The Recent Large Language Models in NLP

Ngoc Tran Khanh Le
S P Jain School Of Global Management
Sydney, Australia
tran.aj23syd008@spjain.org

Nadia Hadiprodjo
S P Jain School Of Global Management
Sydney, Australia
nadia.aj23syd013@spjain.org

Hazem El-Alfy
S P Jain School Of Global Management
Sydney, Australia
hazem.elalfy@spjain.edu.au

Aziz Kerimzhanov
S P Jain School Of Global Management
Sydney, Australia
aziz.aj23syd004@spjain.org

Avtandil Teshebaev
S P Jain School Of Global Management
Sydney, Australia
avtandil.aj23syd017@spjain.org

Faculty of Engineering, Alexandria University
Alexandria, Egypt
ORCID: 0000-0001-7297-9244

2023 22nd International Symposium on Communications and Information Technologies (ISCIT) | 978-1-6654-5731-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCIT57293.2023.10376050

***Abstract* - Over the past few years, Natural Language Processing (NLP) has evolved significantly thanks to the development of large Language Models (LMs). In this paper, we present a survey of four recent language models that we believe have had a significant importance in the NLP field lately: BERT (Google), ELMo (Allen Institute), GPT-3 (OpenAI), and LLaMA (Meta AI). For each model, we analyse its architecture, the dataset on which it was trained, its performance evaluation, as well as the strengths and challenges faced by each. Our paper compares the recent Language Models and their contributions to the field of NLP, and discusses future extensions.**

Keywords - NLP, LLM(s), pre-trained, data, model, architecture, performance, state-of-art, BERT, ELMo, GPT-3, LLaMA, application.

I. Introduction

NLP is the field of computer science that focuses on the interaction between computers and human languages. Language Models are a key component of NLP and are trained on large datasets of natural language text to predict the likelihood of different words or phrases in a given context. Recent developments in language models have revolutionised the field of NLP. In the last few years, we have seen the emergence of several powerful language models, such as ELMo (Embedding for Language Model), Google's BERT (Bidirectional Encoder Representations from Transformers), OpenAI's GPT-3 (Generative Pre-trained Transformer-3), and Meta AI's LLaMA (Large Language Model Meta AI). Those have significantly advanced the state-of-the-art in NLP tasks such as question-answering, sentiment analysis, and text generation.

In this survey, we discuss the latest developments in language models and their impact on the field of NLP. Specifically, we highlight for each model its architecture, how it is trained and evaluated, then infer potential and obstacles of the surveyed language models developed by tech giants, namely ELMo, BERT, GPT-3, and LLaMA.

II. The Recent Large Language Models

A. ELMo (Allen Institute)

ELM (Embedding for Language Model) developed in 2018 by The Allen Institute for Artificial Intelligence is "deep contextualised word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy)" [1]. A deep neural network is to create high-dimensional representations of words that consider their context and capture both syntax and semantics. The network is trained on a large dataset of text and predicts the next word in a sequence given previous words, enabling it to understand complex relationships between words and meanings.

1) Architecture

ELMo uses LSTM (Long Short-Term Memory) in its two-layer bidirectional architecture, along with the CNN (Character-level convolutional neural network) [2].

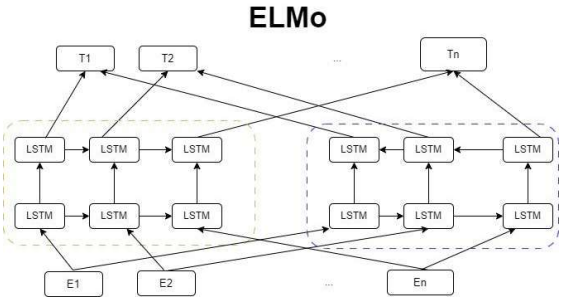

Fig.1. ELMo Architecture [3]

