future internet

MDPI

Review

Large Language Models for Future Internet Ecosystems: A Taxonomy-Based Review of Smart IoT, Edge Intelligence, and Autonomous AI

Sahar Ahmadzadeh *, Gayathri Karthick and Tariq Alsafi

School of Computer Science, York St John University, London E14 2BA, UK; g.karthick@yorksj.ac.uk (G.K.); t.alsafi@yorksj.ac.uk (T.A.)

* Correspondence: s.ahmadzadeh@yorksj.ac.uk

Abstract

Large Language Models (LLMs) are emerging as key enablers of Future Internet ecosystems, supporting intelligent, adaptive, and context-aware services across distributed cyber-physical environments. Beyond traditional natural language processing, LLMs are increasingly integrated into Smart Internet of Things (SIoT) systems to enable semantic interoperability, edge intelligence, multimodal interaction, and autonomous service orchestration. This paper presents a taxonomy-based review of LLM architectures, training paradigms, deployment strategies, and emerging applications within Future Internet infrastructures. The review classifies existing studies by deployment environment, architecture, training strategy, accessibility, and application scope, and analyses the role of LLMs in intelligent IoT environments, with emphasis on edge-based reasoning, agentic AI, human-centric automation, and context-aware decision-making. Key challenges are examined, including scalability, inference latency, privacy, trustworthiness, security, hallucination, and energy efficiency in resource-constrained environments. A comparative analysis of representative LLMs is presented, based on deployment feasibility, multimodal capability, accessibility, and suitability for distributed intelligent services. The originality of the review lies in conceptualizing LLMs as cognitive middleware that provides semantic, reasoning, and coordination capabilities across Smart IoT infrastructures. Finally, future research directions are highlighted, including decentralized AI architectures, digital twins, the Model Context Protocol, retrieval-augmented generation, multimodal sensing, and autonomous agent-based ecosystems.

Keywords: agentic AI; autonomous systems; edge intelligence; future internet; large language models; model context protocol; multimodal intelligence; natural language processing; smart internet of things; transformer models

Academic Editor: Ivan Serina

Received: 26 April 2026
Revised: 9 July 2026
Accepted: 20 July 2026
Published: 26 July 2026

1. Introduction

The field of Natural Language Processing (NLP) has undergone a significant transformation with the emergence of Large Language Models (LLMs). These models are characterized by large-scale parameterization and training on extensive text corpora, enabling advanced capabilities in natural language understanding, content generation, reasoning, code synthesis, and conversational intelligence [1,2]. Recent advancements in transformer-based architecture have further expanded the applicability of LLMs across healthcare, education, software engineering, cybersecurity, intelligent automation, and digital ser-

vices [3]. Figure 1 summarizes the LLM capabilities and application domains that are directly relevant to Future Internet and Smart IoT ecosystems.

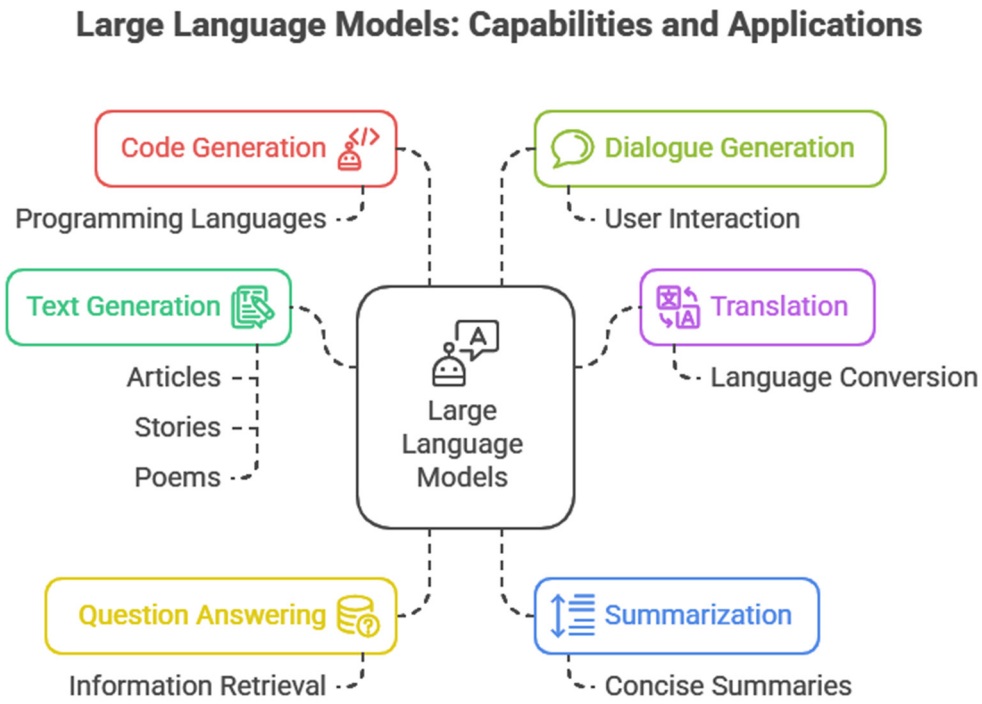

Figure 1. LLM capabilities relevant to Future Internet and Smart IoT ecosystems, including semantic interoperability, edge reasoning, multimodal interaction, autonomous agents, and service orchestration. Source: Authors' own work.

In the context of the Future Internet, LLMs are evolving beyond centralized cloud-based AI systems and becoming integral components of Smart Internet of Things (SIoT) ecosystems. These environments consist of interconnected smart devices, sensors, edge nodes, digital services, and autonomous agents that require intelligent coordination, semantic interoperability, and adaptive decision-making [4]. LLMs provide an effective interface for bridging human intent with machine operations through natural language interaction, contextual reasoning, and intelligent service orchestration. Consequently, LLMs are increasingly recognized as enabling technologies for edge intelligence, digital twins, multimodal sensing, decentralized AI infrastructures, and autonomous intelligent ecosystems.

Despite these opportunities, deploying LLMs in Future Internet and SIoT environments presents several challenges. These challenges include scalability, inference latency, privacy preservation, trustworthiness, security, energy efficiency, hallucination, and the complexity of deploying large-scale models in resource-constrained edge environments. Therefore, Future Internet-oriented research must move beyond general NLP capabilities and examine how LLMs can be classified, evaluated, and integrated into distributed, intelligent, and human-centric digital infrastructures.

Although the literature on LLMs has expanded rapidly, much of the existing research focuses primarily on general-purpose conversational AI, centralized cloud-based deployment, or broad NLP capabilities. Limited attention has been given to how LLMs can support distributed intelligence, Smart IoT systems, edge computing environments, autonomous AI agents, and digital ecosystem orchestration within the Future Internet paradigm. Furthermore, existing studies often lack a structured taxonomy for evaluating LLMs based on deployment feasibility, edge suitability, multimodal capability, privacy preservation, and intelligent service orchestration.

Several review studies have examined LLMs from perspectives such as architecture, applications, training strategies, and ethical challenges. For example, Raiaan et al. [1] and Kumar [2] provide broad overviews of LLM architectures, applications, taxonomies, and open research challenges. Similarly, Shahzad et al. [3] focus on LLM applications within learning environments, while recent studies have explored the integration of LLMs with IoT ecosystems and distributed AI infrastructures [5,6]. However, limited attention has been given to the role of LLMs within Future Internet ecosystems, particularly in relation to Smart IoT integration, edge intelligence, autonomous AI systems, multimodal orchestration, and MCP-enabled infrastructures. A comparison of this paper against existing review studies is provided in Table 1. Therefore, this paper addresses these gaps through a taxonomy-based review focused on Future Internet and autonomous intelligent ecosystems.

Table 1. Comparison of existing LLM review studies.

Survey	Smart IoT	Edge AI	Agentic AI	MCP	Taxonomy
Raiaan et al. [1]	No	No	No	No	Partial
Kumar [2]	No	No	No	No	Partial
Shahzad et al. [3]	No	No	No	No	No
This paper	Yes	Yes	Yes	Yes	Yes

Source: Developed by the authors.

This review is intended for researchers, system architects, IoT developers, edge-AI practitioners, and policy-oriented readers who require a structured understanding of how LLMs can be integrated into Future Internet and Smart IoT infrastructures. The rationale for the review is that most existing LLM surveys remain centered on general NLP capabilities, whereas Future Internet systems require analysis of deployment feasibility, interoperability, latency, privacy, multimodal sensing, and autonomous orchestration.

The study is guided by the following research questions:

1. How can LLMs be taxonomically classified for Future Internet and Smart IoT ecosystems?
2. What are the key technical challenges of deploying LLMs in edge and resource-constrained IoT environments?
3. How can Agentic AI, retrieval-augmented generation, multimodal LLMs, and MCP enhance autonomous Future Internet services?

