ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

A Systematic Review of Large Language Models for Automation in Civil Engineering: Applications, Challenges, and Future Directions

Abhay Kumar[1]; Pawan Kumar[2]; Abhishek Kumar Jha[2]; Akansha Jaiswal[3]; Mohd Zia Hussain[1]; Faiz Akram[4*]

[1]Department of Civil Engineering, Government Engineering College Samastipur, Bihar, India
[2]Department of Civil Engineering, Rashtrakavi Ramdhari Singh Dinkar College of Engineering, Begusarai, Bihar, India
[3]Department of Civil Engineering, Government Engineering College Bhojpur, Bihar, India
[4]Indraprastha Research Laboratory, Indraprastha Institute of Information Sciences Private Limited, New Delhi, India

Corresponding Author: Faiz Akram[4*]

Publication Date: 2026/04/17

Abstract: The swift progress of large language models (LLMs) has generated substantial interest in their capacity to revolutionize automation in diverse fields, such as civil engineering. Although large language models have shown impressive abilities in processing natural language and automating tasks, their potential in civil engineering has not been thoroughly investigated, with research remaining scattered and lacking comprehensive consolidation. This paper presents a thorough systematic review to chart the existing scope of LLM-based automation in civil engineering, with the aim of uncovering primary applications, obstacles, and prospective research avenues. We analyze existing studies across multiple dimensions, such as civil and structural engineering, industrial automation, traffic management, education, scientific research, and software development, then critically evaluate the methodological approaches and practical implementations reported in the literature. The review indicates LLMs hold potential for automating design optimization, construction planning, and decision-making processes, but struggle with issues such as gaps in domain-specific knowledge, poor data quality, and safety risks. Moreover, we pinpoint developing tendencies, such as the merging of LLMs with digital twins and building information modeling (BIM), which may transform automation in the domain. The findings highlight the need for robust evaluation frameworks and interdisciplinary collaboration to address technical and ethical barriers. This review consolidates these insights, establishing a basis for subsequent investigations and the actual implementation of LLMs in civil engineering automation.

Keywords: *Large Language Models, Industrial Automation, Design Optimization, Decision Making, Civil Engineering.*

How to Cite: Abhay Kumar; Pawan Kumar; Abhishek Kumar Jha; Akansha Jaiswal; Mohd Zia Hussain; Faiz Akram (2026) A Systematic Review of Large Language Models for Automation in Civil Engineering: Applications, Challenges, and Future Directions. *International Journal of Innovative Science and Research Technology*, 11(4), 912-926. https://doi.org/10.38124/ijisrt/26apr455

I. INTRODUCTION

Civil engineering traditionally entails intricate, labor-heavy procedures requiring thorough planning, design, and implementation. Traditional methods often rely on manual calculations, iterative design revisions, and heuristic decision-making, which can be time-consuming and prone to human error [1]. With the advent of artificial intelligence (AI), particularly large language models (LLMs), there is growing potential to revolutionize these workflows by introducing automation at unprecedented scales. Large language models, including GPT-4 and BERT, show remarkable proficiency in comprehending and producing text resembling human output, which supports their application in diverse activities from programming to synthesizing academic papers [2]. Their capacity to analyze extensive

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

unstructured datasets and deliver context-sensitive answers renders them exceptionally well-suited for tackling civil engineering problems, as such tasks frequently require the synthesis of diverse data from technical documents, regulatory standards, and live sensor inputs [3].

The adoption of LLMs in civil engineering remains at an early phase, where the majority of implementations concentrate on limited, specialized automation tasks. For example, recent research has examined their application in automating structural design suggestions, improving construction timelines, and producing compliance reports [4]. However, the broader implications of LLMs for automating multidisciplinary workflows, such as integrating geotechnical analysis with architectural planning, remain underexplored. Moreover, while LLMs excel in natural language tasks, their performance in domain-specific applications is often constrained by limited training data, lack of engineering context, and challenges in interpreting technical jargon [5]. These limitations underscore the need for a systematic examination of how LLMs can be tailored to meet the unique demands of civil engineering, where precision, safety, and regulatory compliance are paramount.

A crucial research gap stems from the lack of a unified framework to assess the performance of LLMs in civil engineering automation. Current research often examines specific applications in isolation, neglecting scalability, interoperability, or the connection of LLMs with established digital tools such as Building Information Modeling (BIM) and finite element analysis programs [6]. Furthermore, while LLMs have shown promise in automating routine tasks, their role in high-stakes decision-making, such as structural safety assessments, remains contentious due to concerns about interpretability and reliability [7]. These gaps underscore the need for a thorough examination, which maps existing applications while pinpointing methodological obstacles and prospective research avenues.

This research is driven by the capacity of LLMs to transform civil engineering processes by increasing productivity, lowering expenses, and boosting precision. The automation of repetitive tasks, including document parsing and code generation, with LLMs permits engineers to concentrate on the creative and strategic dimensions of projects [8]. Moreover, their capacity to handle diverse linguistic and multimodal data creates opportunities for worldwide cooperation, which supports uninterrupted knowledge exchange among teams spread across different regions [9]. This research is important because it can connect AI progress with real-world engineering uses, promoting new ideas and dealing with ethical and safety issues.

The remainder of this paper is organized as follows: Section 2 outlines the methodology employed for this systematic review, including the selection criteria and analytical framework. Section 3 presents the results, structured into subsections covering research trends, domain-specific applications, and evaluation challenges. Section 4 examines the consequences of these results, with Section 5 presenting the final observations.

II. METHODOLOGY

- *Review Protocol*

This systematic review follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [10] to uphold methodological precision and clarity. The literature search was conducted across seven major academic databases and search engines, selected based on their relevance to engineering and computer science research. IEEE Xplore was prioritized due to its extensive collection of peer-reviewed conference papers and journals in civil engineering and automation. Scopus and Web of Science were included for their comprehensive coverage of interdisciplinary research, particularly in AI applications. ACM Digital Library and ScienceDirect contained research on software development and industrial automation, and SpringerLink added essential literature on computational methods to these resources. Google Scholar functioned as an additional tool to identify preprints and grey literature in early stages, which may not be included in other databases.

The search queries were constructed to achieve a balance between precision and comprehensiveness, with attention directed toward the overlap of ‘large language models’ and ‘civil engineering automation.’ Different forms of these terms appeared in various databases to accommodate disparities in indexing practices. For example, in IEEE Xplore, the query (("large language model*" AND ("civil engineering automation" OR "automation in civil engineering")) NOT ("review" OR "survey" OR "meta - analysis")) was employed, while arXiv used a title-abstract filter (ti:(large language model*) AND (abs:(civil engineering automation) OR abs:(automation in civil engineering))). Boolean operators and exclusion terms (e.g., “review,” “survey”) were applied consistently to filter out non-empirical studies.

- *Research Dimensions and Analytical Framework*

The review categorizes existing literature into seven thematic dimensions to systematically evaluate the role of LLMs in civil engineering automation. The first dimension examines applications in civil and structural engineering, such as automated design generation and structural health monitoring. The second focuses on industrial and manufacturing processes, where LLMs assist in quality control and supply chain optimization. Automated traffic and mobility applications constitute the third dimension, which includes intelligent transportation systems and urban planning. The fourth dimension pertains to education and assessment, encompassing automated grading and personalized learning tools. The fifth dimension explores scientific research and data analysis, particularly in synthesizing technical literature and experimental data. The sixth dimension investigates software and code development, where LLMs generate or debug domain-specific scripts. The final dimension assesses LLM safety and performance metrics, with a focus on reliability and ethical aspects in engineering settings.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

➢ *Inclusion and Exclusion Criteria*

Studies were included if they (1) explicitly discussed LLM applications in civil engineering or related automation tasks, (2) presented empirical results or case studies, and (3) were published in English. Theoretical models and validation via simulation were included when they yielded practical guidance for application in actual scenarios. Exclusion criteria removed research lacking technical rigor (e.g., opinion articles), concentrating exclusively on fields outside engineering (e.g., generic chatbots), or failing to prioritize automation as a key result. The time frame was unrestricted to document the development of LLM applications, though most included studies appeared after 2020, aligning with progress in transformer-based models.

