Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Large Language Models for Natural Language Processing: Architectures, Training Paradigms, and Real-World Applications – A Systematic Review

Utsha Sarker[1†], Archy Biswas[1†], Navjot Singh Talwandi[1*], Kamaljeet Kaur[1], Dulee Raj Devyani[1], Lalit Vaishnav[1]*

[1]Department of CSE, Apex Institute of Technology, Chandigarh University, Gharuan, Mohali, 140413, Punjab, India.
**Corresponding author(s). E-mail(s): utsha.sarker00775@gmail.com; navjotsingh49900@gmail.com and archyz2021@gmail.com;*
Contributing authors: kamaljeet.e19147@cumail.in; duleedevyani@gmail.com; vlalith7036@gmail.com;

ARTICLE INFO

Received: 28 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

In this manuscript, Large Language Models (LLMs) are praised as a transformative paradigm in the scope of man-made insight, which invigorates tremendous improvement in the capacities of Natural Language Processing (NLP) mechanisms. The triumph of transformer architectures and large scale pre- training has lead to latest LLMs to consistently provide strong performance for a wide range of tasks - from generating text to question answering, translation and reasoning - thus demonstrating their true utility. The ability of these models for assimilation of intricate datasets through linguistic pattern recognition has led to a great deal of progress in both academic research and industrial practice. This investigation presents a holistic and systematic look at the recent scholarship on large language models, and it has a concentrated focus on the topics of architectural evolution, training paradigms, and real-world applications. A systematically structured literature search was carried out from the most prominent academic databases: IEEEXplore database, ACM Digital Library, Scopus, Science Direct and arXiv from 2023 to 2026. The study follows a methodology that is modelled on Preferred Reporting Items for Systematic Reviews and Meta,-Analyses (PRISMA) guidelines. Items for Systematic Reviews and Meta-Analyses (PRISMA), and included the use of well-defined inclusion and exclusion criteria to identify relevant peer reviewed articles and high impact preprints. The review explains important new developments in LLM architectures, which include improvements in transformer architectures, mixture-of-experts architectures, and new alternatives which are focused on efficient sequence modelling. It further challenges the paradigms of training, such as training from a large scale (pre-training), instruction tuning, parameter efficient fine tuning (e.g. LoRA), reinforcement learning from human feedback (RLHF). In addition to this, the paper highlights the growing use of LLMs in other areas such as healthcare, education, software, and scientific research. In spite of their successes, there are significant challenges remaining - hallucination, bias, computational cost and evaluation limitations. The open research problems and possible futures for efficient, reliable, and trustworthy large language models are identified, ending the review.

Keywords: Large Language Models, Natural Language Processing, Transformer Architecture, Pre-training, Fine-tuning, Instruction Tuning, Reinforcement Learning from Human Feedback, Systematic Review.

1 INTRODUCTION

Natural Language Processing (NLP) has experienced a major change in the last decade from traditional statistical models to deep neural models that can model complex language patterns. Early neural approaches, such as Recurrent Neural Networks (RNNs) and Long Short-term Memory (LSTMs) networks, were widely used in sequence modeling tasks, such as machine translation, language modelling and sentiment analysis. Nonetheless, these architectures were limited in their capacity to capture long ranged dependencies and required sequential computation limiting its scalability and efficiency to train such models.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The advent of the transformer architecture by Vaswani et al revolutionised NLP by replacing the idea of recurrence by self attention mechanisms that gives the models the ability to process sequences in parallel and capture global contextual relationships. Transformer based models like BERT, GPT and T5 have proved that large scale pre-training with massive amount of corpus followed by fine-tuning them for a particular task can significantly improve performance in a wide variety of NLP tasks. As computational resources and data grew in scope, researchers increased the scale of these models to billions of parameters, and the so-called Large Language Models (LLMs) were born, which are capable of sophisticated reasoning, knowledge retrieval and generative tasks.

A number of recent studies from the past two years back that scaling down model parameters, training data, and compute resources is what leads to predictable improvements in performance of machine learning models, a phenomenon commonly called scaling laws for language models. These advances have allowed LLMs to achieve state-of-the-art results at tasks such as question answering, summarisation, dialogue systems and code generation. Consequently, LLMs are increasingly being applied to practical applications such as healthcare decision support, educational tutoring systems, scientific research assistance, enterprise knowledge management, and so on.

Despite the speed of change, the body of LLM research has expanded into an intricate web of interweaving research and theoretical advancement.. A multitude of architectural innovations, training paradigms and application domains have emerged, which makes it difficult for researchers and practitioners to achieve a unified understanding of the field. Whilst several surveys have been done on specific aspects of LLMs, e.g. architectural designs, evaluation methodologies, or domains of application, it is still lacking a proper systematic review done on LLMs that examines both the LLMs architectures, the training paradigms, and real-world applications simultaneously and across domains.

To fill this gap, this work is designed as a systematic review of recent literature on Large Language Models with an aim to synthesize current Literature to define current dominant design patterns, to highlight current emerging research directions on Large Language Models through literature review.

Research Questions

This review progresses with the following research questions:

RQ1: What architectures and design patterns dominate in RAs of today?

RQ2: What paradigms for training, e.g. large-scale pre-training, fine-tuning, instruction tuning, alignment methods, etc., are most common for LLM development?

RQ3: What are the main NLP and cross-domain, practical uses of LLMs and what are the limitations, or challenges, when implementing the models in real-world applications?

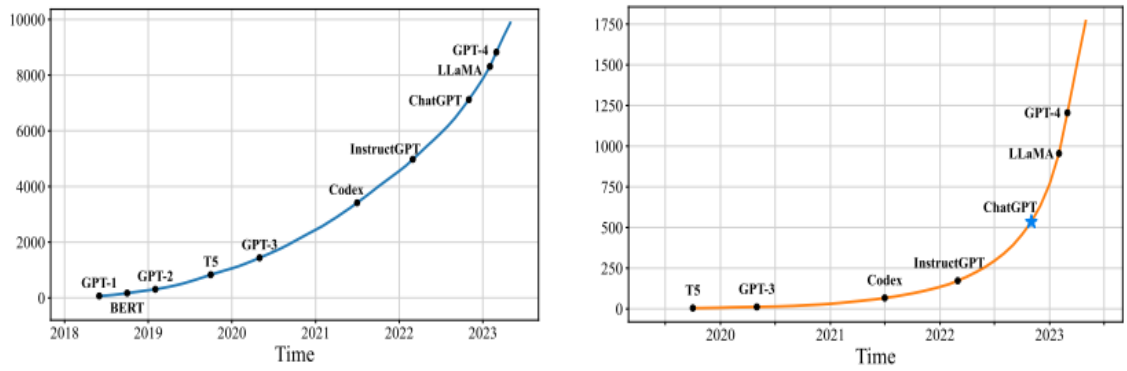
Contributions of This Review

The major contributions of this systematic review are summarised as follows:

- A detailed taxonomy of the architectures for LLMs that includes not only the transformer versions, but also mixture of experts and multimodal versions.
- A broad review of different training approaches, including pre-training approaches, approaches that focus on instruction fine-tuning, implementing parameter-efficient fine-tuning methods, and approaches for alignment (RLHF).
- Mapping out of twin/real life applications of LLMs in a range of fields such as healthcare, education, scientific research as well as software engineering
- A critical analysis of limitations and challenges, e.g. hallucination, bias, safety, computational cost, evaluation problems
- Detection of open research problems and future directions for creating more efficient, trustworthy and domain-adaptable LLM system.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



(a) Query="Language Model" (b) Query="Large Language Model"

Figure 1 - Overview of the Review Scope

As shown in Figure 1, this systematic review focuses on three basic dimensions of studies on LLM: model architectures, training paradigms and real life applications. These sets of interdependent dimensions collectively define the impact of big language models on the NLP systems of today.

