GPT (Generative Pre-trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu

***Abstract*—The Generative Pre-trained Transformer (GPT) represent a notable breakthrough in the domain of natural language processing, which is propelling us toward the development of machines that can understand and communicate using language in a manner that closely resembles that of humans. GPT is based on the transformer architecture, a deep neural network designed for natural language processing tasks. Due to their impressive performance on natural language processing tasks and ability to effectively converse, GPT have gained significant popularity among researchers and industrial communities, making them one of the most widely used and effective models in natural language processing and related fields, which motivated to conduct this review. This review provides a detailed overview of the GPT, including its architecture, working process, training procedures, enabling technologies, and its impact on various applications. In this review, we also explored the potential challenges and limitations of a GPT. Furthermore, we discuss potential solutions and future directions. Overall, this paper aims to provide a comprehensive understanding of GPT, enabling technologies, their impact on various applications, emerging challenges, and potential solutions.**

***Index Terms*—Generative Pre-trained Transformer, Natural language processing, Artificial Intelligence**

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Prabadevi B are with the School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu- 632014, India (Emails: { gokul.yenduri, ramalingam.m, chemmalarselvi.g, supriya.d, praveenkumarreddy, deeptiraj.g2020, prabadevi.b }@vit.ac.in)

Gautam Srivastava is with the Dept. of Math and Computer Science, Brandon University, Canada, and the Research Centre for Interneural Computing, China Medical University, Taichung, Taiwan as well as Dept. of Computer Science and Math, Lebanese American University, Beirut, Lebanon (email: srivastavag@brandonu.ca)

Rutvij H Jhaveri is with the Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, India, (Email: rutvij.jhaveri@sot.pdpu.ac.in).

Weizheng Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China, (E-mail: weizheng.wang@ieee.org).

Athanasios V. Vasilakos is with the Center for AI Research (CAIR),University of Agder(UiA), Grimstad, Norway, (Email: thanos.vasilakos@uia.no).

Thippa Reddy Gadekallu is with the School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India, Lovely Professional University, Phagwara, India, Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon, Jiaxing University , Jiaxing 314001, China, Zhongda Group, China, 314312 (E-mail: thippareddy@ieee.org).

TABLE I
LIST OF KEY ACRONYMS ONLY IF IT IS REPEATED

Acronyms	Description
AI	Artificial Intelligence
AR	Augmented Reality
BERT	Bidirectional Encoder Representations from Transformers
BGN	Boneh–Goh–Nissim
CNN	ConvolutionalNeural Network
DAP	Data Access Point
DLT	Decentralized Ledger Technology
DL	Deep Learning
DRL	Deep Reinforcement Learning
DR	Demand response
EC	Edge Computing
EU	End User
EAPs	Energy Access Points
5G	Fifth-Generation
4G	Fourth-Generation
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HPC	High Performance Computing
HCI	Human Computer Interaction
IoT	Internet of Things
ML	Machine Learning
NLP	Natural Language Processing
NPC	Non Playable Character
PLM	Pre-trained Language Models
PTM	Pre-Trained Models
RNN	Recurrent Neural Network
6G	Sixth-Generation
TL	Transfer Learning
VU	Virtual Reality

I. INTRODUCTION

Language is the cornerstone of human communication and plays a vital role in shaping our interactions with the world. With the advent of NLP, it has revolutionized the way we interact with machines. NLP has become a game-changer in the world of communication, enabling humans to interact with

machines in a more natural way. The evolution of NLP has been fueled by the exponential growth of textual data in the internet. Over the years, NLP has witnessed a significant transformation from simple rule-based systems to complex deep learning-based models. Despite the advances, natural language understanding and generation have long been a challenging problem in the field of NLP, largely due to the complex nature of human language. However, recent advancements have paved the way for the new approaches to tackle these challenges. One such breakthrough in NLP, is the development of the GPT [1]. GPT became famous after the launch of ChatGPT by OpenAI, a research company [2] that focuses on developing AI technologies. GPT is a deep learning model that is pre-trained on large corpora of text data and can be fine-tuned for specific tasks like language generation, sentiment analysis, language modelling, machine translation, and text classification. The transformer architecture used in GPT is a significant advancement over previous approaches to NLP, such as RNN and CNN. It uses a self-attention mechanism to allow the model to consider the context of the entire sentence when generating the next word, which improved the model's ability to understand and generate language. The decoder is responsible for generating the output text based on the input representation [3].

