Stout in Languages, Linguistics and Literature (SLLL)
ISSN: 2979-2800
DOI: https://doi.org/10.61424/mvrk6575
Journal Homepage: https://stoutjournals.org/index.php/SLLL

Natural Language Processing in the Era of Large Language Models: A Systematic Review of Methods and Applications

Chang Wang
Independent Researcher, China
Corresponding Author: Chang Wang, **E-mail**: Wang385@gmail.com

ARTICLE INFORMATION

Submitted: June 05, 2026
Accepted: September 23, 2026
Published: October 22, 2026
Volume: 2
Issue: 1

KEYWORDS

Natural Language Processing; Large Language Models; Transformers; Generative AI; Language Models; Retrieval-Augmented Generation; Artificial Intelligence.

ABSTRACT

Natural Language Processing (NLP) has undergone a substantial transformation with the emergence of large language models (LLMs), which have expanded the capabilities of machines to understand, generate, and interact with human language across increasingly complex tasks. This systematic review examines the evolution of NLP methods in the era of LLMs, with particular emphasis on model architectures, training and adaptation strategies, emerging methodological approaches, and applications across diverse domains. Relevant scholarly literature was systematically reviewed to identify major developments in transformer-based architectures, pre-training and fine-tuning, instruction tuning, reinforcement learning from human feedback, retrieval-augmented generation, prompt engineering, parameter-efficient adaptation, multimodal language processing, and agentic language systems. The findings indicate that LLMs have substantially improved performance in text generation, question answering, summarization, translation, information extraction, dialogue systems, sentiment analysis, and domain-specific language understanding. At the same time, the review identifies persistent challenges involving hallucination, factual reliability, bias, interpretability, computational cost, data privacy, security, evaluation consistency, and the representation of low-resource languages. Applications in healthcare, education, business, scientific research, public services, and software development demonstrate the growing versatility of LLM-based NLP systems, while also highlighting the importance of domain adaptation and human oversight. Overall, the evidence suggests that contemporary NLP is shifting from task-specific language models toward general-purpose, adaptable, and increasingly multimodal language systems. Future research should prioritize reliable and transparent evaluation, efficient model adaptation, multilingual and culturally responsive NLP, responsible deployment, and methods that improve factuality, safety, and human–AI collaboration.

1. Introduction

Natural Language Processing (NLP) has emerged as a central field of artificial intelligence (AI) concerned with enabling computational systems to understand, process, generate, and interact with human language. Its development has progressed from rule-based and statistical approaches to machine-learning methods and, more recently, large-scale neural architectures capable of performing multiple language tasks within unified models. Traditional NLP applications such as text classification, sentiment analysis, machine translation, information extraction, question answering, named-entity recognition, and speech-related language processing have increasingly benefited from advances in representation learning and deep neural networks (Zubiaga, 2024). The transition from task-specific models toward general-purpose language models has consequently transformed both the methodological foundations and practical scope of NLP research.

A major turning point in this evolution has been the emergence of transformer-based architectures. The transformer introduced an attention-based mechanism that enabled models to process relationships among tokens more effectively while supporting large-scale parallel training. Subsequent developments in pretrained language models demonstrated that representations

Copyright: © 2026 The Author(s). This open access article is distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC- BY-NC-SA) License (https://creativecommons.org/licenses/by-nc-sa/4.0/), published by Stout Journals, London, United Kingdom

learned from massive text corpora could be adapted to diverse downstream NLP tasks. Models such as Bidirectional Encoder Representations from Transformers (BERT) established the effectiveness of bidirectional contextual representations, while autoregressive architectures demonstrated increasingly sophisticated capabilities in text generation and language understanding. These developments contributed to a paradigm shift in which large pretrained models became foundational components of NLP systems rather than merely task-specific computational tools….

The emergence of large language models (LLMs) has further accelerated this transformation. LLMs are trained on very large and diverse textual datasets using substantial computational resources and increasingly sophisticated neural architectures. Their scale allows them to acquire broad linguistic and contextual representations and to perform multiple tasks through prompting, instruction following, or task-specific adaptation. Contemporary LLMs can support activities ranging from summarization and translation to question answering, information extraction, dialogue generation, coding assistance, and content creation. This flexibility has challenged the conventional distinction between individual NLP tasks and has encouraged researchers to conceptualize language processing as a more general capability that can be elicited through natural-language instructions.

The methodological landscape of NLP has therefore expanded considerably. Earlier approaches were typically organized around supervised learning, feature engineering, task-specific architectures, and benchmark-specific optimization. In contrast, contemporary NLP increasingly incorporates pretraining, fine-tuning, instruction tuning, prompt engineering, in-context learning, retrieval-augmented generation, parameter-efficient adaptation, and multimodal integration. Transformer-based models have also enabled researchers to investigate increasingly complex relationships between language understanding and generation. Parameter-efficient techniques, including low-rank adaptation and related approaches, have become particularly important because they provide mechanisms for adapting large models without retraining all parameters (Min, 2023). Similarly, retrieval-augmented approaches seek to combine generative language models with external knowledge sources, addressing some limitations associated with relying exclusively on information encoded during pretraining.

Despite these advances, the expansion of LLM-based NLP has introduced substantial methodological and practical challenges. One important concern is the reliability of generated information. LLMs can produce fluent and contextually plausible responses that nevertheless contain factual inaccuracies or unsupported claims. This phenomenon, commonly discussed as hallucination, creates particular challenges for applications in education, healthcare, law, scientific research, and other domains where accuracy and traceability are important. Model performance can also vary according to prompting strategies, datasets, languages, task formulations, and evaluation criteria. Consequently, conventional benchmark performance alone may not adequately capture the reliability, robustness, fairness, or usefulness of contemporary NLP systems.

Another major issue concerns the data and social dimensions of language technologies. Large language models inherit characteristics of the data on which they are trained, including linguistic biases, cultural assumptions, stereotypes, and unequal representation of languages and communities. These issues are particularly significant for low-resource languages, for which training data, linguistic resources, benchmarks, and computational infrastructure remain comparatively limited. The dominance of high-resource languages in NLP research may consequently contribute to unequal technological benefits and may restrict the effectiveness of LLMs across multilingual and culturally diverse contexts. At the same time, concerns surrounding privacy, copyright, transparency, accountability, energy consumption, and the provenance of training data have become increasingly prominent in discussions of responsible NLP development.