2 METHODOLOGY

The following databases have been carefully selected from the list, and this has a reason in that they have gained a reputation for publishing works of high quality that have traversed the fields of artificial intelligence, machine learning, and NLP. The search window was from 2018 to 2025, hence covering the post mirage epoch that saw a fast boom made in modern LLMs.

2.1 Information Sources

A well-ordered study of isolation was planned for big language models and their practical implementation. The search queries combined many keywords related to the architectures of LLMs, their training regimes, and the tasks related to NLP. Sample queries used were:

- IEEE Xplore

- ACM Digital Library

- Scopus

- ScienceDirect

- arXiv

These databases have been selected carefully for the reason that they are jewels in terms of the high-quality of publications crossing the field of artificial intelligence, machine learning and NLP. The search window covered publications published between 2018 and 2025, thus corresponding to the postmip time period that catalysed the development of the modern realm of LLMs.

2.2 Search Strategy

A well-ordered study of isolation was planned for big language models and their practical implementation. The search queries combined many keywords related to the architectures of LLMs, their training regimes, and the tasks related to NLP. Sample queries used were:

- AND "large language model" AND "natural language processing"
- Search for "transformer architecture" AND "NLP"
- "GPT" OR "BERT" AND "pre-training"
- AND "large language model" AND "fine- tuning"

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Few examples of keyword-based searches:

- (or) "instruction tuning" and "alignment" AND ("LLM")
- transformer based language model" AND "applications

Boolean operators (AND, OR) allowed the expansion and refinement of the search results in the relevant databases. Retrieved records were then exported, de-duplicated and combined to curate an initial dataset of the candidate literature.

2.3 Included and Excluded Participants

To maintain the relevance and the methodological robustness of the selected studies, pre-defined inclusion and exclusion criteria were followed rigorously in the screening stage.

Inclusion criteria were based on studies which:

- Research transformer based large language models

- Present comprehensive reviews of architectural design, training strategies or practical uses

- Report empirical evaluation or the results of experiments

Exclusion criteria were applied which eliminated studies that:

- Focus on models of classical NLP with no transformer architectures

- Contribute to pure non technical commentary or opinion pieces

- Being deficient in adequate methodological detail

Dimension	Inclusion Criteria	Exclusion Criteria
Publication Type	Peer-reviewed journal/conference papers and widely cited preprints	Editorials, blogs, opinion articles
Research Focus	Transformer-based large language models for NLP	Classical ML models without transformers
Technical Content	Papers describing architectures, training methods, or applications	Papers without technical contributions
Language	English publications	Non-English papers
Time Range	2018–2025 publications	Older studies unless historically important
Application Scope	NLP or cross-domain LLM applications	Unrelated AI research

Table 1- A short summary of these criteria

2.4 Study Selection Process

The workflow for the selection followed the protocols/compliant to PRISMA:

Identification stage: Initial records were pulled from the designated databases with the crafted search queries.

Screening stage: Duplicate entries were removed and the remaining studies were screened preliminarily on the basis of titles and abstracts to determine the level of relevance.

Eligibility stage: The full texts of the potentially suitable studies were reviewed to ensure that they met the inclusion criteria.

Final inclusion: Only studies that met all stipulated criteria were included for the final review in the data analytics.

This investigation uses a rigorous protocol of systematic literature review, fully compliant with PRISMA 2020 in

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

order to guarantee that transparency, reproducibility, and rigor of methodology. The PRISMA framework provides a structured process for literature identification, thoroughly screening an article, carefully deciding if an article is eligible for inclusion, and including studies in systematic reviews. The general goal of this methodological regime is to carry out a systematic and exhaustive identification, critical appraisal and synthesis of recent scholarly contributions focused on LLM architectures, training paradigms and real world applications for the problem domain of NLP. In depth bibliographic search was conducted in multiple significant academic repositories to ensure that the availability of relevant research in these areas was comprehensive including: The list of repositories would be enumerated here. This stratified approach to filtration made sure that the systematic review only included the good studies that met the focus on the most relevant studies.

2.5 Data Extraction

For each study chosen, a subset of key attributes was retrieved in order to allow for a structured commentary as to the prevailing research trajectories in the LLM field. The extracted elements were:

- Model name & architectural type
- Number of parameters / overall scale of model
- Training paradigm (pre-training, fine-tuning, instruction tuning, RLHF)
- Characteristics of the training set
- Evaluation tasks used and benchmark suites used
- Application domains under consideration
- Performance measures used

These attributes made it possible to build an integrative, comparative overview of divergent architectures of LLM and www training methods.

2.6 Quality Assessment

To assess the methodological quality of individual studies, a quality appraisal framework was used, which was based on a set of criteria:

- Clarity of, and completeness of, architectural descriptions
- Data specification accessibility to training data
- Following the presence of baseline comparisons
- The reproducibility of experimental procedures
- Application of standardized evaluation standards

Studies with low methodological transparency or no empirical evidence were given a lower level of confidence in synthesis.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

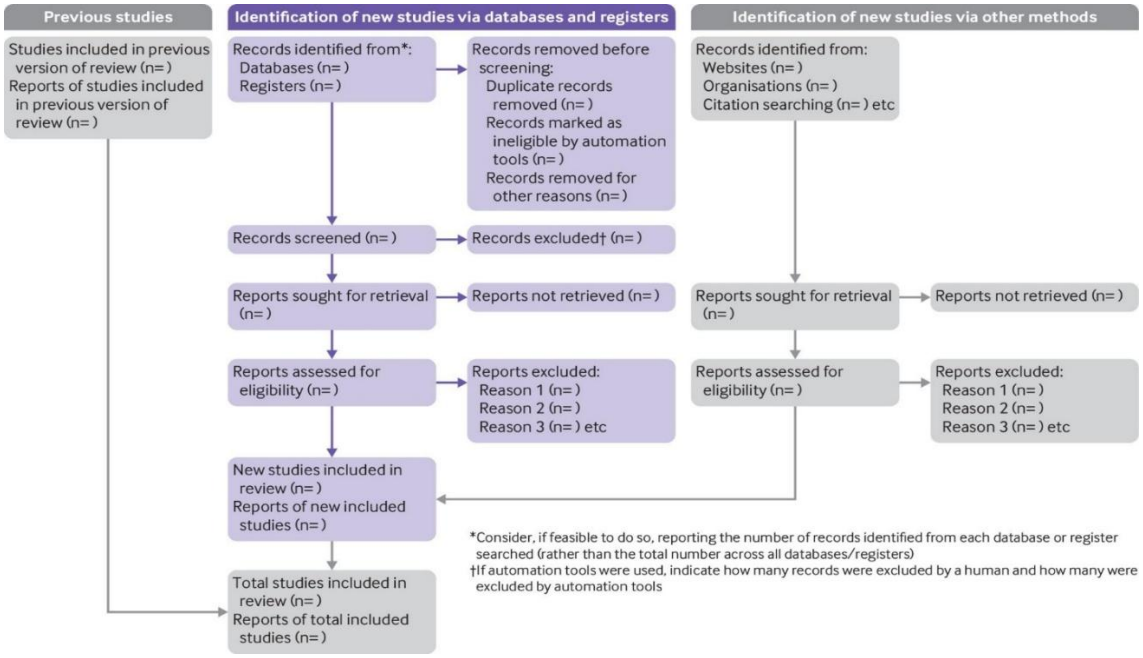


Figure 2 - The PRISMA flow diagram that summarizes the literature selection process adapted from Moher et al.

