Transformer design, positional encoding is employed to convey the relative locations of tokens inside the input embeddings. In certain cases, such as language modeling, masking is employed to guarantee that the model exclusively employs tokens preceding the target word.

During processing, the model is able to zero in on a number of input features thanks to the transformer parts that make up the GPT's base. There is a strong relationship between the output phase and the linear and softmax functions.[3] A probability distribution across a group of output classes is generated by the softmax function, which is commonly used in classification problems.the third To determine the significance of each token, the attention mechanism uses linear transformations to build key, value, and query vectors. By adjusting the output of the multi-head attention layer, the feedforward layers improve predictions using these linear functions. To forecast the next token in the sequence, the last layer uses linear functions.

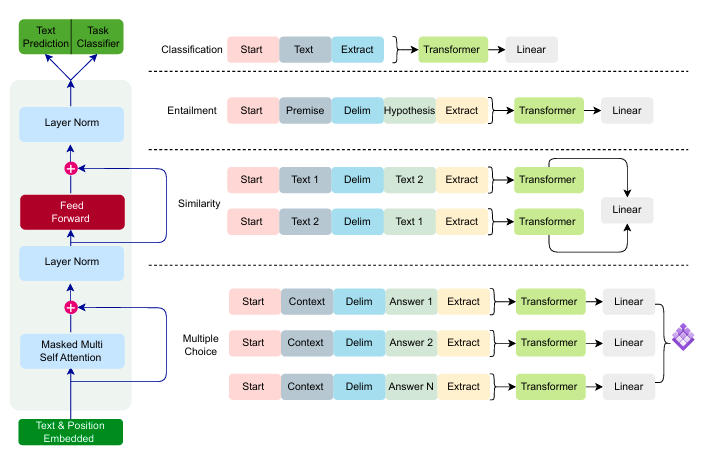


Figure 6: GPT Model with fine tuning

A crucial aspect of GPT is pre-training, which comprises unsupervised model training on a large dataset prior to refining it for specific tasks such as text synthesis or text classification.[2] To improve the pre-trained model's performance on new tasks or datasets, it undergoes further training using task-specific data.

Using the words that have come before it, GPT is able to predict the next word in a sequence, a technique known as language modeling. The model is then able to grasp the interdependencies between words in the training data and their meanings and contexts. Because of its pattern-learning capabilities, GPT is able to produce data that is both coherent and relevant to context, making it an excellent tool for NLP applications.

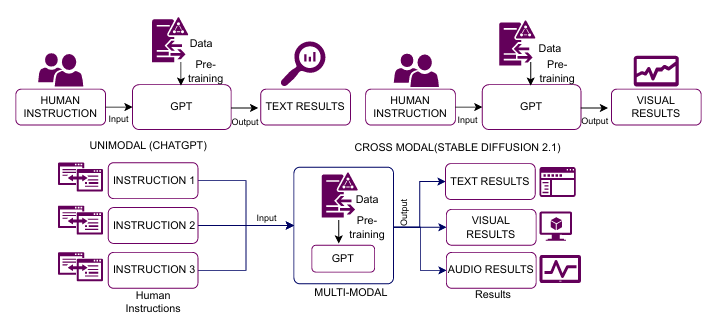


Figure 7: Comparision between Unimodal, crossmodal and multimodal GPT

**Applications of ChatGPT**

1. Education and Research: ChatGPT has significantly transformed education and research by increasing knowledge accessibility, automating academic tasks, and hence improving learning. For researchers, educators, and students, its ability to process vast amounts of data and provide human-like responses is a significant benefit.