This paper aims to provide a taxonomy-based review of LLMs for Future Internet ecosystems. Specifically, this study:

- Reviews the evolution and technical foundations of LLMs;
- Examines the role of LLMs in Smart IoT and edge intelligence environments;
- Proposes a taxonomy for classifying LLMs based on architecture, deployment strategy, accessibility, and application scope;
- Compares representative LLMs from the perspective of edge feasibility, latency, privacy, multimodal support, and IoT suitability;
- Discusses emerging trends including Agentic AI, MCP, retrieval-augmented generation, multimodal reasoning, and autonomous intelligent ecosystems.

The originality of this review lies in positioning LLMs as cognitive middleware for Future Internet systems, providing semantic interoperability, context-aware reasoning, autonomous service orchestration, and adaptive coordination between users, edge devices, autonomous agents, and distributed digital infrastructures. Unlike conventional LLM surveys, this paper focuses specifically on Smart IoT integration, edge intelligence, autonomous orchestration, deployment challenges, and the future role of LLMs in intelligent digital ecosystems.

In this review, cognitive middleware refers to a software layer positioned between heterogeneous IoT devices, sensors, and networks on one side and application-level services on the other, in which LLMs provide four functions that conventional middleware does not: (1) semantic interoperability, translating between heterogeneous data schemas, protocols, and natural-language interfaces; (2) context-aware reasoning, aggregating multimodal sensor input into actionable state; (3) autonomous service orchestration, coordinating tools, agents, and external knowledge sources; and (4) adaptive coordination across distributed edge and cloud nodes [4–6]. An LLM deployment qualifies as cognitive middleware only when it mediates between the device layer and the application layer; standalone conversational use does not.

The remainder of this paper is organized as follows. Section 2 presents the review methodology. Section 3 discusses the historical evolution and technical foundations of LLMs. Section 4 examines training paradigms and deployment-oriented adaptation methods. Section 5 introduces the proposed taxonomy for LLM characterization. Section 6 presents a comparative analysis of representative LLMs. Section 7 discusses Smart IoT and multimodal use cases. Section 8 provides the discussion, including contributions, implications, key takeaways, and limitations. Section 9 highlights major challenges and open issues, while Section 10 outlines future research directions and Section 11 concludes the paper.

2. Review Methodology

This study adopts a taxonomy-based narrative review methodology to examine the role of LLMs within Future Internet ecosystems, particularly in the context of Smart Internet of Things, edge intelligence, autonomous AI systems, and distributed intelligent infrastructures. The primary objective of this methodology is to classify and critically analyze literature according to themes relevant to intelligent Future Internet environments and next-generation digital ecosystems, while maintaining a transparent procedure for literature identification, screening, and thematic coding. The overall taxonomy-based review framework, including literature identification, screening, quality assessment, thematic coding, and synthesis, is illustrated in Figure 2.

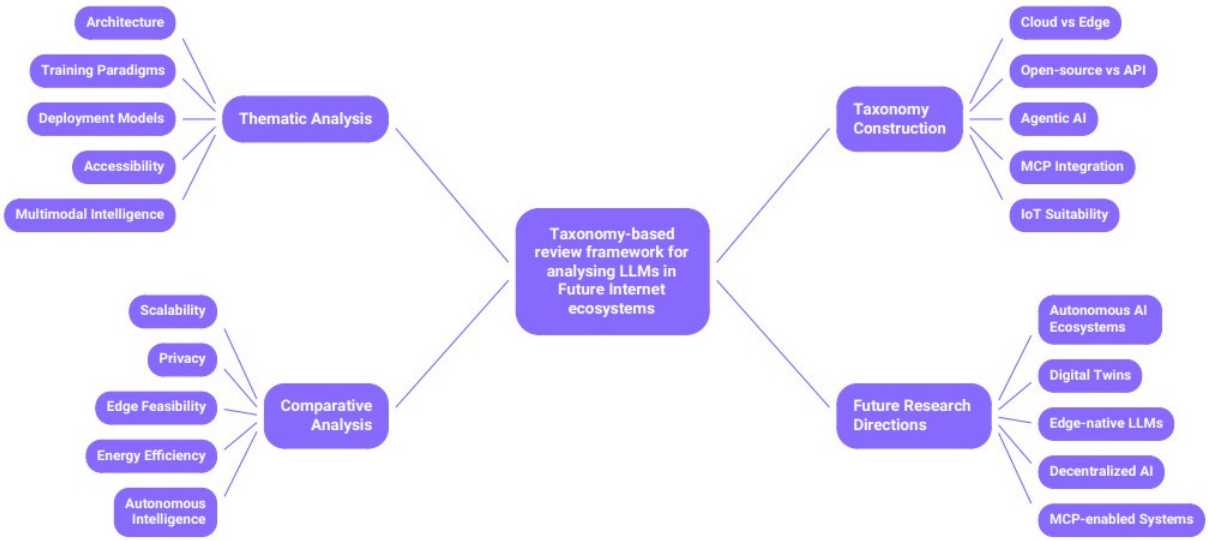

Figure 2. Taxonomy-based review framework for analyzing LLMs in Future Internet ecosystems. Source: Authors' own work.

Relevant literature was searched in IEEE Xplore, ACM Digital Library, Scopus, Web of Science, ScienceDirect, SpringerLink, arXiv, and Google Scholar. The search covered

January 2017 to May 2026. Search terms included combinations of "large language model *", "LLM", "Future Internet", "Internet of Things", "Smart IoT", "edge intelligence", "edge AI", "autonomous agents", "multimodal LLM", "retrieval augmented generation", "digital twins", and "Model Context Protocol". Boolean operators were used to combine LLM-related terms with Future Internet, IoT, edge, multimodal, and autonomous-agent terms.

Studies were included if they addressed one or more of the following areas: LLM architectures, training paradigms, deployment strategies, Smart IoT integration, edge computing, multimodal interaction, autonomous agents, distributed intelligence, digital twins, privacy preservation, trustworthiness, security, energy efficiency, or Future Internet infrastructures. Studies were excluded if they were duplicates, non-English papers, short opinion pieces, or works focused only on general conversational AI without relevance to Future Internet, Smart IoT, or distributed intelligent systems.

The screening procedure involved duplicate removal, title and abstract screening, full-text assessment, and thematic classification. Study quality was assessed using relevance to the research questions, technical depth, methodological clarity, source credibility, and applicability to Future Internet or Smart IoT systems. The final literature set was coded thematically according to architecture, training paradigm, deployment environment, accessibility, application scope, edge feasibility, multimodal capability, privacy/security, and autonomous intelligence. The taxonomy dimensions were selected because they directly reflect the design decisions required when LLMs are deployed in cloud-edge, resource-constrained, privacy-sensitive, and interoperable Future Internet environments.

Because the manuscript is a taxonomy-based review rather than an experimental benchmarking study, the comparison criteria were derived from reported model characteristics and deployment-related literature rather than direct device-level measurements. Where public values are unavailable, the paper reports qualitative deployment implications instead of unsupported numerical claims, as shown in Table 2 below.

Table 2. Search and screening summary.

Element	Description
Databases searched	IEEE Xplore, ACM Digital Library, Scopus, Web of Science, ScienceDirect, SpringerLink, arXiv, and Google Scholar
Date coverage	January 2017 to May 2026
Search strings	("large language model *" OR LLM) AND ("Future Internet" OR "Smart IoT" OR "Internet of Things" OR "edge intelligence" OR "edge AI" OR "autonomous agent *" OR "multimodal LLM" OR "retrieval augmented generation" OR "digital twin *" OR "Model Context Protocol")
Final included studies	66 articles
Inclusion criteria	Studies addressing LLMs, Future Internet, Smart IoT, edge AI, multimodal systems, autonomous agents, privacy, security, energy efficiency, or distributed intelligent infrastructures
Exclusion criteria	Duplicates, non-English papers, short opinion pieces, papers focused only on generic conversational AI, and papers without relevance to Future Internet, SIoT, edge AI, or distributed intelligence
Quality assessment	Relevance, technical depth, methodological clarity, source credibility, and applicability to Future Internet/SIoT systems
Coding procedure	Thematic coding by architecture, training, deployment, accessibility, application scope, edge suitability, multimodality, privacy/security, and autonomous intelligence

Developed by the authors.

Following literature collection and relevance screening, the selected studies were analyzed and organized into thematic dimensions, including architecture, training paradigms, deployment models, accessibility, multimodal intelligence, edge feasibility, privacy and security considerations, Smart IoT integration, and emerging research directions. These dimensions were then used to construct the proposed taxonomy framework and comparative analysis presented in later sections of this paper.