3 BACKGROUND TO LARGE LANGUAGE MODELS AND NATURAL LANGUAGE PROCESSING

Natural language processing has changed dramatically in the last 2 decades and has gone from being a series of rule-based systems to sophisticated deep learning architectures that can capture complex linguistic structure. Early paradigms of NLP heavily relied on the statistical methods such as n-gram language models and feature engineered classifiers. The development of neural networks led to the development of distributed word representations (notably Word2Vec and GloVe) enabling the modelling of semantic relationships between lexical items in the form of dense vector representations. These innovations paved the way for the neural sequence models that could learn contextual language models.

Subsequently, recurrent neural networks (RNNs) and long short - term memory (LSTM) architectures became a notable methodology towards temporal modelling. These models provided excellent performance on tasks such as translation, speech recognition and sentiment analysis. However, their structural sequential processing characteristics hampered parallel processing and made it difficult to undermine the long-sequia incident that long cossmillionaries textual sequences.

It was a watershed moment with the emergence of the transformer architecture; this architecture replaced the recurrent designs with self-attention architectures, giving the models the ability to identify the relationships between tokens regardless of their relative positions. Because of the ease of parallel computation provided by transformers, it greatly improves the scalability and therefore they can also be used to train much larger models with large datasets [1]. This architectural leap catalysed the emergence of today's large language models, i.e., models such as GPT, BERT, T5, etc locus of training in massive quantities of unsupervised or self supervised data, after a pre-training phase accompanied by fine-tuning [2].

Large language models are generally described as deep neural systems for language, made of billions or even trillions of parameters, which is based on transformer architectures, and is trained on mass amounts of text data. These models have the capacity of powerful computing power and large data repositories to enable them to make generalizations about language and digest knowledge about the world. Through pre-training tasks such as masked language modelling or next token prediction, LLMs internalise representations that are generalised and as a result amenable to fine-tuning or prompting for a plethora of downstream tasks [2], [3].

The dominance of LLMs has massively expanded the abilities of NLP systems into many different types of

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

applications. Quite common tasks that have been tackled by LLMs include:

- Text classification which includes sentiment analysis and topic categorisation

- Question answering where models generate accurate responses based on text understanding" sources of knowledge "Recently, Google patented a glue-based interaction model that can answer a question by prompting a human observer to a follow-up prompting

- Machine translation- providing cross-lingual communication

-Text summarisation, creating short digestions of large volumes of documents

- Underlying Dialogue and Conversational agents

- Text generation such as creative writing & code generation

These capabilities have led to widespread adoption of LLMs in many fields including healthcare, education, software engineering and scientific inquiry [3].

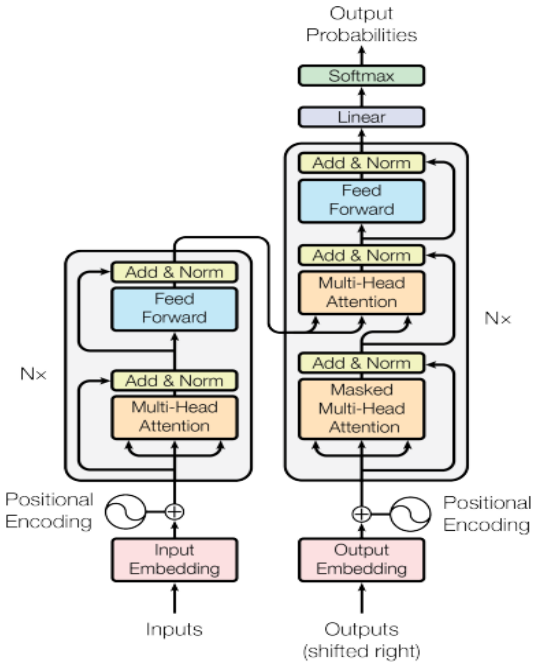


Figure 3. Simplified transformer architecture for modern large language models (modified from Vaswani et al. [1]).

Term	Definition
LLM	Large Language Model trained on large-scale corpora
NLP	Natural Language Processing
MLM	Masked Language Modeling objective
CLM	Causal Language Modeling objective
RLHF	Reinforcement Learning from Human Feedback
LoRA	Low-Rank Adaptation for efficient fine-tuning
MoE	Mixture-of-Experts architecture
PEFT	Parameter-Efficient Fine-Tuning
RAG	Retrieval-Augmented Generation

Table 2. Abbreviations and Key Concepts

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

4 THE LARGE LANGUAGE MODELS ARCHITECTURES

Large Language Models (LLMs) are inherently based on the transformer architecture, which makes use of self-attention based mechanisms for representing the context in the relationship between tokens. Over the years, however, a number of architectural paradigms have developed to solve various NLP tasks and computational needs. These architectures can be broadly classified as encoder only models, decoder only models and encoder decoder models which are optimised to be used for specific types of language processing tasks [1,2].

4.1 Model Families and Architectures

Encoder-Only Architectures

Encoder Venezuela are mainly designed for the language understanding task. To do this, they process the entire sequence of the input at a time and produce contextualised representations for each token. One of the most influential encoder based models is BERT (Bidirectional Encoder Representations from Transformers) which uses a masked language modeling (MLM) objective for pre-training. By means of token masking and token prediction by context tokens, BERT learns bidirectional contextual representations that boost performance significantly on tasks ranging from sentiment analysis, named entity recognition, and question answering [3].

Typical applications for encoder only architectures are cases of:

• Text classification

• Named-entity recognition

• Information extraction

• Natural-language inference

Strengths

Robust contextual tired representing learning

Queen Anne's Media content access is not free; it is safeguarded by the user's own intelligence and senses.It's not Free of Limitations Queen Anne's media content access is not Free of restrictions, because it is judicium of the user's own intelligence and senses.

Limitations

generative capabilities (few in number)

Long form region biosynthesis: Discrepancies in DNA replication discontinued in the activity of vacuoles cell cycle mitochondrion, at lower eukaryotes.Long form region biosynthesis: Discrepancies in DNA replication discontinued in the activity of vacuoles cell cycle mitochondrion at lower eukaryotes.

