Advances in Artificial Intelligence and Machine Learning; Review 5 (4) 4692-4709 Received 29-07-2025; Accepted 08-12-2025; Published online 15-12-2025

Generative AI in Automotive Industry

Arunachalam Thirunavukkarasu arunachalam.thirunavukkarasu@dlr.de
Deutsches Zentrum für Luft-und Raumfahrt e.V. (DLR)
Germany

Domenik Helms domenik.helms@dlr.de
Deutsches Zentrum für Luft-und Raumfahrt e.V. (DLR)
Germany

Christian Ojeda Bernal domenik.helms@dlr.de
Valeo Schalter und Sensoren GmbH
Germany

Sven Mantowsky sven.mantowsky@zf.com
ZF Friedrichshafen AG
Germany

Saqib Bukhari saqib.bukhari@zf.com
ZF Friedrichshafen AG
Germany

Róbert Lajos Bücs robert.buecs@aptiv.com
Aptiv Services Deutschland GmbH
Germany

Corresponding Author: Arunachalam Thirunavukkarasu

Copyright © 2025 Arunachalam Thirunavukkarasu, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The growing popularity of generative artificial intelligence has led to significant changes in various industries, including automotive sector. The industry is undergoing a transformation from traditional design and development of systems to modern AI-based architectures. This development, reminiscent of the transition from simple mobile phones to smartphones, positions generative AI as one of the most important drivers of these changes. This article provides a comprehensive overview of the scalability issues and integration opportunities of generative AI in the automotive industry. We examine the current state and challenges of implementing generative AI in limited hardware environments, such as those typical of automotive systems. In addition, we discuss the evolving E/E architectures in the industry that are paving the way for this technological integration. The article also highlights early examples of generative AI adoption in the automotive sector and provides insight into the industry's ongoing transformation toward more advanced AI-based features.

Keywords: Generative AI, Scalibility challenges, E/E architectures.

Citation: Arunachalam Thirunavukkarasu, et al. Generative AI in Automotive Industry. **Advances in Artificial Intelligence and Machine Learning**. 2025;5(4):260. https://dx.doi.org/10.54364/AAIML.2025.54260

1. INTRODUCTION

The introduction of generative artificial intelligence (AI) signals a new era of innovation for the automotive industry. From disruptive design processes to the development of fully autonomous driving systems, generative AI brings new conceptual opportunities to enhance the performance, safety and quality of customer service of vehicles. As automotive original equipment manufacturer (OEM) continue to incorporate AI to maintain a competitive advantage, it is critical to deploy this technology throughout the production line and fleet.

Generative AI is transforming the development process across all industry sectors, including but not limited to automotive industries. Generative AI is expected to redefine the entire development framework in the production of vehicles for every OEM, supporting and informing all development from concept through to the implementation of built-in features to support the final user throughout the lifecycle of the vehicle.

Generative AI poses potential advantages to the automotive sector but is currently limited by the computational capabilities of the hardware that supports the software aspect of vehicle architectures. In this discussion, we will review the scalability of generative AI in the automotive sectors in two places: in-vehicle functions and functions used for industrial application.

At present, in-vehicle functions consider the hardware limitations of vehicle architecture, even in the most luxurious and basic segment of vehicles. These limitations tend to affect three areas: memory, energy, and computational power. AI functions require a robust hardware performance focus on these three points. Therefore, the rest of the article will primarily focus on the current problems associated with scaling generative AI in in-vehicle functions.

Generative AI has started to change some of the practices the automotive industry follows related to design, manufacturing, maintenance, and even customer-facing practices that contribute to personalised customer experiences.

This paper is organized in the following way: In Section 2, we provide case studies of the early use of generative AI in the automotive pipeline. These examples illustrate the concrete benefits and challenges faced, and demonstrate some of the key lesson learnt that may inform future use cases. In Section 3, we examine the main challenges of scaling generative AI in an automotive context, and discuss hardware constraints, safety concerns, and the complexity of integration that arises from working with generative AI. In Section 4, we discuss the architectural changes needed to enable scalable AI integration in the automotive pipeline, considering hardware and software designs that enable AI driven systems to operate. In section 5, we look forward to future directions and trends related to generative AI in automotive applications, identify areas of emerging research, and consider advancements in technology that could further support generative AI. Finally, Section 6 concludes with a summary of findings and suggestions for the industry.

The current paper is mainly conceptual and review-based. Instead of exploring new experimental results, the report summarizes existing research and industry reports on generative AI applications in the automotive sector. The intention is to draw together evidence from the literature and existing sources in order to understand commonly occurring performance trends, challenges, and opportunities for development. By summarizing empirical findings as reported in existing studies, this

paper sets a framework for quantitative studies and future studies that focus on implementation in automotive AI systems.

2. CURRENT USE OF GENERATIVE AI APPLICATIONS IN AUTOMOTIVE INDUSTRY

Generative AI is bringing about transformative changes in the automotive industry, from design and simulation to in-situ data generation for autonomous systems. Much prior work on generative models has focused on creative applications (image, text), while more recent work has begun exploring its use in mobility engineering. For example, the paper [1] reviews generative models including GANs, diffusion models and generative transformers in applications such as creating static maps, generating dynamic scenarios, and predicting trajectories. In a similar way, paper [2]presents a taxonomy of world models to integrate multi-sensor data, temporal dynamics, and behavior planning. Although multiple projects are showing positive results, few projects have shown large scale deployment in production systems; most projects stay in experimental or simulation settings. This section reviews current applications; later sections analyze the challenges of scaling toward production.