3. Historical Evolution and Technical Foundations of Large Language Models

The evolution of LLMs has been driven by advancements in NLP, deep learning, transformer models, and large-scale computational infrastructures. From early rule-based systems to modern foundation models, LLMs have progressively improved their ability to understand contextual relationships, generate coherent language, perform reasoning tasks, and support multimodal interactions. In this review, the historical overview is presented only to establish the technical foundations required to understand LLM deployment in Future Internet and Smart IoT ecosystems, as illustrated in Figure 3 below.

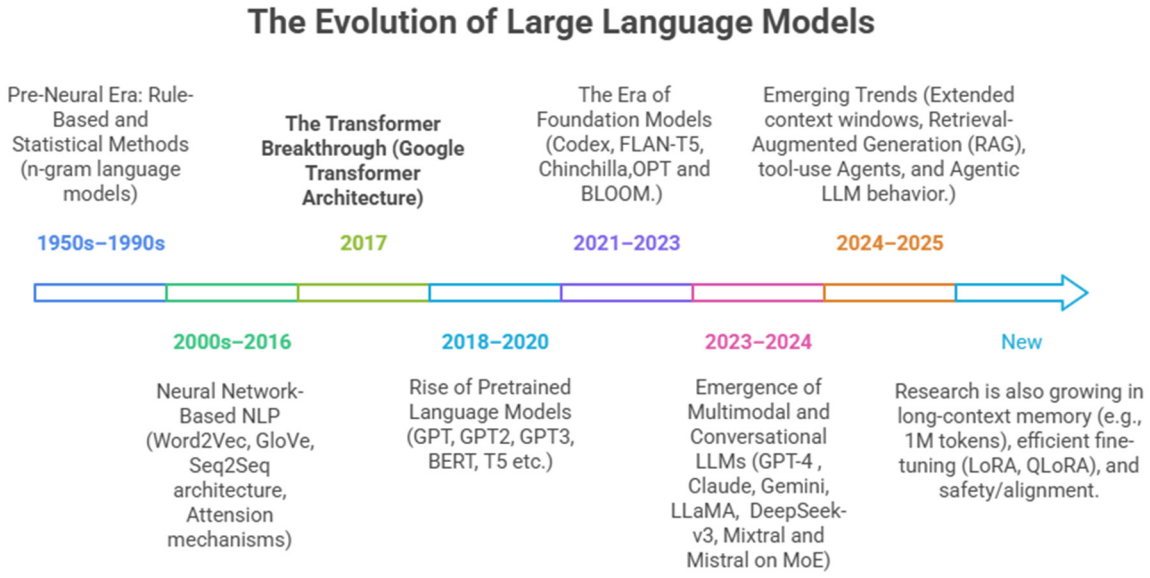

Figure 3. Historical evolution of LLMs from rule-based NLP systems to transformer-based multimodal and agentic AI architectures. Source: Authors' own work.

3.1. Pre-Neural and Statistical NLP Era

Early NLP systems primarily relied on rule-based approaches and symbolic artificial intelligence, where linguistic knowledge was manually encoded using handcrafted grammar, lexicons, and logical rules. During the 1990s, statistical methods such as n-gram models and Hidden Markov Models became dominant for sequence prediction, machine translation, and speech-recognition tasks [7–9]. These approaches improved computational efficiency and probabilistic language modeling; however, they struggled to capture long-range contextual dependencies, semantic relationships, and deeper linguistic meaning.

3.2. Neural Network-Based NLP and Attention Mechanisms

The introduction of neural network-based approaches significantly advanced language modeling and representation learning. Word embedding techniques such as Word2Vec and GloVe enabled semantic vector representations of words within continuous feature spaces, improving contextual understanding and similarity detection [10]. Recurrent Neural Networks, Long Short-Term Memory networks, and Gated Recurrent Units improved

sequential language modeling by capturing contextual dependencies across longer text sequences. Further advancements were achieved through sequence-to-sequence learning and attention mechanisms, which enhanced machine translation, contextual representation learning, and sequence modeling capabilities [11,12].

3.3. Transformers and Foundation Models

The introduction of the Transformer architecture in "Attention Is All You Need" transformed NLP by replacing recurrent architectures with self-attention mechanisms [13]. This development enabled efficient parallel computation, improved contextual representation learning, and scalable sequence modeling. Following the emergence of transformer architectures, a new generation of pretrained language models evolved, including GPT, BERT, XLNet, and ALBERT [14–16]. These models introduced transfer learning paradigms in NLP, where large-scale pretraining on unlabeled corpora was followed by task-specific fine-tuning.

Between 2019 and 2021, LLMs expanded significantly in both scale and capability. GPT-2 demonstrated advanced language-generation performance, while GPT-3 increased model scale to 175 billion parameters and introduced strong zero-shot and few-shot learning capabilities [17,18]. During the same period, models such as T5, ELECTRA, and PaLM explored improved pretraining strategies, multitask learning, and large-scale distributed training approaches [19–21]. These developments established the technological foundation for modern multimodal, conversational, and agentic AI systems, while also highlighting challenges related to scalability, computational cost, alignment, and responsible AI deployment.

3.4. Multimodal and Agentic LLMs

Recent advancements have extended LLMs beyond text-based processing toward multimodal intelligence and autonomous agent-based systems. Modern LLMs are increasingly capable of processing and integrating multiple forms of data, including text, images, audio, video, code, and sensor information. These developments are highly relevant to Future Internet and Smart IoT environments because such systems must interpret heterogeneous data streams while supporting context-aware interaction and autonomous service coordination.

The emergence of agentic AI represents another major shift in LLM development. Agentic LLMs can support planning, reasoning, memory management, and tool execution with varying levels of human supervision. Unlike traditional chat-based systems, agentic AI frameworks can decompose complex tasks into subtasks, interact with external tools, retrieve information dynamically, and adapt actions based on environmental feedback. ReAct combines reasoning and action strategies for interactive problem solving [22], while Toolformer demonstrates how language models can learn to use external tools [23]. Retrieval-Augmented Generation (RAG) improves factual grounding by combining pretrained LLMs with external knowledge-retrieval mechanisms [24–26].

Another emerging concept is the Model Context Protocol, which supports standardized communication between LLMs, external tools, databases, memory systems, and autonomous agents [4]. MCP-based ecosystems are important for Future Internet architectures, edge intelligence, and cyber-physical systems because interoperability, scalability, and real-time adaptation are essential. Nevertheless, multimodal and agentic LLMs introduce challenges related to computational overhead, security risks, hallucination control, privacy preservation, trustworthiness, and safe autonomous reasoning.

4. Training Paradigms and Deployment-Oriented Adaptation

The development of LLMs is supported by several training paradigms that enable linguistic competence, reasoning, task adaptation, alignment, and deployment flexibility. These paradigms include large-scale pretraining, supervised fine-tuning, instruction tuning, reinforcement learning from human feedback, prompting strategies, and efficiency-oriented adaptation methods. Within Future Internet and Smart IoT ecosystems, these approaches are important because LLMs must operate across distributed, privacy-sensitive, and resource-constrained environments while supporting intelligent and context-aware interactions.

4.1. Pretraining and Supervised Fine-Tuning

The foundation of modern LLMs is large-scale pretraining on extensive text corpora. Autoregressive models such as GPT learn through next-token prediction, whereas bidirectional models such as BERT employ masked language modeling to capture contextual relationships within textual data. This pretraining phase enables LLMs to acquire broad linguistic patterns, semantic representations, and general world knowledge from large-scale datasets. Supervised fine-tuning further adapts pretrained models to domain-specific tasks and application environments, including question answering, summarization, code generation, healthcare support, cybersecurity analysis, and IoT data interpretation.

4.2. Instruction Tuning and Human Alignment

Instruction tuning improves the capability of LLMs to follow natural language commands and generate outputs aligned with user intent. Models such as InstructGPT and FLAN-T5 demonstrate how training on diverse instruction-response pairs can improve generalization across multiple tasks [27,28]. This capability is particularly important in Smart IoT environments, where users may interact with intelligent systems through natural language requests related to energy optimization, security monitoring, healthcare assistance, and smart-home automation. Reinforcement Learning from Human Feedback further improves alignment by incorporating human evaluations into the training process through reward modeling and response optimization [28,29].

4.3. Prompting and In-Context Learning for Smart IoT

Prompting techniques enable LLMs to perform downstream tasks without updating model parameters. Approaches such as zero-shot, few-shot, and chain-of-thought prompting allow models to generalize from instructions or examples provided within the input context. In Future Internet environments, prompting strategies support flexible interaction with intelligent systems, where users or autonomous agents issue context-specific requests and receive adaptive responses. However, prompt examples should be evaluated in relation to the deployment context rather than as generic translation, arithmetic, or email-writing tasks. Table 3 illustrates Smart IoT-oriented prompt categories with required context, evaluation rules, and possible failure cases.

