International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 1, 231-237, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I1P138

Original Article

Small Language Models and Neuro-Symbolic AI in Zonal Architectures: Federated Low-Rank Adaptation (Fed-LoRA) for Regional Behavior Modeling

Naresh Kalimuthu
Independent Researcher, USA.

Received On: 20/01/2026 ***Revised On: 20/02/2026*** ***Accepted On: 22/02/2026*** ***Published on: 24/02/2026***

Abstract *- The automotive industry is undergoing a fundamental change, shifting from domain-centric Electrical/Electronic (E/E) architectures to Zonal Architectures designed for the Software-Defined Vehicle (SDV). This shift enables high-performance edge computing but imposes strict constraints on power, thermal management, and resource contention for mixed-criticality tasks. At the same time, autonomous systems must adapt to diverse regional driving habits ranging from strict traffic-law compliance in Western Europe to the chaotic, negotiation-heavy traffic of South Asia posing challenges for existing centralized training approaches. This report introduces a comprehensive framework that combines Small Language Models (SLMs) for semantic reasoning, Neuro-Symbolic AI (NeSy) for safety validation, and Federated Low-Rank Adaptation (Fed-LoRA) for bandwidth-efficient, privacy-preserving ongoing learning. We examine the hardware capabilities of advanced zonal controllers (e.g., NXP S32G3, TI Jacinto 7) and show that, despite having less raw data-processing power than data center GPUs, their NPU-accelerated designs are adequate for quantized SLMs when optimized with techniques such as Federated Silver Bullet (Fed-SB). Additionally, we propose a hierarchical "Regional Behavior" learning model in which a Neuro-Symbolic safety shield ensures that universal traffic rules are followed, while the SLM adapts to local cultural norms via low-rank parameter updates. This hybrid architecture balances the need for local customization with the essential safety guarantees required in next-generation autonomous vehicle mobility.*

Keywords *- Software-Defined Vehicle (SDV), Zonal Architecture, Small Language Models (SLMs), Neuro-Symbolic AI (NeSy), Federated Low-Rank Adaptation (Fed-LoRA), Federated Silver Bullet (Fed-SB), Regional Behavior Modeling, Edge Intelligence, Vehicle Profiling, ISO 26262 Safety Assurance.*

1. Introduction

1.1. The Software-Defined Vehicle (SDV) Evolution

The automotive industry is currently experiencing a convergence of three major trends: architectural transformation, advances in artificial intelligence, and distributed learning models. The concept of the Software-Defined Vehicle (SDV) has developed quickly. We are moving from "SDV 2.0," which separates hardware and software within domain controllers, to "SDV 3.0," a vehicle-wide, service-oriented architecture that operates on consolidated high-performance computing platforms. This evolution is driven by consumer demands for seamless personalization, enhanced automation, and constant connectivity, requiring vehicle architectures that can update functionality after manufacturing via Over-the-Air (OTA) updates.

However, the traditional approach of installing a separate Electronic Control Unit (ECU) for each new feature has reached its physical and logical limits. Modern luxury cars now contain over 100 ECUs and nearly 5 kilometers of wiring, making the wiring harness often the third-heaviest and second-most costly component in the vehicle. To overcome this, the industry is shifting toward Zonal Architectures. In this layout, traditional functional areas such as Powertrain, Body, and Chassis are replaced by geometric zones, such as Front-Left and Rear-Right. A Zonal Control Unit (ZCU) gathers all sensors and actuators within its zone, processing data locally or transmitting it via high-speed Ethernet Time-Sensitive Networking (TSN) backbones to a central system computer.

1.2. The Challenge of Regional Behavior Modeling

A key, often overlooked challenge in deploying Autonomous Driving (AD) and Advanced Driver-Assistance Systems (ADAS) worldwide is the significant variability in driving cultures, regulations, and social norms. Policies trained on structured datasets from cities such as San Francisco or Munich may perform poorly in more aggressive, unstructured traffic environments, such as Naples, Mumbai, or Hanoi. These differences extend beyond simple parameters and are both structural and semantic; for example, in some regions, a turn signal is merely a request, whereas in others it indicates a clear intent.