Decoder-Only Architectures

Decoder only models are mostly made for text generation jobs. They are based on an autoregressive training objective, where they have to predict the next token given previously generated tokens. The family of models GPT (Generative Pre-trained Transformer) is a member of a distinguished actuator with a decree decoder sole. Models such as GPT - 3 and GPT - 4 are having been trained with large scale corpus and next token prediction and are able to produce coherent text for a wide range of tasks including dialogue systems, summarisation and code generation.

Typical use cases include:

• Text generation

• Dialogue systems

• Code generation

• Creative writing

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Strengths

• Strong generative ability

Multiple task prompting (flexible prompting)

Limitations

• Higher computational cost

S:or Susceptible to hallucinations

Encoder-Decoder Architectures

Encoder-decoder architectures combine the strength of the encoding and decoding mechanism. The encoder converts the input sequence to contextual representations and the decoder produces output sequences from the representations. A representative model of this category is T5 (Text-to-Text Transfer Transformer) in which all NLP problems are rephrased as text-to-text problems. This architecture is especially effective for sequence to sequence transformations, such as translation, summarisation and question answering [4].

Such are typical applications include:

• Machine translation

• Text summarisation

• Question answering

Strengths

Flexible sequence- To- Sequence modelling

Generational Natural Language Processing Tasks Verbatim:

Limitations

Increases the free memory in application/on certain applications will be enabled.Increases the computational requirements

Rather, they are franchised. At the very least, rather than is franchised, they are more complex training pipelines.

4.2 Scaling and Design Choices

Modern LLMs are characterised by enormous numbers of parameters, in some cases in the billions or even billions. Scaling laws show that when a model is increased in size, the data, and the compute capabilities, performance improvement will be observed for a large variety of tasks [2].

A number of architectural innovations solve the issues of scalability and efficiency:

It was stated that, "Sparse Attention Mechanisms - simplify computational complexity by limiting attention computation to subsets of tokens, facilitating the processing of longer sequences."

Mixture -of-Experts (MoE) - activate only a subset of specialised neural subnetworks for each input, quickly addressing the issue of scalability without affecting computational efficiency.

Clearly, these tips can help maintain objectivity in machine learning algorithms.

Positonal Encoding Variants So time-series machine Learning can use positional encodings to appreciate order of tokens, transforms are based on order. POS Time-series machine learning. Recent innovations such as Rotary Positional Embeddings (RoPE) and ALiBi positional bias make long-context modelling better.

Another major distinction in the modern development of LLM is the development of open and closed models. Open models like LLaMA and Falcon have the weights open so the community can research it, but the proprietary models

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

like GPT-4 have their weights closed source because of commercial and safety reasons.

4.3 Specialized Architectures

Recent work has also focused on specialised architectures for LMs to cater to specific domains and tasks (LLMs).

Multilingual LLMs

Multilingual language models are trained using data with multiple languages which will enable cross lingual knowledge transfer and better performance on multilingual NLP tasks. Examples include:

• mBERT

• XLM-R

These models support applications such as multi language translation, cross language retrieval and accessing information internationally.

Domain-Specialised LLMs

Domain specific (L)LMs are designed to work in a specialised knowledge domain, such as:

• Biomedical NLP (e.g., BioGPT)

• Legal language models

• Models for code generation (e.g. Code- LLMs)

They include the use of domain-specific data sets and vocabularies to add more accuracy and reliability.

Multimodal LLMs

Multimodal large language models comprise text-based ones with other modalities, such as images, audio, or video, added. These models are capable of tasks such as visual question answering, image captioning and multimodal dialog. While multimodal architectures are an important direction of research, a full discussion of these systems is beyond the scope of this review. A number of dedicated surveys give detailed analyses of multimodal LLMs and their applications.

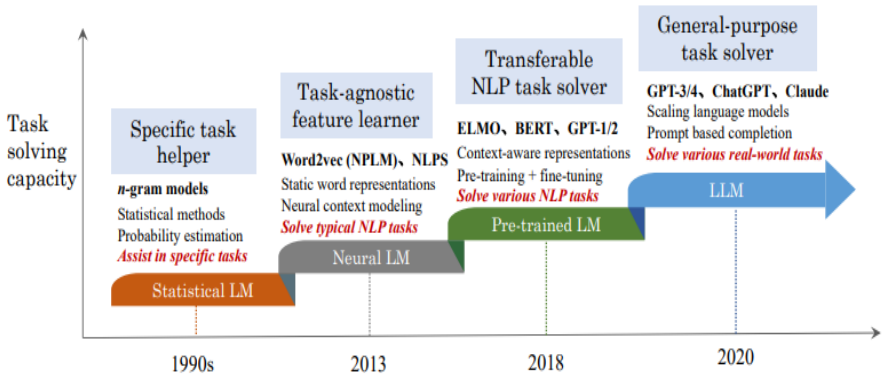


Figure 4- Taxonomy of large language model architectures (adapted from Zhao et al. 2).

Model	Architecture Type	Parameters	Pre-training Data	Release Year	Open / Closed
BERT	Encoder	340M	BooksCorpus + Wikipedia	2018	Open
GPT-3	Decoder	175B	WebText + Common Crawl	2020	Closed
T5	Encoder-Decoder	11B	C4 Dataset	2020	Open

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

PaLM	Decoder	540B	Multilingual web corpus	2022	Closed
LLaMA	Decoder	7B–65B	Public web datasets	2023	Open

Table 3. Representative LLM Architectures

5 LARGE LANGUAGE MODEL TRAINING PARADIGMS

The remarkable performance of large language models is due to the effect of improvements in training paradigms and adaptation strategies. Modern LLM development tends to be a multi stage training pipeline including large scale pre training, task dependent adaptation, alignment for human preference and inference time prompting techniques. The effect of these stages taken together is to allow models to develop a general knowledge of language and to be able to adapt it to a range of real-world applications.

5.1 Pre-training Objectives and Data

The first stage in the development of LLM is self-supervised pre-training on huge text corpus so that models learn about the syntactic structures and semantic relationships and general world knowledge without the need for manually labelled data.

Several popular objectives of pre-training include:

- Causal Language Modelling (CLM) - where the model is trained to predict the next token in a sequence based on the previous tokens. This autoregressive objective is very commonly used in decoder - only architectures like GPT models.

- Masked Language Modelling (MLM) - This is where the system masks tokens within a sentence randomly and teaches it to predict missing tokens, which relies on the context. This objective helps for bidirectional understanding of the context, and it can be frequently applied for encoder-based models such as BERT.

- Couples the property of both causal and bidirectional modelling by having a prefix conditioning sequence and predicting the tokens that follow it.
Span corruption, which is used in models like T5 as a masking strategy (instead of masking individual units, it inserts noise into whole contiguous blocks of tokens), is used to strengthen performance in sequence-to-sequence tasks. Forcing the model to reconstruct missing tokens encourages the model to have a more general, holistic understanding of the structure of the language.

Pre-training datasets are usually aggregations of large-scale web corpora, literary corpus, code repo corpora and domain-specific corpus. Famous examples of these include Common Crawl, Wikipedia, BooksCorpus and GitHub code datasets. The sheer scale and variety of these data sources are a crucial part of improving the findings in terms of the generalisation of models, but they also bring their own possible complications around quality of data and deeply rooted biases, as well as more broadly ethical [2].