4.4. Efficiency-Oriented Training and Adaptation

As LLMs increase in scale and complexity, efficiency-oriented adaptation methods have become essential for practical deployment within edge and IoT environments. Scaling-law research demonstrates that model performance depends on the balance between model size, training data, and computational resources [30]. Compute-optimal training further suggests that increasing parameter size alone is insufficient, emphasizing the importance of efficient data utilization and computational allocation strategies [31,32].

Table 3. Smart IoT-oriented prompt examples for evaluating LLM behavior in context-aware environments.

Smart IoT Category	Example Prompt	Required Context	Evaluation Rule	Possible Failure Case
Smart-home energy management	Analyze today's smart-meter, occupancy, and thermostat data and identify the cause of abnormal energy consumption.	Smart-meter readings, occupancy status, thermostat logs, peak-hour tariff data.	Correctly identifies the device/time cause and provides an actionable recommendation.	Hallucinates a cause that is not present in the sensor logs.
Smart healthcare and wearables	Summarize abnormal heart-rate and sleep-pattern changes and indicate whether escalation is needed.	Wearable sensor data, baseline user profile, clinical thresholds, recent alerts.	Detects abnormal trends without giving an unsupported diagnosis.	Produces medical advice beyond available evidence or user context.
Industrial IoT maintenance	Use vibration, temperature, and maintenance logs to identify whether the motor requires inspection.	Sensor stream, asset identifier, historical faults, maintenance records.	Maps anomaly patterns to inspection recommendations and uncertainty level.	Confuses normal operational variation with equipment failure.
Smart-city traffic and emergency response	Combine traffic-flow, camera, and incident-report data to recommend a rerouting strategy.	Multimodal traffic data, event reports, map constraints, time stamps.	Provides route-level recommendation with explanation and time relevance.	Uses stale data or ignores latency-sensitive constraints.

Developed by the authors.

Techniques such as quantization, pruning, knowledge distillation, Low-Rank Adaptation, and Quantized Low-Rank Adaptation reduce computational overhead and memory requirements, making LLM deployment more feasible for edge devices and resource-constrained systems [33]. These approaches are particularly relevant in Future Internet infrastructures, where latency, energy efficiency, bandwidth limitation, and privacy preservation are critical operational requirements.

4.5. Multimodal, Tool-Augmented, and Retrieval-Based Training

Modern LLMs are increasingly extended with multimodal, tool-augmented, and retrieval-based capabilities. Multimodal models integrate text, images, audio, video, code, and sensor data, enabling richer contextual understanding and adaptive interaction within intelligent environments. This capability is particularly important in Smart IoT ecosystems, where data may originate from cameras, wearable devices, environmental sensors, smart meters, and interconnected cyber-physical systems. Tool-augmented models improve reasoning and factual accuracy by enabling LLMs to interact with external tools such as calculators, APIs, databases, search systems, and software frameworks [22,23]. RAG further improves factual grounding by retrieving relevant external knowledge before response generation [24–26].

Within Future Internet ecosystems, these paradigms support distributed intelligence, real-time decision-making, and autonomous service orchestration across cloud-edge environments. Federated learning, edge-based fine-tuning, and privacy-preserving training are also becoming important because they enable decentralized learning from IoT-generated data while preserving user privacy and reducing communication overhead [5,6].

5. Proposed Taxonomy Framework for LLM Characterization

To move beyond broad parameter-based comparisons and enable a more structured analysis of LLMs, this study proposes a taxonomy framework for characterizing modern LLM ecosystems. As outlined in Table 4, the proposed taxonomy classifies LLMs across five critical dimensions: deployment environment, architecture, training strategy, accessibility, and application scope. These dimensions were selected because they directly influence deployment feasibility, operational reliability, and integration with Future Internet and Smart IoT infrastructures.

Table 4. Taxonomy dimensions for analyzing LLMs in Future Internet and Smart IoT ecosystems.

Dimension	Definition	Future Internet/SIoT Relevance
Deployment environment	Cloud-based, edge-deployed, on-device, or hybrid cloud-edge operation.	Determines latency, bandwidth use, privacy exposure, energy demand, and resilience.
Architecture	Decoder-only, encoder–decoder, multimodal transformer, sparse MoE, or agentic/tool-augmented architecture.	Influences reasoning ability, memory footprint, multimodal processing, and inference efficiency.
Training strategy	Autoregressive pretraining, masked language modeling, instruction tuning, RLHF, RAG, or parameter-efficient tuning.	Controls adaptation to domain tasks, factual grounding, user alignment, and safe interaction.
Accessibility	Open-source/open-weight, API-based commercial model, or restricted-access model.	Affects reproducibility, customization, governance, privacy control, and edge deployment.
Application scope	General-purpose assistant, domain-specific model, multimodal assistant, or autonomous agent.	Connects model capability with Smart IoT workflows, digital twins, service orchestration, and autonomous systems.

Developed by the authors.

Unlike traditional surveys that primarily emphasize model scale and benchmark performance, the proposed taxonomy incorporates edge intelligence, multimodal capability, distributed deployment, accessibility, and autonomous agent-based functionality. These characteristics are increasingly important within Future Internet ecosystems, where intelligent systems operate across cloud-edge infrastructures while supporting scalability, interoperability, privacy preservation, and context-aware service orchestration.

Each taxonomy dimension operationalizes one of the cognitive middleware functions defined in Section 1: semantic interoperability is realized through multimodal and instruction-tuned architectures, context-aware reasoning through retrieval-augmented grounding [24–26], autonomous service orchestration through tool-augmented and agentic mechanisms [22,23], and adaptive coordination through the deployment environment and edge-oriented training dimensions. The classification dimensions are therefore not arbitrary descriptors, but the middleware functions expressed as observable model properties.

The taxonomy provides practical insights for selecting suitable LLM architectures based on deployment constraints, computational resources, privacy requirements, and intelligent service-orchestration needs, as described in Table 5. For example, large proprietary models may provide stronger multimodal and reasoning capabilities, but lightweight open models are more feasible for edge-based and resource-constrained environments. Similarly, RAG and agentic architecture are more appropriate when distributed systems require dynamic information retrieval, tool use, and autonomous task execution.

Table 5. Proposed taxonomy applied to representative LLMs.

Model	Deployment	Architecture	Training/ Alignment Strategy	Accessibility	Application Scope	Future Internet/ SIoT Relevance
GPT-3 [17]	Cloud	Decoder-only Transformer	Autoregressive pretraining; alignment through later instruction/ RLHF systems	API-based	General language generation	Strong reasoning support but low edge feasibility.
GPT-4 [34]	Cloud	Multimodal Transformer	Instruction alignment and multimodal adaptation	API-based	Multimodal assistant	Useful for high -level orchestration, but dependent on cloud connectivity.

Table 5. *Cont.*

Model	Deployment	Architecture	Training/ Alignment Strategy	Accessibility	Application Scope	Future Internet/ SIoT Relevance
PaLM [19]	Cloud	Decoder-only Transformer	Large-scale pretraining and instruction tuning	Restricted/API	General reasoning and multilingual tasks	High capability but low direct edge suitability.
LLaMA/ LLaMA 2 [35,36]	Cloud-edge/hybrid	Decoder-only Transformer	Pretraining and instruction tuning	Open-weight with license conditions	Research and customization	More feasible for compression, adaptation, and edge experimentation.
BLOOM [37]	Cloud/hybrid	Decoder-only Transformer	Causal language modeling	Open-access	Multilingual text generation	Supports reproducibility but has high compute demand.
BioGPT [38]	Cloud	Transformer-based	Biomedical pretraining	Open-source	Biomedical text generation	Illustrates domain-specific specialization for smart healthcare.
Mistral 7B [39]	Cloud-edge/hybrid	Decoder-only Transformer with efficiency optimizations	Efficient pretraining; instruction-tuned variant available	Open-weight	Efficient inference and general tasks	Relevant to edge and resource-aware LLM deployment.
Mixtral [40]	Cloud-edge/hybrid	Sparse Mixture-of-Experts	Sparse MoE pretraining and instruction tuning	Open-weight	Efficient large-scale reasoning	Relevant where sparse activation can reduce active inference cost.

Developed by the authors.

6. Comparative Analysis of Prominent Large Language Models

Several LLMs have emerged with diverse architectural designs, training objectives, accessibility models, and deployment characteristics. Comparative analysis of these models provides important insights into scalability, reasoning capability, multimodal support, openness, computational efficiency, and suitability for distributed intelligent environments. Within Future Internet and Smart IoT ecosystems, model comparison must also consider deployment feasibility in edge and resource-constrained environments, including inference latency, memory footprint, computational cost, energy efficiency, privacy implications, and compatibility with smart-device workflows.