Evaluation has similarly become more complex in the era of LLMs. Traditional metrics designed for specific NLP tasks may not fully reflect the quality of open-ended generated language. Metrics such as BLEU, ROUGE, and accuracy remain useful in appropriate contexts, but researchers increasingly recognize the need for multidimensional evaluation involving factuality, relevance, coherence, safety, robustness, toxicity, bias, and human preferences. Human evaluation remains valuable but can be expensive, difficult to standardize, and susceptible to differences in evaluator expertise and interpretation. The development of more comprehensive evaluation frameworks is therefore necessary to establish meaningful comparisons among models and to determine whether improvements in benchmark scores translate into reliable real-world performance.

The rapid pace of development has also produced a fragmented research landscape. Studies increasingly examine individual components of the NLP ecosystem, including transformer architectures, pretrained models, prompting techniques, retrieval augmentation, fine-tuning strategies, multilingual NLP, domain adaptation, and responsible AI. Although these contributions provide important insights, there remains a need for systematic synthesis that connects methodological developments with the applications, benefits, limitations, and emerging research directions associated with LLM-based NLP. A systematic review can help clarify how the field has evolved, identify dominant methodological approaches, examine the applications in which LLMs have been deployed, and highlight unresolved challenges requiring further investigation.

Against this background, this study systematically reviews the literature on NLP in the era of large language models, with particular attention to the evolution of methods, architectures, adaptation strategies, evaluation approaches, and application domains. The review examines how the transition from traditional NLP to foundation-model-based approaches has altered the design and deployment of language-processing systems. It further considers applications across areas such as information retrieval, text generation, translation, healthcare, education, business, social media analysis, scientific research, and human–computer interaction. In addition, the study synthesizes evidence concerning model limitations, including hallucination, bias, multilingual and low-resource language challenges, computational requirements, privacy, interpretability, and evaluation reliability (Raiaan, 2024). By integrating these dimensions, the review seeks to provide a comprehensive account of the current NLP landscape and identify emerging methodological and application-oriented directions for future research.

2. Methodology

2.1 Research Design

This study employed a systematic literature review design to synthesize the methodological developments, applications, opportunities, and challenges associated with Natural Language Processing (NLP) in the era of Large Language Models (LLMs). A systematic review approach was selected because the rapid expansion of transformer-based architectures, foundation models, instruction-tuned systems, and generative NLP applications has produced a large and multidisciplinary body of research that requires structured synthesis. The review focused on studies examining NLP methods and applications across both traditional and contemporary approaches, with particular attention to the transition from conventional machine-learning and task-specific deep-learning models toward pretrained and large-scale generative language models. The methodology was structured to promote transparency, reproducibility, and consistency in the identification, screening, evaluation, and synthesis of relevant literature.

2.2 Search Strategy and Information Sources

A comprehensive literature search was conducted across major scholarly databases and academic information sources, including Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, PubMed, and Google Scholar. These databases were selected to provide broad coverage of computer science, artificial intelligence, information systems, linguistics, computational social science, healthcare, education, and other domains in which NLP and LLMs are increasingly applied. The search covered literature published from the emergence of modern statistical and neural NLP approaches through the contemporary LLM period, with particular emphasis on studies published between 2017 and 2026 because this period encompasses the widespread adoption of transformer architectures, pretrained language models, and large-scale generative models. Reference lists of highly relevant studies and major review papers were also examined to identify additional publications that may not have been retrieved through database searches.

2.3 Search Terms and Query Development

The search strategy combined keywords associated with NLP, machine learning, deep learning, transformers, pretrained language models, and LLM applications. Search strings were developed using Boolean operators and combinations of terms such as "Natural Language Processing," "NLP," "large language models," "LLMs," "foundation models," "transformer models," "pretrained language models," "generative AI," "BERT," "GPT," "T5," "RoBERTa," "language understanding," "text generation," "text classification," "question answering," "machine translation," "sentiment analysis," "information extraction," and "NLP applications." The search strategy was adapted to the syntax of individual databases while maintaining the conceptual equivalence of the queries. Additional searches combined NLP and LLM terms with application domains such as healthcare, education, finance, customer service, information retrieval, social media, legal services, and scientific research.

2.4 Eligibility Criteria

Studies were considered eligible when they made a substantive contribution to the understanding, development, evaluation, or application of NLP methods, particularly where the work addressed contemporary transformer-based or large language model approaches. Peer-reviewed journal articles, conference papers, systematic reviews, meta-analyses, and influential methodological studies were considered. Studies were included when they investigated NLP architectures, model training, fine-tuning, prompting, retrieval-augmented generation, evaluation, linguistic tasks, or practical applications of NLP and LLMs. Relevant high-impact technical papers describing foundational models were also considered because of their importance to understanding the methodological transition represented by LLMs.

Studies were excluded when they were unrelated to NLP, focused exclusively on non-language machine-learning applications, lacked sufficient methodological or conceptual information, or duplicated another publication. Non-scholarly webpages, marketing materials, opinion pieces, news articles, and publications without sufficient academic or technical substance were excluded from the core synthesis. Where multiple publications reported substantially overlapping findings from the same research project, the most comprehensive or methodologically informative version was prioritized.

2.5 Study Selection and Screening

The study-selection process consisted of identification, duplicate removal, title and abstract screening, full-text assessment, and final inclusion. Retrieved records were initially examined to remove duplicate publications and clearly irrelevant records. Titles and abstracts were subsequently screened against the predefined eligibility criteria. Publications that appeared potentially relevant were retrieved for full-text assessment, during which their methodological relevance, conceptual contribution, and relationship to the review objectives were examined in greater detail. The final corpus consisted of studies that directly contributed evidence concerning the evolution of NLP methods, LLM architectures and techniques, major NLP tasks, application domains, evaluation practices, and limitations. The screening procedure was conceptually aligned with established systematic-review reporting principles, including the use of a transparent identification and selection process.

2.6 Data Extraction and Organization

Information was systematically extracted from the selected publications using a structured analytical framework. Extracted information included publication characteristics, research objectives, NLP methodology, model or architecture used, training or adaptation strategy, dataset characteristics, NLP task, application domain, evaluation approach, principal findings, reported advantages, and identified limitations. For studies focusing on LLMs, additional attention was given to model scale, transformer architecture, pretraining, supervised fine-tuning, instruction tuning, prompting, parameter-efficient adaptation, retrieval augmentation, and human or automated evaluation. The extracted information was organized thematically to facilitate comparison between conventional NLP methods and contemporary LLM-based approaches.