5.2 Adaptation and Fine Tuning Paradigms

Once pre-training has happened large language models are adapted to downstream tasks using a range of fine-tuning approaches.

Standard Fine-Tuning

Standard fine-tuning consists of training the pre-trained model on task-specific labelled data such as classification corpora, question answering benchmarks or summarisation corpora. This procedure is the process of adjusting parameters of a model in order to best address the given task.

Instruction Tuning

Instruction tuning involves training models on datasets that consist of natural language instructions together with desired outputs. This strategy helps the model to better follow user directives in numerous tasks. Instruction tuning data sets will usually contain:

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

- FLAN datasets
- Open instruction datasets
- NLP benchmarks that are task oriented

Parameter - Efficient Fine - Tuning (PEFT)

Fine tuning models of the size of colossuses can be computational prohibitive. To reduce this burden, a number of parameter-efficient fine-tuning methods have been designed. Examples include:

- Adapters - small neural modules that are put between transformer layers
- LoRA (Low shed specified parameter update),
- Prompt tuning - optimises soft prompts in replacement of full model parameters

These techniques put severe limits on the computational needs without requiring loss of performance.

5.3 Alignment and Reinforcement Learning from Human Feedback (RLHF)

With ever-increasing power of LLMs, it has become of central research imperative to get model outputs to align with human preferences and safety desiderata. The current prevalence is Reinforcement Learning from Human Feedback or RLHF. The common phases of the RLHF pipeline include the following:

Collecting of human preference

Human annotators compare different answers of models and choose the most preferable answers.

Reward model training

A reward model is created which operates to predict these human preferences from the curated feedback.

Policy optimisation

The language model is changed using reinforced learning techniques to maximize the reward signal.

Recent investigations have added alternative methods to this post-training such as Direct Preference Optimization (DPO) method that is a simplified way to align models by direct preference comparison optimization and eliminating the iterative loops of reinforcement learning.

5.4 Inference-Time Techniques

Beyond adaptations that have been made during training, LLMs can be used to acquire great performance gains through inference-time strategies.

Prompt engineering

Prompt engineering involves creating well structured prompts to direct model behaviour. Techniques encompass:

- Few-shot prompting
- Chain-of-thought prompting
- Zero-shot prompting

These systems help models to overcome interactive reasoning tasks without further retraining.

Tool use

State-of-the-art LLM systems often utilize external tools, such as calculators, code interpreters, API front ends, etc., to add more to the system than purely generating text.

Retrieval- Augmented Generation (RAG)

RAG combines an external knowledge base into the inference process based on the relevant documents and it

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

conditioned the response of the model on the supplementary knowledge. This mechanism adds to the factual accuracy and minimizes the risks of hallucinations.

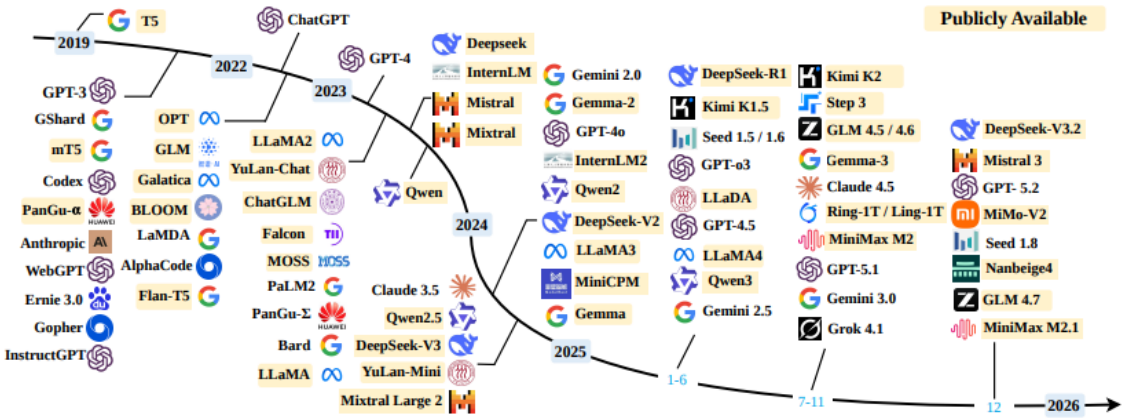


Figure 5. Training and adapting pipeline of large language models (adapted from Zhao et al. [2]).

Paradigm	Objective/Data	Typical Use	Advantages	Limitations
Pre-training	Self-supervised learning on massive corpora	Learn general language representations	Strong generalization	Extremely expensive
Fine-tuning	Task-specific labeled data	Adapt model to tasks	High performance	Requires labeled data
Instruction Tuning	Instruction–response datasets	Multi-task instruction following	Improves usability	Dataset quality dependent
RLHF	Human preference feedback	Alignment and safety	More helpful responses	Expensive annotation
PEFT (LoRA etc.)	Small parameter updates	Efficient adaptation	Lower compute	Slight performance drop

Table 4. Training Paradigms of LLMs

6 REAL - LIFE USE CASES OF LARGE LANGUAGE MODELS IN NLP

Large language models (LLMs) have greatly increased the functional repertory of natural-language-processing systems in a large variety of application domains. By being able to do multiple tasks through pre-training, prompting, and fine-tuning, we can use them as general, purpose land language processing. This section outlines the applications of LLMs in three broad categories: core NLP tasks, industry-based cross domain applications and human facing applications.

6.1 Core NLP Tasks

LLMs are known to always outperform on traditional NLP tasks, which traditionally had to be done with specialised models.

Text classification and Sentiment analysis

LLMs, for example, BERT and RoBERTa have high accuracy when performing classification due to the use of contextual embeddings acquired during pre-training. Fine-tuned LLMs are found to be reliable competitors of the traditional machine learning algorithms and the earlier neural architectures in diverse tasks such as sentiment analysis and topic classification [1].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Information extraction - Named entity recognition (NER)

Encoder-bias architectures dominate NER because of their ability to capture contextual inter token relationship and provide a better development of the entity in complex sentences.

Question answering (QA)

LLMs are supporting both extractive and generative QA. In extractive QA, the models identify the answer spans in the given context while in generative QA, the models make natural-(language) answers. Prompting techniques allow modern LLMs to carry out QA in zero shot settings.

Text summarisation

Sequence - to - sequence models (e.g. T5, BART) are very popular for document summarisation. LLMs are capable of delivering both extractive and abstractive summaries, and thus can be used for news summarisation and document-analysis applications.

Machine translation

Encoder-decoder models continue to be very powerful in translating text from one language to the other using sequence-to-sequence learning.

Information extraction

The support of Trieste for knowledge-graph construction, document indexing, and automated knowledge data client data-extraction workflows LLMs have the ability to distil structured information from unstructured text.

Overall, empirical evidence suggests that LLMs regularly produce state of the art results on a vast number of scale benchmarks [2] for NLP, especially when combined with few shot prompting or task-specific fine-tuning.

6.2 Cross domain and Industrial Applications

Beyond traditional JNLP standards, LLMs are also used more frequently during industry-specific implementation.

Healthcare/ Biomedicalificial Intelligence: healthcare and biomedicalatical NLP

Within the area of Healthcare, LLMs are in use to help with clinical text analysis, biomedicine - literature mining, medical query answering. Domain specific models based on biomedical corpus improves performance on terminology recognition and clinical decision support system.

