Chapter 4

The Development of Large Language Models From Past to Present

Kürşat Arslan[1]

Mehmet Fatih Karaca[2]

Abstract

LLMs (Large Language Models) are a groundbreaking technology for human-computer interaction. LLMs, which are used in many natural language processing areas such as text generation, question-answer systems, translation and coding, show high success in complex language tasks thanks to their transformer architecture and self-attention mechanism. The development process that started with word embedding techniques has made significant progress with models such as BERT and GPT. LLMs trained with large datasets and powerful GPUs have also significantly improved grammar and context learning. Adapter-based fine-tuning methods increase the accessibility of models by reducing training costs. LLMs, which have revolutionized fields such as health, law, finance, education, and content production, can be integrated with different data types with multimodal models. LLMs have potential future uses in areas such as personalized education, autonomous systems and bio-artificial intelligence integration. However, at this point, challenges such as high computational costs and data quality should also be considered. In conclusion, LLMs have revolutionized the field of natural language processing due to their ability to understand and generate human-like language. Efficient algorithms and innovative solutions are needed in the development and dissemination of language models. In the future, it is expected that LLMs will have a wide range of applications and will be more used and visible in many fields. From a Management Information Systems perspective, the effective integration of LLMs into corporate processes is expected to play critical role in decision support, information management, and increasing overall managerial efficiency.

[1] Tokat Gaziosmanpaşa University, Graduate Education Institute, Department of Computer Engineering, kursat.arslan8423@gop.edu.tr, https://orcid.org/0009-0007-4397-5798

[2] Assist. Prof. Dr., Tokat Gaziosmanpaşa University, Erbaa Vocational School of Higher Education, Department of Computer Technologies, mehmetfatih.karaca@gop.edu.tr, https://orcid.org/0000-0002-7612-1437

https://doi.org/10.58830/ozgur.pub1143.c4720

1. Introduction

LLMs (Large Language Models) are large-scale AI (Artificial Intelligence) systems designed and trained to perform NLP (Natural Language Processing) tasks. These models are trained on extensive text corpora to understand and generate human language and to execute various linguistic tasks. Among these tasks is the ability to communicate with humans in a pre-programmed manner through chatbots (Sadıkoğlu et al., 2023). LLMs are AI models based on deep learning architectures and contain hundreds of millions to hundreds of billions of parameters. One of the most important terms related to LLMs is the "token." Tokens are fundamental units that allow input to be divided into smaller, processable segments. Tokens significantly influence factors such as cost and model performance (Liu et al., 2023).

Key characteristics of LLMs include:

- Understanding and generating natural language (context comprehension)
- General-purpose performance across various tasks
- Requirements for large-scale data and computational resources

In the processes of understanding and generating natural language, it is crucial for LLMs to correctly grasp the context regardless of the length of the text. Accurately processing the relationships between words, the overall structure of the text, and its implicit meanings are essential for producing coherent, contextually appropriate, and meaningful responses. Particularly in long texts, the model's ability to conduct in-depth analysis by focusing not only on surface-level word matches but also on meaning relations that span the entire text determines the strength of its language comprehension capability. Therefore, proper processing of context forms the foundation of LLMs' abilities in information synthesis, text generation, and intelligent response production.

LLMs have a wide range of applications, including:

- Text Generation: Creative writing, content creation, and story writing.
- Question-Answering Systems: Chatbots and virtual assistants
- Machine Translation: Cross-lingual text translation
- Text Summarization: Producing concise and meaningful summaries of long texts
- Programming Assistants: Code completion, debugging, and documentation tasks

Examples of these application areas include the use of ChatGPT for text generation, and tools such as Writesonic and Jasper AI for content marketing copy and social media posts. For question-answering systems, applications such as Siri, Alexa, and Google Assistant may be preferred. In machine translation, Google Translate, DeepL Translator, and Microsoft Translator are among the most well-known and widely used tools. For text summarization, SummarizeBot, Google Docs, Microsoft Word AI, and ChatGPT can be utilized. As for programming assistants, Github Copilot, TabNine, Replit AI, and Amazon CodeWhisperer are frequently preferred by users.

2. The History of LLMs

The history of LLMs begins with the concept of semantics-meaning in language-introduced by the French philologist Michel Bréal in 1897. Bréal examined how languages are structured, how they change over time, and how words are interconnected within a language (Fournet, 2011). Since the early 1950s, NLP research has evolved around key tasks including machine translation, information retrieval, text summarization, question answering, information extraction, topic modeling, and, in more recent years, opinion mining (Cambria and White, 2014).

NLP focuses on converting human communication into a form that computers can understand and then converting it back again. Under normal circumstances, computers cannot comprehend written or spoken text. However, through NLP, this has become possible, enabling translation between languages.

In the 1900s, Ferdinand de Saussure made significant contributions to the theoretical framework of NLP by restructuring the fundamental concepts of linguistics at the University of Geneva. Saussure argued that language should be treated as a structural system and proposed, within the framework of sign theory, that every linguistic unit consists of a signifier and a signified. This approach laid the groundwork for modern NLP studies focusing on modeling word meanings and contextual relationships (Saussure, 1916).

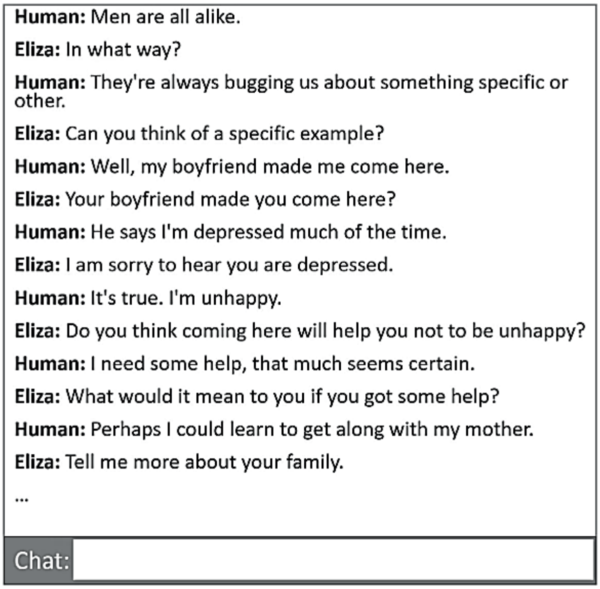
One of the primary goals of NLP is to enable translation. After World War II, the field of NLP gained significant attention due to growing desires for trade, communication, and cultural interaction. However, creating a machine capable of translation proved to be extremely challenging. Unlike humans, who can communicate using complex sentences in their native languages, computers were unable to comprehend such complexity.

The early advances in AI and NLP laid the groundwork for today's LLMs. During this period, the establishment of the fundamental principles of AI and the debates concerning whether machines could possess human-like language abilities played a crucial role in the evolution of technology.