6.1. Model Scale, Architecture, and Accessibility

Prominent LLMs differ significantly in scale, architecture, training objectives, and deployment characteristics. GPT-3 contains 175 billion parameters and follows a decoder-only Transformer architecture, demonstrating strong few-shot and zero-shot learning capabilities [17]. GPT-4 extends these capabilities toward multimodal processing by supporting both text and image inputs, although its detailed architecture and parameter size remain undisclosed [34]. PaLM scales to 540 billion parameters using the Pathways framework for efficient distributed training and enhanced multilingual and multitask performance [19]. In contrast, LLaMA-family models emphasize accessibility and computational efficiency, with smaller variants achieving competitive performance through optimized training strategies [35,36].

Accessibility is a critical consideration in the LLM ecosystem. Commercial API-based models can provide strong reasoning and multimodal capabilities, but they limit reproducibility, local customization, and direct privacy control. Open-weight models such as BLOOM, LLaMA-family models, Mistral, and Mixtral support greater transparency and customization, although their licenses, infrastructure requirements, and deployment constraints must be evaluated carefully [36,37,39,40].

6.2. Deployment-Aware Comparison for Future Internet and Smart IoT

For Future Internet and Smart IoT deployment, the most relevant comparison variables are not only parameter count and model architecture. Edge feasibility, latency sensitivity, quantization potential, bandwidth dependency, retrieval/tool support, security exposure, privacy control, and smart-device compatibility are equally important. Table 6 summarizes deployment-oriented characteristics of representative models. The ratings are qualitative because public reporting of model internals, energy consumption, memory footprint, and latency is inconsistent, and the present review does not perform direct benchmarking on specific hardware.

6.3. Efficiency and Environmental Impact

As LLMs continue to scale in size and complexity, efficiency and environmental sustainability remain critical research concerns. Training and deploying frontier-scale models require substantial computational resources and energy consumption, raising concerns regarding sustainability, accessibility, and environmental impact [41,42]. Techniques such as sparsity, quantization, pruning, knowledge distillation, and parameter-efficient fine-tuning can reduce computational overhead, but the balance among model performance, deployment cost, energy efficiency, and system reliability remains an open problem in Future Internet environments.

The comparative analysis highlights important trade-offs among modern LLMs. Large proprietary models demonstrate strong reasoning and multimodal capabilities but require extensive computational resources and centralized cloud infrastructures. In contrast, lightweight and open-weight models offer improved deployment flexibility, making them more suitable for edge intelligence and Smart IoT applications when combined with compression, retrieval, and safety mechanisms. Therefore, Future Internet-oriented LLM evaluation should move beyond general benchmarks and include deployment-aware metrics such as latency, memory footprint, bandwidth use, privacy control, security robustness, and smart-device compatibility.

Table 6. Deployment-aware comparison of representative LLMs for Future Internet and Smart IoT systems.

Model	Deployment Feasibility	Edge Feasibility	Quantization Support	Memory/compute Implication	Latency Implication	Tool/RAG Support	Privacy/Security Implication	Smart-Device Compatibility
GPT-3 [17]	Cloud/API	Low	Provider-dependent	Very high; not suitable for local edge deployment	Network/API latency; cloud dependent	Can be integrated through external orchestration	Data leaves local environment unless protected by service controls	Suitable for cloud-connected assistants, not direct device inference
GPT-4 [34]	Cloud/API	Low	Provider-dependent	Undisclosed but frontier-scale; cloud dependent	Network/API latency; strong reasoning trade-off	Supports tool and retrieval workflows through applications	Strong capability but privacy depends on deployment governance	Useful for high-level orchestration and multimodal interpretation
PaLM [19]	Cloud/API	Low	Provider-dependent	Very high; centralized infrastructure	Cloud latency; unsuitable for time-critical local control	Can support tool/RAG pipelines through external systems	Centralized processing raises governance and data-control issues	Limited direct compatibility with embedded devices
LLaMA/ LLaMA 2 [35,36]	Local, cloud, or hybrid	Medium to high depending on variant	Commonly quantized in practice	Lower for 7B/13B variants; still hardware dependent	Can be reduced with local inference and compression	Can be combined with RAG, agents, and tools	Local deployment can improve privacy control	Suitable for gateways, edge servers, and constrained deployments after compression
BLOOM [37]	Cloud/hybrid	Medium	Possible but large variants remain heavy	High for 176B-scale model	High without optimized infrastructure	Can be integrated with RAG/tool layers	Open access supports auditability; compute demand remains high	Better suited to research/cloud than direct edge deployment
Mistral 7B [39]	Local, cloud, or hybrid	High relative to large models	Commonly quantized in practice	Lower memory footprint than very large models	Improved by efficient architecture and smaller size	Can be combined with RAG, tool use, and agent frameworks	Local deployment supports stronger data control	Suitable for edge servers and capable gateways after optimization
Mixtral [40]	Local, cloud, or hybrid	Medium to high depending on hardware	Possible; sparse activation supports efficiency	High total parameters but lower active parameters per token	Potential throughput benefits under optimized serving	Can be combined with RAG/tool layers	Open-weight deployment supports governance but requires secure serving	Suitable for edge-cloud and local server settings rather than microcontrollers

Developed by the authors.

7. Smart IoT and Multimodal Use Cases

LLMs should not be understood as replacements for low-level controllers or real-time embedded systems. Instead, their most suitable role is to provide semantic interpretation, natural-language interaction, anomaly explanation, decision-support summaries, workflow orchestration, and coordination among distributed services. In this role, LLMs can act as cognitive middleware between users, smart devices, edge nodes, digital twins, and cloud services.

7.1. Smart-Home and Ambient Assisted Living

Smart-home applications illustrate the need for multimodal sensing rather than text-only interaction. Ambient Assisted Living systems may combine data from Bluetooth beacons, cameras, wearable inertial sensors, motion sensors, door sensors, thermostats, power meters, and environmental sensors to infer context, activities, occupancy, safety risks, or anomalies. Indoor localization is particularly important in elderly care scenarios because fall detection, mobility monitoring, cognitive-assistance support, and caregiver notification require not only recognition of an abnormal event, but also knowledge of where the event occurred inside the home [43,44]. For LLM-based systems, such sensor fusion can provide the grounding needed for more reliable explanations, alerts, and user-facing recommendations. BLE beacon and scanner data can support activity-zone localization, while accelerometer and gyroscope data can provide motion-based context for detecting the user's zone-based indoor position. Such approaches are relevant to smart homes because they connect user activity, movement patterns, and indoor position with practical care scenarios such as fall response, dementia-related monitoring, mobility support, and emergency communication [43].

A broader body of work on non-contact indoor human monitoring further shows that elderly care environments benefit from combining visual, radio-frequency, inertial, ambient, pressure-based, audio, and vital-sign sensing modalities. These modalities support activity recognition, gait analysis, vital-sign measurement, user identification, localization, and 3D modeling of indoor environments. Their combination can improve robustness and adaptability because each modality has different strengths and limitations. Visual sensors provide rich behavioral context but may raise privacy concerns, while radio-frequency and radar-based systems can support less intrusive monitoring in low-light or privacy-sensitive spaces [45].

For LLM-enabled smart homes, these studies provide the sensing-layer foundation. Multimodal sensor fusion can generate the contextual evidence needed for LLM-based explanation, alert summarization, caregiver-facing communication, and user-facing recommendations. For example, an LLM integrated with a smart-home platform could summarize why energy consumption increased, explain whether a detected fall-risk event requires escalation, or translate complex sensor readings into accessible language for older adults and caregivers. However, safety-critical actuation should remain governed by verified control logic, domain rules, and human oversight. The LLM should provide interpretation and orchestration rather than unrestricted autonomous control over physical devices.

7.2. Industrial IoT and Predictive Maintenance

In industrial IoT settings, LLMs can support predictive-maintenance workflows by summarizing sensor anomalies, maintenance logs, and operational constraints. Recent work on compressor monitoring shows how LLM-based predictive-maintenance frameworks can integrate heterogeneous sensor telemetry, maintenance reports, technical manuals, and operational context to support multimodal anomaly detection and maintenance decision support [46]. Such examples are relevant to Future Internet-oriented industrial IoT systems

because they show how LLMs can connect operational data, expert knowledge, and human-facing decision support in complex industrial environments.

When combined with RAG, an LLM can retrieve equipment manuals, historical fault records, and safety procedures before generating recommendations [26,47]. This reduces the risk of unsupported responses and improves traceability. RAG-based industrial troubleshooting is particularly useful where technicians must diagnose faults under data overload, outdated documentation, reliance on scarce expert knowledge, or time-sensitive operating conditions [47]. Nevertheless, deployment must account for latency, cybersecurity, access control, and the need to verify LLM outputs before they affect industrial operations.