2.7 Quality Assessment

The methodological and scholarly quality of the included literature was assessed according to the nature of each publication. Empirical studies were examined in terms of clarity of research objectives, appropriateness of methodology, transparency of datasets and experimental procedures, validity of evaluation measures, reproducibility, and adequacy of reported results. Review articles were assessed according to the transparency of their search and selection procedures, conceptual coherence, breadth of coverage, and quality of synthesis. Technical papers describing NLP and LLM architectures were examined for methodological clarity, empirical validation, and significance to subsequent developments in the field. Greater analytical attention was given to studies that provided transparent experimental evidence, established benchmark comparisons, clearly defined datasets, and reproducible evaluation procedures.

2.8 Data Synthesis and Analysis

A thematic synthesis approach was used to integrate findings across the selected literature. The analysis was organized around the historical and methodological progression of NLP from rule-based and statistical approaches to machine learning, neural networks, recurrent architectures, transformers, pretrained language models, and contemporary LLMs. The synthesis subsequently examined major NLP tasks, including text classification, sentiment analysis, named-entity recognition, information extraction, machine translation, summarization, question answering, dialogue systems, text generation, and information retrieval. Particular attention was given to how LLMs have altered conventional NLP workflows through large-scale pretraining, transfer learning, instruction following, prompting, in-context learning, retrieval augmentation, and generative capabilities.

The literature was further synthesized according to application domains to identify recurring patterns in the use of NLP and LLMs. Comparative analysis considered the methodological strengths and limitations reported across different approaches, including computational requirements, data dependence, generalization, interpretability, robustness, hallucination, bias, privacy, security, and evaluation challenges. Rather than treating individual model architectures as isolated developments, the synthesis interpreted them as part of the broader evolution of NLP toward increasingly general-purpose language technologies.

2.9 Conceptual Framework for the Review

The review was structured around three interconnected dimensions: NLP methods, NLP tasks, and application domains. The methods dimension incorporated traditional computational linguistics, statistical machine learning, deep-learning architectures, transformer-based models, pretrained language models, and LLM-based techniques. The task dimension incorporated language understanding and generation activities such as classification, translation, extraction, summarization, question answering, dialogue, and retrieval. The application dimension considered the deployment of these methods across sectors including healthcare, education, business, finance, public services, research, media, and human-computer interaction. This framework enabled the review to identify methodological continuities and changes while examining how advances in LLMs have expanded the scope of NLP applications.

2.10 Reporting of the Review

The findings were reported through a narrative and thematic synthesis rather than a statistical meta-analysis because the included literature was heterogeneous in terms of research designs, datasets, models, NLP tasks, evaluation metrics, and

application contexts. The review therefore emphasized patterns, methodological developments, comparative findings, and recurring challenges across studies. Where appropriate, influential models and methodological innovations were discussed in relation to their contributions to the evolution of NLP. The overall reporting process was designed to provide a coherent account of how NLP has evolved into the contemporary LLM-driven paradigm and to identify methodological gaps and research priorities for future investigation.

3. Findings and Discussion

The systematic review indicates that the development of Natural Language Processing (NLP) has progressed through a sequence of methodological transformations, moving from manually constructed linguistic rules and statistical representations toward neural architectures, Transformer-based pre-trained models, and increasingly general-purpose large language models (LLMs). Across the reviewed literature, the central change is not simply an increase in model size but a reorganization of the NLP development paradigm around large-scale pre-training, transfer, prompting, instruction following, retrieval, and parameter-efficient adaptation. Recent surveys similarly characterize contemporary LLM-based NLP in terms of parameter-frozen and parameter-tuning approaches, while emphasizing that traditional NLP tasks have increasingly been incorporated into unified language-model frameworks.

3.1 Evolution of Natural Language Processing Methods in the Era of Large Language Models

3.1.1 From Rule-Based and Statistical NLP to Neural Language Models

The findings demonstrate that contemporary LLM-based NLP is the outcome of a cumulative methodological evolution rather than a complete departure from earlier NLP approaches. Early NLP systems relied substantially on manually designed grammatical rules, dictionaries, symbolic representations, and expert-defined linguistic patterns. Statistical NLP subsequently shifted the field toward probabilistic modelling, with approaches such as n-gram language models, hidden Markov models, conditional random fields, and probabilistic classifiers learning patterns from increasingly large corpora. These methods reduced dependence on manually specified rules but remained constrained by sparse representations, limited contextual windows, feature engineering requirements, and difficulties representing long-range linguistic dependencies.

The emergence of distributed word representations and neural networks substantially altered this situation. Neural language models learned continuous representations in which words with related usage patterns could occupy related regions of a representation space. Recurrent neural networks (RNNs) subsequently introduced mechanisms for processing sequences while maintaining information from previous tokens. However, conventional recurrent architectures encountered difficulties when information needed to be retained across long sequences because of vanishing and decaying gradients (Ren, 2024). Long Short-Term Memory (LSTM) networks addressed this limitation through gated memory mechanisms designed to preserve information across extended time intervals. Raza (2025) foundational work demonstrated that LSTM could bridge substantially longer temporal dependencies than conventional recurrent learning mechanisms.

The review further identifies sequence-to-sequence learning as an important intermediate stage between recurrent NLP and modern Transformer models. Encoder-decoder architectures enabled one sequence to be transformed into another, making them particularly influential for machine translation, summarization, dialogue generation, and related generation tasks. Rather than designing a separate architecture for each input-output sequence relationship, sequence-to-sequence learning established a more general framework in which the model learned mappings directly from data. Attention mechanisms were subsequently introduced to alleviate the limitation of compressing an entire input sequence into a single fixed representation. These developments established many of the conceptual foundations subsequently incorporated into Transformer-based systems.

3.1.2 Emergence of Transformer-Based Architectures

A major finding of the review is that the Transformer represents the most consequential architectural transition in the development of modern NLP. Vaswani et al. introduced the Transformer as an architecture based entirely on attention mechanisms rather than recurrence or convolution. The architecture demonstrated strong results in machine translation while offering substantially greater parallelization during training than recurrent sequence models. This architectural characteristic became particularly important as NLP models expanded in parameter size and training-data volume.

The central methodological contribution of the Transformer is self-attention, which enables a model to calculate relationships between elements of a sequence without processing tokens strictly one at a time. This permits representations to incorporate information from different parts of a sequence and makes the architecture particularly suitable for large-scale parallel training. The review shows that the resulting architecture subsequently supported several distinct model families, including encoder-based models such as BERT, decoder-based generative models, and encoder-decoder systems such as T5.

BERT illustrates how the Transformer transformed contextual representation learning. Instead of treating a word as having a single context-independent representation, BERT generated deep bidirectional representations by jointly conditioning on left and right context. Its pre-training and subsequent fine-tuning produced strong results across question answering, language inference, and other NLP tasks without requiring substantial task-specific architectural changes. This finding was important because it demonstrated that a common pre-trained representation could support multiple downstream applications.