As demonstrated in Fig. 1 [3],

- The input to ELMo is the sequence of sentences that goes to CNN.
- CNN creates raw word vectors. This technique extracts meaning from the text data and represents each word as a unique vector. It works as input for Bidirectional language model (BiLM) -LSTM.
- BiLM is used to process the sequence of words in the forward direction and the other in the backward direction. The backward LSTM captures information about the word and its context after it. The forward LSTM captures information about the word and its context before it.
- LSTM is a type of neural network designed to work with sequential data, such as text or speech. It solves the vanishing gradient problem of traditional RNNs by including "memory cells" that store information for long periods and "gates" that control the flow of information. The input gate adds new information, the

forget gate removes information, and the output gate decides what information should be output. LSTM can remember previous inputs and produce outputs dependent on them. It captures dependencies between words or other features in a sequence of inputs.

2) Dataset

The ELMo is trained on a dataset of 5.5B tokens consisting of Wikipedia (1.9B) and WMT 2008-2012. [4]

3) Model Evaluation

ELMo has been evaluated on six benchmark NLP tasks, including question answering using The Stanford Question Answering Dataset (SQuAD), textual entailment with The Stanford Natural Language Inference (SNLI), semantic role labeling (SRL), coreference resolution (Coref), named entity recognition (NER), and sentiment analysis (SST-5). In all six tasks, simply adding ELMo to strong base models has resulted in state-of-the-art performance, with relative error reductions ranging from 5.8% to 24.9%. [1]

Task	Previous SOTA		Our baseline	ELMo + Baseline	Increase (Absolute / Relative)
SQuAD	SAN	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al (2017)	88.6	88.0	88.7 +/- 0.17	0.7 / 5.8%
SRL	He et al (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al (2017)	91.93 +/- 0.19	90.15	92.22 +/- 0.10	2.06 / 21%
Sentiment (5-class)	McCann et al (2017)	53.7	51.4	54.7 +/- 0.5	3.3 / 6.8%

TABLE 1. Evaluation result [4]

4) Highlights

ELMo is the first language model that brought contextualization into focus, allowing for better performance across a multitude of tasks.[2] It represented a significant advancement in NLP by generating dynamic embeddings that capture the meaning of words in context, resulting in better performance. Before ELMo, most word embeddings were generated using static methods such as Word2Vec or Glove where they did not check for the context in words.

5) Challenges

When it was first introduced, it was highly effective in those six tasks. However, it could be computationally expensive and time-consuming, which makes it difficult to use ELMo in real-time applications or on low-resource devices. Large memory footprint, domain-specific training, and lack of interpretability could also make it challenging.

B. BERT (Google)

BERT - is a new language representation model, which stands for Bidirectional Encoder Representations from Transformers, released in 2018 by Google. It is designed to pre-train deep bi-directional representations based on unlabeled text by co-training left and right contexts in all layers. With just an extra output layer, the pre-existing BERT model can be adjusted to create state-of-the-art models for various tasks like language inference and question answering. This can be accomplished without the need for substantial modifications to the architecture of the model specific to the task.

BERT utilises masked language models to facilitate pre-trained deep bidirectional representations. It is the first representation model that uses fine-tuning as a mechanism to achieve state-of-the-art performance for a broad range of tasks, including those at the sentence-level and token-level. This superior performance surpasses that of many models that were explicitly designed for those tasks.

1) Architecture

The model architecture of BERT is a multi-layer bidirectional Transformer encoder.