Financing and Law applications

LLMs are used for analysis of financial documents, risk assessment and processing of legal texts. They assist in the contract analysis, regulatory - compliance monitoring, and summarisation of financial reports.

Customer service systems and enterprise systems

LLM based conversational agents have been widely implemented in customer service platforms. These women and men mark research and which agents bring together LLMs and knowledge bases and retrieval systems to leg on automated, context 3: responses.

Software engineering

LLMs pre-taught on large code bodies are helping with code generation, debug and documentation writing. Coding assistants based on decoder based architectures have the potential to synthesize executable code from natural language descriptions.

Education

Educational tools such as automated tutoring systems, personalised learning assistants and automated grading. LLMs produce explanations, responses to student queries, as well as customized learning paths.

A good deal of industrial implementations are based on hybrid systems that integrate LLMs with peripheral systems,

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

for instance:

- Retrieval augmented generation (RAG)
- LLM + database queries systems
- LLM + knowledge graphs

The architectural subtleties described inside demarcate a degree of additional fidelity to facts alongside model outcomes even as fostering the cohesive synthesis of domain-specific knowledge.

6.3 Human-Facing Applications

Large Language Models have enabled a new wave of self-interactive human-centered AI systems.

Virtual Assistants and Chatbots

Contemporary conversational agents are employing LLMs to support natural dialogic conversations as the basis for contextual understanding and handling multi-turn conversations.

Content Generation Tools

LLMs are widely exploited to perform the task of automatically producing textual content: this includes journalistic articles, marketing materials and creative writing. These tools bring support to tasks ranging from automated report writing to social media post writing.

Coding Assistants

Programmatic aides which are LLMs help developers to write code excerpts and debug software and understand programming concepts.

Workflow Integration

An increasing number of LLMs are now integrated in productivity suites and in enterprise pipelines, thus making it possible to execute tasks within software systems, as a result of natural language directives.

Nonetheless, there are challenges inherent in the widespread deployment of these models such as hallucinations, biases and reliability issues which require careful remediation to ensure safe and trustworthy-applications.

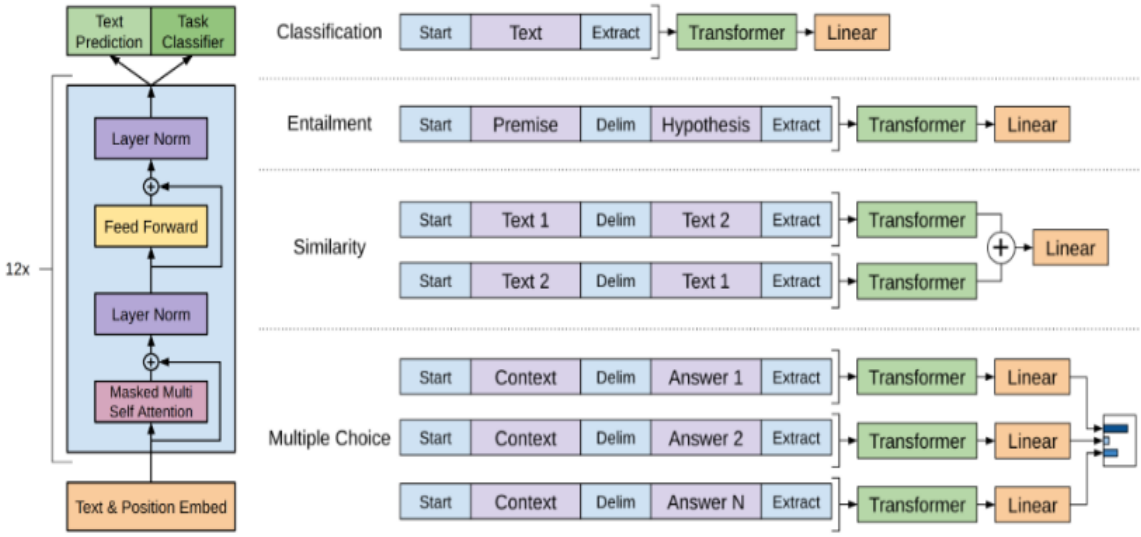


Figure 6. Application landscape for large language models in NLP and Industrial applications

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

(adapted from Minaee et al.Cinema d conscience et intelligence artificielle Acte III, 2022).

Domain/Task	Model	Setup	Dataset	Metrics	Findings
Sentiment Analysis	BERT	Fine-tuning	SST-2	Accuracy, F1	Outperforms classical models
Question Answering	GPT-3	Few-shot prompting	SQuAD	EM, F1	Strong zero-shot ability
Summarization	T5	Fine-tuning	CNN/DailyMail	ROUGE	High-quality summaries
Biomedical NLP	BioGPT	Domain fine-tuning	PubMed	F1	Improved medical knowledge
Code Generation	Codex	Prompting	HumanEval	Pass@k	Effective coding assistance

Table 5. Summary of LLM Applications

7 EVALUATION AND BENCHMARKS PRACTICES

The assessment of large language model research, its rigor and analytical depth are the most crucial factors centering its performance, reliability, and real-life applicability. Owing to the vast capabilities of modern LLMs, evaluation frameworks usually involve several dimensions, such as task effectiveness, ability to reason, robustness, efficiency of computation as well as safety concerns. Over time, many benchmark datasets and evaluation techniques have been developed to embody a procedural assessment of the competencies of LLMs for assorted duties7,8.

7.1 Common Criteria of Evaluation

A spectrum of benchmark suites has gained popular use in determining success of LLMs.

GLUE and SuperGLUE

The General Language Understanding Evaluation (GLUE) benchmark and its successor, SuperGLUE, consist of a series of NLP tasks that have been designed to test for general language understanding. These tasks include natural-language inference, question answering and sentence similarity. SuperGLUE was introduced to pose more challenging problems as follows many models were achieving near human level on the original GLUE benchmark.

MMLU

The Massive Multitask Language Understanding (MMLU) benchmark test is used to rate the ability of a model to perform reasoning and knowledge retrieval tasks in a variety of 57 different fields of study, from mathematics and physics to law and medicine. It is commonly used to assess the general knowledge abilities of large language models.

BIG-Bench

Beyond the Imitation Game Benchmark (BIG-Bench) is a large - scale collaborative benchmark for more than 200 heterogeneous tasks designed to test emergent capabilities for LLMs. These tasks involve interrogating reasoning, knowledge of commonsense and language.

HELM

The Holistic Evaluation of Language Models (HELM) framework examines models across a range of different dimensions - including accuracy, calibration, fairness, robustness and efficiency. HELM puts transparency and an all-round appraisal in realistic scenarios in the forefront.

Safety and Bias Benchmarks

There are several benchmarks that are specifically dedicated to evaluating safety, bias and toxicity in LLM outputs.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

For example, data sets like BOLD manage and examine demographic bias and the dangers of harmful content generation.

7.2 Evaluation Metrics

The performance evaluation of LLMs relies not only on automated evaluation methods but also on methods involving human judgments.