Traditional centralized learning methods, which involve uploading vast amounts of raw sensor data to cloud servers for training, are becoming less feasible for three reasons:

1. Data Gravity and Bandwidth: Transferring petabytes of high-quality video and LiDAR data from millions of vehicles entails significant costs and latency challenges due to limited cellular connectivity bandwidth.
2. Privacy and Regulation: Strict data regulations, such as the GDPR in Europe, the CCPA in California, and China's new data sovereignty laws, limit the transfer of raw data that includes identifiable faces or license plates across borders.
3. The Long Tail of Regional Nuance: A universal model often fails to capture specific, low-probability edge cases unique to a locale, such as "Pittsburgh Lefts" or specific roundabout rules in the UK.

1.3. The Convergence of SLMs, NeSy, and Fed-LoRA

This report explores a novel architecture to address the regional adaptation challenge without risking safety or efficiency:

- Small Language Models (SLMs) at the Edge: Instead of deploying large, generic Large Language Models (LLMs), the industry is shifting towards SLMs (100M–5B parameters) that offer semantic understanding and reasoning directly on the zonal controller. These models can interpret complex traffic situations (e.g., "The policeman is waving me through a red light") that perception stacks alone might miss.
- Neuro-Symbolic AI (NeSy): To reduce the probabilistic errors and hallucinations common in neural networks, NeSy combines deep learning with symbolic logic. A symbolic layer encodes fixed traffic rules and physical constraints, serving as a deterministic "guardian" that verifies SLM outputs before execution.
- Federated Low-Rank Adaptation (Fed-LoRA): This technique enables vehicles to collaboratively fine-tune their SLMs using local data without sharing raw inputs. By keeping the base model unchanged and training only low-rank adapter matrices, Fed-LoRA significantly reduces communication requirements compared with full model sharing.

2. Zonal Architectures: The Hardware Foundation

Switching to Zonal Architecture entails more than merely reconfiguring connections; it fundamentally alters the vehicle's computing system. Zonal controllers function like the car's 'neural ganglia,' processing extensive sensor data and overseeing local control loops. Understanding the capabilities and restrictions of current automotive silicon is crucial for deploying Fed-LoRA and similar technologies SLMs.

2.1. The Geometric Constraint Paradigm

Unlike domain architectures, where ECUs are organized by function, zonal architectures are determined by geometry. This means a single ZCU must handle a diverse range of workloads based solely on proximity. For example, a Front-Left ZCU might manage the driver's door window lift (Body control, low speed), the front-left radar processing (ADAS, high bandwidth), and power delivery to the headlights (Energy management).

This integration imposes strict hardware requirements:

- Mixed-Criticality Isolation: The ZCU must guarantee that non-critical AI tasks (like cabin personalization) do not interfere with safety-critical functions such as braking or steering. This requires advanced virtualization and hardware-enforced isolation, often implemented by hypervisors running on multi-core System-on-Chip (SoC).
- Thermal and Power Constraints: Zonal modules are typically situated in physically limited, challenging environments (e.g., within door panels or wheel wells) where active cooling isn’t feasible. This limits the thermal design power (TDP) for AI computation, making energy efficiency (TOPS/Watt) a vital metric relative to raw performance.
- I/O Concentration: ZCUs serve as gateways, translating legacy signals (CAN, LIN) into high-speed Ethernet data packets. The processor must allocate a considerable amount of silicon area to network acceleration (packet switching), thereby reducing the die area available for AI processing logic.

2.2. Analysis of Zonal Controller Hardware

The ability to run SLMs and Fed-LoRA relies entirely on the capabilities of the Neural Processing Unit (NPU) and Digital Signal Processor (DSP) in the underlying silicon. We evaluate the top platforms that define this domain.

2.2.1. NXP S32G3 Series: The Vehicle Network Processor

The NXP S32G3 is a quintessential zonal gateway processor, widely adopted for its robust networking capabilities.

- Compute Architecture: It features a diverse design with up to four Arm Cortex-M7 lockstep cores dedicated to ASIL-D real-time safety, and up to eight Arm Cortex-A53 cores for more advanced tasks applications.
- Memory Constraints: The device typically integrates ~20 MB of System SRAM. This is a critical bottleneck for Transformer-based SLMs, which require substantial memory for KV (Key-Value) caching during inference. External memory interfaces (Low-Power Double Data Rate 4) are available, but they incur latency penalties.
- AI Capability: The S32G3 excels in network acceleration via its Low Latency Communication Engine (LLCE) and Packet Forwarding Engine (PFE). However, it lacks a massive, dedicated tensor accelerator. AI inference is generally relegated to the Cortex-A53 clusters or specialized DSP libraries, limiting its capacity to very small, highly quantized models (e.g., <500M parameters) or requiring offloading to a central compute node.