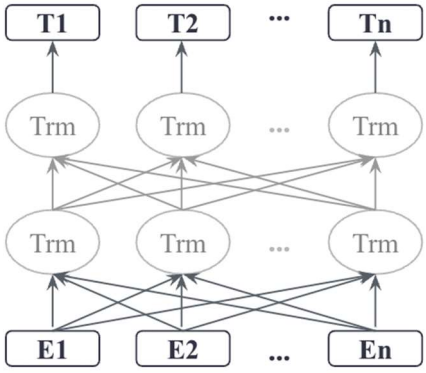

Fig.2. Bidirectional Transformer architecture [5]

2) Dataset

BERT is trained on a large corpus of text from various sources. Google released two pre-training datasets specifically for BERT: BooksCorpus (800 million words) and English Wikipedia (2.500 million words) [5]. As part of the pre-training procedure, the model was trained on two tasks, which are masked language modeling and next-sentence prediction. In masked language modeling, the process involves the masking of particular tokens in the input sequence, followed by predicting their initial values. The subsequent sentence prediction task aims to predict the relationship between two input sentences in terms of whether they are sequential or not.

3) Model Evaluation

The Google BERT is evaluated by many different natural language processing tasks. In many cases, it achieved state-of-art performance on these tasks.

System	MNLI-(m/mm) - 392k	QQP - 363k	QNLI - 108k	SST-2 - 67k	CoLA - 8.5k	STS-B - 5.7k	MRPC - 3.5k	RTE - 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT Base	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT Large	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

TABLE 2. Result table of the model evaluation [5]

4) Highlights

BERT is a highly effective pre-trained language model that has made a significant contribution to the advancement of NLP. Its architecture, pre-training techniques, and remarkable performance across a diverse set of tasks have established the model itself as a widely preferred option for numerous NLP applications.

5) Challenges

However, there are some disadvantages to consider. For example, the large size of the BERT model can make it difficult to store and deploy in certain contexts. In addition, it is expensive, as it requires a lot of computation because of its size. Finally, although BERT has demonstrated impressive performance on a variety of tasks, there is still a need for further research and development to improve the model's ability to handle more complex linguistic phenomena, such as sarcasm or irony [11].

C. GPT-3 (OpenAI)

GPT-3 (Generative Pre-trained Transformer 3) is a state-of-the-art language model developed by OpenAI. It is based on the Transformer architecture and was trained using a massive amount of data to generate human-like natural language responses. It was released on June 11, 2020, and it has received a lot of attention from researchers, industry experts, and media outlets, due to its advanced capabilities and potential for future applications. GPT-3 is the latest version open for public use [21].

The model has 175 billion parameters, making it the largest language model available to date.It has been trained on a wide range of tasks, including language generation, language translation, question answering, sentiment analysis, summarization, chatbots, and text completion [10]. GPT-3 is capable of generating coherent and fluent text that closely resembles human writing [6]. For example: The product ChatGPT will be possibly used for writing scholarly papers [7].

1) Architecture

GPT-3 is based on the Transformer architecture, which is a neural network architecture designed for natural language processing. The Transformer architecture consists of an encoder and a decoder, which work together to generate natural language text. [9]

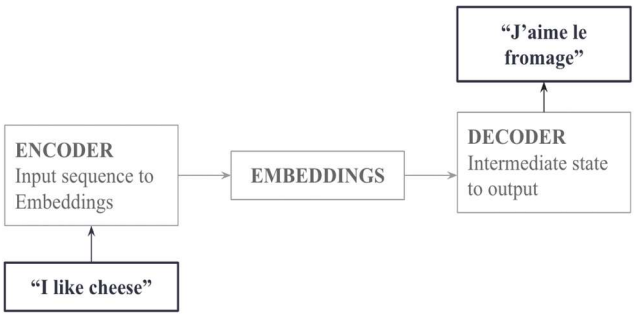

Fig. 3. GPT-3 Transformer Architecture Diagram.[12]

GPT-3 uses a technique called attention, which allows the model to focus on specific parts of the input sequence when generating output. This technique is essential for generating coherent and fluent text, as it allows the model to understand the context of the text it is generating. The GPT-3 model was trained on a massive amount of text data from the internet [21].

It is believed that the model is trained 8 different sizes of the model, ranging from 125 million to 175 billion parameters to study the dependence of machine learning performance on model size. Their aim was to test whether scaling of validation loss is a smooth power law as a function of size, which they did for both validation loss and downstream language tasks [21].