The review also identifies transfer learning as a defining methodological principle of the Transformer era. T5, for example, formulated diverse NLP problems within a unified text-to-text framework and demonstrated strong performance across summarization, question answering, classification, and other tasks (Scherbakov, 2025). Consequently, NLP research increasingly moved away from constructing independent architectures and toward adapting general pre-trained models to particular tasks. This transition also changed the research emphasis from feature engineering and architecture design toward corpus construction, pre-training objectives, model scaling, adaptation strategies, prompting, and evaluation.

3.1.3 Shift from Task-Specific Models to General-Purpose Language Models

The reviewed literature reveals a further transition from task-specific models toward general-purpose foundation models capable of supporting numerous NLP functions within a common architecture. Earlier neural NLP systems were normally optimized for a particular task, such as classification, translation, tagging, or question answering. Transformer-based pre-training weakened this separation by allowing a single model to acquire general linguistic representations before being adapted to downstream applications. The emergence of large-scale autoregressive models extended this paradigm further by demonstrating that sufficiently large models could perform tasks through natural-language instructions or examples rather than requiring a separately trained output architecture for every task.

GPT-3 provided particularly influential evidence for this shift. Its experiments demonstrated zero-shot, one-shot, and few-shot in-context learning, in which task demonstrations could be supplied within the input context rather than changing model parameters. The findings showed that increasing model scale could improve the ability to adapt to tasks from contextual examples. This changed the operational meaning of a language model: rather than functioning solely as a component embedded inside a task-specific NLP system, it could increasingly operate as a general-purpose interface for multiple language tasks.

The review also finds that instruction tuning strengthened this general-purpose paradigm. FLAN demonstrated that instruction-tuning a large pre-trained model across numerous tasks could substantially improve zero-shot performance on previously unseen tasks (Sarker, 2024). Subsequent work showed that scaling the number of instruction-tuning tasks, model size, and reasoning-oriented training examples could further improve performance across diverse benchmarks. Thus, contemporary NLP increasingly distinguishes between the foundation model itself and the methods used to specialize or steer it.

This shift does not mean that task-specific models have become obsolete. The evidence instead indicates a coexistence between general-purpose LLMs and specialized models. Domain-specific systems may remain advantageous where latency, interpretability, computational cost, privacy, or highly controlled outputs are important. Nevertheless, the general trend is toward reusable language foundations that can be adapted through prompting, fine-tuning, instruction tuning, retrieval, or parameter-efficient techniques rather than independently training a complete model for every NLP task.

3.2 Core Methods and Techniques Used in Large Language Model-Based NLP

3.2.1 Pre-Training, Fine-Tuning, and Instruction Tuning

Large-scale pre-training emerged as the foundational method underlying contemporary LLM-based NLP. During pre-training, models learn from very large collections of textual data using self-supervised objectives, most commonly predicting masked or subsequent tokens. The review indicates that this stage provides broad linguistic and world-knowledge representations that can subsequently be transferred to downstream applications. The importance of this approach was demonstrated by models such as BERT and T5, which showed that general pre-trained representations could be adapted to a broad range of NLP tasks.

Fine-tuning subsequently provides a mechanism for adapting a pre-trained model to a particular task or domain. Instead of learning all linguistic representations from a comparatively small task-specific dataset, the model begins with knowledge acquired during pre-training and adjusts its parameters using downstream examples. This approach substantially reduces the amount of task-specific training required and helps explain the rapid expansion of transfer learning in NLP.

The findings further demonstrate the increasing importance of instruction tuning. Instruction tuning involves supervised training on instruction-output pairs so that models learn to interpret natural-language task specifications and generate responses that follow those specifications. Recent research describes instruction tuning as a bridge between the next-token prediction objective

used during pre-training and the user's objective of obtaining useful, instruction-following outputs (Ogayemi, 2026). Empirical evidence from FLAN showed substantial improvements in zero-shot generalization following instruction tuning, while later scaling experiments reported improvements across numerous benchmarks and model families.

The review therefore identifies three interconnected stages: pre-training establishes general language capabilities, fine-tuning specializes those capabilities, and instruction tuning improves the model's ability to apply them in response to natural-language task specifications. Domain adaptation can further specialize these capabilities for fields such as medicine, law, finance, education, and scientific communication. However, adaptation quality remains dependent on training-data quality, domain coverage, and the degree to which specialized data represents the target population and language variety.

3.2.2 Prompt Engineering and In-Context Learning

Prompting represents one of the most important methodological changes associated with LLM-based NLP because it allows researchers and users to influence model behaviour without necessarily modifying model parameters. The review identifies zero-shot prompting, one-shot prompting, few-shot prompting, instruction prompting, and reasoning-oriented prompting as major approaches. Prompt-based learning differs from conventional supervised learning because the task is expressed through a textual context that the language model processes as part of the input. Surveys of prompting research consequently describe prompting as a mechanism through which pre-trained models can perform new tasks with little or no task-specific labelled data.

Zero-shot prompting is particularly significant because it allows a model to receive a task description without explicit examples. One-shot and few-shot prompting provide one or several demonstrations, respectively, enabling the model to infer the expected mapping or response format from contextual examples. GPT-3 provided influential empirical evidence that larger language models could perform a wide range of tasks using such in-context demonstrations. The review suggests that this capability has reduced the dependence on conventional task-specific training for many experimental and practical NLP applications.

Prompt design nevertheless remains an important source of performance variation. Changes in task wording, demonstrations, output constraints, ordering of examples, and contextual information can affect model responses. Consequently, prompt engineering has developed into a methodological discipline concerned with constructing inputs that communicate task requirements clearly and consistently. Chain-of-thought prompting represents a related development in which models are provided with examples containing intermediate reasoning steps. Research by Wei et al. reported improvements on arithmetic, commonsense, and symbolic reasoning tasks when sufficiently large models were prompted with chain-of-thought demonstrations (Mathieu, 2025). However, the review indicates that prompting should not be treated as universally reliable: performance can be sensitive to prompt formulation, task characteristics, language, model scale, and evaluation methodology.

3.2.3 Retrieval-Augmented Generation and Parameter-Efficient Adaptation

Retrieval-Augmented Generation (RAG) emerged in the reviewed literature as an important response to one of the fundamental limitations of purely parametric language models: information encoded in model parameters may be incomplete, difficult to update, or insufficiently precise for knowledge-intensive tasks. RAG combines a generative language model with an external retrieval mechanism, allowing relevant documents or passages to be retrieved and incorporated into the generation process. Hou (2024) demonstrated that RAG could improve performance on knowledge-intensive NLP tasks and produce more specific, diverse, and factual language than a comparable parametric-only baseline.