- Role in Fed-LoRA: The S32G3 is ideal for lightweight data collection and preprocessing tasks, rather than serving as a training node. It can run the "Symbolic" logic layer of a NeSy architecture on its real-time cores effectively.

2.2.2. TI Jacinto 7 (TDA4VM / DRA829V): The Edge AI Specialist

Texas Instruments' Jacinto 7 platform features a unique architectural approach by incorporating significant AI acceleration directly into the gateway/ADAS processor.

- Compute Architecture: It merges dual Arm Cortex-A72 cores for versatile processing, alongside C7x DSPs and specialized Matrix Multiplication Accelerators (MMA).
- AI Performance: The deep learning accelerators, which deliver approximately 8-32 TOPS depending on the variant, operate at low power levels of 5-20 W, eliminating the need for active cooling. The C7x DSP combined with MMA is specifically optimized for the matrix-vector operations essential to deep neural networks.
- SLM Suitability: Benchmarks show that specialized NPUs like the MMA significantly outperform general-purpose CPUs such as the Cortex-A72 in SLM inference, especially in throughput and energy efficiency. The TDA4VM is capable of processing quantized SLMs with 1B-3B parameters, making it suitable for performing Fed-LoRA updates locally.
- Safety Integration: Like the S32G, it includes Cortex-R5F microcontroller islands tailored for ASIL-D safety, allowing physical separation of the "Neural" component (on MMA) from the "Symbolic" component (on R5F) architecture.

2.2.3. NVIDIA DRIVE Thor: The Centralized Brain

Although frequently regarded as a central computer, NVIDIA Thor influences zonal strategies by supporting "Centralized Training, Zonal Execution."

- Compute Power: Thor integrates the Blackwell GPU architecture, delivering up to 2000 TOPS.
- Transformer Engine: Importantly, it includes a built-in Transformer Engine that supports FP8 precision and is specifically engineered to accelerate the training and inference of Transformer models, which form the foundation of this architecture (SLMs).
- Architectural Implications: In a vehicle with a Thor central node and S32G zonal nodes, the computationally intensive Fed-LoRA aggregation can occur on Thor, while the S32G nodes handle sensor ingestion and lightweight symbolic filtering.

2.2.4. Renesas R-Car S4: The Communicator

The R-Car S4 emphasizes high-bandwidth communication and integrated security.

- Specs: It features Cortex-A55, Cortex-R52, and RH850 cores, along with 8MB of SRAM and a 3-port Ethernet TSN switch.
- Application: The "Whitebox SDK" supports open development, but its main strength is serving as a secure gateway. It is well-suited for managing the networking tasks in Federated Learning, such as encrypting and transmitting model updates, but it may find it challenging to handle the computational demands of backpropagation for larger models SLMs.

2.3. Hardware Comparison for Fed-LoRA Deployment

Table 1: Comparison of Automotive Compute Platforms for AI and SLM Deployment

Feature	NXP S32G3	TI Jacinto 7 (TDA4)	NVIDIA Thor	Renesas R-Car S4
Primary Role	Zonal Gateway / Networking	ADAS / Vision / Gateway	Central Compute / AI	Gateway / Body Control
AI Accelerator	Limited (CPU bound)	C7x DSP + MMA (Matrix Accel.)	Blackwell Tensor Cores	CNN IP
Est. Compute (TOPS)	< 5 TOPS	~8 - 32 TOPS	~2000 TOPS	Medium
Memory	~20MB SRAM + LPDDR4	Shared L2/L3 + LPDDR4	High Bandwidth LPDDR5X	8MB SRAM + LPDDR4
SLM Viability	Quantized Micro-models (<500M)	Quantized SLMs (1B-3B)	Full SLMs/LLMs (>7B)	Quantized Micro-models
Fed-LoRA Role	Data Collection / Symbolic Logic	Client (Inference + Update)	Server / Aggregator / Training	Secure Transmission
Safety Core	Cortex-M7 (Lockstep)	Cortex-R5F (Lockstep)	Safety Islands	Cortex-R52 / RH850

2.4. Software-Defined Resource Management

The complexity of these System-on-Chips requires a sophisticated software stack, with hypervisors playing a key role in resource partitioning. For example, a "Safety Partition" on the Cortex-R cores might run the symbolic logic engine (ASIL-D), while a "Performance Partition" on the Cortex-A cores or NPU manages the SLM (QM/ASIL-B). Research into "Zonal Architecture" resource contention highlights the need for dynamic scheduling to avoid SLM activity bursts during a Fed-LoRA update from affecting the latency of safety-critical messages on the Time-Sensitive Networking (TSN) bus.