LLMs constitute one of the most significant components of NLP and machine learning within the field of AI. The development of LLMs began with the formation of the foundational concepts of AI and has progressed substantially over the years. In this process, the contributions of mathematician Alan Turing were particularly influential. While working to decipher the Nazis' Enigma machine during World War II, Turing laid the foundations of AI by posing the question, "Can machines think like humans?". From that point onward, research in this field continued at full speed.

In 1950, Alan Turing introduced what is now known as the Turing Test. The Turing Test refers to a machine's ability to participate in human conversations in a way that creates a human-like impression. Turing initially named this the "imitation game." Fundamentally, the imitation game consisted of three participants: Person A, Machine B, and an interrogator. The interrogator's aim was to determine, using simple questions, whether the respondent was Person A or Machine B. The machine's goal was to convince the interrogator that it was a human rather than a machine, while Person A was present to help reveal which participant was the actual machine (Ciesla, 2024).

One result of these early studies was ELIZA, a predecessor of modern systems such as ChatGPT. Emerging in the 1960s, ELIZA was unable to pass the Turing Test, although it could partially convince users that it was human. However, Google's LaMDA and OpenAI's ChatGPT successfully passed the Turing Test in 2022 and 2023, respectively (Ciesla, 2024). An example of a conversation with ELIZA, presented in Figure 1, illustrates that the AI produced responses that were partially human-like.



Human: Men are all alike.

Eliza: In what way?

Human: They're always bugging us about something specific or other.

Eliza: Can you think of a specific example?

Human: Well, my boyfriend made me come here.

Eliza: Your boyfriend made you come here?

Human: He says I'm depressed much of the time.

Eliza: I am sorry to hear you are depressed.

Human: It's true. I'm unhappy.

Eliza: Do you think coming here will help you not to be unhappy?

Human: I need some help, that much seems certain.

Eliza: What would it mean to you if you got some help?

Human: Perhaps I could learn to get along with my mother.

Eliza: Tell me more about your family.

...

Figure 1. An Example of a Conversation with ELIZA (Weizenbaum, 1966)

In the history of NLP, the identification of systems and programs has often been accompanied by many unrealistic assumptions. Considering that existing approaches to requirements engineering are critical and costly, the possibility of creating a system capable of automatically understanding end-user needs appears highly attractive. Many projects propose developing systems through natural or near-natural conversational methods with customers (Ryan, 1993).

By the 2020s, the results of Alan Turing's work began to manifest. Initially, computers were not proficient at making predictions, and creating extensive dictionaries held little significance for them. However, human languages are inherently chaotic, and consecutive expressions often carry multiple meanings. Elements such as context and humor are generally extremely difficult for our devices to interpret (Eloundou et al., 2023).

2.1. The Use of Rule-Based Methods

Rule-based methods have played an important role in the development of LLMs and the field of NLP. These methods were particularly used in early NLP studies to understand the structure of language and enable computers to process text.

To better understand rule-based methods, the topic of sentiment analysis can be considered. Sentiment analysis, also referred to as opinion mining, has emerged as an important research domain within NLP and AI. This field primarily aims to identify and extract subjective information from text to reveal the emotions conveyed by individuals or groups. As digital communication platforms continue to grow rapidly, the need for robust sentiment analysis tools to examine and make sense of the immense volume of user-generated data has increased. Traditionally, sentiment analysis was conducted using manually crafted rules and lexicons; today, it is supported by machine learning and deep learning methods, offering more detailed and scalable solutions (Gupta et al., 2024).

Rule-based systems date back to the early 2000s. The initial systems relied on handcrafted lexicons and patterns matching-i.e., rules-to detect emotions. As machine learning advanced, sentiment analysis techniques began to rely increasingly on statistical models and feature-driven classifiers (Gupta et al., 2024). Rule-based approaches refer to the use of manually created lexicons during periods when technology was insufficient and computers were unable to extract the desired information from text.

2.2. Statistical Approaches in Language Modeling

The limitations of rule-based methods necessitated a transition to more advanced approaches in language modeling. This shift occurred particularly with the widespread adoption of statistical methods in the field of language processing. Unlike rule-based systems, statistical approaches offered a data-driven and flexible methodology, enabling a better modeling of the complexity of language. Therefore, the move from rule-based methods to statistical approaches marked a significant turning point in NLP.

As nearly all aspects of society have become digitized, data analysis has emerged as an indispensable tool across various industries. For example, financial institutions use data analysis to make informed decisions about stock trends, hospitals monitor patients' health conditions, and companies develop strategic plans through data-driven insights (Sun et al., 2024).

The general data analysis workflow usually consists of several essential steps. First, data are collected from studies or extracted from databases and imported into tools such as Excel. Next, software like Excel or programming languages such as Python and R are employed to clean and analyze the data in order to derive meaningful insights. For more advanced tasks, including statistical inference and predictive analysis, statistical methods and machine learning models are typically applied. This process generally encompasses data preprocessing, feature engineering, modeling, evaluation, and other related steps (Sun et al., 2024).

However, statistical approaches have a notable disadvantage: a systematic lack of statistical training. As a result, individuals without a background in statistics may struggle to determine which types of analyses are appropriate, even when data are provided. As data and models grow in complexity, developing a thorough understanding of established statistical methods typically necessitates graduate-level training in statistics (Sun et al., 2024).

2.3. Deep Learning

Despite the success of statistical approaches, the pursuit of more complex and flexible solutions in language modeling has led to the development of deep learning techniques. Deep learning aims to better model the intricate structure of language using feature-based neural networks. A standard neural network consists of many simple, connected processors called neurons, each producing a series of real-valued activations. Input neurons are activated through sensors that perceive the environment, while other neurons are activated via weighted connections from previously active neurons. Depending on the problem and the network's connectivity, such behaviors may require long causal chains of computation. At each stage, the network transforms the total learning activation (Schmidhuber, 2015).

2.4. The Word2Vec and GloVe Techniques

With the widespread adoption of deep learning in language modeling, word embedding techniques such as Word2Vec and GloVe have gained importance. These techniques address challenges in text processing or NLP, such as feature extraction from unstructured text. Feature extraction plays a critical role in text classification by converting text into a structured form that learning algorithms can process. The chosen feature extraction technique directly affects classification performance, which has led to extensive research aimed at improving performance in this area. Transforming text into a vector space representation using a term frequency matrix is a commonly used method in NLP. This approach compresses unstructured text into a more organized and analyzable form (Dharma et al., 2022).

2.4.1. The Word2Vec Technique

The word embedding technique represents each word as a point in space by converting alphanumeric characters into vectors. This approach positions words with similar contexts or meanings close to each other in the vector space, thereby capturing both the semantic and syntactic aspects of words.

In 2013, a Google team led by Tomas Mikolov developed and published the Word2Vec technique for word embedding. This technique consists of two models (Dharma et al., 2022):

- Skip-gram
- CBOW (Continious Bag of Word)

The Skip-gram model aims to predict context words based on a given input word. Its core idea is to estimate the context of a word when the word itself is provided. This approach, based on word embeddings, is an enhanced version of the N-gram model, which attempts to understand context by skipping words at certain intervals rather than using consecutive words (Sonkar et al., 2020).