The ability of ChatGPT to provide personalized learning experiences is one of its most significant advantages for classroom application. ChatGPT offers students prompt explanations, summaries, and methodical solutions for challenging subjects, functioning as a virtual tutor. This flexibility enables students to learn at their own pace and attain a deeper understanding of subjects without the constraints of traditional classroom attendance. ChatGPT may generate flashcards, quizzes, and practice questions tailored to specific learning needs, hence enhancing information retention. ChatGPT is an excellent resource for educators, facilitating the automation of routine tasks such as content creation, lesson planning, and assessment grading. It enables educators to create interactive exercises, formulate assignments, and develop artificial intelligence-driven chatbots for classroom engagement. Minimizing administrative burdens enables educators to focus more on student engagement, curriculum development, and innovative teaching methodologies.

ChatGPT assists in academic writing and literature reviews by synthesizing research papers, identifying key concepts, and suggesting potential avenues for further research. Its functionalities enable researchers to swiftly analyze extensive datasets, discern patterns, and generate structured reports. ChatGPT assists with abstracts, introductions, and conclusions, so maximizing and conserving time in academic writing. Proofreading and linguistic assistance represent other essential applications of ChatGPT in the educational setting. Many students and scholars find it challenging to organize their essays, theses, or research papers. ChatGPT can enhance writing clarity, grammatical precision, and sentence structure. Translating books, elucidating complex concepts, and improving linguistic proficiency render academic writing more accessible to a global audience, significantly benefiting non-native English speakers. Additionally, ChatGPT contributes to the cultivation of critical thinking and innovation. While it provides rapid information, it also presents multiple perspectives on a topic, so facilitating scholarly and student engagement in critical discourse. As a collaborator in brainstorming, it can generate concepts for artistic endeavors, thesis topics, and research proposals, thereby fostering intellectual curiosity and innovation.

ChatGPT raises ethical and reliability concerns in research and education notwithstanding its advantages. Institutions must address issues such as plagiarism, misinformation, and excessive reliance on AI-generated content. Colleges and educators should implement procedures that prioritize fact-checking and human oversight in academic work to ensure the ethical usage of AI. Moreover, due to the potential for artificial intelligence algorithms to generate inaccurate or biased responses, students and researchers must critically evaluate the provided information.

2. Health-care: Patient data privacy is a critical issue, requiring rigorous security measures to ensure confidentiality and compliance with medical legislation like HIPAA and GDPR. The primary application of ChatGPT in the healthcare sector is the provision of immediate medical information. ChatGPT enables people to pose general health inquiries, understand symptoms, and receive guidance on prevalent illnesses. While it does not replace expert medical advice, it provides information on sickness prevention, medication usage, and lifestyle recommendations, serving as an initial resource for health education. ChatGPT assists medical professionals with clinical decision support by consolidating patient cases, analyzing medical literature, and suggesting potential diagnoses based on symptoms. Rapid access to relevant studies and treatment guidelines enables nurses and physicians to stay abreast of the latest medical research. ChatGPT can assist medical personnel in generating electronic health records, discharge summaries, and medical reports, thereby reducing their documentation burden. ChatGPT enhances virtual consultations in telemedicine by serving as an AI-driven assistant, hence facilitating more effective interactions between doctors and patients. Prior to a medical appointment, it can arrange appointments, prioritize patient inquiries, and provide first evaluations. This enhances healthcare accessibility, especially for individuals residing in remote areas with limited access to medical services. ChatGPT's ability to analyze extensive data, condense study findings, and assist in hypothesis generation benefits medical research. Researchers can leverage insights from scientific literature to predict health risks, analyze disease breakout trends, and facilitate drug discovery. It can also facilitate systematic reviews, funding applications, and research paper authorship, so enhancing the research process. ChatGPT's impact on mental health support is yet another significant aspect. AI-powered chatbots can provide emotional support, regulate stress and anxiety, and offer self-care guidance. While not a replacement for therapy, ChatGPT can be integrated into mental health applications to offer immediate guidance and guide users to professional assistance when necessary. While ChatGPT offers benefits, its application in healthcare raises ethical and reliability concerns. Erroneous or deceptive medical information can have significant consequences; therefore, content generated by artificial intelligence must be evaluated by medical professionals. A significant worry is patient data privacy, necessitating stringent security procedures that ensure anonymity and compliance with HIPAA and GDPR.