7.3. Smart Cities and Urban Intelligence

Smart-city applications require integration across traffic sensors, environmental monitoring, public-safety systems, energy grids, and citizen-facing services. LLMs can support city operators by summarizing incidents, interpreting heterogeneous data sources, assisting emergency-response coordination, and explaining optimization decisions. In these contexts, the Future Internet value of LLMs lies in semantic interoperability: the ability to translate between human requests, machine-readable data, service APIs, and autonomous agents. This is particularly important in IoE-enabled smart cities, where people, processes, data, devices, services, and applications must interoperate across heterogeneous systems. Semantic technologies such as ontologies, linked open data, knowledge graphs, ontology alignment, and automated reasoning provide a foundation for making such integration more machine-interpretable and reusable [48].

Recent work on urban LLM agents extends this role from passive decision support to agentic coordination across the cyber-physical-social space of cities. Urban LLM agents can combine urban sensing, memory management, reasoning, execution, and learning, and can be applied across urban planning, transportation, environmental monitoring, public safety, and citizen-facing social services [49]. Empirical work on smart-city management also shows that multi-agent LLM systems integrated with city document databases and service APIs can improve the routing and quality of responses to heterogeneous urban queries, especially when RAG is used to incorporate local regulations and city-specific knowledge [50]. However, LLM-based urban intelligence should remain connected to verified data pipelines, explicit governance rules, auditability, and human oversight, because planning, safety, mobility, and public-service decisions affect physical infrastructure and citizens directly.

7.4. Digital Twins and Autonomous Agents

Digital twins provide virtual representations of physical systems and can be used to simulate conditions, monitor system state, and support predictive analytics. LLMs can enhance digital twins by enabling natural-language querying, explanation of simulated outcomes, and orchestration of tool-based workflows. In industrial digital-twin environments, this role is closely connected with semantic interoperability because physical assets, information models, simulation models, and service interfaces must exchange information in consistent and machine-interpretable forms. Recent work shows that LLM agents can support the generation of Asset Administration Shell models from unstructured technical datasheets, thereby reducing manual modeling effort and improving standardized information exchange in Industry 4.0 digital twins [51].

Agentic AI can extend this capability by allowing LLM-driven agents to retrieve information, call external tools, coordinate services, and update plans according to environmental feedback. Multi-agent LLM systems have been proposed for digital-twin simulation

parametrization, where specialized observation, reasoning, decision, and summarization agents interact with digital-twin data and control interfaces to search for feasible simulation parameters [52]. Related work on cyber-physical systems also shows that LLM-based planning can translate high-level human prompts into usage plans, but that these plans must be grounded in physical-system dynamics and checked for safety before they are executed [53]. However, such agentic systems must be constrained by explicit permissions, audit logs, safety policies, and fallback mechanisms.

8. Discussion

8.1. Main Objective of the Review

This review addresses the lack of a Future Internet-oriented synthesis of LLMs. Existing LLM surveys mainly emphasize model architecture, benchmark performance, training strategies, or general NLP applications. However, Future Internet and Smart IoT ecosystems require a different analytical focus because LLMs must operate across distributed, heterogeneous, latency-sensitive, privacy-sensitive, and resource-constrained environments. The problem addressed in this paper is therefore how LLMs can be classified, compared, and evaluated as enabling components of Smart IoT, edge intelligence, multimodal sensing, autonomous agents, and distributed digital infrastructures.

8.2. Contributions to the Literature

This study contributes to the literature in four ways. First, it reframes LLMs as cognitive middleware for Future Internet ecosystems rather than as general-purpose conversational systems. Second, it proposes a taxonomy that classifies LLMs according to deployment environment, architecture, training strategy, accessibility, and application scope. Third, it connects LLM capabilities with Future Internet requirements such as semantic interoperability, edge feasibility, privacy preservation, tool use, retrieval support, and autonomous orchestration. Fourth, it identifies deployment-oriented challenges and research directions for integrating LLMs into Smart IoT and edge-cloud infrastructures.

The evidential basis for the four middleware functions is uneven, and this asymmetry is itself a finding of the review. Semantic interoperability and context-aware reasoning are well supported by the reviewed literature on multimodal models and retrieval-augmented generation, and autonomous service orchestration is increasingly supported by work on tool use, agentic mechanisms, and the Model Context Protocol. Adaptive coordination across distributed edge and cloud nodes remains the least supported function: the deployment-aware comparison shows that most representative models are evaluated in centralized settings, and federated or edge-native adaptation is discussed more often as a requirement than demonstrated as a capability. Claims that LLMs already act as full cognitive middleware for Future Internet ecosystems should therefore be read as conditional on progress in this fourth function.

8.3. Practical, Social, and Research Implications

The practical implication is that system designers can use the proposed taxonomy to select LLM architecture according to deployment constraints, including latency, memory footprint, energy consumption, privacy requirements, and compatibility with edge or cloud-edge infrastructures. For Smart IoT developers, the review highlights that LLMs are most useful when combined with retrieval systems, tools, multimodal sensing, and domain-specific safety controls. The social implication is that LLM-enabled Future Internet systems may improve accessibility, automation, and human-centric interaction, but they also increase risks related to privacy, bias, hallucination, surveillance, and unsafe autonomous decision-making. The research implication is that future work

should move beyond model-size comparisons and evaluate LLMs using deployment-aware criteria, including edge feasibility, security, interoperability, energy efficiency, and trustworthy reasoning.

8.4. Key Takeaways

The target readers of this review include researchers in LLMs and Future Internet systems, IoT and edge-AI engineers, smart-city and smart-home system designers, cybersecurity researchers, and decision-makers evaluating AI-enabled digital infrastructures. The key takeaway is that LLMs can support Future Internet ecosystems only when their integration is designed around deployment constraints, multimodal context, privacy, trustworthiness, and interoperability. Large cloud-based models offer strong reasoning and multimodal capability, but lightweight, open, compressed, or hybrid models are more realistic for edge and Smart IoT environments.

9. Challenges, Limitations, and Open Issues in LLMs

Despite the remarkable advancements and widespread adoption of Large Language Models (LLMs), several technical, ethical, computational, and societal challenges remain unsolved. As these models continue to increase in scale and complexity, concerns related to computational cost, energy consumption, hallucination, bias, interpretability, privacy, and security have become increasingly significant. Furthermore, deploying LLMs in distributed environments such as Smart Internet of Things (SIoT), edge computing, and Future Internet architecture introduces additional constraints related to latency, bandwidth, resource limitations, energy availability, and decentralized data management. These challenges are particularly important in resource-constrained environments where real-time decision-making, privacy preservation, and system reliability are essential. This section discusses the major limitations and open research challenges associated with modern LLMs, highlighting the need for more efficient, transparent, secure, and trustworthy AI systems for Future Internet and Smart IoT ecosystems.

a. *Computational and Energy Challenges*

Modern Large Language Models (LLMs) require enormous computational resources for both training and deployment. State-of-the-art models such as GPT-4, PaLM, and Claude are trained using thousands of GPUs or TPUs across distributed cloud infrastructures, resulting in substantial financial and environmental costs [17,19]. Training large-scale transformer-based models consumes significant electrical energy and produces a considerable carbon footprint, raising important concerns regarding sustainability, accessibility, and environmental impact [41,42]. In addition to training complexity, inference operations for real-time applications require high memory capacity, computational throughput, and low-latency processing capabilities. These requirements significantly limit the deployment of LLMs on edge devices and resource-constrained Smart Internet of Things (SIoT) environments. Future Internet applications such as autonomous vehicles, smart healthcare systems, industrial IoT, and intelligent urban infrastructures require efficient real-time processing closer to the data source, making computational optimization a critical challenge. Although techniques such as quantization, pruning, knowledge distillation, sparse architectures, and parameter-efficient fine-tuning have improved deployment efficiency [33,39,40], balancing model performance, computational cost, energy consumption, and deployment scalability remains a major open research problem within Future Internet and distributed intelligent environments.

b. *Hallucination and Reliability Issues*

One of the major limitations of Large Language Models (LLMs) is their tendency to generate hallucinations, where models produce factually incorrect, misleading, or fabricated information while presenting it with high confidence. Hallucinations may arise due to limitations in training data, probabilistic text generation, outdated knowledge, insufficient contextual grounding, or incomplete reasoning processes. This issue raises serious concerns in critical domains such as healthcare, finance, legal systems, cybersecurity, and autonomous Smart IoT environments, where inaccurate outputs may lead to harmful consequences or unsafe decision-making. Although techniques such as Retrieval-Augmented Generation (RAG), external knowledge integration, fact verification, tool-augmented reasoning, and reinforcement learning from human feedback have improved response reliability, ensuring factual consistency, explainability, and trustworthy reasoning remains a major open research challenge [22–26,28,29].