The current use of generative AI in automotive applications can be categorized into several key areas:

Synthetic Data Generation: Generative AI plays a crucial role in generating synthetic datasets where specific conditions are scarce or challenging to reproduce. For instance, the WEDGE dataset [3], which consists of 3360 images in 16 extreme weather conditions, was constructed using generative vision-language models, provides diverse weather scenarios essential for training autonomous driving systems. The possibility of using realistic synthetic data is of great importance for improving autonomous driving capabilities in extreme scenarios where little or no real data exists, such as uncommon situations encountered during a trip or extreme weather conditions. Recent works such as SynDiff-AD [4] and Exploring the Effects of Synthetic Data Generation [5] further support this by showing measurable performance gains in semantic segmentation and end-to-end driving tasks through diffusion-based data augmentation. Latent diffusion models are used to generate synthetic images for rare weather or lighting scenarios, improving performance of semantic segmentation and end-to-end driving models in both CARLA and real dataset benchmarks [4]. The paper [5] deals with a case study on autonomous driving for semantic segmentation, examines how synthetic per-pixel labelled virtual scenes affect training, and identifies key factors such as texture realism, domain gap, and diversity of scene composition.

Super-Resolution of Sensor Data: Another recently explored use of generative AI in the automotive industry is presented in [6], where the authors employed it to improve the resolution of images captured by vehicle sensors. This type of application can increase the safety of other applications such as recognition systems and object detection. Complementing this, [7] introduced a deep unrolling super-resolution network tailored for LiDAR automotive scenes—demonstrating improved resolution and fidelity while keeping computational load manageable for embedded vehicle hardware.

World Model Simulation for Autonomous Driving: Generative AI has shown great capabilities as simulators for autonomous driving through the so-called world models, for example, in the project GAIA-1 [8] as it creates realistic driving environments. These simulators can be used for training and validating autonomous driving algorithms facilitating a safe development testing. Recent advances such as GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving extend this approach with multi-camera and controllable scene synthesis [9]. It presents a latent diffusion world model that can generate spatiotemporally consistent, multi-camera video conditioned on environment, agent configurations, and vehicle dynamics. Additionally, [2] offers a wider perspective on the development of world models along with approaches for predicting future scenes, using multiple modalities (image, LiDAR, radar), and behavioral planning. Furthermore, currently there are numerous commercial efforts demonstrating how generative simulation is coming from laboratory research to industry.

Vehicle Design Assistance: Generative AI is also reshaping the vehicle design process. The Toyota Research Institute has developed a generative AI tool that is capable of accelerating and improving the phases of vehicle design [10]. This tool can identify new designs to suggest, or optimize existing designs, vastly reducing the time and costs of vehicle development. Likewise, Audi has explored generative models for wheel rim design to invent photorealistic concepts or to recombine concepts from existing designs, improving form and functional performance [11].

In summary, current literature indicates that while early work has demonstrated the feasibility of generative AI in vehicle-related and automotive end tasks, most solutions continue to be isolated research projects or proofs-of-concept. Few works address full production pipelines, and even fewer address production at scale under automotive safety, hardware and regulatory constraints. This paper aims to address that gap by synthesizing the perspectives of academia and industry and focusing on what is needed to transition from experimental study settings to robust scalable generative AI systems for automotive applications.

2.1 Comparitive Emprical Evidence From Existing Studies

Despite the fact that this is primarily a review paper, this paper captures the essential empirical findings in recent literature to enhance the analytical view point of how generative AI is already providing measurable benefits for automotive use cases. TABLE 1 provides a summary of select quantitative outcomes across a variety of fields related to generative AI, including synthetic data, improvements in perception, simulation, and vehicle design. Most of the obtainable comparison data indicates that generative models are starting to demonstrate measurable benefits, relative to accuracy and efficiency, even if most work is still in research or simulation only applications.

Table 1: Comparative Empirical Evidence from Recent Generative AI Studies in Automotive Applications

Application Area	**Quantitative Results / Empirical Evidence**	**Identified Limitation**
Synthetic Data Generation [4]	+1.2% (Mask2Former), +2.3% (SegFormer) on Waymo; up to 20% driving performance boost in CARLA.	Domain gap between synthetic and real data still limits generalization.
Synthetic Scene Diversity [5]	SegFormer achieved 60.65 mIoU (+8 over DeepLabv3+); geometry simplification < 2 mIoU impact; realistic lighting and tone mapping yield +1–2 mIoU gains on Cityscapes and BDD.	Domain gap to real data (15–20 mIoU), high rendering cost (1470 GPU h), and dataset-specific sensitivity limit generalization.
Synthetic High-Quality 3D Point Clouds [8]	Mean squared error (MSE) between predicted and true point clouds: 0.36 mm; IoU for object detection: 85.4%; AP for detection: 92.3%.	Generated point clouds and semantic masks may not perfectly match true values due to LiDAR noise and 3D reconstruction uncertainty.
Generative World Modelling [9]	Outperforms state-of-the-art with PSNR at 16.4 dB and IoU for object detection at 86.5% vs. 79.2% for the next best method.	Requires large volumes of training data and complex multimodal supervision.
AI-Assisted Vehicle Design [10]	Drag-guided generation produces vehicle designs with optimized aerodynamic coefficients, balancing aesthetics and performance.	Relies on surrogate drag models that may introduce inaccuracies if insufficiently trained or lacking physical constraints.