➢ *Study Selection Process*

The initial search yielded 859 records, reduced to 548 after removing duplicates and irrelevant entries (e.g., non-English publications). During the title and abstract screening process, 305 studies failing to meet the inclusion criteria were excluded. Full-text assessment of the remaining 166 records further excluded 71 studies due to insufficient technical detail or misalignment with the research dimensions. The PRISMA flowchart (Figure 1) displays the inclusion of 95 studies in the final review.

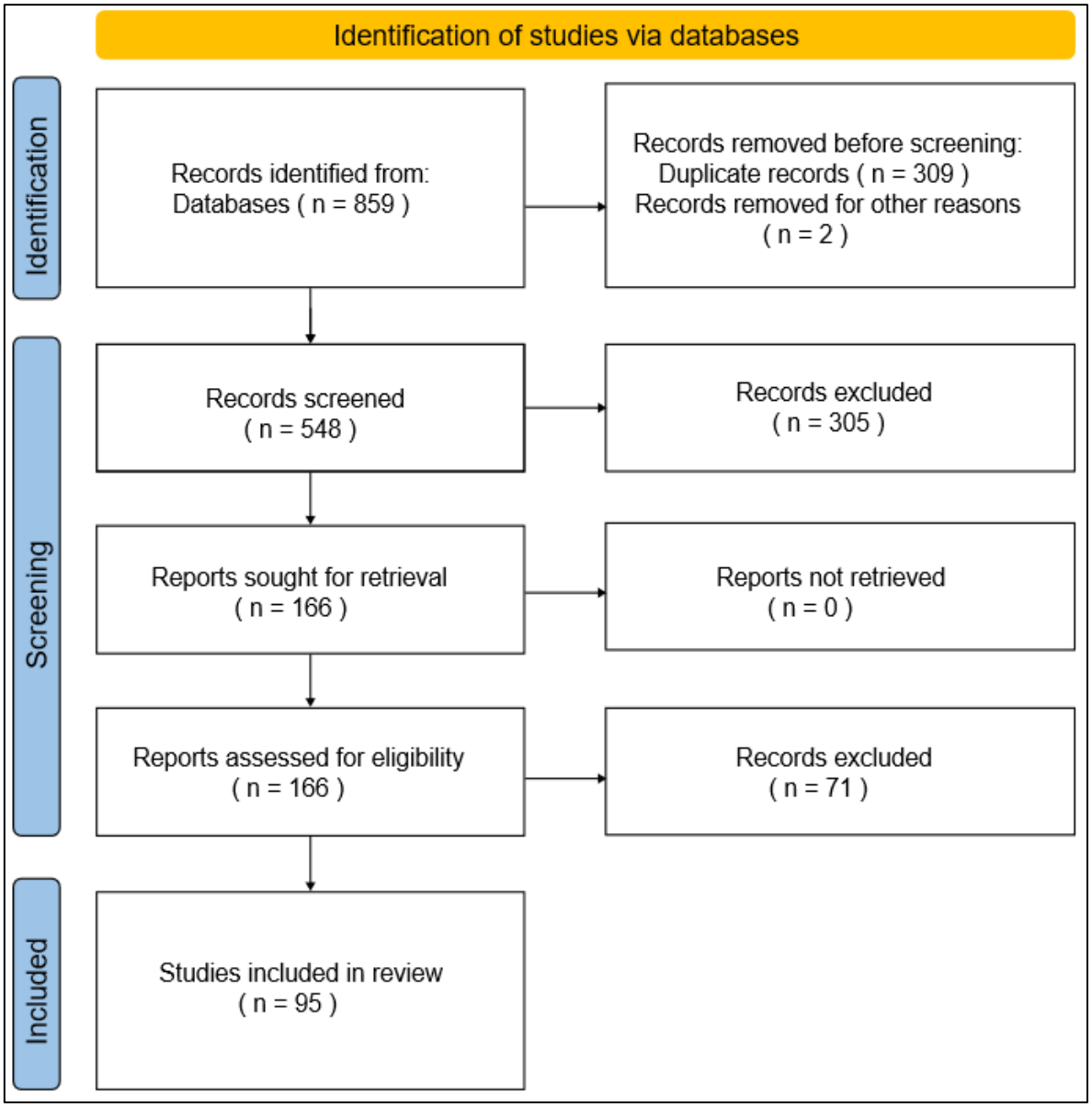

Fig 1 PRISMA Flowchart of the Study Selection Process

The selection process may exhibit biases, such as favoring recent publications at the expense of earlier foundational contributions and depending on English-language studies, potentially neglecting pertinent non-English research. To address these issues, backward snowballing was conducted on pivotal papers to uncover supplementary sources, and the search was not limited by geographical boundaries. Varied study designs, which included controlled experiments and field deployments, were addressed by thematically synthesizing findings instead of employing quantitative methods.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

III. RESULTS

➢ *Research Trends*

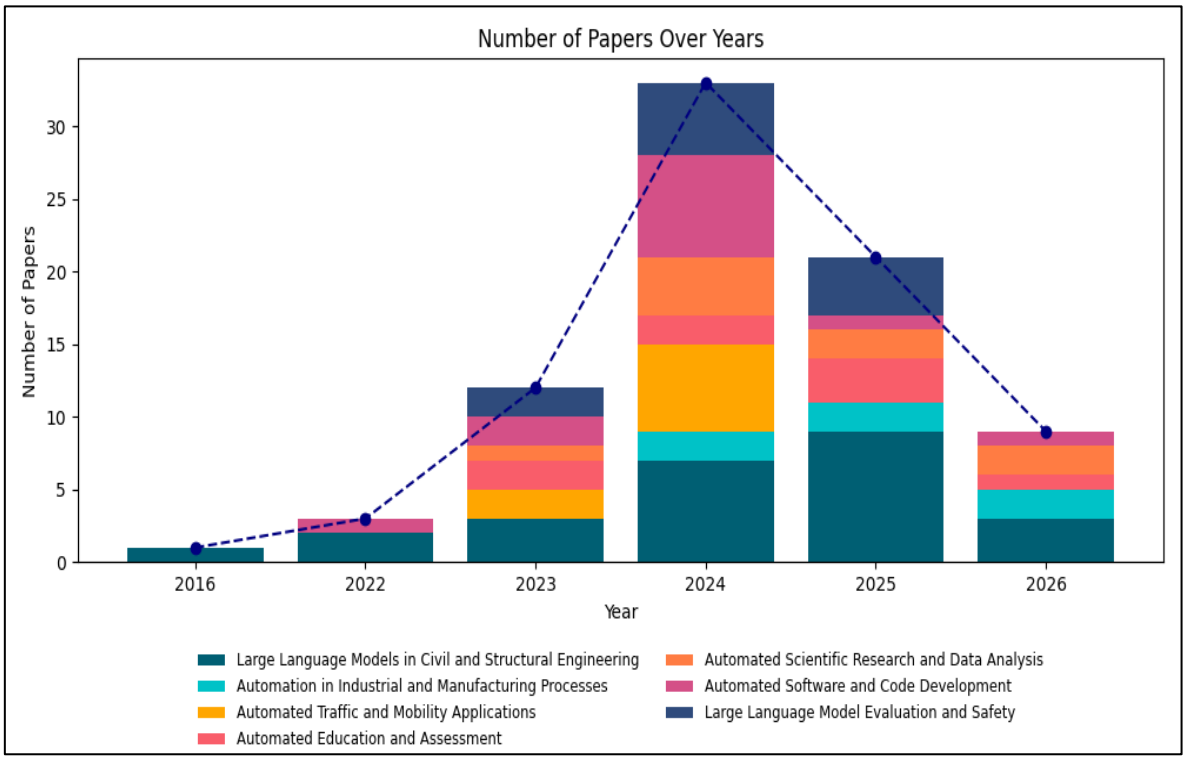

Fig 2 Research Trends in the Domain of Large Language Models for Automation in Civil Engineering

Examining publication trends shows a distinct increase in scholarly attention directed at large language models (LLMs) for civil engineering automation, especially post-2023. Prior to 2022, only isolated studies explored this intersection, with a single publication in 2016 and three in 2022. The discipline experienced notable progress in 2023, marked by 12 publications, and then witnessed a sharp increase to 33 in 2024. This surge coincides with broader advancements in transformer-based architectures and their adaptation to domain-specific tasks. The subsequent decline to 21 and 9 publications in 2025 and 2026, respectively, suggests a consolidation phase where researchers shifted focus from exploratory applications to refining methodologies and addressing implementation challenges.