GPT can perform a wide range of tasks in NLP. One of its key strengths is in natural language understanding (NLU), where it can analyze and comprehend the meaning of text, including identifying entities and relationships in sentences. It's also proficient in natural language generation (NLG), which means it can create text output, such as writing creative content or answering questions in a comprehensive and informative way. Alternatively, GPT is also code generator, where it can write programming code in various languages, such as Python or JavaScript. GPT can also be utilized for question answering, which means it can provide summaries of factual topics or create stories based on the input text. Additionally, GPT can summarize a piece of text, such as providing a brief overview of a news article or research paper, and it can be used for translation, which makes it possible to translate text from one language to another. Overall, GPT's ability to perform a wide range of NLP tasks with high accuracy and precision, makes it an invaluable tool for various industries, including finance, healthcare, marketing, and more. As NLP technology continues to advance, we can expect GPT and other language models to become even more sophisticated and powerful, enabling us to communicate with machines more naturally and effectively.

A. Motivation

GPT has become a transformative technology in the field of NLP, enabling the rapid development and growth of a wide range of industries and applications. Despite its wide adoption and numerous potential applications, there is still much to be explored and understood about GPT's capabilities. Although there are studies on GPT in the literature related to academia and libraries [4], education [5], GPT models [6], banking and corporate communication [7], advancements in chatGPT and its version [8], and on generative AI's [9], no existing reviews are dedicated to providing a comprehensive survey on GPT. Therefore, there is a need for a comprehensive review that focuses on GPT's architecture, enabling technologies, potential applications, emerging challenges, interesting projects and future directions. These limitations motivated us to conduct this review. Hence, this review will not only help researchers and practitioners in this field to gain a better understanding of GPT but also provide valuable insights into its potential applications and major limitations when conducting the research.

B. Related Surveys and Contributions

The GPT model is a type of DL model that uses self-supervised learning to pre-train massive amounts of text data, enabling it to generate high-quality language output. The recent advancements in GPT model research can be attributed to the continual improvement of its architecture, increased availability of computing power, and the development of novel techniques to fine-tune the model for specific tasks. These advancements have led to the creation of larger and more powerful GPT models, enabling them to perform a wider range of NLP tasks with unprecedented accuracy and fluency. These GPT models have demonstrated great potential in transforming various industries like healthcare [10], customer service [11], financial industry [12] and so on. These applications are enabled by the generation of high-quality and diverse data like large-scale corpora of text data with different fast-growing enabling technologies [13], [14]. There are numerous survey papers published to provide a comprehensive overview of the latest developments in GPT models, insights into the different architectures, training methods, evaluation metrics, and highlight the challenges and future directions of this field. This literature survey aims to review and analyze the key findings and contributions of the most recent survey papers published on GPT models, to provide a comprehensive and up-to-date understanding of the state-of-the-art in this exciting and rapidly evolving field.

Lund et al. [4] presents the potential effects of AI and GPT models, specifically ChatGPT, on academia and libraries. They discussed the capabilities of ChatGPT in generating human-like responses and its potential applications. They examine how AI-powered chatbots and virtual assistants based on GPT models can enhance student learning experiences, assist with research tasks, and improve library services. They also address concerns regarding data privacy, biases, and the need for ethical guidelines. Overall, this survey paper highlighted the transformative potential of AI and GPT models while emphasizing the importance of responsible deployment and human oversight.

Kasneci et al. [5] have reviewed the potential opportunities and challenges of using large language models, specifically ChatGPT, for educational purposes. They highlighted the benefits and limitations of using such models by discussing their implications for teaching and learning. In addition, a defined strategy and pedagogical approach with a heavy focus on critical thinking and fact-checking are required while using such large language models in educational institution. Thus,