Together, these studies provide a quantitative snapshot of how generative AI contributes measurable value in perception, simulation, and design tasks. Yet in doing so, they underscore the gap between research-grade success and actual industrial–or commercial–deployment; the interplay of these two very different domains is discussed further in the following section on scalability and integration.

3. CHALLENGES OF SCALING GENERATIVE AI IN THE AUTOMOTIVE INDUSTRY

The automotive space presents several challenges surrounding the scalability of generative AI. First, challenges relating to generative AI, including data management and high demanding computational resources, especially related to in-vehicle functions, as there is limited hardware within vehicle architectures. Second, challenges that are specific to the automotive industry represent further additive challenges, including strict safety standards and the seamless integration of AI with traditional engineering approaches. These challenges must be addressed upfront in order to realize scalable implementation of generative AI in the automotive space. This paper focuses on inference as a larger area of opportunity or challenge. FIGURE1 presents an overall framework to categorize these challenges along with their interdependencies through the AI lifecycle — from training, inference, and through to the deployment stage related to vehicle architectures.

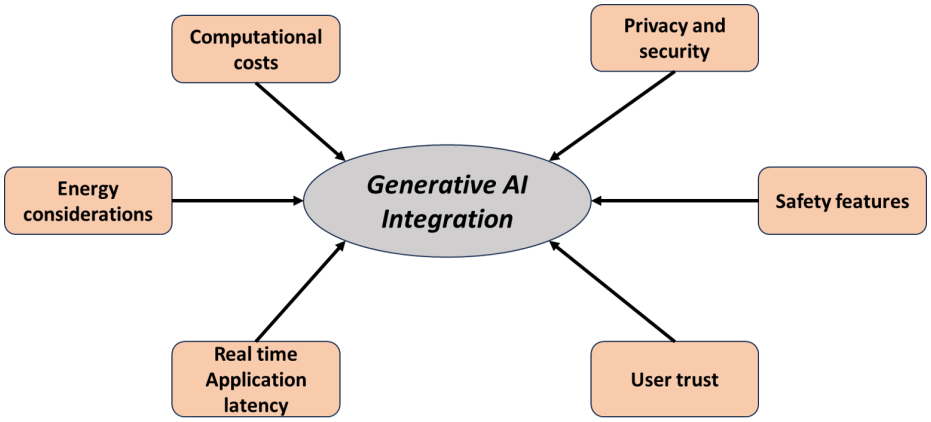

Figure 1: Key challenges in scaling generative AI integration within the automotive industry.

3.1 Computational Resource Demands and Inference Costs

Deploying generative AI models can be extremely expensive due to the need of significant computing resources [12]. Existing deployments typically rely on traditional cloud computing resources that can introduce latency and security concerns due to transmitting data across the network [13]. According to McKinsey [14], the demand for computing hardware used for AI inference in a data center is twice as large as that of AI training , and up to four times larger in the edge market, and continues to grow. Furthermore, [15] says that NVIDIA believes that inference can make up as much as 80-90% of the total cost of a neural network.

3.2 Data Management, Privacy and Security Concerns

Making generative AI systems scalable involves collecting and processing large amounts of data, which raises privacy and security issues, as well as questions of ownership of the rights to the data [16]. Data protection processes must be put in place to provide protection from unauthorized access and breaches. Privacy considerations become especially important with any user-generated data, like speech recognition input or perception data from sensors [17].

3.3 User Acceptance and Trust

For scaling it is necessary to acquired large user acceptance and confidence in the AI permissioned functions [18, 19]. For generative AI to be usable for potential users we need transparent and reliable models with simple interfaces in place making it available to the regular customer . Another aspect that we need to pay attention in the adoption of models and influence acceptance is model interpretability, which allows user to understand how a decision was made , and confirm the system behaves safely and consistently [20].

3.4 Latency and Real-Time Performance

The automotive sector is a domain that has inherently required its algorithms to operate in real-time settings [12]. Hence , an important aspect to consider is the inference latency associated with using those models . In generative models , depending on the architecture 's complexity , the inference latency could vary widely when producing the necessary output once the data is sent . If those inferences exceed the time limits imposed by the industry, the safety of the user may be in jeopardy . Therefore, it is essential for applications such as advanced driver assistance systems (ADAS) and Autonomous driving to maintain low latency in performance [20]. This is even more important in scenarios in which the functions being completed by the AI models correspond to safety-critical applications. Because of this , those models need to be able to process large input data , in real - time patterns for safe operation of the vehicle; otherwise, the time lost processing input signals may result in dangerous scenarios.

This is inherently connected to one of the greatest troubles faced in the move to vehicles containing generative AI for their variety of functions: computing power in their architectures. Unlike data centers with numerous AI processing units (GPUs , TPUs , etc .) where generative AI is typically deployed, vehicle systems with much more limited processing units. Furthermore , there is little diversity in computing devices commonly designed for very specific functions in the vehicle , resulting in an impediment to deploying generative AI in a general sense . Ideally, the models used should be as universally compatible with most of the hardware used in the industry.