3. Small Language Models (SLMs) at the Edge

Deploying large LLMs such as GPT-4 or Llama-3-70B on edge devices isn't feasible because of memory bandwidth

and latency limits. As a result, Small Language Models (SLMs) Transformer models with 100M to 5B parameters—are now the best choice for enabling semantic intelligence in vehicles edge.

3.1. The Case for SLMs in Automotive

SLMs offer a unique value proposition for regional behavior modeling:

- Latency-Critical Inference: A Zonal Control Unit (ZCU) needs to make driving decisions within milliseconds. Relying on the cloud for inference adds unpredictable network latency (RTT), which is unsuitable for safety-critical functions. Running SLMs locally on the NPU provides consistent and deterministic responses times.
- Privacy and Compliance: Processing cabin audio/video (for driver monitoring) or exterior camera feeds locally ensures that sensitive biometric and location data remain within the vehicle, thereby complying with GDPR and other privacy regulatory frameworks.
- Contextual Reasoning: Unlike traditional Convolutional Neural Networks (CNNs), which classify objects like "This is a pedestrian," SLMs are capable of reasoning about intent and context, such as "The pedestrian is looking at their phone and walking towards the curb, implying a high risk of stepping out."

3.2. ptimization Strategies for Resource-Constrained Zones

Achieving effective SLMs on the hardware described in Section 2 requires aggressive optimization.

3.2.1. Quantization and Precision

Standard FP32 or FP16 models are too large for zonal controllers, so quantization to INT8 or INT4 is essential. Research indicates that INT4 quantization can decrease memory usage by 4 times with minimal accuracy reduction for many applications tasks.

- Hardware Alignment: Specialized NPUs, such as the RaiderChip or TI MMA, demonstrate 50-70% performance gains over general-purpose CPUs when executing low-precision formats like F16/Q4K. For example, the Energy-Delay Product (EDP) of NPUs running quantized models can be up to 140% better than GPUs, positioning them as the only feasible choice for battery-constrained devices EVs.

3.2.2. Architectural Efficiencies

Modern SLMs are implementing architectural updates to optimize edge inference: Grouped-Query Attention (GQA) replaces Multi-Head Attention (MHA), reducing the KV cache size and alleviating the main-memory bottleneck in long-context inference. Additionally, KV Cache Compression methods are crucial for zonal controllers with limited SRAM, such as the 20MB restriction on S32G3.

3.2.3. Performance: NPU vs. CPU

The selection of the backend is crucial. Empirical studies show that ARM CPUs, commonly used in ZCUs, are energy-efficient but struggle to achieve sufficient throughput for real-time SLM inference. x86 CPUs perform poorly in both efficiency and throughput within this context. Specialized NPUs and DSPs are the leading architectures, delivering the highest tokens-per-second and superior energy efficiency. Consequently, effective deployment of Fed-LoRA depends on ZCUs equipped with dedicated NPU IP.

4. Neuro-Symbolic Ai: The Safety Assurance Layer

While SLMs excel at pattern recognition and generating semantic content, they are also prone to randomness, known as "hallucinations," and often lack interpretability. In critical areas such as autonomous driving, relying on a black-box model that cannot explain its lane-change decisions poses risks. Neuro-Symbolic AI (NeSy) tackles this issue by combining the learning prowess of neural networks with the logical reasoning strengths of symbolic systems.

4.1. The Hybrid Architecture

NeSy systems in automotive contexts typically follow a hierarchical architecture:

- The Neural Module (Perception & Prediction): This component employs Deep Learning techniques, such as CNNs, LSTMs, and Transformers, to convert high-dimensional raw data, such as pixels and LiDAR points, into meaningful state representations. For instance, a CNN-LSTM network might identify spatial features and track the movement patterns of nearby vehicles.
- The Symbolic Module (Reasoning & Safety): This uses formal logic systems, including Answer Set Programming (ASP), First-Order Logic (FOL), or Prolog, to encode explicit rules. These rules encapsulate traffic laws, physics constraints, and safety invariants—for example, Distance(Car, Pedestrian) < Threshold -> Action(Brake).
- The Fusion Layer: Techniques such as "Attention Gates" and "Logic Tensor Networks" integrate these modules. The symbolic module acts as a "Safety Shield," overriding neural predictions when they conflict with logical constraint rules.