The distribution across thematic areas highlights uneven growth trajectories. Civil and structural engineering applications are predominant, comprising 25 studies, and show consistent expansion between 2016 and 2025. This reflects the discipline's early recognition of LLMs' potential for design automation and structural analysis. In contrast, industrial and manufacturing automation appeared later, with all six studies published from 2024 to 2026, which suggests a nascent yet increasing focus on production optimization. Traffic and mobility applications reached their highest point in 2024 (six studies), probably due to smart city projects, whereas research focused on education showed steady production starting from 2023. Scientific research automation and code development display similar growth trajectories, reaching their highest points in 2024, while LLM evaluation studies became more prominent starting in 2023, reflecting growing concerns about reliability in engineering applications.

The temporal patterns suggest three evolutionary phases: initial exploration (2016–2022), rapid diversification (2023–2024), and selective maturation (2025–2026). Early studies primarily tested LLMs' basic capabilities in interpreting engineering documents, while later works integrated them with simulation tools and digital twins. The reduction in publication output after 2024 could indicate a shift from initial proof-of-concept presentations to addressing scalability and interoperability issues, with ongoing attention to assessment and safety supporting this observation. This trajectory mirrors the Gartner Hype Cycle, where early enthusiasm gives way to pragmatic development as limitations become apparent.

➢ *Applications of Large Language Models in Civil and Structural Engineering*

The application of large language models (LLMs) in civil and structural engineering has shown transformative potential in various areas, including design automation and safety compliance. A systematic analysis of the included studies reveals distinct clusters of applications, each addressing specific challenges in the field. The following taxonomy categorizes these applications based on their primary focus areas, methodologies, and implementation strategies.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

Table 1 Taxonomy of LLM Applications in Civil and Structural Engineering

Application Domain	Specific Task	Methodology/Approach	Sources
Design Automation	BIM Generation	Multi-agent LLM frameworks	[11], [12], [13], [14]
	Code-Compliant Design	Rule-based LLM integration	[12], [14]
	Foundation Design	Router-driven multi-agent systems	[15], [16]
Structural Analysis	Automated Analysis	LLM-driven workflows	[17], [18]
	Reliability and Robustness	Agent-based LLM systems	[18]
Construction Planning	Cost Budgeting	Generative AI	[19]
	Activity Sequencing	Small language models	[20]
Safety and Compliance	Fire Safety Planning	Decision support systems	[21], [22]
	Building Regulation Interpretation	NLP-based analysis	[23]
Document Processing	Construction Document NER	Fine-tuned LLMs	[24]
	BIM Data Retrieval	Multi-agent LLM integration	[25]
Multimodal Applications	Hazard Identification	Vision-language frameworks	[26]
	Crack Image Classification	Multimodal LLMs	[27]
Human-Robot Collaboration	Construction Workflows	Multimodal VR + LLM	[28]

Design automation stands out as the foremost application, where multi-agent LLM systems produce Building Information Modeling (BIM) data based on natural language inputs [11]. These systems often employ hierarchical task decomposition, where an orchestrator LLM delegates subtasks to specialized agents for geometry generation, material specification, and compliance checking. For instance, [12] illustrates the automation of code-compliant concrete structure design via rule-based LLM application by interpreting regulatory documents and producing parameterized models. Foundation design, an essential but repetitive process, gains advantages from router-based multi-agent systems that categorize design issues and choose specialized experts in the field [15].

Structural analysis workflows have been augmented through LLM-driven automation, particularly in preliminary design stages. Research including [17] illustrates that large language models can transform engineering inquiries into inputs for finite element analysis, thereby decreasing the need for manual preprocessing. Agent-based validation layers that cross-check results against first principles improve the dependability of these systems [18]. However, the interpretability of LLM-generated analyses remains a challenge, as these models often function as black boxes.

Construction planning applications employ large language models for generative cost estimation and schedule optimization. [19] employs generative AI to synthesize cost databases into budget proposals, while [20] uses small language models for predicting missing tasks in construction sequences. Another essential domain is safety compliance, where LLMs aid in analyzing intricate building regulations [23] and modeling fire evacuation situations [21]. These applications highlight the models' ability to process heterogeneous regulatory documents and extract actionable insights.

Multimodal LLM applications bridge the gap between textual and visual data in civil engineering. Vision-language systems support automated hazard detection by linking site photographs to safety guidelines [26], and multimodal approaches categorize crack formations in steel bridges by merging visual data with inspection documents [27]. Human-robot collaboration systems, exemplified by the ones in [28], merge LLMs with virtual reality interfaces to support natural language interaction between construction teams and robotic assistants.

A number of investigations do not fit within the main classification yet yield important findings. [29] investigates workflow automation in structural engineering by employing LLM agents to oversee iterative design-analysis cycles. [30] examines LLMs in the role of BIM interpreters, with the aim of deriving IFC data to automate quantity takeoffs. [31] and [32] establish benchmarks for assessing LLM performance in engineering documentation, which resolves the absence of standardized evaluation metrics. These studies collectively highlight the growing influence of LLMs in civil engineering, while identifying ongoing difficulties in adapting to the field and ensuring dependable performance.

Wide-ranging applications illustrate both the adaptability of LLMs and the necessity for adjustments tailored to specific contexts. Although general language capabilities are advantageous for design automation and document processing, structural analysis and safety applications demand close alignment with numerical simulations and domain-specific knowledge repositories. This division implies upcoming studies ought to focus on mixed frameworks merging large language models and symbolic reasoning approaches.

- *Large Language Models in Industrial and Manufacturing Automation*

Large language models (LLMs) are transforming industrial and manufacturing automation by tackling obstacles in control systems, process optimization, and knowledge management. Recent research illustrates the ways in which LLMs improve judgment, optimize design procedures, and support human-robot interaction in industrial environments.

A hierarchical categorization of the included studies reveals distinct functional roles and application domains, as shown in Table 2. Industrial automation applications

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

primarily focus on control systems and ontology-based knowledge augmentation. For instance, [33] introduces an LLM-based framework for controlling industrial automation systems, where natural language commands are translated into executable control logic. This method diminishes the necessity for specialized programming skills while preserving the dependability of the system. In a related approach, [34] merges ontologies with knowledge-augmented LLMs to direct human-robot cooperation, which results in intuitive task assignment and mistake rectification in manufacturing processes.

Table 2 Taxonomy of LLM Applications in Industrial and Manufacturing Automation

Application Domain	Functional Role	Specific Task	Sources
Industrial Automation	Control and Decision-Making	Human-Robot Collaboration Guidance	[34]
		Industrial System Control	[33]
	Knowledge and Ontology Support	Capability Ontology Generation	[35]
Manufacturing Automation	Process Optimization	Manufacturing Operations Automation	[36]
	Design and Engineering	Mechanical Design	[37]
	Testing and Validation	Analog Circuit Testbench Generation	[38]
General Manufacturing	Broad Applications	LLM Integration in Manufacturing	[39]

Manufacturing automation employs large language models in design, performance evaluation, and operational refinement. [37] illustrates the capacity of LLMs to operate as mechanical designers by producing CAD-compatible specifications based on textual descriptions. This capability accelerates prototyping by automating preliminary design iterations. In analog circuit testing, [38] proposes a framework in which LLMs produce testbenches automatically, thereby diminishing the need for manual work in verification processes. Process optimization studies, such as [36], highlight LLMs' role in automating manufacturing operations through real-time data analysis and adaptive scheduling.

Connecting LLMs to specialized knowledge repositories continues to be a pivotal driver for industrial implementations. [35] investigates the application of LLMs in creating capability ontologies, which structure manufacturing constraints and resource dependencies. These ontologies improve the compatibility between LLMs and current industrial systems by resolving issues related to semantic ambiguity. The need for thorough verification of LLM-produced results in safety-sensitive settings is emphasized by [39], which examines wider obstacles in applying LLMs within industrial systems.

Recent developments indicate a movement toward multimodal LLM frameworks merging textual, visual, and sensor inputs for comprehensive automation. Although the included studies do not directly address this direction, it corresponds to industrial demands for real-time quality inspection and predictive maintenance. The lack of research on large language models in supply chain automation suggests a neglected domain where natural language processing could improve logistics optimization and demand prediction.