The evidence indicates that RAG is particularly useful in settings where information changes frequently or where domain-specific knowledge is not adequately represented in the original training corpus. Examples include organizational knowledge bases, scientific literature, institutional documents, technical manuals, and question-answering systems over specialized collections. RAG also provides an architectural pathway toward improved traceability because retrieved documents can be exposed as supporting evidence. Nevertheless, retrieval does not automatically guarantee factual correctness; irrelevant, incomplete, outdated, or poorly ranked documents can still propagate errors into generated responses.

Parameter-efficient adaptation provides a complementary solution to the computational costs associated with full model fine-tuning. Adapters and Low-Rank Adaptation (LoRA) allow a large pre-trained model to retain most of its original parameters while training only a comparatively small number of additional parameters. Hu et al. reported that LoRA can dramatically reduce the number of trainable parameters and GPU memory requirements while achieving performance comparable to or better than conventional full fine-tuning in several evaluated settings (Omar, 2025). The review therefore identifies RAG and parameter-efficient adaptation as complementary strategies: RAG primarily expands a model's access to external information, whereas methods such as LoRA make it more computationally practical to adapt model behaviour to particular domains or tasks.

3.3 Applications of Large Language Models Across NLP Tasks

3.3.1 Text Classification, Information Extraction, and Sentiment Analysis

The reviewed evidence demonstrates that LLMs have increasingly been applied to conventional language-understanding tasks, including sentiment analysis, topic classification, intent detection, named-entity recognition (NER), relation extraction, and text categorization. Earlier systems typically required task-specific labelled datasets and specialized architectures. LLMs can instead formulate many of these tasks as natural-language instructions, for example asking a model to identify the sentiment of a review, classify a document into predefined categories, or extract entities and relationships from a passage.

This development is particularly significant for low-data and multi-task settings. Prompting permits a single model to perform several classification or extraction functions without separately training an independent model for every task. Fine-tuned or instruction-tuned models can also be constrained to produce structured outputs, making them useful for extracting names, organizations, locations, events, concepts, or relationships from unstructured text. The broader literature confirms that pre-trained Transformer models substantially altered conventional NLP by providing transferable representations for classification and language-understanding tasks. BERT, for example, achieved strong results across multiple benchmarks while requiring comparatively limited task-specific architectural modification.

However, the findings also indicate that LLM performance on structured tasks should not automatically be equated with unrestricted superiority over specialized models. Traditional supervised systems can remain highly competitive where training data are abundant, label definitions are stable, output formats are strictly constrained, and computational efficiency is important. LLMs may also produce plausible but incorrectly extracted entities or classifications when instructions are ambiguous (Hou, 2024). Consequently, their principal methodological contribution is flexibility: the same model can address multiple classification and extraction tasks, rather than eliminating the need for task-specific evaluation.

In sentiment analysis, LLMs can additionally interpret nuanced linguistic phenomena such as contextual polarity, implicit sentiment, mixed opinions, and domain-specific expressions. For example, a statement such as "the update is ambitious but unusable in practice" contains both positive and negative evaluative components that may be difficult to capture with simple lexical polarity rules. LLMs can use broader context to interpret such expressions, although their outputs may vary with prompting and model configuration. The review therefore identifies contextual understanding and task flexibility as major advantages while retaining concerns about consistency, reproducibility, and calibration.

3.3.2 Machine Translation, Text Summarization, and Question Answering

Machine translation represents one of the historical application areas in which neural sequence modelling and Transformer architectures demonstrated particularly strong effects. The original Transformer was evaluated directly on English-German and English-French translation and achieved strong translation performance while enabling substantially greater training parallelization than recurrent alternatives. Contemporary LLMs extend this development by treating translation as one component of a broader multilingual language capability. Rather than maintaining entirely separate architectures for every translation direction, multilingual and general-purpose models can perform translation through shared representations and natural-language instructions.

The review also identifies text summarization as a major application of generative LLMs. Transformer-based encoder-decoder systems established a strong foundation for abstractive summarization, while modern generative models can produce summaries of different lengths, styles, and levels of technicality according to instructions. T5's unified text-to-text formulation, for example, demonstrated strong results across summarization and other language tasks. LLMs can therefore move beyond extracting individual sentences toward generating condensed representations of documents while preserving their central meaning.

Question answering has similarly benefited from the combination of contextual representations, generation, and retrieval. BERT demonstrated strong extractive question-answering performance, while subsequent generative and retrieval-augmented approaches expanded the range of questions that could be addressed. RAG is particularly relevant to open-domain question answering because the system can retrieve external passages before generating an answer (Hadi, 2023). The original RAG research reported state-of-the-art results on several open-domain QA tasks and showed advantages over purely parametric sequence-to-sequence models.

Despite these advances, the findings indicate that translation, summarization, and question answering remain sensitive to factuality and evaluation problems. A fluent translation can contain subtle semantic errors; a coherent summary can omit important information; and a grammatically convincing answer can contain unsupported claims. Consequently, contemporary

evaluation increasingly needs to consider factual consistency, semantic adequacy, source attribution, multilingual robustness, and domain-specific correctness rather than relying exclusively on surface-level similarity metrics.

3.3.3 Text Generation, Dialogue Systems, and Conversational AI

The strongest transformation identified in the review concerns open-ended text generation and conversational interaction. Earlier dialogue systems commonly depended on manually designed response rules, retrieval-based response selection, or relatively narrow task-specific neural models. LLMs have expanded this paradigm by generating responses dynamically from conversational context. This allows the same underlying model to perform functions such as question answering, explanation, rewriting, brainstorming, dialogue management, summarization, and content generation.

The emergence of in-context learning and instruction tuning has been particularly important for conversational AI. GPT-3 demonstrated that large autoregressive models could adapt to new tasks from contextual examples, while instruction-tuned models subsequently improved their ability to follow natural-language requests. Recent research further conceptualizes instruction tuning as a mechanism for improving both usability and controllability by aligning model outputs with explicit human instructions. These developments have enabled conversational systems to maintain broader contextual continuity and respond flexibly to heterogeneous user requests.

The findings nevertheless indicate that linguistic fluency should be distinguished from factual reliability and genuine understanding. LLMs can generate grammatically coherent and contextually appropriate responses while still producing inaccurate information, unsupported assertions, or inconsistent statements. This limitation is particularly important in open-domain conversational systems because the system is not restricted to a predefined set of verified responses (Zubiaga, 2024). Retrieval augmentation can partially address knowledge limitations by supplying external evidence, while instruction tuning and structured prompting can improve adherence to task requirements. However, neither technique completely eliminates the possibility of erroneous generation.