If we desire to increase the use of generative AI solutions in the automotive sector, the integration of generative AI into vehicle systems and the infrastructure is required [17]. This is an immense challenge, as integration will require overcoming issues of compatibility with existing and reliable technologies and specifying how generative AI will interact with these existing technologies while maintaining the performance of the systems.

3.5 Safety Validation

The possible application of generative AI for ADAS functions presents a challenge for the automotive industry , requiring significant validation in order to ensure the safety and effectiveness of the system [17]. Validation considers hundreds of thousands of hours of real-world driving data to verify all possible scenarios. MLOps could help with validation and the requirements of validation by implementing automatic validation and scalable monitoring pipelines , which would allow testing and deployment to be more robust and efficient.

3.6 Energy Efficiency

The issue of energy use efficiency, a challenge affecting most industries, has become an important topic of conversation for the automotive industry in recent years as a consequence of the transition to electric vehicles and an ongoing rising number of sensors employed in electric vehicles [21].

As stated above, generative AI (Artificial Intelligence) models are recognized to consume high computational power, which translates to high energy consumption today when these work in practice.

Therefore, efficient energy use is critical to preserve the performance of the vehicle , and in the case of electric vehicles, to extend their operational life-span. To scale the use of generative AI in the automotive industry, technical and operational challenges must be addressed . Among them are, limitations of the in-vehicle hardware, data privacy and security issues, user acceptance, real time performance needs, need for extensive validation, and high energy consumption.

Successful , scaled implementation of generative AI will be dependent on resilient integration within the existing in - vehicle systems and infrastructure which maintains a high use of performance and safety. If the sector can respond to these challenges , we can then unlock the significant transformational potential of AI to supplement vehicle functionalities and ultimately the user experiences.

3.7 Summary of Key Challenges and Mitigation Strategies

To clarify the avove mentioned relationships, TABLE 2 summarizes the main challenge categories, their corresponding impacts, and possible mitigation approaches.

Table 2: Summary of Challenges in Scaling Generative AI in the Automotive Industry

Challenge Category	**Primary Impact**	**Underlying Causes**	**Potential Mitigation Strategies**
Computational Demand	High cost of deploying Generative AI models	Large demand for compute hardware, growing inference costs	Model compression, quantization, edge–cloud hybrid inference
Data Privacy & Security	Risk of data leakage or IP violation	Complex sensor data, cross-border data sharing	Robust data protection, encryption, anonymization
User Trust & Acceptance	Low adoption of AI-driven functionalities	Limited transparency, complex interfaces	Explainable models, user-friendly design, transparent testing
Latency & Real-Time Constraints	Safety risks due to processing delays	Network bottlenecks, limited compute power	Hardware acceleration, optimized architectures, on-device inference
Safety Validation	Slow certification and testing cycles	Insufficient test coverage, lack of automation	Synthetic validation datasets, automated MLOps testing
Energy Efficiency	High energy demand, reduced EV lifespan	Power-hungry AI models	Energy-aware design, dynamic scheduling, efficient cooling
Model Robustness & Explainability	Unsafe behavior in rare driving scenarios	Biased data, opaque models	Domain adaptation, uncertainty estimation, interpretable design

Continued on next page

Table 2 (continued)

Challenge Category	Primary Impact	Underlying Causes	Potential Mitigation Strategies
Regulatory & Ethical Compliance	Deployment delays, unclear accountability	Absence of safety certification pathways	Standardized documentation, model auditing, traceable decision logs

3.8 Discussion

As shown in TABLE 2, the majority of the challenges are not isolated; rather they are interconnected. For example, model compression to improve latency also enables energy efficiency, but may compromise interpretability. Improved validation and monitoring pipelines can provide safety assurance to improve user trust. Thus, any sustainable scaling of generative AI technology in automotive systems will require a holistic design approach that addresses all of the aspects, including efficient architectures, secure data management, and user interface transparency.

4. ADAPTING AUTOMOTIVE ARCHITECTURES FOR SCALABLE AI INTEGRATION

To further broaden the utilization of generative AI in the automotive sector, explicitly in the role of driving assistant, perhaps the most imperative element will be its physical direct integration of generative AI into the vehicle architecture itself. Recently, work has centered on creating vehicle architectures that incorporate more generic AI models for self-driving and various specific driving functions like detection of vehicles, pedestrians, and lane segmentation. Even if these are not "generative" models, the core theories, design factors, and implications of using a model like that are all highly relevant. Some principles for factoring a generative model into a vehicle architecture include, but are not limited, centralization of the architectural design, multi-core accelerators, update/change the systemically, overall architecture pertaining to security and virtualize procedures, communication relationship between data and system, and software architecture. With the usage of these principles, it is apparent that generative models can provide a wider spectrum of capabilities and allow the vehicle to practice smarter contextualized choices that are also personalized.