According to OpenAI, the training data consisted of over 45 terabytes of text data, which is equivalent to several thousand books' worth of text. OpenAI used a variety of techniques to clean and pre-process the data, including filtering out low-quality content, removing duplicates, and standardising the formatting of the text. In addition to attention, GPT-3 uses other techniques such as dropout, layer normalisation, and residual connections, which help to improve the performance and stability of the model.

Dropout: Dropout is a neural network regularisation technique that involves randomly dropping out some neurons during training to prevent overfitting. In GPT-3, dropout is applied to the output of each layer.

Layer Normalisation: Layer normalisation is used to normalise layer outputs in a neural network, improving stability and training speed by reducing the impact of input feature distribution differences. In GPT-3, it's applied to the output of each layer.

Residual Connections: Residual connections prevent the vanishing gradient problem in deep neural networks by adding the input of a layer to its output, bypassing the layer's weights. In GPT-3, residual connections are used between each layer pair.

GPT-3 uses other advanced **NLP techniques** to generate human-like text. Some of them are:

- Transformer Architecture
- Attention Mechanism
- Pre-training with Unsupervised Learning
- Fine-tuning on Specific Tasks
- Large-Scale Training

2) Dataset

The Dataset was collected from a wide range of sources, including books, articles, and web pages. The dataset used to train GPT-3 is not publicly available, as it contains copyrighted and sensitive information. However, OpenAI has released a smaller version of the dataset called the "GPT-3 Playground" dataset, which contains a subset of the original training data and can be used for research and experimentation purposes.

Dataset	Quality of tokens	Proportion
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

TABLE 3. GPT-3 training data [21]

Common Crawl is a nonprofit organisation that crawls the web and freely provides its archives and datasets to the public. Common Crawl's web archive consists of petabytes of data collected since 2011. It completes crawls generally every month [11]. It can be seen as most of the training source for the latest models would from this source.

3) Model Evaluation

GPT-3 is evaluated on a number of NLP tasks, including language modeling, cloze tasks, question-answering, translation, Winograd-style tasks, reading comprehension, and synthetic and qualitative tasks. Brown and his team also discuss the SuperGLUE benchmark and natural language inference tasks, and demonstrate that GPT-3 shows strong performance on many of these benchmarks under specific context such as reading comprehension, German translation, Reversed Words and Anagrams, PhysicalQA, Winograd [21].

Another recent scientific study has evaluated the cognitive abilities of OpenAI's GPT-3 language model by cognitive psychology. The study tested how well GPT-3 performed on a series of standard experiments that measure its abilities to make decisions, search for information, think through problems, and understand cause-and-effect relationships. The study finds out that GPT-3 performs well on some tasks but not as well on others, such as the Cognitive Reflection Test. It draws a conclusion that while GPT-3's performance was human-like in some areas, it could not pass as a human in a cognitive psychology experiment [13].

4) Highlights

High-quality text generation: GPT-3 can generate highly coherent and natural-sounding text, making it useful for applications such as chatbots, language translation, and content creation.

Large pre-trained model: GPT-3 is pre-trained on a massive amount of data, which allows it to perform well on a wide range of natural language processing tasks without requiring much additional training.

Zero-shot learning: GPT-3 has the ability to perform certain tasks with no additional training, which is known as zero-shot learning. This makes it easy to use for a variety of applications. [8]

5) Challenges

Limited training data: GPT-3 relies heavily on pre-training data, which can limit its performance in certain applications where the data is highly specific or limited [21].

Limited understanding of context: GPT-3 can struggle with understanding the broader context of a given task, which can lead to inaccurate or irrelevant output [21].

Large size and computational cost: it can limit its usage and make it inaccessible for many researchers or practitioners [21].