3.4 Performance, Benefits, and Limitations of LLM-Based NLP

3.4.1 Accuracy, Generalization, and Contextual Understanding

The systematic review indicates that large language models (LLMs) have substantially expanded the performance capabilities of natural language processing by providing a common language representation that can be transferred across multiple tasks. Unlike conventional NLP systems that frequently require task-specific architectures, manually engineered features, or separately labelled datasets, LLMs can perform classification, summarization, translation, question answering, information extraction, sentiment analysis, and text generation through a shared pretrained model. The literature consistently associates this development with the use of large-scale pretraining followed by prompting, instruction tuning, or parameter-efficient adaptation. Transformer-based architectures have been particularly important because self-attention enables models to consider relationships among words and phrases across relatively long contexts, improving their ability to interpret linguistic dependencies that may be difficult for traditional bag-of-words, n-gram, or feature-engineered approaches to capture.

The reviewed evidence further suggests that LLMs demonstrate strong transferability between tasks and domains. A model initially trained on broad collections of text can often be adapted to a new task with comparatively limited task-specific data. This represents an important departure from conventional supervised NLP, where performance is often strongly dependent on the availability of labelled examples for each individual application. For example, an instruction-tuned LLM may be able to summarize a research article, classify a customer complaint, extract clinical concepts from a medical narrative, or answer questions about a legal document without requiring an entirely new model architecture for each task. The finding supports the broader argument in previous NLP research that pretrained language representations can substantially reduce the dependence on task-specific feature engineering and labelled data.

Contextual understanding is another major benefit identified in the reviewed literature. LLMs are capable of using surrounding linguistic information to distinguish meanings that depend on context, discourse, and relationships between entities. This is particularly valuable for ambiguous words, conversational interactions, long-form documents, and complex instructions. For instance, the meaning of the word “bank” can be interpreted differently depending on whether a sentence concerns financial services or a river. Similarly, a question-answering system can use preceding conversational turns to interpret pronouns and omitted information. However, the review also indicates that apparent contextual understanding should not automatically be equated with human-like comprehension. Models can perform impressively on contextual tasks while still failing on questions requiring precise world knowledge, complex logical relationships, temporal reasoning, or robust understanding of implicit assumptions.

The evidence also shows that generalization remains uneven. LLMs may perform strongly on familiar tasks represented in their training data but experience substantial performance degradation when confronted with distribution shifts, unusual inputs, specialized terminology, novel situations, or poorly represented populations. This limitation is particularly important because benchmark performance can overstate real-world capability when test datasets closely resemble training distributions (Min, 2023). Consequently, the literature increasingly emphasizes evaluation under realistic conditions, including adversarial inputs, domain shifts, long-context tasks, and out-of-distribution examples. The findings therefore indicate that LLMs have improved the breadth and flexibility of NLP substantially, but generalization remains dependent on model architecture, training data, prompting strategy, domain adaptation, and evaluation conditions.

Multilingual performance represents a similarly mixed finding. Major LLMs increasingly support numerous languages and can facilitate translation, cross-lingual information retrieval, multilingual question answering, and text generation. Nevertheless, the review identifies a persistent imbalance between high-resource languages and low-resource languages. Languages with extensive digital resources, standardized writing systems, and substantial representation in training corpora generally receive better model support than languages with limited digitized text. Consequently, multilingual capability should not be interpreted as equivalent performance across languages. For many African, Indigenous, and other low-resource languages, problems involving limited corpora, orthographic variation, code-switching, morphological complexity, and insufficient evaluation datasets remain significant. This finding reinforces earlier research demonstrating that the benefits of pretrained language models are distributed unevenly according to the quantity and quality of available linguistic resources.

3.4.2 Hallucination, Bias, Toxicity, and Reliability

Despite their impressive linguistic capabilities, the literature identifies hallucination as one of the most persistent limitations of LLM-based NLP. Hallucination refers to the generation of information that is presented fluently and confidently but is unsupported, inaccurate, or fabricated. This problem is particularly consequential in applications where factual precision is essential. For example, an LLM may generate a plausible-looking academic citation that does not exist, attribute an incorrect finding to a real study, or provide an apparently authoritative answer to a question for which the model lacks reliable evidence. The review therefore finds that linguistic fluency and factual reliability are distinct dimensions of model performance. A response can be grammatically coherent and contextually appropriate while nevertheless containing substantive factual errors.

The reviewed studies further indicate that hallucination can arise from several interacting factors, including incomplete or noisy training data, probabilistic generation, insufficient grounding in external knowledge, ambiguous prompts, and weaknesses in reasoning and retrieval. Retrieval-augmented generation (RAG), tool use, external knowledge bases, citation mechanisms, and post-generation verification have consequently emerged as important approaches for improving factuality. Nevertheless, these techniques do not eliminate the problem. Errors can still occur when retrieved information is incomplete, irrelevant, outdated, or incorrectly interpreted (Raiaan, 2024). The findings therefore suggest that high-stakes NLP applications require verification mechanisms rather than treating model-generated text as inherently trustworthy (Adeniyan, 2025).

Bias is another major limitation identified across the literature. Because LLMs learn statistical patterns from large-scale human-generated datasets, they can reproduce or amplify social, cultural, demographic, and linguistic biases contained in those data. Bias can appear in generated descriptions, occupational associations, sentiment classification, toxicity detection, translation, and conversational responses. For example, a model may associate particular occupations or behavioural characteristics disproportionately with certain demographic groups. Such patterns are especially problematic when NLP systems are used for recruitment, education, healthcare, law, public services, or other contexts in which unequal treatment can have consequential effects.

The review also identifies toxicity as an important concern. LLMs can reproduce offensive, discriminatory, abusive, or otherwise harmful language when such patterns are present in their training data or elicited through particular prompts. Safety alignment and content filtering have reduced some undesirable outputs, but the evidence indicates that these mechanisms can be inconsistent across languages, cultural contexts, and types of prompts. A model may therefore demonstrate different safety behaviour depending on the language in which a request is presented. This problem is particularly relevant to multilingual NLP because safety resources and evaluation datasets are often concentrated in high-resource languages.

Reliability also remains constrained by inconsistent reasoning. LLMs can sometimes produce correct explanations for relatively difficult problems but fail on simpler variations of the same problem, particularly when wording or contextual conditions change. This inconsistency challenges the assumption that strong benchmark performance necessarily reflects stable reasoning competence (Ren, 2024). The literature therefore increasingly distinguishes between task accuracy, factuality, robustness, calibration, and reasoning consistency. These findings indicate that evaluating LLMs solely through aggregate benchmark scores provides an incomplete picture of their reliability in practical NLP applications.