4.2. Neuro-Symbolic Reinforcement Learning (NSRL)

For modeling regional behavior, the system must adopt policies. NSRL integrates Deep Q-Networks (DQN) with symbolic logic.

- Mechanism: The DQN determines the best action from the state vector, while the Symbolic Module assesses this action according to the rule set.
- Reward Shaping: If an action breaks a rule, like running a red light, the symbolic layer gives a large negative reward (-1). When the action is safe and legal, it gets a positive reward (+1).
- Outcome: This directs the neural agent to develop policies that optimize efficiency within safety boundaries. Testing on the Lyft Level 5 Motion Prediction dataset demonstrates that this method attains 98% accuracy in scenario-based decision-

making, surpassing methods that rely solely on deep learning models.

4.3. Regional Logic and Rule Alignment

A key challenge is that traffic rules are not universal. "Right on Red" is legal in the US but illegal in most of Europe.

- Symbolic Ontology: The Symbolic Module depends on a localized ontology. When a vehicle crosses a border, such as from France to Germany, the ZCU loads the appropriate "Rule Set" into the symbolic system engine.
- FedNSL (Federated Neuro-Symbolic Learning): Research indicates using Federated Learning to address rule heterogeneity. Systems such as LR-XFL (Logic-Reasoning Explainable FL) or FedNSL enable vehicles to infer new logical rules from data and resolve conflicts between local rules and global models. This allows the "Symbolic Shield" to be updated dynamically according to current regulations environment.

5. Federated Low-Rank Adaptation (Fed-LoRA)

Federated Learning (FL) enables privacy-preserving collaborative learning, but transmitting full model updates is bandwidth-prohibitive. Low-Rank Adaptation (LoRA) offers a solution by freezing the pre-trained model weights (W_0) and optimizing only low-rank matrices (A and B). Fed-LoRA combines these concepts to enable efficient edge training.

5.1. The Fed-LoRA Mechanism

In a standard LoRA configuration, the weight update is parameterized as $\Delta W = BA$, where $A \in R^{rxd}$ and $B \in R^{dxr}$ are low-rank matrices ($r \ll d$).

- Local Training: Each ZCU fine-tunes A and B on its local data (e.g., disengagements, human driver traces).
- Aggregation: Instead of sending the full ΔW(millions of parameters), the ZCU sends only A and B.
- Global Update: The server aggregates these matrices. However, simple averaging does not mathematically correspond to the average of the full weight updates because of the nonlinearity of the product. This leads to "aggregation error".

5.2. Advanced Aggregation: Fed-SB (Silver Bullet)

To mitigate aggregation errors and lower communication expenses, the Fed-SB (Federated Silver Bullet) method is presented as a cutting-edge solution.

- Mechanism: Fed-SB initializes the adapter matrices and through singular value decomposition of a proxy update and then freezes them. It subsequently inserts a small, learnable square matrix between the two, satisfying the condition.
- Training: Only the small matrix is trained locally on the ZCU.
- Exact Aggregation: Because both are fixed and shared among all clients, the update aggregation is linear in the number of servers that just calculate. This approach is mathematically precise, removing the aggregation noise that arises in methods such as Fed-LoRA and similar FedEx-LoRA.
- Communication Efficiency: Communication cost scales proportionally, leading to up to a 230x reduction in communication overhead relative to standard federated fine-tuning methods, while remaining independent of the number of clients.

5.3. Handling Heterogeneity: VPFLA

Vehicles in various regions encounter distinct data distributions (Non-IID). Combining a "Rural" adapter with an "Urban" one reduces performance for both.

- Vehicle Profiling-aware Federated Low-rank Adaptation (VPFLA): This framework features a "Vehicle Profiling" module that creates a profile vector capturing the vehicle's particular driving traits, such as traffic density and road geometry location.
- Clustered Learning: The server groups vehicles into clusters using these profiles. It creates specialized "Regional Adapters" by aggregating updates exclusively from relevant peers. This method, called "Personalized Federated Learning," has demonstrated a 95.97% accuracy rate in IoV simulations and reduces communication overhead by 99.97%.