Overall, GPT-3 is a highly complex and sophisticated language model that uses a range of advanced techniques to generate human-like natural language text. Its impressive performance is a testament to the power of deep learning and the potential for AI to transform the way we interact with language and technology. GPT-3 represents a significant advance in the field of NLP and demonstrates the power of advanced machine learning techniques for natural language processing.

D. LLaMA (Meta AI)

LLaMA, introduced in February 2023, stands for Large Language Model Meta AI - Meta's LLM of AI. It works as an open source, completely free, allowing researchers, government organisations, and even ordinary users to access it for free. According to Meta's announcement, the model has a maximum of 65 billion parameters which implements predicting the next word in a sequence of text which is trained on text from 20 languages in Latin and Cyrillic alphabets [14]. It is stated that the largest model trained on 1.4 trillion tokens and the smallest on 1 trillion tokens [14]. This model carries a lot of promise in generating text, chatting, summarising written documents, and more complex tasks like solving parametric mathematical theorems [15].

1) Architecture

LLaMA is applied to the transformer architecture as their fundamental in which Encoder encodes the input into a vector representing the semantics while Decoder receives this represented vector and converted it into the target language demonstrated in Figure 4. [16]

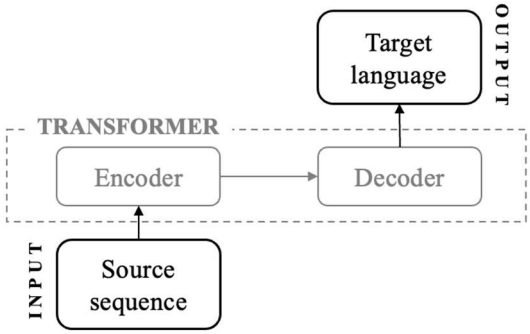

Fig. 4. The overall architecture of transformer

In addition to the original structure, Meta AI has some adaptations in terms of model namely:

- **Root Mean Square Layer Normalisation (RMSNorm)** is to stabilise the training process as well as generalisation capability [17].
- **Swish-Gated Linear Unit (SwiGLU)** activation function leads to faster convergence during training and better generalisation performance because of its improved non-linearity and gating properties [18].

- **Rotary Positional Embedding (RoPE)** allows the network to take into account the order of the input data and better capture the relationships between the different elements of the sequence [19].

All models are trained by the algorithm so-called AdamW optimizer. It is known as a powerful optimizer that helps to prevent overfitting and improve generalisation performance thanks to the weight decay term [20]. In this case, Meta AI applied the indicators as following a weight decay of 0.1 and gradient clipping of 1.0 with a cosine learning rate schedule.

2) Dataset

The training occurs via roughly 1.4 trillion tokens two thirds of it come from Common Crawl. The remaining sources of the dataset are C4, Github, Wikipedia, Books, ArXiv, and Stack Exchange (Table 4).

Dataset	Sampling proportion	Disk size
Common Crawl	67%	3.3 TB
C4	15%	783 GB
Github	4.5%	328 GB
Wikipedia	4.5%	83 GB
Books	4.5%	85GB
ArXiv	2.5%	92 GB
Stack Exchange	2%	78 GB

TABLE 4. Distribution of training data including number of epochs and disk size on 1.4 trillion tokens [15]

3) Model Evaluation

To evaluate model performance, Meta AI has conducted a comparison between a few non-publicly available Large Language Models (LLMs) including GPT3, Gopher, Chinchilla, and Pathways Language Model (PaLM) [15].

As Meta AI research team records [15], LLaMA's benchmarks cover 20 tasks consisting of zero-shot and few-shot task that involves text completion, question answering, and language modelling on long-range dependencies:

- **By zero-shot tasks**, the examination is fed with a textual description of a task and a test example.
- **By few-shot tasks** (from 1 to 64 shots), the examination is fed with a few examples of the task along with a test example.

The outcome, in general, is presented as a ranking of suggested answers or the answer itself.