At the foundation of all embedding models lies the idea that "a word is defined by the words around it" (Firth, 1957). For example, Word2Vec's CBOW model attempts to predict a randomly selected word by using the other words in a sentence as context. These models generally treat context words with equal weight. However, it is evident that some context words are more influential than others in predicting the masked word and therefore should be given greater weight (Sonkar et al., 2020).

2.4.2 The GloVe Technique

GloVe is a technique that combines two different approaches: count-based methods (e.g., Principal Component Analysis) and direct prediction methods like Word2Vec. While Word2Vec relies solely on information from local context windows, the GloVe algorithm also incorporates word co-occurrence information and global statistics to capture semantic relationships between words. GloVe employs a global matrix factorization method that represents the presence or absence of words in a corpus.

Word2Vec is a feedforward neural network model. Therefore, Word2Vec is often referred to as a "neural word embedding," whereas GloVe is a log-bilinear model and is generally classified as a "count-based model". GloVe learns relationships between words by analyzing how frequently words co-occur within a corpus. This analysis, which leverages word occurrence ratios, enhances performance in tasks such as generating meaningful embeddings and solving word analogies (Dharma et al., 2022).

3. Fundamental Features of LLMs

The fundamental features of LLMs constitute the distinguishing elements that set them apart from other AI and language processing techniques. These features directly influence the model's design, learning processes, and performance. The key features of LLMs can be summarized as follows:

- Transformer Architecture
- Encoder-Decoder Structure
- Self-Attention Mechanism
- Data and Training Process
- Scale of Parameters
- Generalization Capability
- Token Structure

One of the important hyperparameters used during the training of LLMs is the batch size. Batch size refers to the number of data samples processed by the model in each training step. Its significance in the training process is particularly critical for efficiently utilizing memory and computational resources, especially in large-scale models (Goyal et al., 2017).

3.1. Transformer Architecture

In recent years, the development of transformer-based models such as BERT, GPT, and their variants has driven significant advances in the field of NLP. These models have achieved notable success in challenging tasks such as understanding and generating human language. They have particularly revolutionized tasks in NLU (Natural Language Understanding) and NLG (Natural Language Generation), including sentiment analysis and document summarization. Additionally, transformers have proven effective in other domains, such as computer vision and autonomous driving (Huang et al., 2023).

The main reasons why transformer architecture is so effective in the field of NLP are as follows (Vaswani et al., 2017):

- Parallel Processing: Unlike traditional RNNs (Recurrent Neural Networks), transformers process data in parallel rather than sequentially, speeding up training and inference.
- Learning Long-Term Dependencies: The self-attention mechanism excels at capturing relationships between words in long sentences.

- Scalability: The encoder-decoder structure can be adapted for a wide range of tasks, from machine translation to language modeling.

The transformer architecture and the Multi-Head Attention layer, which is responsible for understanding input texts, are illustrated in Figure 2. Within the Multi-Head Attention layer, the Scaled Dot-Product Attention layer is included. This layer is a fundamental component of the self-attention mechanism and helps extract contextual meaning by determining the relationships between elements in the input sequence.

Scaled Dot-Product Attention is divided into three main components:

- Q (Query): Query vectors
- K (Key): Key vectors
- V (Value): Value vectors

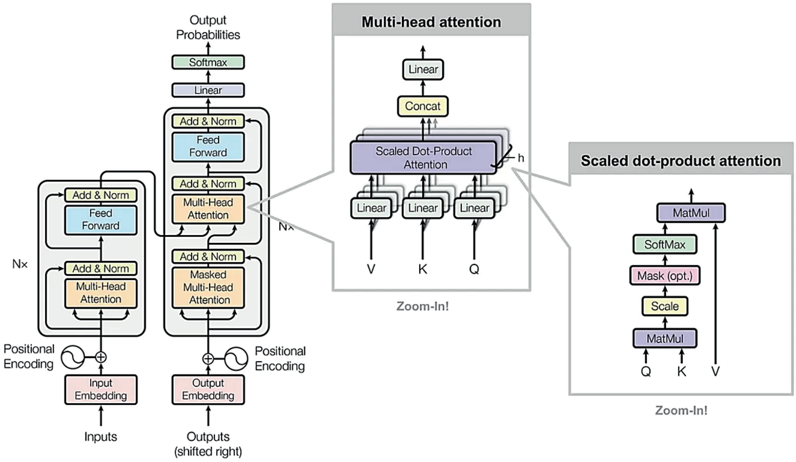


Figure 2. Transformer Model Architecture and Multi-Head Attention Layer Components (Vaswani et al., 2017)

3.2. Encoder-Decoder Structure

The transformer architecture consists of two main components: the encoder and the decoder. The encoder learns the context between words using Multi-Head Attention (MHA) and feedforward network layers to understand the input text. The decoder generates the target sequence by employing both MHA and cross-attention mechanisms that process information from the encoder. Attention masks in the decoder ensure sequential processing. The interaction between the encoder and decoder forms the foundation of transformers' success in text understanding and generation (Huang et al., 2023).

The primary difference between the encoder and decoder is that the decoder contains an additional cross-attention layer. This layer interacts with the outputs from the encoder to generate the target sequence more effectively (Huang et al., 2023).

3.3. Self-Attention Mechanism

Self-attention evaluates the relationships of each element in a sequence with all other elements, providing a significant advantage in determining the context of words. For example, it can correctly identify whether the word "bank" in a sentence refers to a financial institution, a fog bank, or the bank of a river by examining the surrounding words (Vaswani et al., 2017).

The self-attention mechanism plays a crucial role in modeling long-term dependencies and language relationships by effectively learning contextual connections in areas such as language modeling and speech recognition. As a core component of the transformer architecture, it offers faster and more parallel processing capabilities compared to traditional LSTM (Long Short-Term Memory) models. Self-attention determines the importance of each element in a sequence relative to others while still requiring information about the order of elements. Therefore, positional encodings are typically used in transformer models.

This mechanism provides an effective solution in speech recognition and language modeling by reducing WER (Word Error Rate) and improving overall performance. As a result, self-attention, with its ability to extract contextual information and learn long-term dependencies, occupies a central role in modern NLP applications (Irie et al., 2019).

Table 1 presents the maximum path lengths, per-layer complexity, and minimum sequential operations for different types of layers. The computational complexity and sequential processing requirements of the self-attention mechanism show significant differences compared to other model types.

Table 1. Comparison of Self-Attention and Other Layer Types (Vaswani et al., 2017)

Layer Type	Per-Layer Complexity	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 . d)$	$O(1)$	$O(1)$
Recurrent	$O(n . d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k . n . d^2)$	$O(1)$	$O(\log_k(n))$
Self-Attention (restricted)	$O(r . n . d)$	$O(1)$	$O(n/r)$

3.4. Data and Training Process

LLMs are developed to deliver human-level performance in language understanding and generation (Sadıkoğlu et al., 2023). This process encompasses several subtopics, including the training of LLMs. Key aspects to consider include large and diverse datasets, the high computational power required for training, and the roles of distributed systems and GPUs.