Table 2 : Comparison of FL Aggregation Strategies

Method	Parameters Transmitted	Aggregation Accuracy	Client Scalability	Heterogeneity Support
FedAvg (Full Fine-Tuning)	Full Model (d^2)	Exact	Poor (Bandwidth heavy)	Standard
Standard Fed-LoRA	Adapters A,B (2dr)	Approx/Noisy	Medium	Medium
FedEx-LoRA	A, B + Error Matrix	Improved	Linear cost with clients	High
Fed-SB (Silver Bullet)	Matrix R (r^2)	Exact	Excellent (Constant cost)	High (via padding)

6. Regional Behavior Modeling: A Unified Framework

We propose a unified Neuro-Symbolic Fed-LoRA architecture for Zonal SDVs. This framework leverages the hardware of the zonal controller to run a safety-guarded SLM that continuously adapts to regional norms via Fed-SB.

6.1. The "Guardian-Agent" Architecture

The system is composed of two primary parallel processes running on the ZCU:

6.1.1. The Agent (SLM on NPU)

- Input: Tokenized scene description (for example, from CLIP processing camera feeds) combined with Context History.
- Model: A quantized SLM, such as Phi-2, equipped with a loaded "Regional LoRA" Adapter."
- Task: Generates the targeted movement or action, such as "Nudge forward to signal intent" merge").

6.1.2. The Guardian (Symbolic Logic on Safety Core)

- Input: High-confidence object list + State Vector.
- Knowledge Base: Regional Traffic Rules (loaded via GPS geofencing).
- Task: Validates the Agent's action: verify (Action, Rules). If deemed Safe, proceed to the Actuator; if Unsafe, block and execute Fallback_Maneuver.

6.2. The Learning Loop (Fed-SB)

When the Guardian vetoes an action (or the human driver intervenes), a "Learning Trigger" is activated.Data Capture: The sequence of events leading to the intervention is stored.Local Fine-Tuning: During idle/charging periods, the ZCU fine-tunes the local R matrix of the Fed-SB adapter on this data. The objective is to align the SLM's prediction with the safe/human-preferred action.Aggregation: The updated R matrix is transmitted to the Regional Cloud.Distribution: An updated Global R matrix (specific to that region) is pushed back to the fleet, effectively propagating the "lesson learned" to all vehicles in the zone.

6.3. Case Study: The "Merging" Problem

Consider the behavior of merging onto a congested roundabout.

- Region A (Germany): Rules are strict. The Guardian enforces Yield_If_Car_In_Circle. The SLM learns a conservative policy
- Region B (Vietnam): Traffic flows continuously. A strict yield means the car never enters. The human driver forces a merge.
- Adaptation: The local Fed-SB updates in Region B shift the SLM's latent space to favor "assertive" actions. The Symbolic Guardian in Region B is configured with a relaxed constraint (e.g., Yield implies Avoid_Collision, not Wait_Until_Empty), allowing the assertive behavior to pass verification. The combination enables the vehicle to operate naturally in Hanoi without violating fundamental safety principles.

7. Conclusion

The merging of Zonal Architectures, Small Language Models, and Neuro-Symbolic AI marks a new phase in automotive development, transforming vehicles from fixed, hardware-based machines into adaptable, software-driven agents. Deploying SLMs at the edge equips vehicles with the semantic understanding needed to handle complex human environments. Encasing these models within a Neuro-Symbolic safety framework allows manufacturers to maintain strict safety compliance, despite AI's probabilistic nature. Additionally, using Federated Low-Rank Adaptation—such as advanced, bandwidth-efficient methods like Fed-SB enables fleets to learn continuously and adapt to various global driving cultures without risking user privacy or incurring excessive data costs. The future vehicle is thus more than just a sensor platform; it becomes a collaborative, reasoning, edge node within a worldwide learning network.

References

[1] Santos, Afonso & Martins, José & Sousa, João & Rodríguez, Manuel & Pinto, Sandro. (2025). Let's Get Physical: Rethinking the Static Partitioning Hypervisor Architecture for an MMU-Less Memory Model. IEEE Access. PP. 1-1. 10.1109/ACCESS.2025.3636061.

[2] Hossain, Md Sanowar & Jesser, Alexander. (Jan 2026). Reinforcement Learning-Based Adaptive Wire Gauge Selection for Zonal Automotive Harness Design Under Dynamic Driving Scenarios. IEEE Access. PP. 1-1. 10.1109/ACCESS.2026.3650809..