Common metrics comprise:

- **Accuracy and F1 -** score is a standard indicator for the accuracy of the classification task and for question-answering benchmark.
- **BLEU (Bilingual Evaluation Understudy)** is a measure of the similarity between the generated and the reference text, and is commonly used in the evaluation of machine translation.
- **ROUGE (Recall Oriented Understudy for Gisting Evaluation):** This is used to measure the overlap between generated and reference summaries and hence makes it easy to measure the quality of summarization.
- **Perplexity** is used to measure the likelihood of correctly guessing the next token in language -- modelling tasks.

Human evaluation: the evaluators analyze the model's outputs for characteristics such as fluidity, relevance, factual correctness and helpfulness.

7.3 Shortcomings of Existing Methods of Evaluation

Notwithstanding the availability of numerous benchmarks, the evaluation of LLMs suffers from difficulties.

Benchmark Saturation - It's not unusual for models now to have near-perfect results on popular benchmarks such as GLUE which lowers performance for the capacity to discriminate incremental improvements.

Prompt Sensitivity - LLM performance may vary considerably depending on prompt phrasing and framing of the context, making it difficult to replicate results of evaluation.

Data Contamination - training corpora can be so large that it becomes astronomically huge, which potentially means that this can lead to data leakage and overestimate performance metrics.

Reproducibility Issues - differences in the evaluation setup, prompting strategies and the utilisation of a model can lead to inconsistent outcomes across studies.

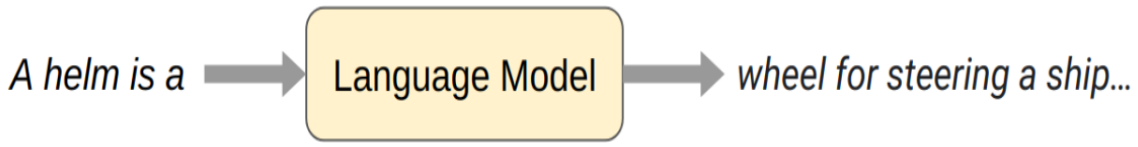


Figure 7. Important dimensions of large language model evaluation including capability, robustness, efficiency, safety and user-centric factors (adapted from Liang et al. [3]).

Benchmark	Task Types	Size	Metrics	Observations
GLUE	Language understanding tasks	9 tasks	Accuracy	Nearly saturated
SuperGLUE	Harder NLP tasks	8 tasks	Accuracy, F1	Designed for stronger models
MMLU	Multi-domain reasoning	57 subjects	Accuracy	Tests knowledge breadth

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

BIG-Bench	Emergent abilities	200+ tasks	Task-specific	Shows reasoning ability
HELM	Holistic evaluation	Hundreds scenarios	Accuracy, fairness metrics	Focus on transparency

Table 6. Evaluation Benchmarks

8 PROBLEMS, RISKS, AND PENDING ISSUES

Despite the incredible progress that has been made using Large Language Models, there remains a constellation of technical, operational and societal issues. These obstacles impinge upon both the reliability of the outputs of LLM models, as well as with regards to the responsible use of these systems in an authentic setting. This section considers the main issues associated with model behaviour, constrains of deployment and wider societal consequences.

8.1 Model and Data Challenges

Hallucinations and Errors of Fact

Hallucination has been identified as an exception to the most common limitations reported with LLMs, in which tokens generated by the model are palliative but incorrect or fabricated. Given that LLMs are trained on the statistical distribution of tokens, and not on mechanisms of factual verification, they can be expected to generate output that looks credible but is wrong.

Bias and Toxicity

Training sets (corpora) obtained from large web corpora often contain feelings, prejudices, and offensive terms which this has been shown to model in its output. The existence of bias and toxicity makes anxieties for fairness and calls for scrutiny of ethical AI implementation, especially in sensitive areas such as recruitment systems or in healthcare as a decision aid.

Data Leakage & Data Contamination

Large scale training data sets sometimes include benchmarking data for evaluation. This phenomenon is generally known as data contamination and can produce unrealistically high performance scores which compromises the integrity of evaluation metrics.

Long-Context Limitations

Although in recent time models architecture allowed to have more context window size, however, it notes that still many LLMs tend to show difficulty condition that face long documents or long chain of reasoning thereby reducing performance on tasks of processing long sequences.

8.2 Deployment Challenges

Computational Cost and IO (Latency)

The process of training and deploying LLMs requires large amounts of computational resources which often requires utilisation of specialised hardware like GPU's or Tera Flops Processing Unit. Increased inference latency and infrastructural costs can be limitations of deploying or using real time inference within a resource limited environment.

Privacy and Security Problems

LLM driven systems may inadvertently suffer from information leakage caused by the information derived by the model from the sensitive information contained in the training data set or the user prompt. In addition, vulnerabilities like prompt-injection attacks can affect the behaviour of a model in such a way that the security of the system is compromised.

Interpretability and Monitoring

LLMs are complicated neural architectures that have low interpretability. Understanding how models came to produce certain outputs poses a major challenge and therefore makes debugging, continuous performance

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

monitoring and adherence to regulatory standards challenging.

Maintenance and Update of a Model

Models trained using static datasets are at risk of obsolescence, in the face of new information coming in. Consequently, processes to continuously update and vigilantly sense-check to ensure that the models remain accurate and relevant, are imperative.

8.3 Societal and Ethical Issues

Misinformation and Content Manipulation

Large language models have the power of producing voluminous quantities of remarkably realistic text, which understandably produces serious concerns about proliferation of misinformation and automation of content manipulation.

Academic Integrity: Plagiarism

The skill of LLMs to draft essays, reports, and code also raises fears about plagiarism and academic dishonesty, which are problems worth reviewing.

Job Displacement and Work Force Changes

The automation that can come with the power of LLMs promises to dramatically transform a number of different professions, specifically those that use a lot of text work such as writing, translating and customer support.

Regulatory and governance Issues

Governments and regulatory bodies are increasingly examining the societal impacts of AI systems which has led to the development of new policies and regulatory frameworks to ensure the responsible deployment of AI systems.

8.4 Limitations of the Current Research

Current studies on LLMs have a slew of methodological shortcomings.

-Training data and model architectures

- A lack of transparency with regard to training data and model architectures.

- New algorithms to evaluate environmental health risks

- An over-reliance upon benchmark datasets.

- An insufficient amount of human-centred evaluation.

- Difficulties in replicating results because of proprietary models.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Question: What is the name of the Russian mercenary group?
Answer: Wagner group.

Question: Where is the Wagner group?
Answer: In Syria.

Question: Who is the leader of the Wagner group?
Answer: A Russian citizen, Dmitry Utkin. He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine. According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.

Figure 8. Risk- Benefit matrix to show the varying deployment scenarios for large language models (adapted from Bender et al. [2]).

Challenge	Example Issues	Impact	Mitigation
Hallucination	Incorrect generated facts	Misleading information	Retrieval-augmented models
Bias	Social stereotypes	Unfair decisions	Dataset filtering
Data leakage	Benchmark contamination	Inflated evaluation	Transparent dataset reporting
Long context limits	Reduced reasoning ability	Poor long-document processing	Long-context transformers
Deployment cost	GPU/TPU requirements	Expensive inference	Model compression

Table 7. Challenges and Mitigation

9 FUTURE RESEARCH DIRECTIONS

Despite the fantastic progress of large language models to date, remarkable opportunities exist in the future for the development of their architectures, training regimes, applications, and evaluation procedures. (Partnered with Artificial Intelligence vs Machine learning) Future efforts by academics and researchers efforts are likely to focus on improving efficiency, reliability, adaptability and human-centered integration of LLM systems.