The evaluation is taken place under diverse aspects that are believed to appropriately measure the efficiency of a model [15], such as Common-Sense Reasoning, Closed-book Question Answering, Reading Comprehension, Mathematical Reasoning, Code Generation, and Massive Multitask Language Understanding (MMLU). Overall, LLaMA has surpassed the other benchmarks in most contexts. It is stated that the model firmly improves itself. However, the model still suffers a slight setback in terms of reasoning mathematical, and understanding massive multitask language due to the limitation of training sources [15].

4) Highlights

On one hand, LLaMA has demonstrated competitive performance with some of the best existing LLMs on standard language NLP benchmarks. Moreover, LLaMA's few-shot learning capabilities enable it to perform well on new tasks with little training data, such as question answering, classification, and language modeling on long-range dependencies.

5) Challenges

On the other hand, there are growing concerns of LLaMA about the potential for bias and toxicity in the datasets used to train these models, as well as ethical concerns around the use of language models for automated content generation or manipulation. Researchers have proposed various methods to address these issues, such as carefully curating training datasets to reduce bias, using bias-detection algorithms to identify and mitigate bias in the models, or designing models that are more interpretable and transparent in their decision-making processes [15].

III. Which Model Functions Better?

We conduct a comparison of the selected models with regards to their strength on one axis (training effort and understanding capability) and their computational complexity and contextual sense on the other axis (Figure 5).

On one hand, it can be seen that although ELMo is the pioneer in contextualization, the model shares the same computational complexity of its contemporary model, BERT, which came to the field only a few months later. However, BERT has performed better with regards to its capability of understanding the feeding, thanks to its use of the masked language model.

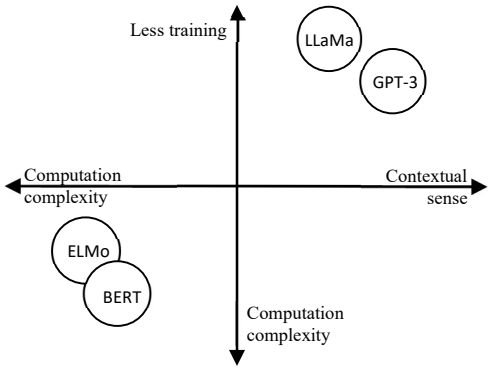

Fig. 5. Proposition map of the studied language models

On the other hand, GPT-3 and LLaMA seem to have improved in that they both require less training, thanks to their use of zero-shot and few-shot learning. In terms of their shortcomings, those two models rely heavily on their training data to generate output. Hence, their responses occasionally fall out of expectation, such as giving irrelevant, biased or misleading answers due the dataset they are fed.

IV. Conclusion

Our paper surveys some of the recent Large Language Models that were developed by tech giants over the latest few years, namely BERT, ELMo, GPT-3, and LLaMA. While ChatGPT has become lately popular and controversial, our objective is to shed more light on other similar technologies and compare them to each other. We focus on each of the aforementioned models in terms of their overall concept and operation to explore their potential and limitations. In particular, we discuss the backbone of the model itself, i.e., the underlying model and algorithm. To offer a fair judgment, we describe the datasets used to train each model in the existing context and the methods used in evaluating them for accuracy and efficiency.

As a whole landscape, the LLMs have impeccably demonstrated their versatility that is ready for numerous applications in the real-world context in accordance with natural language instructions. The advancement gradually mends the gap between artificial intelligence and human beings which changes the way how people can access and process information. Although the rising hope of LM being able to play a critical assisting role for the user's sake, limitations in training data and memory majorly cause bottlenecks in the innovating way of LLMs. This means if the hinder-walls are lowered as the upcoming advancement, LLMs can surpass their blind spot, not just enhance their resilience and alleviate identified problems.

For that reason, we anticipate the expansion and development of larger and more complex language models in the near future, such as the GPT-3 model with 175 billion parameters. Those models have shown significant improvements in their ability to generate coherent and human-like text, and we expect this trend to continue with even larger models being developed.