3.Industry: ChatGPT is transforming industries by enhancing automation, streamlining processes, and refining decision-making. AI-powered chatbots in customer service respond to inquiries, reduce wait times, and personalize user interactions. Medical recordkeeping, patient engagement, and AI-assisted diagnosis contribute to healthcare improvement. The finance industry use ChatGPT for customized banking solutions, risk assessment, and fraud detection. Predictive maintenance, quality assurance, and supply chain optimization all enhance artificial intelligence in manufacturing. Content creation and marketing employ artificial intelligence to provide customer insights, SEO content, and advertisements. Academic assistance and AI-enhanced learning instruments enhance both research and instruction. Artificial intelligence is utilized in the legal and human resources sectors for automating recruitment, conducting compliance assessments, and analyzing documents. To ensure the ethical deployment of artificial intelligence, challenges such as data security, bias, and misinformation must be addressed. ChatGPT's function will evolve as enterprises use artificial intelligence (AI), hence enhancing efficiency, inventiveness, and accessibility across many business sectors.

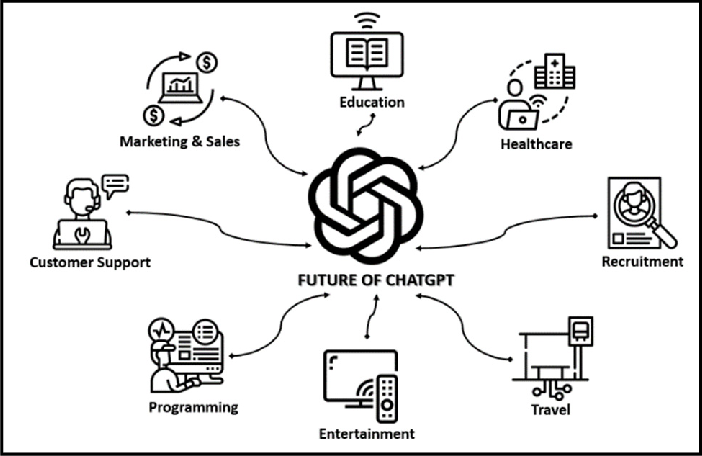


Figure 8: Applications of ChatGPT

**Conclusion**

GPT and other large language models significantly alter our relationships with technology and each other. These models advance various fields through customized guidance, client support, language translation, and content generation. Benchmark translation problems demonstrate that transformer-based architectures, such as GPT, offer significant efficiency improvements compared to traditional recurrent and convolutional networks in translation tasks. However, as these technologies advance, ethical and social concerns like as biases in training data, privacy safeguards, and their impact on human creativity and employment must also be considered. Optimizing the benefits of these models necessitates meticulous and contemplative application that mitigates potential risks. The ethical implementation and ongoing evaluation will enable us to maximize the potential of GPT and similar models, fostering a more inclusive, efficient, and innovative digital future.

**References**

1. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends, and challenges. Springer Nature.

2. Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., B., P., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). GPT (Generative Pre-trained Transformer)– A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. arXiv preprint arXiv:2305.10435v2. <https://doi.org/10.48550/arXiv.2305.10435>

3. “Introducing OpenAI”. [Accessed on 01.02.2025]. [Online]. Available: https://openai.com/blog/introducing-openai

4. Amatriain, X., Sankaran, A., Bing, J., Bodigutla, P. K., Hazen, T. J., & Kazim, M. (n.d.). Transformer models: An introduction and catalog.

5. Vaswani, A., Jones, L., Shazeer, N., Gomez, A. N., Parmar, N., Uszkoreit, J., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Proceedings of NeurIPS 2017. <https://doi.org/10.48550/arXiv.1706.03762>

6. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. arXiv. https://arxiv.org/abs/1409.3215v3

7. Schmidt, R. M. (2019). Recurrent neural networks (RNNs): A gentle introduction and overview. arXiv. https://arxiv.org/abs/1912.05911v1