4.1 From Distributed to Centralized Architectures

Today , E/E (electrical/electronic) architectures in vehicles include more than 100 Electronic Control Units (ECUs) [22–24]. In this type of architecture , dubbed a decentralized architecture, the software functions execute independently of each other on a separate ECU. This architecture offers a number of advantages to manufacturers, such as modularity and redundancy, yet has limitations that hinder scalability. For example , the traditional flexibility - reducing one-to-one mapping of functions to ECUs limits the flexibility that is needed to meet the demands of the modern vehicle [25]. Furthermore , decentralized architectures are not able to meet the increasing demands of

future vehicles [26]. Currently , domain-based E/E architectures are becoming more common , in which robust domain controller units replace multiple traditional ECUs, enabling more elaborate functionalities and centralized - based zonal architectures through a TCUs (central compute unit) . McKinsey reported that by 2032 , it is anticipated that 30 percent of all vehicles globally will employ zonal controllers , and 40 percent of vehicles will have domain based architectures with a centralized high performance computing unit . In contrast, 29 percent of vehicles will still use a distributed architecture with the traditional ECUs [27]. A key driver in these architectural changes is the widespread integration of AI applications such as generative AI and autonomous driving systems, which require computational and centralized power levels [22]. Hardware constraints are an important limitation for deploying sophisticated AI models for use in vehicular contexts , particularly due to the amount of computational resources that generative models with billions of parameters demand . There has been an explosion of AI accelerators by manufacturers in recent years to facilitate AI functionality in automobiles . A review of AI accelerators released by manufacturers in recent years [28, 29] reflects that these devices can be categorized based on peak power (watts) and peak performance (GOps/sec), yielding five different device clusters . These classifications can vary from low-power chips for sensor processing to full- blown data center systems. AI accelerators are typically positioned in the mid-range of power and performance for autonomous vehicles : most AI accelerators (for intelligent decision making during driving) utilize Int8 precision for inference , but some accelerators utilize FP16 for training (where more precision is needed) with power ranging from 10 - 100 watts The suggested designs in , for example [30] and [31], show varying integration of multicore System-on-Chips (SoCs) into vehicle systems. The first option uses a domain-based E/E architecture that uses dedicated domain controllers with AI accelerators that provide domain sub-ECUs acting to manage domain functions. The second option brings dedicated cores from the different vendors to a centralized ECU for each function offered by that vendor . This is called the integrated architecture.

4.2 System Updates

Typically, a new vehicle's life span is more than ten years [25]. Over that life span, technologies have changed rapidly, especially AI technology. It is reasonable to expect technology to continue to evolve regarding AI application in the vehicle capacity. The new software updates will also scale to a generalized generative AI application that supports one or more of the use cases of the vehicle capacity. In fact, the updates will need to function with little to no downtime and very limited interruption. Only the current E/E architecture available today trend to the vehicles cannot support dynamic online software updates while the vehicle software functions [24]. To address this limitation, [24] discusses a container-based software update mechanism. The author claims to provide containers as a mechanism to deliver the scale and isolation to support AI feature development in the new vehicle architecture. Containers provide the possibility of dynamic update and resource management that prolong the life period for advanced AI development capabilities, if integrated into future vehicles.

4.3 Virtualization

The technology of hardware virtualization is fundamentally important to combine safety-critical domains with non-safety-critical domains in a single computing unit while optimizing the hardware resources [28]. As the automotive industry converges on centralized E/E architectures, not only the hardware and software will need to be restructured, but also the processes for the development of future automotive systems will require enormous implications. In distributed architectures, functional partitioning is easier because of the physical separation of functionalities across numerous ECUs, which will make it much more challenging to keep partitioned functional features and integrate new functional features in the centralized architecture [25]. The authors of [32], point out that utilizing modern software development paradigms like containers to encapsulate a reusable service has a lot of benefits including a more flexible approach to functional partitioning that can allow containers to move between ECUs, and the speed at which software updates can be delivered to the field and vehicle [33].

4.4 Security

Software is becoming the primary enabler for new functionalities as vehicles shift from hardware-oriented architectures to software-defined architectures [22]. With more and more safety features and connected functionalities, they also become more vulnerable to security threats [34]. Existing reference architectures are unable to address security issues in modern system designs, e.g., AI-enabled functionalities (i.e., autonomous driving), do not provide the necessary precision or details in defining security requirements [34]. Further, automotive software design has a recognized gap where safety perspectives are not adequately integrated, and new architectural solutions are required to satisfy safety and security requirements [23].Existing systems architecture proposals have focused on fail-safe system designs, but current architectural designs must be fail-operational [25]; i.e., systems must provide not only safety in the instance of a failure, but additionally, must provide a high level of functionality, so that the system continues to function during failures. Additionally, with software being regularly updated, it is vital that safety and security risks can be managed, while also delivering additional new functionalities after point of sale [24]. The architecture mentioned in [30] includes redundancy of safety-critical tasks by using duplicate tasks on different ECUs. This redundancy is meant to ensure at least one ECU continues to perform safety-critical services if another ECU fails for some reason. Alternatively, [31] suggests a typical approach for achieving safety in centralized ICT architecture is to extend it to include parallel processing units and to combine the components in a redundant manner. Future safety standards will need to evolve with new technologies that are developing quickly, such as AI model technologies that don't fit into any of the functional safety criteria defined at this time, such as ISO 26262. AI models as black-box algorithms represent a radical departure from the traditional assumption of predictability of system behaviors, and we need to find new ways to govern safety regulation with regard to these technologies in the automotive space [25].