c. *Bias, Fairness, and Ethical Concerns*

Large Language Models are trained on massive internet-scale datasets that often contain societal and demographic biases related to gender, race, religion, culture, and political viewpoints. Consequently, LLMs may reproduce or amplify harmful stereotypes, discriminatory behavior, and offensive content within generated responses [54]. Sheng et al. further demonstrate that seemingly simple prompts can trigger biased, toxic, or socially harmful outputs in language generation systems [55]. Although mitigation strategies such as adversarial data augmentation, bias-controlled fine-tuning, dataset filtering, and reinforcement learning-based alignment have been explored [56], eliminating bias remains difficult due to the scale and diversity of training data. These ethical concerns become particularly critical in sensitive application domains such as hiring, education, healthcare, law enforcement, and autonomous decision-making systems, where biased outputs may directly affect individuals and communities.

d. *Robustness and Adversarial Vulnerabilities*

Large Language Models can be highly sensitive to adversarial prompts, malicious instructions, or small input perturbations, which may lead to incorrect, unsafe, or manipulated outputs. Prompt injection attacks, jailbreak prompting, and adversarial trigger strategies expose important vulnerabilities in real-world LLM deployments. Wallace et al. demonstrate that carefully crafted trigger sequences can manipulate model behavior and significantly degrade output reliability [57]. These vulnerabilities are particularly concerning in Future Internet and Smart IoT environments, where LLMs may support autonomous agents, decision-making systems, intelligent assistants, and cyber-physical infrastructures. Malicious prompt manipulation in such environments could potentially influence automated workflows, bypass safety mechanisms, or generate harmful actions. Therefore, improving robustness, secure alignment, adversarial defense mechanisms, and trustworthy reasoning remain an important research direction for future LLM systems.

e. *Privacy and Security Challenges*

Modern LLMs raise significant privacy and security concerns due to their reliance on massive datasets, external interactions, and distributed infrastructures. Sensitive confidential information may unintentionally appear in generated outputs if models memorize private data during training. Furthermore, prompt injection attacks, data poisoning, model extraction, unauthorized access, and malicious API exploitation represent major security risks in real-world deployments. These concerns become even more critical in distributed Smart Internet of Things (SIoT) and Future Internet ecosystems, where LLMs interact with decentralized devices, sensor networks, cloud-edge infrastructures, and autonomous intelligent systems. Emerging solutions such as federated learning, privacy-preserving

training, secure model alignment, differential privacy, and encrypted inference mechanisms are being explored to improve the trustworthiness, resilience, and security of LLM-based intelligent systems [6].

f. *Edge Deployment and Scalability Challenges*

Deploying Large Language Models (LLMs) in edge computing and Smart Internet of Things (SIoT) environments remains a major challenge due to limited computational resources, memory capacity, bandwidth availability, and energy constraints on edge devices.

Most state-of-the-art LLMs are designed for large-scale cloud infrastructures and require substantial GPU resources for both training and inference. However, Future Internet applications such as autonomous vehicles, smart healthcare systems, industrial IoT, and smart cities require low-latency and real-time decision-making capabilities closer to the data source. These requirements create significant scalability and deployment challenges when integrating LLMs into distributed and resource-constrained environments. To address these limitations, researchers are exploring lightweight architectures, model compression, quantization, pruning, knowledge distillation, sparse computation, and edge-native AI frameworks. In addition, federated learning and distributed inference mechanisms are being investigated to support collaborative intelligence across decentralized IoT networks while preserving privacy and reducing communication overhead. Despite these advancements, achieving scalable, energy-efficient, privacy-preserving, and real-time LLM deployment across heterogeneous edge environments remains an important open research problem within Future Internet ecosystems.

10. Future Directions

Future research on LLMs must increasingly align with the vision of Future Internet and Smart IoT ecosystems, where intelligence is distributed, context-aware, adaptive, and continuously interconnected across cloud, edge, and device layers. Within this paradigm, LLMs are expected to function as cognitive engines for autonomous systems, intelligent digital twins, multimodal analytics, and adaptive service orchestration across interconnected cyber-physical environments.

Several Future Internet and Smart IoT research directions, as outlined in Table 7, are particularly important. Edge-native LLMs should be designed for resource-constrained devices, gateways, and edge servers. RAG should be adapted for IoT knowledge bases, technical manuals, sensor histories, and local operational rules. Multimodal sensing should be integrated with LLM reasoning so that models can interpret data from cameras, microphones, smart meters, environmental sensors, wearables, and industrial devices. Privacy-preserving learning should support local adaptation without unnecessary movement of sensitive data. MCP and agentic AI should be explored as mechanisms for interoperable tool use, service orchestration, and multi-agent collaboration.

Table 7. Future research roadmap for LLM-enabled Future Internet and Smart IoT ecosystems.

Research Direction	Current Gap	Needed Method	Future Internet/SIoT Value
Edge-native LLMs	Most frontier models remain cloud-centered.	Compression, quantization, pruning, sparse activation, and hardware-aware inference.	Lower latency, better privacy, and reduced bandwidth dependency.
RAG for IoT knowledge bases	LLM outputs may be ungrounded or outdated.	Retrieval from manuals, sensor logs, digital-twin states, and local policies.	More reliable and explainable decision support.
Multimodal sensing	Many LLM systems remain text-focused.	Fusion of camera, audio, wearable, environmental, and smart-meter data.	Context-aware reasoning for smart homes, healthcare, industry, and cities.

Table 7. *Cont.*

Research Direction	Current Gap	Needed Method	Future Internet/SIoT Value
Federated and privacy-preserving adaptation	IoT data is sensitive and distributed.	Federated learning, differential privacy, secure aggregation, and encrypted inference.	Local learning without unnecessary data exposure.
MCP and agent interoperability	Tool and agent ecosystems are fragmented.	Standardized context exchange, permissioning, memory, and audit trails.	Interoperable autonomous services across cloud-edge-device layers.
Digital-twin integration	Physical systems lack explainable AI interfaces.	LLM interfaces for simulation, monitoring, explanation, and control-policy support.	Human-centric management of cyber-physical infrastructures.
Trustworthy safety monitoring	Autonomous outputs may be unsafe or biased.	Guardrails, verification, constrained actions, logging, and human-in-the-loop escalation.	Safer deployment in critical and semi-autonomous environments.
Energy-aware inference	LLM deployment may be environmentally costly.	Energy profiling, adaptive model routing, green AI reporting, and workload scheduling.	Sustainable scaling of intelligent Future Internet services.

Developed by the authors.

The emergence of open-weight and efficient models further reinforces the importance of transparency, accessibility, reproducibility, and collaborative innovation in the responsible development of next-generation intelligent systems. Overall, future LLM research must balance intelligence, scalability, sustainability, and trustworthiness to support the evolving requirements of the Future Internet and Smart IoT ecosystems.

11. Conclusions

This paper presented a taxonomy-based review of LLMs for Future Internet and Smart IoT ecosystems. The review classified LLMs according to deployment environment, architecture, training strategy, accessibility, and application scope, and used these dimensions to explain how LLMs can function as cognitive middleware across distributed digital infrastructures. The analysis shows that LLMs are relevant to Future Internet systems not only because of their language-generation ability, but because they can support semantic interoperability, multimodal interpretation, contextual reasoning, tool use, retrieval-based grounding, and autonomous service orchestration.

The study also identified the main limitations that hinder LLM deployment in edge and resource-constrained IoT environments. These include computational cost, energy consumption, inference latency, memory footprint, hallucination, privacy risk, security vulnerability, bias, and limited robustness. These challenges are amplified in Smart IoT environments because intelligent systems must operate close to physical devices, interact with sensitive sensor data, and support reliable real-time or near-real-time decision-making.

Future research should therefore focus on lightweight and energy-efficient LLM architectures, trustworthy and explainable AI systems, federated and privacy-preserving learning, multimodal sensing, RAG-based grounding, agentic AI ecosystems, MCP-enabled interoperability, and secure edge-native deployment strategies. The integration of LLMs with Smart IoT and Future Internet infrastructures represents an important step toward autonomous, adaptive, scalable, and human-centric digital ecosystems, but this integration must be governed by deployment-aware design, transparent evaluation, and strong safety controls.

Author Contributions: S.A.: Conceptualization, Supervision, Writing—review and editing. G.K.: Investigation, Methodology, Formal Analysis, Writing—original draft. T.A.: Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement: Data sharing is not applicable. No new data were created or analyzed in this study.