9.1 Architectural Innovations

Emerging architectures of LLM are meant to favor the efficiency and the modular design. Present models often require very large amounts of computers, and thus lack accessibility and scalability in their applications. Investigations relating to the development of efficient transformer variants, sparse attention mechanisms and mixture-of-experts architectures collided for the purpose to reduce computational overhead at the same time as achieving robust performance (see reference [1]).

Another great path forward is retrieval-augmented architectures, in which LLMs have access to external sources of

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

knowledge (search engines, knowledge graphs, databases) to use during the inference process. This strategy can reduce the occurrence of hallucinations and preserve up-to date knowledge. Such a strategy can moderate the hallucinations and preserve modern-day knowledge.

Moreover, multimodal large language models are expected to play an ever-greater role that combines text, images, audio, and video in order to accomplish tasks such as visual reasoning, multimodal question answering, and image captioning.

9.2 Developments in Training Paradigms

Future training strategies will presumably focus on fine-tuning alignment, efficiency and adaptability.

A critical research avenue relates to the improvements of alignment methods, including reinforcement learning from human feedback (RLHF) and direct preference optimisation (DPO) in order to ensure that the output of models is valued by humans and protects requirements.

Another salient direction is low resource adaptation, to allow for large models to perform well on areas where there is limited labelled data. Techniques such as parameter efficient fine tuning (PEFT), Prompt tuning, and in adapter-based training, organisations can customise LLMs without performing extensive retraining activities.

Researchers are equally delving into constant learning approaches that allow LLMs to incorporate new data without the need for a full retraining.

Learning model: This facilitates the relevance in the model in a dynamically changing knowledge environment.

9.3 Emerging Applications

Future applications of LLMs are expected to move past the general NLP applications and into highly specialised domain systems.

Examples include:

Domain specific - LLMs that are custom made for healthcare, finance, and legal analysis Coding assistants for software development. Coding Assistants for software development.

Everything follows - Scientific research assistants who can synthesise literature.

In many of these scenarios, LLMs will work in human-AI collaborative models, where users interact with AI tools to multiply their productivity and make better decisions. Designing effective interfacing and flow of work to support such a collaboration is still a challenging area for researchers.

Furthermore, ensuring safety, transparency and regulatory compliance will be critical with deploying LLMs in high - stake setting such as healthcare or legal decision - support systems.

9.4 Frameworks of Evaluation in the Future

Current approaches to evaluation rely heavily on static evaluation, which may or may not properly reflect the real world. Future scholarship is expected to focus on task orientated and user centered evaluation methods.

Some of the potential enhancements include:

so-called "Dynamic benchmarks" that evolve over time.

E.g. "evaluation frameworks that weave in human feedback".

Transparency - Clear reporting standards for data on training and standards for evaluating its quality.

It will help to more accurately reflect a realistic appraisal of LLM capabilities and limitations.

Research Area	Future Direction	Benefits	Challenges
Efficient architectures	Sparse transformers, MoE	Lower compute cost	Maintaining performance

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Retrieval-augmented LLMs	Integration with search/databases	Better factual accuracy	Latency
Multimodal models	Text + image + audio	Richer reasoning	Training complexity
Alignment	Improved RLHF, DPO	Safer AI systems	Scaling feedback
Continual learning	Updating models over time	Up-to-date knowledge	Catastrophic forgetting

Table 8. Future Research Directions

10 CONCLUSION

Large language models have revolutionized natural language processing and equipped models with powerful language comprehension and generation capabilities in a wide range of tasks. The development of transformer architecture, large-scale pre-training and alignment has allowed LLMs today to demonstrate extraordinary performance in tasks such as question answering, summarisation, translation and conversational AI. The scaling of model parameters and training data sets very rapidly has also led to significant advances in generalisation and reasoning capability making LLMs a key paradigm in current AI research (see references [1], [2]).

This systematic review summarises recent progress in the development of the architectures of LLMs, the training paradigms and the applications of these models in the real world. The identification of emerging trends brings a few highlights to Vorderwallner's important observations: That modern models tend to prefer transformer based architectures using encoder only, decoder only, encoder decoder setups to fulfill different conceptual tasks this t ime.Today's models tend to be drawn to the same kinds of transformer-based designs, and use encoder only, decoder only and encoder decoder architectures to fulfill a range of task requirements. Training strategies look and evolves not only to classical "pre training/fine-tuning" but includes several sophisticated pipelines including instruction tuning, RLHF, parameter efficient adaptation, etc. LLM's applications has grown.

Despite these breakthroughs, there are still a number of persistent challenges, including hallucinations, prejudicial biases, computation cost and limitations of current evaluation methodologies. Addressing these issues will require additional study of sound evaluation frameworks, transparent reporting of practices and sophisticated alignment methods. As the technology under the LLM umbrella continues to evolve, there is a need to ensure these systems are deployed appropriately, with correct ethical considerations, and to be assessed through a human-centred lens, to maximise the benefit with minimum risk.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017). https://doi.org/10.48550/arXiv.1706.03762

[2] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019). https://doi.org/10.48550/arXiv.1810.04805

[3] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21(140), 1–67 (2020)

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 1877–1901 (2020). https://doi.org/10.48550/arXiv.2005.14165

[5] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems (NeurIPS) (2022). https://doi.org/10.48550/arXiv.2203.02155

[6] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al.: A survey of large language models. ACM Computing Surveys 56(9), 1–38 (2023). https://doi.org/10.1145/3589335

[7] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J.: Large language models:

A survey. arXiv preprint (2024). https://doi.org/10.48550/arXiv.2402.06196
[8] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al.: Holistic evaluation of language models. arXiv preprint (2022). https://doi.org/10.48550/arXiv.2211.09110
[9] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 610–623 (2021). https://doi.org/10.1145/3442188.3445922
[10] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al.: On the opportunities and risks of foundation models. arXiv preprint (2021). https://doi.org/10.48550/arXiv.2108.07258
[11] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research 25(70), 1–53 (2024)
[12] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., et al.: LLaMA: Open and efficient foundation language models. arXiv preprint (2023). https://doi.org/10.48550/arXiv.2302.13971
[13] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., et al.: PaLM: Scaling language modeling with pathways. arXiv preprint (2022). https://doi.org/10.48550/arXiv.2204.02311
[14] Srivastava, A., Rastogi, A., Rao, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint (2022). https://doi.org/10.48550/arXiv.2206.04615
[15] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint (2020). https://doi.org/10.48550/arXiv.2009.03300
[16] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: International Conference on Learning Representations (ICLR) (2019). https://doi.org/10.48550/arXiv.1804.07461
[17] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
[18] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
[19] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
[20] Kaplan, J., McCandlish, S., Henighan, T., Brown, T., Chess, B., Child, R., et al.: Scaling laws for neural language models. arXiv preprint (2020). https://doi.org/10.48550/arXiv.2001.08361