4.5 Data Transmission

The shift in vehicle architecture is associated with the growing demand for data transmission powered by AI models and the growing input throughput due to the increasing number of sensors needed to enable these models. Hence, this has contributed to an increase in demand for data bandwidth [29]. For many years, vehicle communication has been implemented and based upon the controller area network (CAN) protocol [31]. The CAN protocol has been the norm over the last several decades and gradually is becoming inadequate for today's needs based on data rates alone. The complexity and data requirements are starting to push the industry to newer standards of vehicle communication, such as automotive Ethernet, which is quickly establishing itself as the new standard of vehicle communication [22]. Current data transmission standards and the data transmission performance they offer are adequate for non-future vehicle communication needs, but those data transmission standards will soon not be sufficient. The volume of data overall and the number of interconnections between the ECUs will quickly exceed the data transmission performance capabilities in advanced vehicle architecture [24]. As vehicles continue to evolve, there will be a necessity to use communication protocols that offer higher bandwidth and scalable communication for AI-based functionality and complex sensor networks found in modern vehicles.

4.6 Software Architecture

The transition to advanced automotive technology like AI-based systems and fully autonomous systems mean that our software architecture approaches for vehicles requires rethinking. While software architectures have served us well and acceptable for legacy systems and will continue to do so, they are insufficient when considering the increasing complexity, scalability, and flexibility for vehicles. We will discuss about two architectural paradigms (layered architecture, and service-oriented architecture with orchestration) that will lead the way for the next generation of software-defined-vehicles and how the architectural paradigms are well-suited for intelligent and autonomous systems.

4.6.1 Layered approach

In [34], the authors note that it is almost always true that a set of intelligent features over an existing platform only brings intelligence, as an additional layer, to process the data, plan for the appropriate response, and make decisions. The downfalls of traditional software architectures cannot live up to the expectations of robust requirements for the AI features alone. In response to this observation, they present an architecture for software that has been built from the ground up based upon software requirements in a technical standard from the SAE J3016. This architecture is a recommended best practice for the development of autonomous vehicles and makes use of literature in robotics. A different example is [31] where-in a layered architecture for autonomous driving is presented. Layered designs have advantages for purposes of software development as the design allows for any layer to be replaced or modified without breaking the logic of the system. The layers can remain flexible and update one portion of the system without breaking any of the logic in the other layers. In their architecture, the authors create 3 layers: an OS layer, data communication layer, and an algorithm layer.

4.6.2 Service-oriented architecture (SOA)

As indicated by [32], current signal-based architectures are reaching their maximum capability to handle the complexity of internal communication of modern vehicles and hence they recommend evolving to service-oriented architectures (SOA) where the services are encapsulated in containers, which are processed in a data-centric publish-subscribe approach. The cited approach is based on modern software development practices and addresses the requirement for fast software deployment. The work in [22] underscored the importance of acknowledging the possibility of adding new software feature during the automobile vehicle's life-cycle to the similarities with their expectation of today architectures. To address the issue, the authors claim that converting to a SOA should be challenging but essential for converting a vehicle into a software defined platform. At the same time, this study identifies potential security risks associated with transitioning to SOA. One possibility proposed by the authors would be hybrid architecture where traditional signal-based communication coexists with service-oriented communications.

4.6.3 Software-defined vehicles (SDVs)

A discussion regarding the important shift in vehicle software architecture occurred recently across manufacturers. In this shift, the architecture allows the vehicle to be programmed to control vehicle functions via software - as opposed to hardware - also referred to as the software-defined vehicle (SDV). Like modern E/E (Electrical/Electronic) architectures the SDV utilizes a centralized computing unit model where software is built with capabilities of services, that can be updated or overwritten without hardware modifications. For the SDV to become the standard in automotive applications new frameworks for the development of residual software is required (i.e., the AUTOSAR Adaptive Platform). The AUTOSAR Adaptive Platform provides key components of SDV concept including - dynamic/scalable software architecture, Over-the-Air (OTA) updates, and the ability to integrate high-powered systems connected via high-bandwidth communications protocol. Although designed with more modern features, the AUTOSAR Adaptive Platform still does not support all automotive developer requirements - with safety and timing rates among strengths of the original AUTOSAR platform. For those reasons, a hybrid solution that utilizes both platforms is recommended within a service oriented architecture [22]. The hybrid solution provides future automotive system flexibility and performance while maintaining the safety level.

5. FUTURE DIRECTIONS AND TRENDS

Generative AI's chief potential for scale in the automotive industry is likely to evolve alongside a number of emerging trends and technologies. These anticipated trends address existing constraints in generative AI including cost, computational efficiency, and design optimization and increases the possibilities for AI deployment with vehicles.

Specialized Hardware: Presently, the use of deep learning accelerators presents a cost-benefit ratio. The hardware used most often is composed of GPUs and TPUs, which improve execution times for both the training and inference phases of running AI deep learning models. In most situations, this hardware type almost always comes with memory constraints and an overall higher

cost to process a problem than a more general-purpose (to say CPUs) processing unit. For the future, the focus must shift to develop an accelerator that could work at an optimal cost to energy ratio for the autonomous vehicle context. Neuromorphic chips and domain-specific architectures show potential as pathways to minimising latency and energy consumption, which are better suited to autonomous vehicle real-time generative AI inference in an embedded environment.

Efficient Training and Inference Techniques: Future research focuses on developing more efficient training and inference processes of AI models. One potential approach for future research is the recent work in adaptive inference which provides a mechanism for models to adjust their compute resources during inference in accordance with the application and input features of the task at-hand. The use of model compression techniques and mixed-precision compute complement this process as they decrease compute overhead directly while still maintaining model performance . The applied benefit of these approaches is enabling AI technology to run on constrained in-vehicle hardware in an industry application. And in addition to the applied benefits of these techniques, there are also academic benefits to studying the theoretical performance vs. energy consumption vs. inference latency trade-offs.