The datasets used for LLMs play a critical role in their success. LLMs acquire cross-linguistic understanding and contextual knowledge primarily from the datasets on which they are trained. These datasets serve as the foundational infrastructure supporting model development and directly influence their language comprehension and generation capabilities.

LLM datasets are classified into five main categories based on their intended use (Liu et al., 2024):

1. Pretraining Corpora: Large-scale text collections used to acquire general language knowledge.
2. Instruction Fine-Tuning Datasets: Used to customize models for improved performance on specific tasks.
3. Preference Datasets: Designed to generate results that better align with user preferences.
4. Evaluation Datasets: Created to measure model accuracy, effectiveness, and overall performance.
5. Traditional NLP Datasets: Classic datasets commonly used in the field of NLP.

For LLM training, access to large datasets is important, but the quality and diversity of these datasets are equally critical. Research has highlighted the challenges associated with these datasets and suggested potential directions for future studies (Liu et al., 2024).

LLMs are developed through training on various datasets at different stages. This process is typically divided into four main categories-pretraining, instruction fine-tuning, preference datasets, and evaluation datasets-to enhance the model's general capabilities, optimize it for specific tasks, and improve the end-user experience (Liu et al., 2024).

Datasets used in the pretraining phase include large-scale internet texts, academic articles, books, and code repositories. Instruction fine-tuning datasets consist of human-created instruction-response pairs, designed to improve the model's performance on specific tasks. Preference datasets are developed to ensure that the model's outputs align with human preferences. Evaluation datasets are used to measure LLM performance, testing the model's accuracy, reliability, and ethical behavior.

Figure 3 illustrates the datasets used in the development process of LLMs, categorized along a timeline. Pretraining, instruction fine-tuning, preference, and evaluation datasets are shown in different colors, clearly highlighting the role of each dataset at various stages of model development.

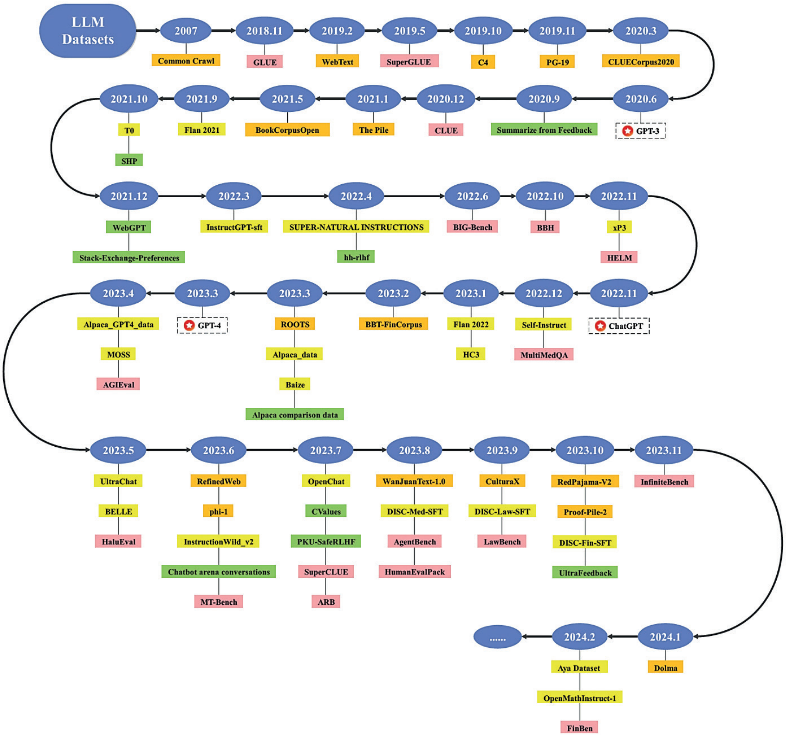


Figure 3. Timeline of LLM Datasets. Orange: Pretraining, Yellow: Instruction Fine-Tuning, Green: Preference Datasets, Pink: Evaluation Datasets (Liu et al., 2024)

In conclusion, large and diverse datasets play a vital role in enhancing the language learning capacity of LLMs. These datasets improve the models' ability to understand and generate both grammar and context more effectively.

Training LLMs require an extremely high memory capacity. To address this challenge, current approaches often use a combination of CPUs and GPUs during training, such as ZeRO-Offload (Yang et al., 2024). While ZeRO-Offload makes model training more accessible, it is generally inefficient in memory management and requires expert intervention.

The demand for high computational power during training is closely tied to the infrastructure and connectivity technologies of GPU-centric clusters. This involves short-range communication protocols like NVLink (NVIDIA Link) and long-range protocols such as RDMA (Remote Direct Memory Access)-enabled NICs (Network Interface Cards). However, large-scale RDMA networks face issues such as lockups, performance degradation, and high costs. Scaling challenges in Clos network architectures and over-subscription strategies used by data center providers to reduce costs further exacerbate performance losses. Therefore, in addition to powerful hardware, optimized network architectures and protocols are critical for training large-scale models (Wang et al., 2023).

The intensive computational power required for training and inference of LLMs makes GPUs a fundamental component of their computing resources. Early GPU programming and execution models laid the groundwork for modern GPU architecture. These cores are optimized to accelerate complex matrix operations during LLM training. Additionally, the memory systems of modern GPUs are specifically designed to efficiently perform tensor operations.

However, due to the heterogeneous nature of GPU architecture, users must be prepared to handle hardware diversity. This requires developing software that can adapt to different GPU designs and optimizing systems accordingly. The variety and advanced capabilities of GPUs play a critical role in distributed systems, enabling the efficient training and deployment of large-scale models (Zhang and Zijan, 2024).

3.5. Scale of Parameters

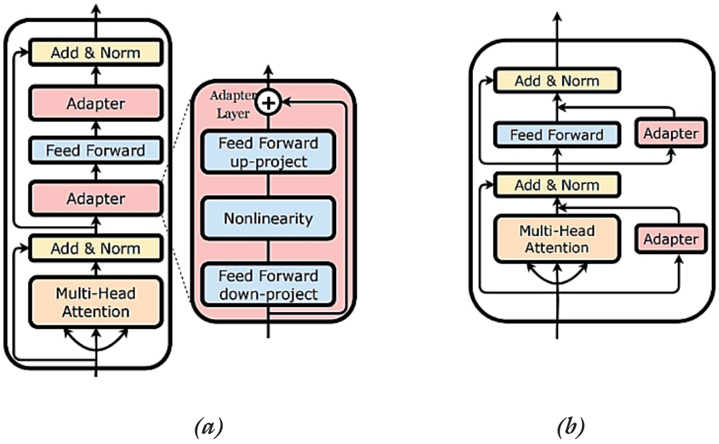
Increasing the number of parameters in LLM development has a significant impact on performance. A larger number of parameters enable models to achieve superior results in language understanding, reasoning, and various tasks. However, this increase also demands high costs and computational resources during training and fine-tuning. To address these challenges, adapter-based PEFT (Parameter-Efficient Fine-Tuning) methods have been developed (Hu et al., 2023).