The examined applications together illustrate the capacity of LLMs to diminish human involvement in monotonous tasks while increasing flexibility in changing industrial settings. Subsequent studies ought to tackle scalability issues, especially when linking large language models with outdated industrial control frameworks and guaranteeing resilience against hostile data.

- *Automated Traffic and Mobility Applications*

Incorporating large language models (LLMs) into traffic and mobility systems has introduced novel paradigms for scenario generation, decision-making, and autonomous driving. These applications employ the contextual comprehension and generative functions of LLMs to tackle intricate problems in urban transportation, safety assessment, and automated vehicle systems. The examined research presents a range of approaches, from task-specific adjustments to architectures merging multiple data types, all addressing different facets of traffic control and transportation solutions.

Table 3 Taxonomy of LLM Applications in Traffic and Mobility

Application Domain	Method/Approach	Specific Focus	Sources
Traffic Safety	Domain-Specific LLM Tuning	Transportation safety expertise	[40]
	General LLM Evaluation	Feasibility for traffic safety research	[40]
Traffic Simulation	LLM-Assisted Scenario Generation	Urban mobility simulation (SUMO)	[41]
Traffic Management	LLM-Based Decision Making	Mixed traffic scenarios at intersections	[42]
	Multi-Agent Systems	Smart urban mobility integration	[43]
Autonomous Driving	Reinforcement Learning Optimization	Human-centric reward shaping	[44]
	Policy Learning Workflow	Urban driving automation	[45]
	Vision-Language Model Integration	Autonomous driving and scene understanding	[46], [47]
	Knowledge-Driven Multi-Agent Systems	Autonomous driving with LLMs	[48]

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

Traffic safety applications illustrate the customization of LLMs to meet domain-specific needs. [40] adapts a pre-trained model to operate as a transportation safety specialist, which supports real-time risk evaluation and adherence verification. This approach addresses the challenge of interpreting heterogeneous safety regulations, which often vary across jurisdictions. Conversely, [40] assesses broad-application LLMs such as ChatGPT for traffic safety applications, identifying shortcomings in managing specialized terminology and situation-specific limitations. These studies collectively highlight the trade-off between adaptability and specialization when deploying LLMs in safety-critical domains.

Traffic simulation benefits from LLMs' generative capabilities, particularly in creating diverse and realistic scenarios for testing. [41] automates traffic scenario generation for the Simulation of Urban Mobility (SUMO) platform, where natural language prompts are translated into parameterized simulation configurations. This reduces the manual effort required for edge-case testing while improving scenario diversity. Traffic management systems, exemplified by the approaches in [42], apply LLMs to improve signal timing and routing choices in mixed-traffic settings, where conventional rule-based methods fail to handle unpredictability. Implementing multi-agent frameworks, suggested by [43], improves adaptability by permitting dynamic coordination among vehicles, infrastructure, and control centers.

Autonomous driving stands as the most challenging field of application, demanding a close interconnection of perception with decision-making and execution. [44] refines reinforcement learning policies for autonomous vehicles by applying reward functions derived from LLMs, which are aligned with human preferences. This method addresses the issue of reward misalignment often observed in simulated training settings. Policy learning frameworks, exemplified by [45], employ LLMs to automate driving policy creation and improvement, which diminishes the need for manually programmed regulations. Vision-language models, as shown in [46] and [47], connect textual instructions with visual scene comprehension, which results in more intuitive human-vehicle interaction. Knowledge-driven architectures, such as the framework proposed in [48], merge LLMs with structured knowledge graphs to increase decision-making transparency in unclear driving situations.

The alignment of these approaches highlights the revolutionary capacity of LLMs in traffic and transportation networks. Nevertheless, difficulties remain in achieving real-time performance, interpretability, and safety assurance, especially when models function in open-world settings with incomplete data. Subsequent studies ought to focus on hybrid frameworks merging large language models with symbolic reasoning approaches to improve reliability and transparency.

- *Automated Education and Assessment in Civil Engineering*

The adoption of large language models (LLMs) in civil engineering education and assessment has introduced innovative methods for generating feedback, automating evaluations, and upholding academic honesty. These applications employ the models' abilities to comprehend and generate natural language to tackle persistent issues in scalable assessment and individualized education. The selected studies illustrate a range of approaches, including domain-adapted models and comparisons of automated assessment methods.

Table 4 Taxonomy of LLM Applications in Automated Education and Assessment

Application Domain	Evaluation Focus	Specific Task	Sources
Automated Feedback and Grading	Conceptual Understanding	Engineering problem-solving feedback	[49]
		Conceptual question evaluation	[50]
	Mathematical Assessment	Constructed response scoring	[51]
Automated Writing Evaluation (AWE)	Essay Scoring	Reliability and fairness	[52]
		Comparative study of approaches	[53]
	Foreign Language Assessment	Automated essay scoring	[54]
Text Analysis in Education	Online Learning Research	Automated text coding	[55]
Academic Integrity and Ethics	AI in Education	Post-pandemic challenges	[56]

Automated feedback and grading systems mark a major progress in engineering education. [49] illustrates that carefully adjusted large language models can deliver precise evaluations for intricate engineering problem-solving activities, which lessens the burden on educators without compromising educational standards. The study highlights the importance of domain-specific adaptation, as general-purpose models often fail to capture nuanced engineering principles. Similarly, [50] explores scalable grading solutions for conceptual questions, where LLMs evaluate student responses against predefined knowledge graphs. Mathematical assessment presents distinct difficulties owing to the requirement for symbolic reasoning, which [51] tackles by employing hybrid architectures merging LLMs and computer algebra systems to grade constructed responses in math assessments.

Automated writing evaluation (AWE) has seen substantial innovation through LLMs, particularly in assessing technical reports and research papers. [52] conducts a critical analysis of the dependability of essay scoring by large language models, uncovering disparities in how grading standards are applied among various systems. This research highlights the necessity for calibration frameworks to guarantee fairness, particularly in high-stakes evaluations. Studies including [53] compare LLMs with rule-based and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

conventional machine learning methods, showing that LLMs achieve better results when dealing with unclear or imaginative answers. The evaluation of foreign language skills gains advantages from the multilingual capacities of large language models, as evidenced by [54], which illustrates how systems such as ChatGPT can assess non-native linguistic competence and deliver constructive guidance for progress.

Text analysis applications extend beyond grading to support research in educational methodologies. [55] introduces a systematic calibration framework for coding student discussions in online learning environments, which makes possible large-scale analysis of engagement patterns. This method supports data-informed adjustments in curriculum development and intervention approaches. Academic integrity issues, with a focus on the improper employment of LLMs for plagiarism or unapproved aid, are analyzed in [56], which proposes detection methods and educational strategies designed for the post-pandemic academic environment.

The aggregated results indicate LLMs present unmatched scalability in education and evaluation, yet their efficacy hinges on deliberate alignment with established pedagogical principles. Specialized adaptation for particular domains, mixed frameworks uniting symbolic and neural methods, and rigorous verification procedures stand out as essential elements for effective implementation. Future research should address the interpretability of automated evaluations and develop standards for model transparency in academic settings.

- *Automated Scientific Research and Data Analysis*

Employing large language models (LLMs) in scientific inquiry and data examination has transformed knowledge acquisition, experimental procedure development, and cross-disciplinary data synthesis. These models show outstanding abilities in analyzing intricate technical literature, performing repetitive research tasks automatically, and generating insights from diverse datasets. The examined studies display a varied array of approaches, spanning from multimodal learning frameworks to independent research systems, each tackling unique obstacles in scientific automation.

Table 5 Taxonomy of LLM Applications in Automated Scientific Research and Data Analysis

Application Domain	Task	Methodology	Sources
Material Science and Engineering	Knowledge extraction from literature	Automated text mining for material properties	[57], [58], [59]
	Protocol generation	LLM-based automated synthesis of machine-readable protocols	[60]
Chemical and Biomedical Research	Reaction mining	Multimodal LLM-based analysis of electrosynthesis data	[61]
	Data annotation and integration	Automated single-cell RNA-seq data processing	[62]
	Toxicity data extraction	LLM-based parsing of unstructured radiology reports	[63]
Mathematical and Computational Discovery	Theorem discovery	Program search with LLMs	[64]
Autonomous Research	Hypothesis generation and experimentation	LLM-driven autonomous chemical research	[65]

Applications in material science and engineering illustrate the capacity of LLMs to expedite research by automating the extraction of knowledge from extensive scientific literature. [57] establishes a framework to derive mechanical constitutive models from research papers, which markedly diminishes the need for manual curation in cultural heritage conservation. In an analogous approach, [58] employs text mining methods in the study of metal-organic frameworks (MOFs), which makes possible the extensive derivation of material properties from diverse sources. The automation of LLMs aids in database creation, as evidenced by [59], where systematic extraction of magnetic material properties fills structured data repositories. Protocol generation stands as another essential application, with [60] introducing ProtoCode, a tool that converts PCR protocols from publications into machine-readable forms, thereby improving reproducibility in life sciences.