Integration of AI with Other Technologies: Generative artificial intelligence will increasingly operate in conjunction with related technologies such as 5G, blockchain, and hybrid edge-cloud computing. 5G could enable low-latency, high-bandwidth communication to facilitate cooperative driving, while blockchain could ensure secure and auditable data management. Further research is needed to develop distributed inference frameworks that balance latency, privacy, and bandwidth in complex vehicle networks.

Sustainable and Green AI: Currently, high energy consumption is one of the key characteristics of AI not only requiring high energy production but also raising responsible sourcing concerns. On the science side, efforts have prompted additional research on sustainable solutions. The research into such things as green training protocols, energy-efficient architectures, and renewable energy integration should be effective in terms of the reduced impact on the environment, while seeking sustainable options for an economy.

Enhanced Data Privacy and Security: Future work should pay attention to privacy-preserving, scalable methods for multimodal automotive data, and be cognizant of the evolving regulations surrounding the EU AI Act helping us to improve data privacy and security. Techniques like federated learning, where models are trained across decentralized devices without sharing raw data, can help protect sensitive information [13].

Human-Centered and Trustworthy AI: User acceptance will remain critical for the widespread adoption of generative AI. Transparent and explainable AI systems can foster trust and improve user experience. Future studies should explore how explainable AI, ethical design, and human-in-the-loop systems can enhance user confidence in AI-assisted automotive technologies.

6. CONCLUSION

This article gave an overall idea of the potential applications of AI in the automotive industry. Various challenges realted to scaling of the AI onto embedded systems were discussed in detail followed by the emphasis of adapting different automotive architectures. We reviewed its current applications, identified the main obstacles to scaling-such as computational limits, safety validation, data privacy-and outlined possible directions for improvement.

Although, this paper does not include new experimental validation, it consolidates quantitative evidence and comparative findings from existing research and industrial studies. The aim was to synthesize how prior empirical efforts have demonstrated measurable improvements in perception accuracy, design efficiency, and simulation realism, while also highlighting the remaining barriers to practical implementation. We believe, this literature-based synthesis provides a structured foundation for future experimental and benchmarking work in the field.

From an industrial standpoint, the automotive sector is still at the early stages of generative applications, largely underpinned by developments in electrical and electronic (E/E) architectures that can accommodate the high computational demands of generative models. Largely applications are focused on the early design, simulation, and visualization process, with challenges still remaining for real time in vehicle generative systems that can address hardware constraints, while also managing current vehicle systems. By utilising specific hardware, optimising neural architectures, increasing training and inference efficiency, managing integration with other new technologies (the Metaverse), building sustainable AI practices, and bolstering data privacy the industry can start entering a phase of more scalable and efficient GenAI use. This is expected to coincide with many of the same aspects, as there was increased trends across the industry to more interpretable and user-centric design methodologies, as the next wave of innovation in automotive intelligence occurs.

Academically speaking, continuity is needed with further research in areas including: model optimisation for embedded devices, standardised testing for safety-critical AI, or approaches that could help connect the development of AI to automotive engineering. Research in these areas can provide an elemental groundwork for increased reliability and explanation of AI systems in vehicles.

In brief, generative AI has the ability to significantly affect how we design, engineer and operate vehicles. However, the achievement of this potential can only be made through sustained collaboration between researchers, developers and policy-makers. If these efforts align, the next decade could see a gradual but meaningful shift toward AI-enabled mobility systems that are safer, more efficient and adaptive to user needs.

7. ACKNOWLEDGEMENT

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project “NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung”. The authors would like to thank the consortium for the successful cooperation.

References

[1] Winter K, Vivekanandan A, Polley R, Shen Y, Schlauch C, et al. Generative AI for Autonomous Driving: A Review. 2025. ArXiv preprint: https://arxiv.org/pdf/2505.15863.

[2] Feng T, Wang W, Yang Y. A Survey of World Models for Autonomous Driving. 2025. ArXiv preprint: https://arxiv.org/pdf/2501.11260.

[3] Marathe A, Ramanan D, Walambe R, Kotecha K. WEDGE: A Multi-Weather Autonomous Driving Dataset Built From Generative Vision-Language Models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE. 2023:3318-3327.

[4] Goel H, Narasimhan SS, Akcin O, Chinchali S. SynDiff-AD: Improving Semantic Segmentation and End-To-End Autonomous Driving With Synthetic Data From Latent Diffusion Models. 2025. ArXiv preprint: https://arxiv.org/pdf/2411.16776v2.

[5] Silva M, Seoane A, Mures OA, López AM, Iglesias-Guitian JA. Exploring the Effects of Synthetic Data Generation: A Case Study on Autonomous Driving for Semantic Segmentation. Vis Comput. 2025;41:7379-7397.

[6] Hüsem H, Orman Z. A Survey on Image Super-Resolution With Generative Adversarial Networks. Acta Infologica. 2020;4:139-154.

[7] Gkillas A, Lalos A, Ampeliotis D. An Efficient Deep Unrolling Super-Resolution Network for Lidar Automotive Scenes. Proceedings of the IEEE International Conference on Image Processing. Image Process (ICIP). IEEE. 2023:1840-1844.