Acknowledgments: The authors would like to thank the researchers, reviewers, and contributors whose work and feedback assisted in the development of this review paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Raiaan, M.A.K.; Mukta, M.S.H.; Fatema, K.; Fahad, N.M.; Sakib, S.; Mim, M.M.J.; Ahmad, J.; Ali, M.E.; Azam, S. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access* **2024**, *12*, 26839–26874. [CrossRef]
2. Kumar, P. Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artif. Intell. Rev.* **2024**, *57*, 1. [CrossRef]
3. Shahzad, T.; Mazhar, T.; Tariq, M.U.; Ahmad, W.; Ouahada, K.; Hamam, H. A comprehensive review of large language models: Issues and solutions in learning environments. *Discov. Sustain.* **2025**, *6*, 27. [CrossRef]
4. Duan, Q.; Lu, Z. AI Agent Communications in the Future Internet-Paving a Path Toward the Agentic Web. *Future Internet* **2026**, *18*, 171. [CrossRef]
5. Sarhaddi, F.; Nguyen, N.; Zuniga, A.; Khan, M.; Hui, P.; Tarkoma, S.; Flores, H.; Nurmi, P.; Thi Nguyen, N. LLMs and IoT: A Comprehensive Survey on Large Language Models and the Internet of Things. *IEEE Open J. Commun. Soc.* **2026**, *7*, 3585–3616. [CrossRef] [PubMed]
6. Wu, Y.; Li, Z.; Guo, B.; He, S.; Liu, B.; Liu, X.; He, S.; Guo, D. New Paradigm of Distributed Artificial Intelligence for LLM Implementation and Its Key Technologies. *Comput. Sci. Rev.* **2026**, *59*, 100817. [CrossRef]
7. Jelinek, F. *Statistical Methods for Speech Recognition*; MIT Press: Cambridge, MA, USA, 1998. Available online: https://mitpress.mit.edu/9780262546607/statistical-methods-for-speech-recognition/ (accessed on 19 July 2026).
8. Rosenfeld, R. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE* **2000**, *88*, 1270–1278. [CrossRef]
9. Mor, B.; Garhwal, S.; Kumar, A. A systematic review of hidden Markov models and their applications. *Arch. Comput. Methods Eng.* **2021**, *28*, 1429–1448. [CrossRef]
10. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781. [CrossRef]
11. Bharadi, V.A.; Alegavi, S.S. Using Luong and Bahdanau Attention Mechanism on the Long Short-Term Memory Networks: A COVID-19 Impact Prediction Case Study. In *Smart Computing and Self-Adaptive Systems*; CRC Press: Boca Raton, FL, USA, 2021; pp. 1–25. [CrossRef]
12. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany, 8–10 August 2016; pp. 207–212. Available online: https://aclanthology.org/P16-2034/ (accessed on 19 July 2026).
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008. [CrossRef]
14. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
15. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2019; Volume 32. [CrossRef]
16. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricu, R. ALBERT: A Lite BERT for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942. [CrossRef]
17. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. [CrossRef]
18. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
19. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. PaLM: Scaling language modeling with pathways. *J. Mach. Learn. Res.* **2023**, *24*, 1–113.

20. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
21. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555. [CrossRef]
22. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing reasoning and acting in language models. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. [CrossRef]
23. Schick, T.; Dwivedi-Yu, J.; Dessi, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 68539–68551. [CrossRef]
24. Li, Z.; Wang, Z.; Wang, W.; Hung, K.; Xie, H.; Wang, F.L. Retrieval-augmented generation for educational application: A systematic survey. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100417. [CrossRef]
25. Klesel, M.; Wittmann, H.F. Retrieval-Augmented Generation (RAG). *Bus. Inf. Syst. Eng.* **2025**, *67*, 551–561. [CrossRef]
26. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kuttler, H.; Lewis, M.; Yih, W.; Rocktaschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474. [CrossRef]
27. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, L.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **2024**, *25*, 1–53.
28. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744. [CrossRef]
29. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv* **2022**, arXiv:2204.05862. [CrossRef]
30. KKaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D.; et al. Scaling laws for neural language models. *arXiv* **2020**, arXiv:2001.08361. [CrossRef]
31. Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L.A.; Welbl, J.; Clark, A.; et al. Training compute-optimal large language models. *arXiv* **2022**, arXiv:2203.15556. [CrossRef]
32. Besiroglu, T.; Erdil, E.; Barnett, M.; You, J. Chinchilla scaling: A replication attempt. *arXiv* **2024**, arXiv:2404.10102. [CrossRef]
33. Kim, S.; Shen, S.; Thorsley, D.; Gholami, A.; Kwon, W.; Hassoun, J.; Keutzer, K. Learned token pruning for transformers. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 784–794. [CrossRef]
34. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Alvarez, R.; Anderson, K.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
35. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
36. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
37. Workshop, B. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv* **2023**, arXiv:2201.02199. [CrossRef]
38. Luo, R.; Sun, X.; Xia, F.; Qin, B.; Liu, T. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Brief. Bioinform.* **2022**, *23*, bbac409. [CrossRef] [PubMed]
39. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825. [CrossRef]
40. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; et al. Mixtral of Experts. *arXiv* **2024**, arXiv:2401.04088. [CrossRef]
41. Verdecchia, R.; Sallou, J.; Cruz, L. A systematic review of Green AI. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2023**, *13*, e1507. [CrossRef]
42. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for modern deep learning research. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13693–13696. [CrossRef]
43. Thakur, N.; Han, C.Y. Multimodal Approaches for Indoor Localization for Ambient Assisted Living in Smart Homes. *Information* **2021**, *12*, 114. [CrossRef]
44. Ranieri, C.M.; MacLeod, S.; Dragone, M.; Vargas, P.A.; Romero, R.A.F. Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors. *Sensors* **2021**, *21*, 768. [CrossRef] [PubMed]
45. Nguyen, L.N.; Susarla, P.; Mukherjee, A.; Cañellas, M.L.; Casado, C.Á.; Wu, X.; Silvén, O.; Jayagopi, D.B.; López, M.B. Non-contact multimodal indoor human monitoring systems: A survey. *Inf. Fusion* **2024**, *110*, 102457. [CrossRef]
46. Palma, G.; Cecchi, G.; Rizzo, A. Large Language Models for Predictive Maintenance in the Leather Tanning Industry: Multimodal Anomaly Detection in Compressors. *Electronics* **2025**, *14*, 2061. [CrossRef]

47. Narimani, A.; Klarmann, S. Integration of Large Language Models for Real-Time Troubleshooting in Industrial Environments based on Retrieval-Augmented Generation (RAG). In *Proceedings of the 7th European Conference on Industrial Engineering and Operations Management, Augsburg, Germany, 16–18 July 2024*; IEOM Society International: Southfield, MI, USA, 2024; pp. 287–298. [CrossRef]
48. Pliatsios, A.; Kotis, K.; Goumopoulos, C. A systematic review on semantic interoperability in the IoE-enabled smart cities. *Internet Things* **2023**, *22*, 100754. [CrossRef]
49. Han, J.; Ning, Y.; Yuan, Z.; Ni, H.; Liu, F.; Lyu, T.; Liu, H. Large Language Model Powered Intelligent Urban Agents: Concepts, Capabilities, and Applications. *arXiv* **2025**, arXiv:2507.00914. [CrossRef]
50. Kalyuzhnaya, A.; Mityagin, S.; Lutsenko, E.; Getmanov, A.; Aksenkin, Y.; Fatkhiev, K.; Fedorin, K.; Nikitin, N.O.; Chichkova, N.; Vorona, V.; et al. LLM Agents for Smart City Management: Enhancing Decision Support Through Multi-Agent AI Systems. *Smart Cities* **2025**, *8*, 19. [CrossRef]
51. Xia, Y.; Xiao, Z.; Jazdi, N.; Weyrich, M. Generation of Asset Administration Shell with Large Language Model Agents: Toward Semantic Interoperability in Digital Twins in the Context of Industry 4.0. *IEEE Access* **2024**, *12*, 84863–84877. [CrossRef]
52. Xia, Y.; Dittler, D.; Jazdi, N.; Chen, H.; Weyrich, M. LLM experiments with simulation: Large Language Model Multi-Agent System for Simulation Model Parametrization in Digital Twins. *arXiv* **2024**, arXiv:2405.18092. [CrossRef]
53. Banerjee, A.; Maity, A.; Kamboj, P.; Gupta, S.K.S. CPS-LLM: Large Language Model based Safe Usage Plan Generator for Human-in-the-Loop Human-in-the-Plant Cyber-Physical System. *arXiv* **2024**, arXiv:2405.11458. [CrossRef]
54. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: Barcelona, Spain, 2016; Volume 29. [CrossRef]
55. Sheng, E.; Chang, K.W.; Natarajan, P.; Peng, N. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3407–3412. [CrossRef]
56. Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; Chang, K.W. Learning gender-neutral word embeddings. *arXiv* **2018**, arXiv:1809.01496. [CrossRef]
57. Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; Singh, S. Universal adversarial triggers for attacking and analyzing NLP. *arXiv* **2019**, arXiv:1908.07125. [CrossRef]