In the near future, we expect language models to become more adept at understanding and processing natural language with contextual information. This means that language models will be able to better understand the meaning of words and sentences based on the context in which they are used, leading to more accurate and nuanced language processing.

References

[1] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, “Deep contextualised word representations”, ArXiv: 1802.05365 [cs.CL], 2018.

[2] J. Wei, “ELMo: Why it’s one of the biggest advancements in NLP”, https://towardsdatascience.com/, 2021.

[3] P. Joshi, “A Step-by-Step NLP Guide to Learn ELMo for Extracting Features from Text“, https://www.analyticsvidhya.com/ , 2019.

[4] The Allen Institute for AI, “AllenNLP - About ELMo“, https://allenai.org/allennlp/software/elmo , 2018.

[5] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, “*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”,* ArXiv: 1810.04805v2 [cs.CL], 2019.

[6] R. Imamguluyev, “The Rise of GPT-3: Implications for Natural Language Processing and Beyond”, Available at: nternational Journal of Research Publication and Reviews, ISSN: 2582-7421, 2023.

[7] B.D.Lund, T. Wang, “Chatting about ChatGPT: How may AI and GPT impact academia and libraries?”, ISSN: 0741-9058, 2023.

[8] R. Gruetzemacher, F. Barton, “Deep Transfer Learning & Beyond: Transformer Language Models in Information Systems Research”, In *ACM Computing Surveys*, vol. 54, no. 204, pp. 1-35, doi: 10.1145/3505245, 2022.

[9] AltexSoft, “Language Models, Explained: How GPT and Other Models Work.”, www.altexsoft.com/blog/language-models-gpt., 2023.

[10] N.S. Chauha, “OpenAI GPT-3: Understanding the Architecture.” The AI Dream, https://www.theaidream.com/post/openai-gpt-3-understanding-the-architecture , 2022.

[11] Project Pro,“BERT NLP Model Explained for Complete Beginners.”, www.projectpro.io/article/bert-nlp-model-explained/558, 2023.

[12] R. Patel, Space-O AI, “Revolutionising AI: OpenAI GPT-3 Architecture.”, www.spaceo.ai/blog/openai-gpt-architecture, 2023

[13] R. Cami, “AI Study Evaluates GPT-3 Using Cognitive Psychology”, https://www.psychologytoday.com/au/blog/the-future-brain/202303/ai-study-evaluates-gpt-3-using-cognitive-psychology, 2023

[14] Meta AI, “Introducing LLaMA: A foundational, 65-billion-parameter large language model”, https://ai.facebook.com/blog/large-language-model-llama-meta-ai/, 2023

[15] H.Touvron, et al., “LLaMA: Open and Efficient Foundation Language Models”, ArXiv: 2302.13971v1 [cs.CL], 2023.

[16] Verma, N., “Components of Transformer Architecture”, https://lih-verma.medium.com/components-of-transformer-architecture-748f74a1a40d, 2021.

[17] B. Zhang, and R. Sennrich, “Root Mean Square Layer Normalisation”, ArXiv: 1910.07467v1 [cs.LG] , 2019.

[18] C. Raffel, et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, ArXiv: 1910.10683v3 [cs.LG], 2020.

[19] J. Su, et al., “Reformer: Enhanced Transformer With Rotary Position Embedding”, ArXiv:2104.09864v4 [cs.CL], 2022.

[20] I. Loshchilov and F.Hutter, “Decoupled Weight Decay Regularization”, ArXiv: 1711.05101v3 [cs.LG] , 2019.

[21] T.B. Brown, et al. "Language models are few-shot learners." Advances in neural information processing systems 33, ArXiv: 2005.14165, 2020.

[22] M. Matthew, D. Laura, “An Analysis of Deep Contextual Word Embeddings and Neural Architectures for Toponym Mention Detection in Scientific Publications.”, in *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, page 48-56, Minneapolis, Minnesota. Available at ACL Anthology: W19-2607, 2019.