Chemical and biomedical studies employ large language models for various analytical purposes. [61] merges multimodal learning and LLMs to extract electrosynthesis reactions, establishing connections between textual procedures and experimental data for thorough reaction analysis. In single-cell genomics, [62] shows how scExtract automates RNA-seq data annotation and cross-dataset merging, tackling scalability challenges in omics research. Clinical applications encompass [63], which analyzes unstructured radiology reports to derive post-SBRT toxicity information, thereby improving the efficiency of outcomes evaluation in radiation oncology. These applications highlight LLMs' ability to bridge domain-specific knowledge gaps while handling technical jargon and ambiguous expressions inherent in scientific writing.

Mathematical discovery constitutes a cutting-edge application, where [64] applies large language models in program exploration to uncover new mathematical theorems. This method illustrates the capacity of models to go further than identifying patterns by generating novel theoretical perspectives. Autonomous research systems extend this

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

frontier, with [65] creating an LLM-powered platform capable of independently designing, conducting, and refining chemical experiments. Such systems challenge traditional research paradigms by integrating literature synthesis with experimental design and hypothesis generation.

These methodologies coming together highlights the revolutionary impact of LLMs in scientific inquiry, yet issues remain concerning strict validation and cross-disciplinary applicability. Future work should focus on developing standardized benchmarks for scientific automation tasks and improving model interpretability in high-stakes research applications.

- *Automated Software and Code Development in Civil Engineering*

Incorporating large language models (LLMs) into civil engineering software and code development has introduced major progress in program repair, code generation, and quality assurance. These models show outstanding abilities in automating tasks that typically require extensive manual effort, including debugging, generating test cases, and performing verification, while also tackling specialized challenges in the development of engineering software.

Table 6 Taxonomy of LLM Applications in Automated Software and Code Development

Application Area	Task	Approach	Sources
Program Repair and Debugging	Automated Program Repair	Parameter-efficient fine-tuning of LLMs	[66]
		LLM-assisted formal verification for AI code repair	[67]
	Zero-shot Vulnerability Repair	Evaluating LLMs for vulnerability fixes without fine-tuning	[68]
	Self-Correction and Critiquing	Tool-interactive critiquing for LLM self-correction	[69]
Code Generation and Synthesis	Multi-turn Program Synthesis	Open LLM for iterative code generation	[70]
	Automated Proof Synthesis	Leveraging LLMs for formal proof generation in Rust	[71]
	Loop Invariant Generation	Enhancing invariant generation for complex programs using LLMs	[72]
Code Quality and Consistency	Comment Inconsistency Detection	Detecting Rust code-comment mismatches with LLMs	[73]
	Conflicting Requirements Detection	Satisfiability-aided LLMs for requirement conflict identification	[74]
	Declarative Specification Repair	Repairing software specifications using LLMs	[75]
Testing and Benchmarking	Automated Test Case Generation	Directional ensemble LLMs with Retrieval-Augmented Generation (RAG)	[76]
	Holistic Code Evaluation	Contamination-free benchmarking of LLMs for code tasks	[77]
	Code Efficiency Benchmarking	Evaluating LLMs on code efficiency metrics	[78]
Patent and Legal Automation	Automated Patent Claim Refinement	Novel LLM-based framework for patent claim improvement	[79]

Program repair and debugging constitute a crucial domain in which LLMs markedly diminish the need for manual intervention in detecting and correcting software errors. [66] investigates parameter-efficient fine-tuning methods to adjust large language models for automated program repair, showing better results in fixing syntax and logical mistakes with reduced computational costs. Formal verification, long recognized as a challenging endeavor in engineering software, gains advantages from LLM application as evidenced in [67], where models aid in correcting AI-produced code by comparing it with prescribed specifications. Zero-shot methods, exemplified by the research in [68], show that large language models can identify and fix software vulnerabilities without prior task-specific training, but their performance depends on the intricacy of the vulnerabilities. Self-correction mechanisms, illustrated by [69], permit large language models to progressively improve their outputs with tool-assisted evaluation, which increases dependability in safety-critical contexts.

Code generation and synthesis applications exploit the capacity of LLMs to comprehend and generate executable code based on abstract descriptions. [70] presents an openly available large language model that supports iterative program synthesis through multi-turn dialogues with users, producing increasingly refined outputs, which is especially beneficial for accelerating prototyping in civil engineering software creation. The amalgamation of rigorous techniques, illustrated by [71], exhibits how LLMs can construct proofs for Rust programs, thereby connecting informal specifications with implementations capable of verification. Loop invariant generation, which poses a substantial difficulty in program verification, is improved by LLM-based methods in [72], where neural generation and symbolic validation are jointly applied to address complex engineering algorithms.

Code quality assurance stands as another essential area where LLMs play a role in upholding strong engineering software. [73] tackles the ongoing issue of code-comment divergence in Rust projects by employing LLMs to identify and correct discrepancies between code and its accompanying documentation. Requirement analysis benefits from satisfiability-aided LLMs, as shown in [74], which identify conflicting specifications in engineering software

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

projects by modeling requirements as logical constraints. Research in [75] investigates declarative specification correction, showing how large language models can fix mistakes in structured software descriptions, thereby lowering risks in subsequent implementation stages.

Testing and benchmarking tools have developed to support code produced by large language models. [76] proposes a method where multiple LLMs are jointly employed with retrieval-augmented generation (RAG) to generate thorough test cases for civil engineering software. Evaluation methodologies have similarly advanced, with [77] introducing contamination-free benchmarks to assess LLM performance on code-related tasks without data leakage concerns. Code efficiency, an essential metric for engineering applications, is methodically quantified in [78], where standardized benchmarks are defined to evaluate computational resource consumption in LLM-produced implementations.

Specialized applications in patent automation, including [79]'s ClaimBrush framework, illustrate the capacity of LLMs in legal-technical fields by automating patent claim refinement, which demands exact comprehension of both technical and legal terminology. This capability could streamline intellectual property processes for civil engineering innovations.

The wide range of these applications emphasizes the revolutionary influence of LLMs on civil engineering software development methods, yet also draws attention to ongoing difficulties in validation, performance, and specialization. Subsequent studies ought to focus on hybrid frameworks merging neural methods with symbolic logic to improve dependability in systems crucial for engineering.

- *Evaluation and Safety Considerations for Large Language Models in Engineering Applications*

Implementing large language models (LLMs) in civil engineering automation requires strict assessment frameworks and safety measures to guarantee dependability in mission-critical applications. Recent research has established tailored approaches to evaluate model efficacy, identify weaknesses, and address hazards linked to the incorporation of LLMs in engineering processes.

Table 7 Taxonomy of LLM Evaluation and Safety Approaches

Evaluation Dimension	Methodology	Application Context	Sources
Safety Benchmarking	Automated safety evaluation (S-EVAL)	Comprehensive risk assessment	[80]
	Disinformation capability analysis	Adversarial attack resilience	[81]
	Trustworthiness assessment (TrustLLM)	Model alignment and reliability	[82]
Hallucination Detection	Automated detection feasibility	Output veracity analysis	[83]
Automated Verification	Proof debugging with LLMs (Laurel)	Formal verification assistance	[84]
Adversarial Robustness	Universal attack vectors	Alignment vulnerability analysis	[85]
Domain-Specific Evaluation	Time series anomaly detection	Aerospace software monitoring	[86]
	News summarization benchmarks	Content generation quality	[87]
Instruction Compliance	Task adherence evaluation	Prompt following capability	[88]
Risk Taxonomy	Systemic risk classification	LLM deployment risk mitigation	[89]
Evaluation Infrastructure	Automated evaluation (XFinder)	Reliable assessment frameworks	[90]

Safety benchmarking arises as a pivotal assessment criterion, with [80] proposing S-EVAL, a framework designed for automation that methodically evaluates LLM outputs in diverse hazard domains such as the production of injurious material and prejudiced judgment formation. This approach addresses the challenge of scalable safety testing in engineering applications where models process technical documentation and generate design recommendations. The analysis of disinformation capacities in [81] shows that LLMs may unintentionally spread incorrect technical details due to flawed training data, posing a major issue for systems handling engineering documentation. The TrustLLM framework [82] establishes multidimensional trustworthiness metrics, assessing aspects such as truthfulness and safety alignment in domain-specific applications.