PEFT reduces cost and resource requirements by optimizing only a small set of external parameters instead of retraining the entire large model. This approach not only offers low-cost training but also provides strong performance even in smaller-scale models. In fact, it can achieve results comparable to models with 175 billion parameters and, in some cases, even surpass them (Hu et al., 2023).

LLMs utilize adapter-based PEFT methods, which enable pretrained models to be applied more effectively and efficiently across different tasks. Examples of adapter types include (Hu et al., 2023):

- Sequential Adapter: These adapters incorporate learnable modules in a sequential manner within a specific sublayer.
- Parallel Adapter: These adapters aim to include learnable modules in parallel across different sublayers of the backbone model.

Figure 4 visualizes adapter-based PEFT methods in LLMs. The figure highlights architectural differences between sequential and parallel adapters, comparing how adapters are integrated within the backbone model. Sequential adapters strengthen the learning process through sequentially added modules within specific layers, while parallel adapters operate concurrently with different sublayers, providing a more flexible and efficient fine-tuning mechanism.



(a) *(b)*

Figure 4. Detailed Adapter Architecture: (a) Sequential Adapter, (b) Parallel Adapter (Hu et al., 2023)

In conclusion, increasing the number of parameters enhances the performance of large models, while methods such as PEFT allow these advantages to be achieved more efficiently. These approaches provide more sustainable solutions in terms of cost and resource usage, making LLMs more accessible (Hu et al., 2023).

3.6. Generalization Capability

LLMs are highly powerful due to their zero-shot and few-shot learning abilities. By leveraging these methods, LLMs can demonstrate general cognitive capabilities without requiring task-specific training data. Techniques such as "chain-of-thought reasoning" allow models to achieve high performance even in complex logical reasoning and multi-step tasks with zero-shot learning (Kojima et al., 2022).

Zero-shot learning is an approach in which a model completes a task using only a prompt or instruction, without encountering any task-specific training examples. This means the model can make predictions directly on tasks it has never seen during training, without additional examples or fine-tuning. Such learning evaluates the model's ability to understand and solve diverse tasks based on the general knowledge it acquired during training (Kojima et al., 2022).

Few-shot learning refers to a model's ability to learn a task from only a few examples. In this approach, the model shows how to perform the task with just a limited number of examples, without requiring extensive data or a large training set. This enables the model to achieve high accuracy on a specific task while being trained on only a small number of examples (Kojima et al., 2022).

In conclusion, zero-shot and few-shot learning are crucial methods that reveal the capabilities and broad cognitive potential of LLMs. These approaches make LLMs more efficient, flexible, and applicable across a wide range of tasks.

3.7. Token Structure

This section discusses tokens, one of the most critical elements of LLMs, highlighting their advantages and disadvantages. Tokens affect factors such as cost and model performance. The computational cost of LLMs generally depends on the number of input and output tokens. In commercial applications, adding extra text to expand context increases the token count, thereby raising costs. A higher number of tokens not only increases cost but also extends processing time.

Token compression techniques optimize the balance between cost and performance, enabling efficient use in retrieval-augmented models. These techniques include:

- Summarization Compression: Summarization models can be used to allow LLMs to learn the same information from shorter inputs.
- Semantic Compression: By removing non-essential words, the model can operate with shorter inputs that retain the same meaning.

Token compression combined with effective context management increases usability in information-dense tasks while maintaining model accuracy. To implement this approach, certain rules are followed:

- LLMs operate with a defined maximum token limit.
- When working with long contexts, the model must be optimized using early truncation or sliding strategies to prevent information loss.

In conclusion, token efficiency is a critical factor that directly affects the cost, speed, and accuracy of LLMs. Developers can enhance model effectiveness by employing techniques such as summarization, semantic compression, and context management.

While the use of tokens offers many advantages for LLMs, it also has some drawbacks. LLMs operate with a defined maximum token limit. When processing long documents or very large contexts, the model may perform early truncation, resulting in information loss.

The costs of LLMs increase with the number of tokens processed. As the number of input and output tokens grows, computation time and expenses rise. Large token counts can particularly increase commercial costs when using APIs.

The tokenization process, which splits text into tokens, can sometimes divide words into meaningless or incomplete segments. This can lead to loss of meaning, especially in low-resource languages or in words with complex structures. In agglutinative languages such as Turkish, incorrect tokenization can cause semantic distortions.

Token-based systems cannot always capture subtle nuances in text. For example, irony, metaphors, or cultural contexts may be misinterpreted by the model. During summarization, some important information may be overlooked, and semantic compression can lead to information loss and potential misinterpretations.

Token-based processing is one of the fundamental elements that make LLMs operational. However, it also brings challenges such as length limitations, cost, information loss, and context management. To mitigate these disadvantages, techniques like token compression, context management, and more advanced memory mechanisms have been developed.

4. Development and Future Perspective of LLMs

LLMs have revolutionized the field of NLP, performing a wide range of tasks such as human-like text generation, translation, code writing, and information synthesis. Prominent models in this domain include ChatGPT, LLaMA, Gemini, and DeepSeek, each developed using technologies with similar or distinct features.

ChatGPT, developed by OpenAI, is an extension of the GPT (Generative Pretrained Transformer) series. First released in 2022, it demonstrated significant advancements with GPT-3.5 and GPT-4, attracting widespread attention for its ability to engage in human-like dialogue. With the release of GPT-4 in 2023, features such as enhanced contextual understanding, multimodal processing (text and visual input), and an extended context window were introduced. In particular, the GPT-4 Turbo version offers faster, optimized responses at lower costs, reflecting OpenAI's goal of improving model performance.

LLaMA, developed by Meta, was designed to meet the growing demand for open-source LLMs. LLaMA 1 (2023) aimed to deliver high performance with a smaller number of parameters, while LLaMA 2 improved scalability and efficiency. The LLaMA 3 version, released in 2024, is reported to be trained on larger datasets and to possess significantly enhanced contextual understanding capabilities. One of the main advantages of LLaMA models is their open-source nature, providing a suitable infrastructure for academic research and enterprise customization.

Google's earlier language model, Bard, was rebranded as Gemini in 2023 and developed by Google DeepMind. Gemini 1.0 (2023) introduced multimodal capabilities, handling both visual and text inputs, differentiating it from competitors. The Gemini 1.5 (2024) version features an expanded context window, enabling better comprehension of long-term dependencies. Gemini's most notable distinction is its advanced ability to interpret visual content compared to other models. Google has optimized this model particularly for search engines, AI-powered assistants, and content generation.

DeepSeek was founded in 2023 in Hangzhou, China, by information and electronics engineer Liang Wenfeng. Liang had previously supported AI-focused projects through the incubator program of the High-Flyer fund, which he established in 2015. The company's vision is to achieve AGI (Artificial General Intelligence) capable of matching or surpassing human performance across various domains.