[3] Lu, Zhenyan & Li, Xiang & Cai, Dongqi & Yi, Rongjie & Liu, Fangming & Liu, Wei & Luan, Jian & Zhang, Xiwen & Lane, Nicholas & Xu, Mengwei. (2025). Demystifying Small Language Models for Edge Deployment. 14747-14764. 10.18653/v1/2025.acl-long.718.

[4] Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., & Molchanov, P. (2025). Small Language Models are the Future of Agentic AI. *ArXiv*. https://arxiv.org/abs/2506.02153.

[5] H. Sun, H. Tian, W. Ni, J. Zheng, D. Niyato and P. Zhang, "Federated Low-Rank Adaptation for Large Models Fine-Tuning Over Wireless Networks," in IEEE Transactions on Wireless Communications, vol. 24, no. 1, pp. 659-675, Jan. 2025, doi: 10.1109/TWC.2024.3497998.

[6] Singhal, R., Ponkshe, K., Vartak, R., Varshney, L. R., & Vepakomma, P. (2025). Fed-SB: A Silver Bullet for Extreme Communication Efficiency and Performance in (Private) Federated LoRA Fine-Tuning. ArXiv. https://arxiv.org/abs/2502.15436.

[7] S. Xiao, X. Huang, M. Zhou, C. Liang and Q. Chen, "Vehicle Profiling-Aware Personalized Federated Low-Rank Adaptation in IoVs," in IEEE Communications Letters, vol. 30, pp. 159-163, 2026, doi: 10.1109/LCOMM.2025.3633387.

[8] Zhao, Shijun & Zhang, Qianying & Qin, Yu & Feng, Wei & Feng, Dengguo. (2019). SecTEE: A Software-

based Approach to Secure Enclave Architecture Using TEE. 1723-1740. 10.1145/3319535.3363205.

[9] Xing, P., Lu, S., & Yu, H. (2023). Federated Neuro-Symbolic Learning. ArXiv. https://arxiv.org/abs/2308.15324

[10] Zhang, Y., & Yu, H. (2023). LR-XFL: Logical Reasoning-based Explainable Federated Learning. ArXiv. https://arxiv.org/abs/2308.12681.

[11] Salah, Islam & Son, Junggab & Robila, Stefan & Kim, Daeyoung. (2026). Evaluating Small Language Models for Intrusion Detection on Automotive Embedded Platforms. 1-7. 10.1145/3769002.3769959.

[12] G. Elinoff, "Zonal Architecture: The Next Phase for Software-Defined Vehicles," Electropages, May 2025. [Online]. Available: https://www.electropages.com/blog/2025/05/zonal-architecture-next-phase-software-designed-vehicles.

[13] Design World Staff, "Addressing Zonal Architecture Challenges in the Automotive Industry," Design World, Jan. 2024. [Online]. Available: https://www.designworldonline.com/addressing-zonal-architecture-challenges-in-the-automotive-industry/.

[14] NVIDIA, "Federated Learning in Autonomous Vehicles Using Cross-Border Training," NVIDIA Developer Blog, 2023. [Online]. Available: https://developer.nvidia.com/blog/federated-learning-in-autonomous-vehicles-using-cross-border-training/

[15] Semiconductor Engineering, "Designing Vehicles Virtually," Semiconductor Engineering, Jan. 2025. [Online]. Available: https://semiengineering.com/designing-vehicles-virtually/.

[16] NVIDIA, "DRIVE Thor: The Centralized Brain for Autonomous Vehicles," NVIDIA Blog, Sept. 2022. [Online]. Available: https://blogs.nvidia.com/blog/drive-thor/.

[17] NXP Semiconductors, "S32G3 Data Sheet," Rev. 4, Sept. 2025. [Online]. Available: https://www.nxp.com/docs/en/data-sheet/S32G3.pdf.

[18] Texas Instruments, "TDA4VM Jacinto™ 7 Processors for Driver Assistance," Data Sheet, Rev. K, April 2024. [Online]. Available: https://www.ti.com/lit/ds/symlink/tda4vm.pdf.

[19] Renesas Electronics, "R-Car S4: Automotive System-on-Chip (SoC) for Car Server/Communication Gateway," 2024. [Online]. Available: https://www.renesas.com/en/products/r-car-s4?srsltid=AfmBOopZvz2qJO6bYyM7AqZ56ht3kHBnTfM3_MbaZs_vdft52iots1TV