3.4.3 Computational Cost, Data Requirements, and Environmental Impact

The systematic review identifies computational requirements as one of the most significant structural limitations of contemporary LLM development. Large-scale models require substantial computational infrastructure during pretraining, including high-performance accelerators, large memory capacity, high-speed networking, extensive storage, and sophisticated distributed-computing systems. The computational burden increases with model size, training-token volume, sequence length, and the number of training iterations. Consequently, cutting-edge LLM development has increasingly become concentrated among organizations and research groups with access to substantial financial and computational resources.

Training is not the only source of computational cost. Inference can also be resource intensive, particularly when models contain billions of parameters or are required to process long contexts. Applications involving large numbers of users may therefore generate considerable recurring infrastructure costs. Latency can become another practical concern because complex models require more computation to produce responses. These constraints are especially relevant for real-time applications such as conversational systems, intelligent tutoring, clinical decision-support interfaces, and customer-service platforms where rapid responses are expected.

The review also finds that data requirements constitute a critical component of LLM development. Large models depend on enormous text corpora, but increasing the quantity of training data does not automatically guarantee improved performance. Data quality, diversity, representativeness, duplication, language distribution, copyright status, and temporal relevance can all influence model behaviour. Training corpora may contain misinformation, stereotypes, personal information, outdated knowledge, or systematic underrepresentation of particular linguistic communities (Raza, 2025). Thus, the central challenge is increasingly shifting from simply acquiring more data toward developing higher-quality, better-curated, more representative, and appropriately governed datasets.

Environmental implications have consequently received greater attention. Training and operating large models consume electricity, while associated infrastructure also involves cooling, hardware production, storage, and data-centre resources. The environmental burden varies considerably according to model architecture, hardware efficiency, energy source, training duration, inference volume, and geographic location. The findings therefore caution against treating model size as an uncomplicated indicator of progress. Smaller models, knowledge distillation, quantization, pruning, parameter-efficient fine-tuning, sparse architectures, and optimized inference can provide alternative pathways for improving efficiency. This creates an important research direction in which computational effectiveness and environmental sustainability become explicit evaluation criteria alongside accuracy.

3.5 Emerging Trends and Future Directions in NLP Research

3.5.1 Multimodal, Multilingual, and Domain-Specific Language Models

The review identifies multimodality as a major emerging direction in NLP. Earlier language models primarily operated on text, whereas contemporary systems increasingly integrate text with images, audio, video, documents, and other information modalities. Multimodal models can therefore perform tasks such as describing images, interpreting charts, answering questions about documents, analysing visual-textual information, and combining spoken and written language. This development expands NLP from isolated text processing toward more comprehensive systems for interpreting heterogeneous information. In education, for example, multimodal systems can potentially interpret written questions alongside diagrams; in medicine, they can integrate textual clinical information with medical images; and in scientific research, they can combine textual descriptions with tables, graphs, and other representations (Watara et al., 2025).

Multilingual and cross-lingual NLP also remain important areas of development. The reviewed evidence points toward increasing efforts to build models that support a broader range of languages and linguistic varieties rather than concentrating primarily on English and other high-resource languages. These efforts include multilingual pretraining, cross-lingual transfer, translation-based data augmentation, community-developed datasets, and language-specific adaptation. However, the persistence of resource disparities indicates that simply adding languages to a model does not ensure equitable performance (Scherbakov, 2025). Future multilingual research must therefore consider language-specific morphology, sociolinguistic variation, dialects, code-switching, culturally specific meanings, and locally appropriate evaluation criteria.

Domain-specific LLMs represent another important trend. General-purpose models offer broad capabilities, but specialized models can be adapted to terminology, documentation conventions, regulatory requirements, and reasoning patterns within particular fields. Medical LLMs, for example, can be adapted to clinical notes and biomedical literature; legal models can process statutes, judgments, and legal documents; educational models can support tutoring, feedback, assessment, and instructional design; and scientific models can assist with literature analysis and technical knowledge discovery. The findings nevertheless indicate that domain adaptation introduces its own challenges, particularly concerning data privacy, expert validation,

hallucination, accountability, and the consequences of incorrect outputs. Specialized models should therefore be evaluated not only according to general language performance but also according to domain-specific safety and validity requirements.

3.5.2 Responsible, Explainable, and Human-Centred NLP

Responsible NLP has emerged as a central research priority as LLMs move from experimental environments into real-world applications. The literature increasingly emphasizes fairness, transparency, privacy, accountability, safety, and human oversight throughout the model lifecycle. Responsible development begins with data collection and curation but extends through model training, evaluation, deployment, monitoring, and eventual retirement. This lifecycle perspective is important because harmful outcomes can arise not only from the model itself but also from inappropriate data, system design, user interaction, deployment context, or organizational practices.

Explainability is particularly important in high-stakes applications. Users may need to understand why an NLP system generated a particular classification, recommendation, summary, or answer. However, the review shows that explaining an LLM remains challenging because its internal representations and generation processes are highly complex. Techniques such as attention analysis, feature attribution, probing, chain-of-thought alternatives, retrieval citations, confidence estimation, and post-hoc explanation can provide useful information, but none should automatically be treated as a complete representation of the model's internal reasoning. Consequently, future explainability research should focus on explanations that are meaningful to the intended users and empirically validated rather than merely plausible-looking narratives.

Human-centred NLP places users and affected communities at the centre of system development. This involves participatory design, user feedback, human-in-the-loop decision-making, accessibility, culturally sensitive evaluation, and mechanisms for contesting or correcting model outputs. In educational and healthcare contexts, for example, LLMs may be most appropriately positioned as decision-support or augmentation technologies rather than unquestioned replacements for teachers, clinicians, researchers, or other professionals (Sarker, 2024). The evidence therefore supports a shift from evaluating whether an LLM can complete a task independently toward assessing whether human–AI collaboration produces reliable, equitable, and beneficial outcomes.

Privacy represents an additional dimension of responsible NLP. Large language systems may process sensitive documents, conversations, educational records, clinical information, or proprietary organizational data. Risks can arise from inappropriate data collection, retention, model memorization, insecure deployment, or unintended disclosure. Privacy-preserving training, controlled data access, anonymization, secure deployment, and stronger governance mechanisms are therefore becoming increasingly important. The review suggests that responsible NLP requires technical safeguards to be combined with institutional policies, regulatory compliance, documentation, auditing, and clearly defined human accountability.