DeepSeek's first model, DeepSeek Coder, was launched in November 2023. The model has gained attention as a language model particularly focused on code generation and technical documentation. Sub-projects like DeepSeek Coder aim to facilitate large-scale code production. Compared to other models, DeepSeek's primary advantage is its superior performance in mathematical computations and code-based tasks. The general comparisons of the above mentioned LLMs are given in Table 2.

Table 2. General Comparison of LLMs

Model	Developer	Year	Open Source	Key Strengths of the Model
ChatGPT	OpenAI	2022	No	Dialogue-based interaction, large dataset
LLaMA	Meta	2023	Yes	Open source, low hardware requirements
Gemini	Google	2023	No	Multimodal learning, advanced visual-text integration
DeepSeek	DeepSeek AI	2023	No	Code generation and technical computations

LLMs have recently gained prominence in machine learning research due to their rapid advancements, driving transformative changes across various fields such as NLP, biomedical analysis, software development, and content creation. LLMs like ChatGPT, Gemini, LLaMA, and DeepSeek now span a wide range of applications, from conversational chatbots for user interaction to analytical tools that assist in scientific research. Current studies focus on enhancing these models' performance and exploring their extensive application potential across diverse domains. The success of advanced general-purpose LLMs relies on two key factors:

1. Developing a robust model architecture with a large set of parameters,
2. Training this architecture on vast and comprehensive datasets.

For instance, OpenAI's GPT-4 Turbo leverages millions of parameters to utilize an extensive knowledge base, while Google DeepMind's Gemini 1.5

model demonstrates significant advancements in multimodal capabilities by integrating visual and textual data. Meta's LLaMA 3 provides optimized lightweight models for the open-source community, and DeepSeek stands out in technical domains such as programming and code generation.

Current applications of LLMs span several key domains, including healthcare and biomedical research, law and finance, education and academic research, as well as content creation and media. These examples illustrate how LLMs are transforming both professional and creative workflows across diverse sectors:

- Healthcare and Biomedical: LLMs are used to analyze medical reports, support clinical diagnoses, and review literature in biomedical research. For example, Med-PaLM 2 assists doctors in making more accurate diagnostic and treatment decisions.
- Law and Finance: LLMs help analyze legal documents, accelerate legal research, and make financial predictions. Harvey AI is employed by law firms to optimize legal analysis processes.
- Education and Academic Research: LLMs are effective in automated assignment grading, academic writing support, and summarizing scientific papers. Elicit AI aids researchers in accelerating literature reviews.
- Content Creation and Media: LLMs are actively used for news summarization, generating marketing copy, and creative writing. ChatGPT and Gemini support users in creative processes, serving a broad range of applications in media and content production.

Among the potential future applications of LLMs are the broader integration of advanced AI assistants into individuals' daily lives, the development of autonomous systems with more intelligent and independent decision-making capabilities, and the utilization of personal AI assistants to organize daily activities, manage health data, and guide educational processes. Furthermore, within the scope of bio-artificial intelligence integration, these models are expected to contribute to more advanced predictive processes in personalized medicine by combining genetic analysis with biotechnology.

The large scale of LLM models and their datasets will result in significant computational costs during the pre-training phase. Therefore, there is an increasing need for innovative architectures and algorithms that reduce pre-training costs and optimize data usage to enable the more sustainable development and widespread deployment of LLMs in the future.

In this context, techniques such as adapter-based PEFT methods help make large-scale models more accessible by reducing their associated costs. Simultaneously, the development of smaller and more customizable models will facilitate broader adoption of LLMs, particularly for corporate organizations and individual users.

4.1. Multimodal Models

Although MM-LLMs (MultiModal Large Language Models) have made significant progress today, most existing systems can only provide multimodal understanding on the input side. However, some models possess the ability to generate content across different modalities (text, image, video, and audio). For instance, ChatGPT's GPT-4 Turbo model can understand text- and image-based inputs, while the Gemini 1.5 model can generate content with advanced visual and text integration. DeepSeek excels particularly in code generation, whereas LLaMA models incorporate developments for multimodal learning within the open-source ecosystem. While humans use these modalities together to understand their environment and communicate, MM-LLMs capable of receiving and generating content in any modality play a critical role in advancing human-level AI development (Wu et al., 2023).

Research efforts in this area continue. One such example is a system called Next-GPT (Wu et al., 2023). A notable feature of Next-GPT is its utilization of existing high-performance encoders and decoders. This allows the system to be fine-tuned with only 1% additional parameters while offering a low-cost training process, making it easier to integrate new modalities into the system. The architecture of the proposed system is illustrated in Figure 5.

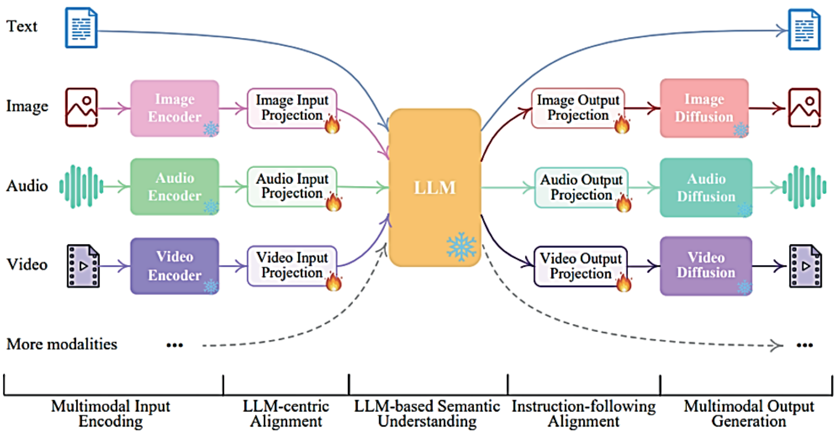


Figure 5. Proposed Next-GPT Model Architecture for MM-LLMs (Wu et al., 2023)

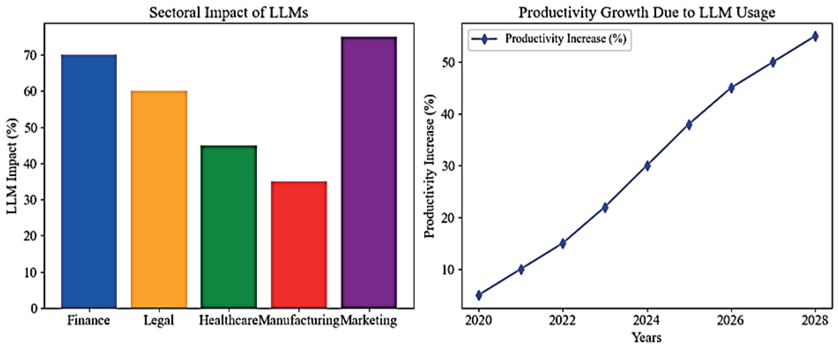
4.2. Changes Brought by LLMs in the Labor Market

LLMs, particularly technologies such as GPT, have significant transformative potential in the business world. Research indicates that these models could impact a substantial portion of the workforce in the United States. For instance, a certain percentage of employees' tasks and up to half of some workers' duties are expected to be affected by these models (Eloundou et al., 2023).