Hallucination detection presents unique challenges in engineering contexts where factual accuracy is paramount. [83] examines the core constraints of automated hallucination detection and shows that existing approaches face difficulties in differentiating between plausible fabrications and valid technical outputs without specialized knowledge. This finding underscores the need for hybrid verification systems in civil engineering applications. The approach of Laurel [84], which employs verification assistance, illustrates how LLMs can aid engineers in correcting errors in complex proofs for industrial codebases, although the research emphasizes that human supervision remains necessary for assertion validation in critical systems.

Research on adversarial robustness uncovers inherent weaknesses in aligned large language models. [85] identifies universal attack vectors that circumvent safety filters by means of meticulously designed suffixes, underscoring possible security vulnerabilities in scenarios where LLMs handle engineering specifications originating from unreliable sources. Specialized evaluation frameworks, exemplified by [86]'s examination of anomaly detection in aerospace software time series data, deliver customized assessment approaches for technical domains where standard benchmarks fall short. The quantitative evaluation of content generation quality in [87] employs news summarization benchmarks, which yield adaptable findings for automated report creation in engineering projects.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

Instruction compliance evaluation, examined in [88], establishes task-specific metrics to assess the degree to which LLMs adhere to intricate engineering instructions, which is essential for design automation systems. [89] proposes a thorough risk classification framework for LLM systems, which identifies domain-specific dangers including erroneous material property estimations and faulty structural analyses. Automated assessment systems such as XFinder [90] permit dependable large-scale evaluation by employing large language models as judges, yet this self-referential method demands precise adjustment to avoid the perpetuation of bias.

The accumulated results stress that although LLMs hold revolutionary promise for automating civil engineering tasks, their secure implementation demands evaluation approaches with multiple layers, merging automated assessments, specialized knowledge from the field, and structured validation techniques. Subsequent studies must focus on establishing safety standards tailored to engineering applications and creating dependable validation processes to guarantee that model results adhere to the strict reliability demands of infrastructure systems.

IV. DISCUSSION

An analysis of the reviewed studies shows consistent patterns in how large language models (LLMs) are applied in civil engineering automation. Collectively, the research indicates large language models hold considerable promise for streamlining repetitive activities, including handling documents, producing code, and making initial design suggestions [11] [12] [66]. Nevertheless, their effectiveness in safety-critical contexts shows inconsistency, as research indicates divergent levels of dependability in structural assessment and adherence verification [18] [23]. This contrast indicates large language models perform well in boosting human efficiency for clearly outlined subtasks, but their independent judgment abilities need more development prior to application in critical engineering contexts.

A central theoretical insight arising from this review is the necessity for hybrid frameworks merging neural language models and symbolic reasoning mechanisms. Multiple studies [15] [48] [84] consistently show pure LLM methods face difficulties in performing exact numerical calculations and adhering strictly to compliance checks, which are essential in civil engineering. This limitation points to an important gap in current AI paradigms, where the statistical nature of LLMs conflicts with the deterministic demands of engineering practice. Subsequent theoretical structures ought to focus on merging neural and symbolic approaches, possibly by employing modular designs in which large language models process linguistic interactions and specialized computational tools perform task-specific analyses.

The examined applications show large language models can greatly cut down time required for documentation, debugging code, and initial design cycles [19] [70] [73]. For engineering firms, this translates to measurable efficiency gains, particularly in the early project phases where rapid prototyping and requirement analysis are critical. However, the implementation challenges are non-trivial. Multiple investigations [33] [40] [57] highlight the necessity of adapting to specific domains by refining models or employing retrieval-augmented generation (RAG) to attain satisfactory outcomes in specialized applications. This implies off-the-shelf LLM deployments could produce less than ideal outcomes unless extensively adapted to civil engineering terminologies and knowledge repositories.

The methodological constraints of this analysis necessitate thorough examination. Although the PRISMA-guided method guaranteed methodical inclusion of principal databases, the swift progress in LLM studies implies that certain recent developments might not be entirely encompassed. Proof-of-concept studies outnumber longitudinal field evaluations [28] [42], potentially fostering undue optimism about practical applicability. Furthermore, the review's focus on English-language publications may overlook valuable contributions from non-English research communities, particularly in regions with active smart infrastructure initiatives. These limitations imply that the documented results could indicate the maximum extent of present abilities, while actual applications confront further obstacles in deployment.

Subsequent investigations ought to focus on key unresolved issues highlighted in this review. There is a pressing need for standardized evaluation benchmarks tailored to civil engineering tasks, as current assessments often rely on general NLP metrics ill-suited to technical domains [31] [77]. Creating such benchmarks would make it possible to compare studies meaningfully and speed up advances in optimizing models for specific domains. Another understudied area is the human-AI collaboration dynamics in engineering teams, where LLMs could reshape traditional workflows but may also introduce new coordination overhead [34] [43]. Research tracking these sociotechnical elements over time would yield important knowledge for organizational implementation.

New prospects are found where LLMs are coupled with digital twin technologies and IoT-supported infrastructure [25] [41]. The capacity to analyze live sensor information together with archived records may lead to automation systems that are more dynamic and adaptable. However, this direction requires advances in multimodal reasoning and temporal modeling capabilities beyond current LLM architectures. Similarly, the application of LLMs for sustainability optimization in civil engineering, such as automated life-cycle assessment and circular design, remains largely unexplored despite its growing importance [59] [65].

The issues of safety and ethics related to the application of LLMs in civil engineering require ongoing focus. Although a number of studies propose evaluation frameworks [80] [82], empirical data on failure modes in operational contexts remains scarce. Future work should investigate the robustness of these models against adversarial inputs in engineering contexts, where malicious alterations to specifications or sensor data could have catastrophic

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

consequences [85]. Furthermore, the ecological consequences of extensive LLM implementation in engineering automation necessitate evaluation, with special attention to the carbon emissions associated with training and inference compared to conventional approaches.

The examined body of work indicates a shift in LLMs from experimental instruments to functional aids in the automation of civil engineering. Nevertheless, their function is expected to stay supplementary rather than replacing in the near term, as human knowledge remains essential for verification and choices. The discipline would gain from greater cooperation between artificial intelligence scholars and professional engineers to align technological progress with practical demands and limitations. As the technology advances, the development of optimal guidelines for ethical implementation will be critical to maximize its advantages while reducing threats to infrastructure security and societal confidence.

V. CONCLUSION

This systematic review has explored the changing function of large language models (LLMs) in automating civil engineering tasks, drawing together findings from various applications such as structural design, construction planning, and traffic management. The results indicate large language models can improve productivity by automating routine documentation, code creation, and initial analyses, but their dependability in safety-critical areas is still limited by shortcomings in deterministic reasoning and specialized knowledge. The review establishes that effective deployments generally feature hybrid architectures that merge neural language models with symbolic systems, which tackle core deficiencies in numerical accuracy and adherence to regulations.

The real-world consequences go further than improvements in productivity, indicating that LLMs may broaden the availability of engineering knowledge via conversational interfaces but also require strict verification procedures. Subsequent studies ought to focus on creating assessment standards tailored to engineering and explore the merging of digital twins and IoT systems across multiple modalities. The discipline must also tackle rising issues in adversarial robustness and environmental sustainability to guarantee ethical implementation. Through the identification of these prospects and constraints, this review establishes a basis for progressing LLM applications in civil engineering while upholding the field's exacting criteria for safety and dependability.