3.5.3 Future Research Priorities and Research Gaps

A major research gap identified by the review concerns the need for more robust and comprehensive evaluation frameworks. Existing benchmarks remain useful for measuring progress, but they do not always reflect the complexity of real-world NLP. Future evaluation should therefore assess multiple dimensions simultaneously, including accuracy, factuality, robustness, fairness, toxicity, calibration, interpretability, efficiency, privacy, and human usefulness. Evaluations should also incorporate distribution shifts, adversarial conditions, multilingual inputs, domain-specific scenarios, and long-context tasks. Such multidimensional evaluation would provide a more realistic understanding of model capabilities and limitations than single-score benchmark comparisons.

Factuality and hallucination remain particularly important unresolved problems. Future research should develop stronger grounding methods, reliable retrieval systems, uncertainty estimation, source verification, citation validation, and automatic and human factuality assessment. An important gap concerns the ability of models to recognize when they do not possess sufficient information to answer a question. Rather than maximizing response generation, reliable NLP systems should be capable of expressing uncertainty and directing users toward appropriate verification mechanisms when confidence is insufficient.

Interpretability and transparency also require further investigation. Although numerous explanation techniques have been proposed, there remains limited consensus concerning what constitutes a sufficiently faithful explanation of an LLM's behaviour. Future research should distinguish between explanations that merely make outputs understandable to users and explanations that accurately reflect the mechanisms producing those outputs (Ogayemi, 2026). Greater transparency regarding training data characteristics, model development procedures, evaluation conditions, limitations, and known failure modes would also improve reproducibility and accountability.

Low-resource language processing constitutes another substantial gap. The review demonstrates that improvements in mainstream multilingual models do not necessarily translate into equivalent benefits for underrepresented languages. Future research should prioritize community-created corpora, culturally appropriate annotation practices, language documentation, open datasets, multilingual transfer, synthetic-data validation, and evaluation benchmarks designed specifically for low-resource settings. This is particularly important for languages that have limited digital representation and for which conventional data-intensive approaches are difficult to apply.

Efficiency is likewise likely to become a defining research priority. Instead of continuously increasing model size, future NLP research is expected to place greater emphasis on smaller specialized models, parameter-efficient adaptation, quantization, pruning, knowledge distillation, sparse architectures, retrieval-based systems, and hardware-aware optimization. These approaches may reduce computational and financial barriers while making advanced NLP more accessible to universities, smaller organizations, public institutions, and researchers in resource-constrained settings.

Finally, the review identifies a need for longitudinal and real-world evaluation. Many LLM studies assess models at a particular point in time using static benchmarks, whereas deployed systems operate in changing linguistic, social, technological, and informational environments. Model performance can change as prompts, users, knowledge sources, software versions, and societal contexts evolve (Mathieu, 2025). Longitudinal studies are therefore needed to examine reliability, bias, factuality, user behaviour, and unintended consequences over extended periods. Standardized reporting protocols and interoperable benchmarks could facilitate meaningful comparisons across models and applications. Overall, the findings indicate that the next phase of NLP research should move beyond demonstrating increasingly capable language generation toward developing systems that are accurate, efficient, transparent, culturally inclusive, robust, safe, and demonstrably useful in real-world human contexts.

4. Conclusion

This systematic review concludes that Natural Language Processing (NLP) has undergone a fundamental transformation with the emergence of Large Language Models (LLMs), shifting the field from predominantly task-specific, feature-engineered and supervised approaches toward increasingly general-purpose, context-aware and instruction-following language systems. The reviewed evidence indicates that LLMs have substantially expanded NLP capabilities across text classification, information extraction, machine translation, summarization, question answering, dialogue systems, sentiment analysis, information retrieval, content generation and domain-specific applications, while approaches such as prompting, in-context learning, fine-tuning and retrieval-augmented generation have enabled greater adaptability across tasks and contexts. At the same time, the findings demonstrate that LLM-based NLP is not without limitations, particularly regarding hallucination, factual reliability, bias, interpretability, computational and financial costs, privacy, data dependence, and challenges in evaluating outputs consistently across languages and domains. Consequently, the future development of NLP requires a balanced emphasis on model capability, reliability, transparency, efficiency and responsible deployment. Continued research should therefore focus on improving factual grounding, multilingual and low-resource language performance, domain adaptation, evaluation frameworks, human oversight and resource-efficient architectures so that LLM-based NLP can deliver broader and more dependable benefits while addressing its technical, ethical and societal challenges.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Adeniyan T. (2025). Cyber Risk Assessment Frameworks for Protecting U.S. Critical Infrastructure Against Emerging Threats. Iconic Research And Engineering Journals, 8(8).
[2] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, *3*, 312.
[3] Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., ... & Wang, H. (2024). Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, *33*(8), 1-79.
[4] Mathieu, B., Anas, N., Thomas, P., Robert, P., & Samir, L. (2025). Exploring the applications of natural language processing and language models for production, planning, and control activities of SMEs in industry 4.0: a systematic literature review. *Journal of Intelligent Manufacturing*, *36*(8), 5263-5283.
[5] Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., ... & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1-40.

[6] Ogayemi, A. (2026). From obligation to enforcement: mapping EU AI act and CRA cybersecurity requirements to technical controls for LLM-based autonomous agents. *International Journal of Information Security*, *25*(5), 164.
[7] Omar, M., Nassar, S., Sharlf, K., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2025). Emerging applications of NLP and large language models in gastroenterology and hepatology: a systematic review. *Frontiers in Medicine*, *11*, 1512824.
[8] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, *12*, 26839-26874.
[9] Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports*, *15*(1), 13755.
[10] Ren, M. (2024). Advancements and applications of large language models in natural language processing: A comprehensive review. *Applied and Computational Engineering*, *97*(1), 55-63.
[11] Sarker, A., Zhang, R., Wang, Y., Xiao, Y., Das, S., Schutte, D & Xu, H. (2024). Natural language processing for digital health in the era of large language models. *Yearbook of Medical Informatics*, *33*(01), 229-240.
[12] Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., & Lenert, L. A. (2025). The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, *32*(6), 1071-1086.
[13] Watara, S. A., Moreira , G., & Anyah , V. (2025). Enhancing Supply Chain Efficiency Through Machine Learning: A Predictive Analytics Approach to Risk Identification and Timely Deliveries. *International Journal of Scientific Research and Modern Technology*, *4*(11), 196–206. https://doi.org/10.38124/ijsrmt.v4i11.1286
[14] Zubiaga, A. (2024). Natural language processing in the era of large language models. *Frontiers in artificial intelligence*, *6*, 1350306.