The influence of LLMs is not limited to low-skilled jobs but is also likely to be noticeable in high-income and more complex professions. This can lead to profound changes in productivity and business processes. The ability of LLMs to perform work tasks faster while maintaining the same quality can enhance productivity, generating broader economic benefits. In summary, LLMs like GPT can be regarded as general-purpose technologies, capable of creating large-scale impacts on the economy, society, and the business environment. These impacts include altering workforce dynamics, increasing efficiency, promoting inclusivity, and revolutionizing education and learning processes. The advancement of these models carries the potential to redefine work processes and fundamentally transform labor market dynamics (Eloundou et al., 2023).

Based on data from firms such as McKinsey, the World Economic Forum, PricewaterhouseCoopers, Harvard Business Review, Accenture, and BCG, an estimated sector-wise impact of LLMs and productivity increases is illustrated in Figure 6. The figure presents two different charts related to LLM effects (Schmidhuber, 2015; Schmidt, 2023; Shukla, 2024; Mayer et al., 2025; World Economic Forum, 2025).

As shown in Figure 6(a), the sector most affected by LLMs is marketing, with an estimated impact of 75%. The least affected sector is manufacturing, with a projected impact of 40%. These estimates are shaped by the level of process automation within the respective sectors and the potential use of NLP models. Figure 6(b) illustrates the effect of LLM adoption on productivity growth. Since 2020, with the integration of LLMs into business processes, a continuous increase in productivity has been observed. This increase, which was around 5% in 2020, is expected to reach 50% by 2028. This trend highlights the positive impact of AI-supported automation on workforce productivity.



(a) *(b)*

Figure 6. LLM Impacts: (a) Sector-wise Effects of LLM (Schmidhuber, 2015; Mayer et al., 2025; World Economic Forum, 2025), (b) Sector-wise Productivity Growth Associated with LLM Adoption (Schmidt, 2023; Shukla, 2024)

4.3. Broader Accessibility

Although it is difficult to fully predict the future applications of LLMs, current studies and the rapid pace of technological advancement provide some insights. At present, LLMs have proven effective in supporting content creation processes, such as social media post generation (Tanudin, 2024), and in assisting researchers by providing rapid access to information in academic studies (Rane et al., 2023). Furthermore, in the coming years, LLMs are expected to play a role in more complex processes, including the development of personalized educational platforms, enhancing customer service efficiency, and supporting diagnostics and patient education in healthcare. However, to fully understand the true potential of these technologies and make precise predictions, further research and practical implementations are required.

Conclusion

This study examined LLMs and advancements in the field of NLP. Initially, the historical development of LLMs and the evolution of language processing techniques from the 1950s to the present were explored. In particular, the progression from word embedding techniques such as Word2Vec and GloVe to transformer-based models like BERT and GPT was detailed. The study emphasized how these models, trained on large datasets with GPU infrastructures requiring high computational power, enhance language knowledge and contextual understanding.

The fundamental characteristics of LLMs and their capabilities in understanding and generating natural language were discussed. Technical details such as the transformer architecture, self-attention mechanism, and encoder-decoder structure were explained. Additionally, the broad applications of LLMs-including text generation, question-answering systems, machine translation, and programming assistance-were highlighted. A comparative analysis of popular models such as ChatGPT, LLaMA, Gemini, and DeepSeek were also presented.

The study addressed the training processes of LLMs and the critical role of large datasets used during these processes. Techniques that optimize cost and resource usage, such as adapter-based PEFT methods, were explained in terms of their contribution to making LLMs more accessible. Moreover, the impact of token structures and token compression techniques on model performance and operational costs was examined.

The potential future applications of LLMs and their transformative impact on sectors such as healthcare, law, finance, education, and content creation were discussed. The development of multimodal models and their integration of diverse data types-including visual, textual, audio, and video-was highlighted as a means to expand their range of applications. Additionally, the effects of LLMs on the labor market and their potential to enhance productivity were evaluated.

In forward-looking assessments, it is anticipated that the challenges encountered in the development of LLMs will not be limited solely to technical constraints (such as computational costs, data quality, and model scalability). During the integration of these models into organizational structures, multidimensional managerial issues that fall within the scope of Management Information Systems-such as compatibility with existing information systems, the redesign of organizational processes, user acceptance, data governance, and ethical responsibilities-are expected to emerge. In particular, the integration of LLM outputs into decision support systems necessitates the redefinition of human-machine interaction and the reassessment of managerial control mechanisms. In this context, the success of LLMs will depend not only on algorithmic advancements but also on the extent to which these technologies can be effectively positioned at the strategic, managerial, and organizational levels from an MIS (Management Information Systems) perspective.

In conclusion, LLMs have driven revolutionary changes in the field of NLP, and significant advancements are anticipated in the future. However, it is important to consider challenges such as high computational costs and the quality and diversity of datasets in the development and deployment of these models. Consequently, innovative solutions such as adapter-based PEFT methods and more efficient algorithms are expected to contribute to the sustainable development of LLMs in the future.

Managerial Implications of LLMs

The transformer architecture and self-attention mechanism enable LLMs to analyze long and complex texts while preserving contextual coherence. From a business perspective, this capability allows strategic decision support systems to produce more accurate and consistent outputs. Particularly in the analysis of large volumes of qualitative data-such as customer feedback, call center records, and social media data-this architectural structure provides managers with deeper insights.

The token-based structure directly affects the cost and performance dimensions of LLM usage. Especially in LLM services offered via APIs, the number of input and output tokens has become a critical factor determining firms' monthly operational costs, making it necessary for organizations to consider scalability and cost-benefit trade-offs in their artificial intelligence investments. This situation increases the importance of model selection, context management, and data summarization strategies for MIS managers.

Owing to the contextual awareness provided by the self-attention mechanism, organizations can holistically analyze heterogeneous data originating from different departments, thereby strengthening cross-functional decision-making processes. Moreover, the general-purpose nature of transformer-based LLMs allows a single model to be utilized across multiple functions within the MIS domain, including reporting, knowledge management, customer relationship management, and operational analytics.

Finally, approaches such as token efficiency and PEFT enable large-scale artificial intelligence solutions to be adopted not only by large technology firms but also by small and medium-sized enterprises; this contributes to the development of a more inclusive and sustainable structure in digital transformation processes from an MIS perspective.

Strategic Recommendations

Managers and decision-makers planning to adopt LLMs should consider these technologies not merely as tools for enhancing operational efficiency, but as core components of a long-term digital transformation strategy. First, LLM use cases should be clearly defined and aligned with measurable business objectives in areas such as customer relationship management, decision support systems, knowledge management, and process automation.