[8] Hu A, Russell L, Yeo H, Murez Z, Fedoseev G, et al. GAIA-1: A Generative World Model for Autonomous Driving. 2023. ArXiv preprint: https://arxiv.org/pdf/2309.17080.

[9] Russell L, Hu A, Bertoni L, Fedoseev G, Shotton J, et al. GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving. 2025. ArXiv preprint: https://arxiv.org/pdf/2503.20523.

[10] Arechiga N, Permenter F, Song B, Yuan C. Drag-Guided Diffusion Models for Vehicle Image Generation. 2023. ArXiv preprint: https://arxiv.org/pdf/2306.09935.

[11] https://www.audi-mediacenter.com/en/press-releases/reinventing-the-wheel-felgan-inspires-new-rim-designs-with-ai-15097.

[12] Vuruma SKR, Margetts A, Su J, Ahmed F, Srivastava B. From Cloud to Edge: Rethinking Generative AI for Low-Resource Design Challenges. 2024. ArXiv preprint: https://arxiv.org/pdf/2402.12702v1.

[13] Xu M, Du H, Niyato D, Kang J, Xiong Z, et al. Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services. IEEE Commun Surv Tutor. 2024;26:1127-1170.

[14] https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies/.

[15] https://www.forbes.com/sites/moorinsights/2019/05/09/google-cloud-doubles-down-on-nvidia-gpus-for-inference/.

[16] Garikapati DP, Shetiya SS. Autonomous Vehicles: Evolution of Artificial Intelligence and the Current Industry Landscape. Big Data Cogn Comput. 2024;8:42.

[17] Katiyar N, Shukla A, Chawla N, Singh MR. S- Kumar Singh, et al. AI in Autonomous Vehicles: Opportunities Challenges and Regulatory Implications. Educ Admin Theory Pract. 2024;30:6255-6264.

[18] Stappen L, Dillmann J, Striegel S, Vögel HJ, Flores-Herr N, et al. Integrating Generative Artificial Intelligence in Intelligent Vehicle Systems. In: Proceedings of the 2023 IEEE 26th international conference on intelligent transportation systems (ITSC). IEEE. 2023:5790-5797.

[19] Ebel P. Generative AI and Attentive User Interfaces: Five Strategies to Enhance Takeover Quality in Automated Driving. 2024. ArXiv preprint: https://arxiv.org/pdf/2402.10664.

[20] Yan H, Li Y. A Survey of Generative AI for Intelligent Transportation Systems. 2023. ArXiv preprint: https://arxiv.org/pdf/2312.08248v1.

[21] Abdelhamid S, Hassanein HS, Takahara G. Vehicle as a Mobile Sensor. Procedia Comput Sci. 2014;34:286-295.

[22] Rumez M, Grimm D, Kriesten R, Sax E. An Overview of Automotive Service-Oriented Architectures and Implications for Security Countermeasures. IEEE Access. 2020;8:221852-221870.

[23] Antinyan V. Revealing the Complexity of Automotive Software. In: Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering. ACM. 2020:1525-1528.

[24] Ayres N, Deka L, Passow B. Virtualisation as a Means for Dynamic Software Update Within the Automotive E/E Architecture. In: IEEE SmartWorld ubiquitous intelligence computing advanced trusted computing scalable computing communications cloud big data computing Internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE. 2019:154-157.

[25] Bandur V, Selim G, Pantelic V, Lawford M. Making the Case for Centralized Automotive E/E Architectures. IEEE Trans Veh Technol. 2021;70:1230-1245.

[26] Askaripoor H, Hashemi Farzaneh M, Knoll A. E/E Architecture Synthesis: Challenges and Technologies. Electronics. 2022;11:518.

[27] https://www.mckinsey.com/industries/semiconductors/our-insights/advanced-semiconductors-for-the-era-of-centralized-e-e-architectures

[28] Reuther A, Michaleas P, Jones M, Gadepally V, Samsi S, et al. AI Accelerator Survey and Trends. In: IEEE High Performance Extreme Computing Conference (HPEC). IEEE. 2021:1-9.

[29] Reuther A, Michaleas P, Jones M, Gadepally V, Samsi S, et al. Lincoln AI Computing Survey (LAICS) Update. In: IEEE High Performance Extreme Computing Conference (HPEC). IEEE. 2023:1-7.

[30] Kugele S, Hettler D, Peter J. Data-Centric Communication and Containerization for Future Automotive Software Architectures. In: IEEE International Conference on Software Architecture (ICSA). IEEE. 2018:65-6509.

[31] Obermaisser R, El Salloum C, Huber B, Kopetz H. From a Federated to an Integrated Automotive Architecture. IEEE Trans Comput-Aided Design Integr Circuits Syst. 2009;28:956-965.

[32] Cebotari V, Kugele S. On the Nature of Automotive Service Architectures. In: IEEE International Conference on Software Architecture Companion. ICSA-C. IEEE. 2019:53-60.

[33] Helms D, Uven P, Grüttner K. Modular Over-The-Air Software Updates for Safety-Critical Real-Time Systems. Insight. 2022;25:85-88.

[34] Stähle H, Mercep L, Knoll A, Spiegelberg G. Towards the Deployment of a Centralized Ict Architecture in the Automotive Domain. In: 2nd Mediterranean Conference on Embedded Computing. MECO. IEEE. 2013:66-69.