REFERENCES

[1]. H. Khairulzaman and F. Usman, "Automation in civil engineering design in assessing building energy efficiency," *Unable to determine the complete publication venue*, 2018.
[2]. M. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, *et al.*, "A survey on large language models: Applications, challenges, limitations, and practical usage," *Authorea*, 2023.
[3]. K. Du *et al.*, "LLM-MANUF: An integrated framework of fine-tuning large language models for intelligent decision-making in manufacturing," *Advanced Engineering Informatics*, 2025.
[4]. S. Qin *et al.*, "Intelligent design and optimization system for shear wall structures based on large language models and generative artificial intelligence," *Journal of Building Engineering*, 2024.
[5]. J. Saad-Falcon, O. Khattab, K. Santhanam, *et al.*, "UDAPDR: Unsupervised domain adaptation via LLM prompting and distillation of rerankers," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023.
[6]. G. Lee, S. Jang, and S. Hyun, "A generalized LLM-augmented BIM framework: Application to a speech-to-BIM system," arXiv preprint arXiv:2409.18345, 2024.
[7]. Y. Hu, Y. Goktas, D. Yellamati, *et al.*, "The use and misuse of pre-trained generative large language models in reliability engineering," in *2024 annual reliability and maintainability symposium*, 2024.
[8]. Y. Xiong, J. Wang, B. Li, Y. Zhu, and Y. Zhao, "Self-organizing agent network for llm-based workflow automation," arXiv preprint arXiv:2508.13732, 2025.
[9]. R. Agbareia, M. Omar, O. Zloto, N. Chandala, T. Tai, *et al.*, "The role of prompt engineering for multimodal LLM glaucoma diagnosis," medRxiv, 2024.
[10]. M. Page, J. McKenzie, P. Bossuyt, *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.
[11]. C. Du, S. Esser, S. Nousias, *et al.*, "Text2BIM: Generating building models using a large language model-based multiagent framework," *Journal of Computing in Civil Engineering*, 2026.
[12]. J. Chen and Y. Bao, "A multi-agent large language model (llm) framework for code-complying design automation of concrete structures," *Available at SSRN 5193679*, 2025.
[13]. S. Jang and G. Lee, "Interactive design by integrating a large pre-trained language model and building information modeling," *Computing in civil engineering 2023*, 2023.
[14]. J. Chen and Y. Bao, "Artificial intelligence copilot for automated design of buildings using knowledge graphs and numerical models," *Available at SSRN 6330100*, 6330.
[15]. S. Youwai, D. Phim, V. Murcia, and R. Onas, "Large language model-based multi-agent systems for automated foundation design: Router-driven task classification and expert selection framework," *AI in Civil Engineering*, 2026.
[16]. S. Youwai, D. Phim, V. Murcia, and R. Onas, "Investigating the potential of large language model-based router multi-agent architectures for foundation design automation: A task classification and expert …," arXiv preprint arXiv:2506.13811, 2025.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

[17]. H. Liang, M. Kalaleh, and Q. Mei, "Integrating large language models for automated structural analysis," arXiv preprint arXiv:2504.09754, 2025.
[18]. J. Liu, Z. Geng, R. Cao, L. Cheng, P. Bocchini, *et al.*, "A large language model-empowered agent for reliable and robust structural analysis," *Unable to determine the complete publication venue*, 2026.
[19]. P. Parsafard, O. Elezaj, D. Ekundayo, *et al.*, "Automation in construction cost budgeting using generative artificial intelligence," in *Proceedings of the 2024 dubai conference*, 2024.
[20]. A. Singh, A. Pal, and S. Hsieh, "A two-phase AI-driven approach to automated construction planning using small language models for activity sequencing and missing task prediction," *Unable to determine the complete publication venue*, 2025.
[21]. D. Durmus, S. Isaac, A. Carbonari, *et al.*, "Knowledge-based systems in the era of large language models: A case study in fire safety management," in *International symposium on automation and robotics in construction (isarc)*, 2025.
[22]. D. Durmus, A. Giretti, O. Ashkenazi, *et al.*, "The role of large language models for decision support in fire safety planning," *Unable to determine the complete publication venue*, 2024.
[23]. S. Fuchs, M. Witbrock, J. Dimyadi, and R. Amor, "Using large language models for the interpretation of building regulations," arXiv preprint arXiv:2407.21060, 2024.
[24]. J. Zhou and Z. Ma, "Named entity recognition for construction documents based on fine-tuning of large language models with low-quality datasets," *Automation in Construction*, 2025.
[25]. D. Liu, X. Zhou, and Y. Li, "Enhancing natural language retrieval of BIM data through integration of large language models with multi-agent systems," in *Proceedings of CAADRIA*, 2025.
[26]. Q. Chen and X. Yin, "Tailored vision-language framework for automated hazard identification and report generation in construction sites," *Advanced Engineering Informatics*, 2025.
[27]. X. Wang, Q. Yue, and X. Liu, "Crack image classification and information extraction in steel bridges using multimodal large language models," *Automation in Construction*, 2025.
[28]. S. Park, C. Menassa, and V. Kamat, "Integrating large language models with multimodal virtual reality interfaces to support collaborative human–robot construction work," *Journal of Computing in Civil Engineering*, 2025.
[29]. H. Liang, Y. Zhou, M. Kalaleh, and Q. Mei, "Automating structural engineering workflows with large language model agents," arXiv preprint arXiv:2510.11004, 2025.
[30]. S. Jin, D. Kim, J. Lee, and D. Lee, "Language models as BIM interpreters: Unlocking IFC data for automation in construction informatics," *Available at SSRN 5563440*, 5563.
[31]. P. Bazrafshan, K. Melag, and A. Ebrahimkhanlou, "Semantic and lexical analysis of pre-trained vision language artificial intelligence models for automated image descriptions in civil engineering," *AI in civil engineering*, 2025.
[32]. A. Doris, D. Grandi, R. Tomich, *et al.*, "Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation," *Journal of Computing and Information Science in Engineering*, 2025.
[33]. Y. Xia, N. Jazdi, J. Zhang, C. Shah, *et al.*, "Control industrial automation system with large language model agents," *Unable to determine the complete publication venue*, 2025.
[34]. J. Oyekan, C. Turner, M. Bax, and E. Graf, "Applying ontologies and knowledge augmented large language models to industrial automation: A decision-making guidance for achieving human-robot …," arXiv preprint arXiv:2505.18553, 2025.
[35]. L. da Silva, A. Kocher, F. Gehlhoff, *et al.*, "On the use of large language models to generate capability ontologies," in *Ieee international conference on emerging technologies and factory automation*, 2024.
[36]. S. Katragadda, "Utilizing LLM models for advanced automation, manufacturing operations," *Journal of Mechanical, Civil and Industrial Engineering*, 2026.
[37]. Y. Jadhav and A. B. Farimani, "Large language model agent as a mechanical designer," *Journal of Engineering Design*, 2026.
[38]. W. Chen, C. Liu, W. Huang, J. Lyu, M. Yang, *et al.*, "Analogtester: A large language model-based framework for automatic testbench generation in analog circuit design," in *IEEE/ACM international conference on computer aided design*, 2025.
[39]. Y. Li *et al.*, "Large language models for manufacturing," arXiv preprint arXiv:2410.21418, 2024.
[40]. O. Zheng, M. Abdel-Aty, D. Wang, Z. Wang, *et al.*, "ChatGPT is on the horizon: Could a large language model be suitable for intelligent traffic safety research and applications?" arXiv preprint arXiv:2303.05382, 2023.
[41]. S. Li, T. Azfar, and R. Ke, "Chatsumo: Large language model for automating traffic scenario generation in simulation of urban mobility," *IEEE Transactions on Intelligent Vehicles*, 2024.
[42]. S. Masri, H. Ashqar, and M. Elhenawy, "Leveraging large language models (LLMs) for traffic management at urban intersections: The case of mixed traffic scenarios," arXiv preprint arXiv:2408.00948, 2024.
[43]. H. Xu *et al.*, "Genai-powered multi-agent paradigm for smart urban mobility: Opportunities and challenges for integrating large language models (llms) and retrieval-augmented …," arXiv preprint arXiv:2409.00494, 2024.
[44]. Z. Zhou *et al.*, "Human-centric reward optimization for reinforcement learning-based automated driving using large language models," arXiv preprint arXiv:2405.04135, 2024.
[45]. Z. Peng, Y. Wang, X. Han, L. Zheng, and J. Ma, "Learningflow: Automated policy learning workflow for urban driving with large language models," arXiv preprint arXiv:2501.05057, 2025.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