During the model selection process, it is critical to conduct a comprehensive comparison between cloud-based API solutions and on-premises deployments in terms of cost, data security, and scalability. Taking token-based pricing models and increasing processing volumes into account, strategies for context management and data summarization should be developed.

Moreover, the adoption of adapter-based PEFT approaches enables the integration of LLMs into organizational processes without requiring substantial hardware investments, thereby reducing the total cost of ownership. From a human resources perspective, the formation of multidisciplinary teams that include not only technical staff but also business units will support the accurate interpretation and effective use of LLM outputs.

Finally, issues such as data quality, ethical use, transparency, and legal compliance should be established as integral components of the LLM strategy. When these factors are considered collectively, the informed and strategic adoption of LLMs can be regarded as a significant MIS investment that provides organizations with a sustainable competitive advantage.

References

Cambria E., White B. Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine 2014; 9(2): 48-57. https://doi.org/10.1109/MCI.2014.2307227

Ciesla R. *The book of chatbots: From ELIZA to ChatGPT.* Springer 2024. https://doi.org/10.1007/978-3-031-51004-5

Dharma EM., Gaol FL., Warnars HLHS., Soewito B. The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. Journal of Theoretical and Applied Information Technology 2022; 100(2): 349-359.

Eloundou T., Manning S., Mishkin P., Rock D. GPTs are GPTs: An early look at the labor market impact potential of large language models 2023. https://doi.org/10.48550/arXiv.2303.10130

Firth JR. A Synopsis of Linguistic Theory, 1930-1955. In Studies in Linguistic Analysis. Special Volume of the Philological Society, Oxford 1957; 1-32.

Fournet A. Michel Bréal (1832-1915), A forgotten precursor of enunciation and subjectivity. ReVel 2011; 9(16): 201-213.

Goyal P., Dollár P., Girshick R., Noordhuis P., Wesolowski L., Kyrola A., Tulloch A., Jia Y., He K. Accurate, large minibatch SGD: Training imagenet in 1 hour 2017. https://doi.org/10.48550/arXiv.1706.02677

Gupta S., Ranjan R., Singh SN. Comprehensive study on sentiment analysis: From rule based to modern LLM based system 2024. https://doi.org/10.48550/arXiv.2409.09989

Hu Z., Wang L., Lan Y., Xu W., Lim EP., Bing L., Xu X., Poria S., Lee RKW. LLM-Adapters: An adapter family for parameter-efficient fine-tuning of large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 2023; 5254-5276. https://doi.org/10.18653/v1/2023.emnlp-main.319

Huang Y., Xu J., Lai J., Jiang Z., Chen T., Li Z., Yao Y., Ma X., Yang L., Chen H., Li S., Zhao P. Advancing transformer architecture in long-context large language models: A comprehensive survey 2023. https://doi.org/10.48550/arXiv.2311.12351

Irie K., Zeyer A., Schlüter R., Ney H. Language modeling with deep transformers. Proceedings of the 20th Annual Conference of the International Speech Communication Association 2019; 3905-3909. https://doi.org/10.21437/Interspeech.2019-2225

Kojima T., Gu SS., Reid M., Matsuo Y., Iwasawa Y. Large language models are zero-shot reasoners 2022. https://doi.org/10.48550/arXiv.2205.11916

Liu J., Li L., Xiang T., Wang B., Qian Y. TCRA-LLM: Token compression retrieval augmented large language model for inference cost reduction. Findings of the Association for Computational Linguistics: EMNLP 2023; 9796-9810. https://doi.org/10.18653/v1/2023.findings-emnlp.655

Liu Y., Cao J., Liu C., Ding K., Jin L. Datasets for large language models: A Comprehensive survey 2024. https://doi.org/10.48550/arXiv.2402.18041

Mayer H., Yee L., Chui M., Roberts R. Superagency in the workplace: Empowering people to unlock AI's full potential 2025. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

Rane NL., Tawde A., Choudhary SP., Rane J. Contribution and performance of ChatGPT and other Large Language Models (LLM) for scientific and research advancements: A double-edged sword. International Research Journal of Modernization in Engineering Technology and Science 2023; 5(10): 875-899. https://doi.org/10.56726/irjmets45213

Ryan K. The role of natural language in requirements engineering. Proceedings of the IEEE International Symposium on Requirements Engineering 1993, 240-242. https://doi.org/10.1109/ISRE.1993.324852

Sadıkoğlu E., Gök M., Mıjwıl MM., Kösesoy İ. The evolution and impact of large language model chatbots in social media: A comprehensive review of past, present, and future applications. Veri Bilimi 2023; 6(2): 67-76. https://dergipark.org.tr/en/pub/veri/issue/81532/1400734

Saussure FD. Cours de linguistique générale. Payot 1916.

Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks 2015; 61: 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

Schmidt E. Eric Schmidt: This is how AI will transform how science gets done. MIT Technology Review 2023. https://www.technologyreview.com/2023/07/05/1075865/eric-schmidt-ai-will-transform-science/

Shukla M. What companies do now will determine their future in the Intelligent Age. World Economic Forum 2024. https://www.weforum.org/stories/2024/12/companies-determine-future-in-intelligent-age/

Sonkar S., Waters AE., Baraniuk RG. Attention word embedding. Proceedings of the 28th International Conference on Computational Linguistics 2020; 6894-6902. https://doi.org/10.18653/v1/2020.coling-main.608

Sun M., Han R., Jiang B., Qi H., Sun D., Yuan Y., Huang J. A survey on large language model-based agents for statistics and data science 2024. https://doi.org/10.48550/arXiv.2412.14222

Tanudin D. Can LLMs be used to create social media posts on LinkedIn? A study of communication 2024; 1-18. https://kth.diva-portal.org/smash/get/diva2:1894799/FULLTEXT01.pdf

Vaswani A., Shazerr N., Parmar N., Uszkoreit J., Jones L., Gomez, AN., Kaiser L., Polosukhin I. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems 2017; 6000-6010. https://doi.org/10.48550/arXiv.1706.03762

Wang W., Ghobadi M., Shakeri K., Zhang Y., Hasani N. Rail-only: A low-cost high-performance network for training LLMs with trillion parameters 2023. https://doi.org/10.48550/arXiv.2307.12169

Weizenbaum J. ELIZA—A computer program for the study of natural language communication between man and machine. Communications of the ACM 1966; 9(1): 36-45. https://doi.org/10.1145/365153.365168

World Economic Forum. Industries in the intelligent age white paper series 2025. https://www.weforum.org/publications/industries-in-the-intelligent-age-white-paper-series/

Wu S., Fei H., Qu L., Ji W., Chua TS. NExT-GPT: Any-to-any multimodal LLM 2023. https://doi.org/10.48550/arXiv.2309.05519

Yang H., Zhou J., Fu Y., Wang X., Roane R., Guan H., Liu T. ProTrain: Efficient LLM training via adaptive memory management 2024. https://doi.org/10.48550/arXiv.2406.08334

Zhang Z. Understanding GPU architecture implications on LLM serving workloads 2024.