[46]. X. Tian *et al.*, "Drivevlm: The convergence of autonomous driving and large vision-language models," arXiv preprint arXiv:2402.12289, 2024.
[47]. Y. Zhang and Y. Nie, "Interndrive: A multimodal large language model for autonomous driving scenario understanding," in *Proceedings of*, 2024.
[48]. K. Jiang *et al.*, "Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models," *IEEE Transactions on Intelligent Transportation Systems*, 2024.
[49]. P. Larrondo, J. Ortiz, and B. Frank, "Work-in-progress: Fine-tuning large language models for automated feedback in complex engineering problem-solving," in *Asee annual conference*, 2024.
[50]. R. Gao, X. Guo, X. Li, A. Narayanan, N. Thomas, *et al.*, "Towards scalable automated grading: Leveraging large language models for conceptual question evaluation in engineering," arXiv preprint arXiv:2411.03659, 2024.
[51]. W. Morris, L. Holmes, J. Choi, and S. Crossley, "Automated scoring of constructed response items in math assessment using large language models," *International Journal of Artificial Intelligence in Education*, 2025.
[52]. S. Rony, T. Fei, and S. Arsovski, "Educational justice. Reliability and consistency of large language models for automated essay scoring and its implications," *Unable to determine the complete publication venue*, 2025.
[53]. S. Yeung, "A comparative study of rule-based, machine learning and large language model approaches in automated writing evaluation (AWE)," in *Proceedings of the 15th international learning analytics and knowledge conference*, 2025.
[54]. A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," *Research Methods in Applied Linguistics*, 2023.
[55]. X. Niu and J. Zhang, "Enhancing automated text coding in online learning research: A systematic calibration framework for large language models," *IEEE Transactions on Learning Technologies*, 2026.
[56]. M. Perkins, "Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond," *Journal of University Teaching and Learning Practice*, 2023.
[57]. R. Hu, Y. Wu, T. Su, Y. Wang, S. Hu, and J. Huang, "Automated extraction of mechanical constitutive models from scientific literature using large language models: Applications in cultural heritage conservation," arXiv preprint arXiv:2602.16551, 2026.
[58]. S. Bae, M. Jeon, and H. Moon, "Text mining in MOF research: From manual curation to large language model-based automation," *Chemical Communications*, 2025.
[59]. Y. Zhang, S. Itani, K. Khanal, E. Okyere, G. Smith, *et al.*, "Gptarticleextractor: An automated workflow for magnetic material database construction," *Journal of Magnetism and Magnetic Materials*, 2024.
[60]. S. Jiang, D. Evans-Yamamoto, D. Bersenev, *et al.*, "ProtoCode: Leveraging large language models (LLMs) for automated generation of machine-readable PCR protocols from scientific publications," *SLAS technology*, 2024.
[61]. S. Leong, S. Pablo-García, Z. Zhang, *et al.*, "Automated electrosynthesis reaction mining with multimodal large language models (MLLMs)," *Chemical Science*, 2024.
[62]. Y. Wu and F. Tang, "scExtract: Leveraging large language models for fully automated single-cell RNA-seq data annotation and prior-informed multi-dataset integration," *Genome Biology*, 2025.
[63]. J. Pijanowski, Y. Mezgueldi, A. Lee, D. Moghanaki, *et al.*, "Automated extraction of unstructured post-SBRT toxicity data from radiology reports using large language models," arXiv preprint arXiv:2602.23492, 2026.
[64]. B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, *et al.*, "Mathematical discoveries from program search with large language models," *Nature*, 2024.
[65]. D. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous chemical research with large language models," *Nature*, 2023.
[66]. G. Li, C. Zhi, J. Chen, J. Han, and S. Deng, "Exploring parameter-efficient fine-tuning of large language model on automated program repair," in *Proceedings of the 39th ieee/acm international conference on automated software engineering*, 2024.
[67]. Y. Charalambous, E. Manino, and L. Cordeiro, "Automated repair of AI code with large language models and formal verification," arXiv preprint arXiv:2405.08848, 2024.
[68]. H. Pearce, B. Tan, B. Ahmad, R. Karri, *et al.*, "Examining zero-shot vulnerability repair with large language models," in *IEEE symposium on security and privacy*, 2023.
[69]. Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, *et al.*, "Critic: Large language models can self-correct with tool-interactive critiquing," arXiv preprint arXiv:2305.11738, 2023.
[70]. E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, *et al.*, "Codegen: An open large language model for code with multi-turn program synthesis," arXiv preprint arXiv:2203.13474, 2022.
[71]. J. Yao, Z. Zhou, W. Chen, and W. Cui, "Leveraging large language models for automated proof synthesis in rust," arXiv preprint arXiv:2311.03739, 2023.
[72]. R. Liu, M. Chen, L. Wu, J. Ke, and G. Li, "Enhancing automated loop invariant generation for complex programs with large language models," *Science of Computer Programming*, 2025.
[73]. Y. Zhang, Z. Liu, Y. Feng, and B. Xu, "Leveraging large language model to assist detecting rust code comment inconsistency," in *Proceedings of the 39th ieee/acm international conference on automated software engineering*, 2024.
[74]. M. Fazelnia, M. Mirakhorli, and H. Bagheri, "Translation titans, reasoning challenges: Satisfiability-aided language models for detecting conflicting requirements," in *Proceedings of the 39th*

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr455

ieee/acm international conference on automated software engineering, 2024.

[75]. M. Hasan, J. Li, I. Ahmed, and H. Bagheri, "Automated repair of declarative software specifications in the era of large language models," arXiv preprint arXiv:2310.12425, 2023.

[76]. W. Sisomboon, J. Kaewyotha, and W. Songpan, "Automated software test case generation using directional partially weighted ensemble large language models with retrieval-augmented generation (RAG)," *IEEE Access*, 2026.

[77]. N. Jain *et al.*, "Livecodebench: Holistic and contamination free evaluation of large language models for code," arXiv preprint arXiv:2403.07974, 2024.

[78]. M. Du, L. Tuan, B. Ji, Q. Liu, *et al.*, "Mercury: A code efficiency benchmark for code large language models," in *Advances in neural information processing systems*, 2024.

[79]. S. Kawano, H. Nonaka, and K. Yoshino, "Claimbrush: A novel framework for automated patent claim refinement based on large language models," in *2024 IEEE international conference on artificial intelligence and knowledge engineering*, 2024.

[80]. X. Yuan, J. Li, D. Wang, Y. Chen, X. Mao, *et al.*, "S-eval: Towards automated and comprehensive safety evaluation for large language models," *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2025.

[81]. I. Vykopal, M. Pikuliak, I. Srba, R. Moro, *et al.*, "Disinformation capabilities of large language models," in *Proceedings of the 62nd annual meeting of the association for computational linguistics*, 2024.

[82]. Y. Huang *et al.*, "Trustllm: Trustworthiness in large language models," arXiv preprint arXiv:2401.05561, 2024.

[83]. A. Karbasi, O. Montasser, J. Sous, and G. Velegkas, "(Im) possibility of automated hallucination detection in large language models," arXiv preprint arXiv:2504.17004, 2025.

[84]. E. Mugnier, E. Gonzalez, N. Polikarpova, *et al.*, "Laurel: Unblocking automated verification with large language models," in *Proceedings of the 44th ACM SIGPLAN international conference on programming language design and implementation*, 2025.

[85]. A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Kolter, *et al.*, "Universal and transferable adversarial attacks on aligned language models," arXiv preprint arXiv:2307.15043, 2023.

[86]. Y. Liu, Y. Luo, X. Li, X. Dong, B. Gu, *et al.*, "Evaluating large language models for time series anomaly detection in aerospace software," *Unable to determine the complete publication venue with the given information*, 2025.

[87]. T. Zhang, F. Ladhak, E. Durmus, P. Liang, *et al.*, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, 2024.

[88]. J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, *et al.*, "Instruction-following evaluation for large language models," arXiv preprint arXiv:2311.07911, 2023.

[89]. T. Cui *et al.*, "Risk taxonomy, mitigation, and assessment benchmarks of large language model systems," arXiv preprint arXiv:2401.05778, 2024.

[90]. Q. Yu *et al.*, "Xfinder: Large language models as automated evaluators for reliable evaluation," arXiv preprint arXiv:2405.11874, 2024.