RESEARCH-ARTICLE

Foundation Models for Autonomous Driving Systems: An Initial Roadmap

XIONGFEI WU

MINGFEI CHENG

XIAONING REN

QIANG HU

JIANLANG CHEN

YUHENG HUANG

View all

Published: 07 April 2026
Accepted: 27 March 2026
Revised: 02 February 2026
Received: 01 October 2025

Foundation Models for Autonomous Driving Systems: An Initial Roadmap

XIONGFEI WU, University of Luxembourg, Luxembourg
MINGFEI CHENG, Singapore Management University, Singapore
XIAONING REN, University of Science and Technology of China, China
QIANG HU*, Tianjin University, China
JIANLANG CHEN, Kyushu University, Japan
YUHENG HUANG, The University of Tokyo, Japan
MAXIME CORDY, University of Luxembourg, Luxembourg
YAO ZHANG, Tianjin University, China
XIAOFEI XIE, Singapore Management University, Singapore
LEI MA, The University of Tokyo & University of Alberta, Japan & Canada
YVES LE TRAON, University of Luxembourg, Luxembourg

Recent advances in foundation models (FMs), including large language models (LLMs), vision-language models (VLMs), and world models, have opened new opportunities for autonomous driving systems (ADSs) in perception, reasoning, decision-making, and interaction. However, ADSs are safety-critical cyber-physical systems, and integrating FMs into them raises substantial software engineering challenges in data curation, system design, deployment, evaluation, and assurance. To clarify this rapidly evolving landscape, we present an initial roadmap, grounded in a structured literature review, for integrating FMs into autonomous driving across three dimensions: FM infrastructure, in-vehicle integration, and practical deployment. For each dimension, we summarize the state of the art, identify key challenges, and highlight open research opportunities. Based on this analysis, we outline research directions for building reliable, safe, and trustworthy FM-enabled ADSs.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Software and its engineering** → *Software verification and validation*; • **Computer systems organization** → **Embedded and cyber-physical systems**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Foundation Model, Autonomous Driving System, Roadmap, V2X

1 Introduction

Autonomous driving systems (ADSs) operate in open, dynamic, and safety-critical environments. Traditional approaches, while effective in controlled environments, often struggle with unseen situations, unexpected

* Corresponding author: Qiang Hu.
Authors' Contact Information: Xiongfei Wu, University of Luxembourg, Luxembourg; Mingfei Cheng, Singapore Management University, Singapore; Xiaoning Ren, University of Science and Technology of China, China; Qiang Hu*, Tianjin University, China; Jianlang Chen, Kyushu University, Japan; Yuheng Huang, The University of Tokyo, Japan; Maxime Cordy, University of Luxembourg, Luxembourg; Yao Zhang, Tianjin University, China; Xiaofei Xie, Singapore Management University, Singapore; Lei Ma, The University of Tokyo & University of Alberta, Japan & Canada; Yves Le Traon, University of Luxembourg, Luxembourg.

ACM 1557-7392/2026/4-ART
https://doi.org/10.1145/3807455

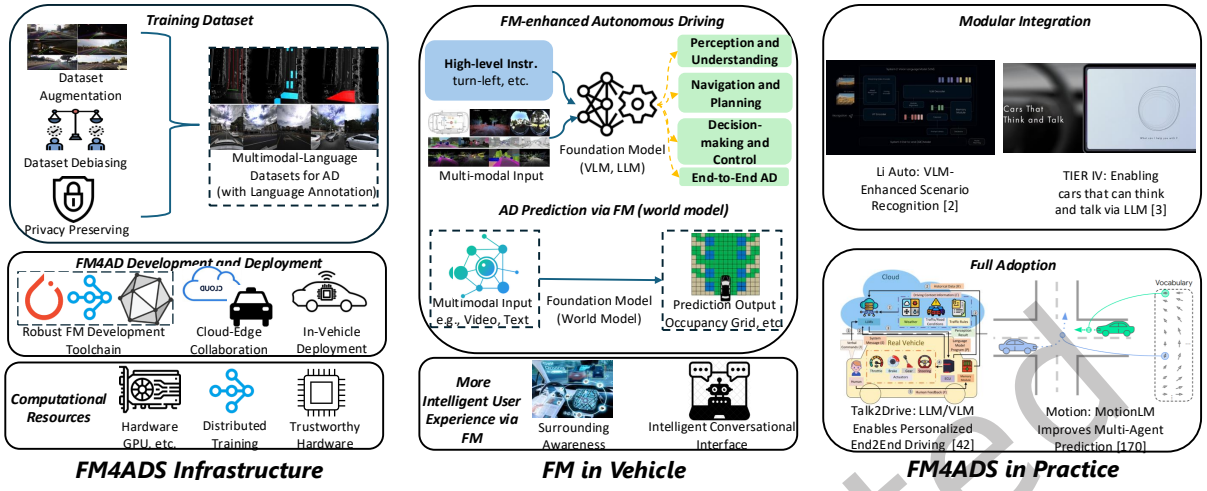

Fig. 1. Overview of the roadmap.

obstacles, and dynamic interactions that characterize real-world driving conditions [216]. A key limitation is their reliance on predetermined rules and supervised learning over finite labeled datasets, which cannot fully capture the diversity and long-tail nature of scenarios.

The emergence of foundation models (FMs) trained on vast and diverse datasets has demonstrated unprecedented capabilities in reasoning and generalization across various domains [19]. These models have exhibited remarkable abilities in understanding context, reasoning about the context, and generating appropriate responses. Their success in natural language processing [47, 194] and computer vision [119, 147, 159] tasks suggests promising applications in addressing the fundamental challenges of autonomous driving.

The autonomous driving ecosystem stands to benefit significantly from the integration of FMs, which can enhance real-world scenario understanding, improve decision-making, and facilitate robust system development. For instance, FMs can leverage their broad knowledge base to interpret complex traffic scenarios, understand natural language instructions from passengers, and make informed decisions in previously unseen situations [245]. This integration could bridge the gap between traditional autonomous driving systems' capabilities and the requirements for truly robust autonomous operation in diverse real-world conditions. However, despite their promising enhancement for autonomous driving, FMs also have certain problems. Due to their increasing complexity, FMs heavily rely on *data* and are especially hard to *manage*. This becomes a significant obstacle to integrating them into autonomous driving systems.

To help researchers better understand the role of foundation models in autonomous driving, considerable efforts have been made to broaden the community's perspective on their potential contributions. Song *et al.* [178] and Petrovic *et al.* [156] surveyed the use of generative AI, particularly LLMs, for testing autonomous driving systems. Other studies have explored how foundation models can advance autonomous driving more broadly [64, 216, 227]. While these works provide valuable insights and highlight important challenges and opportunities, most focus on specific topics or model families, for example, using LLMs to generate testing scenarios for ADS. To move beyond such fragmented perspectives, we propose a structured roadmap for integrating FMs into autonomous driving. Our roadmap spans three dimensions: FM infrastructure, its integration across autonomous driving system modules, and their practical real-world applications, as shown in Figure 1. For each dimension, we review the current research progress, identify existing challenges, and highlight research gaps that need to be

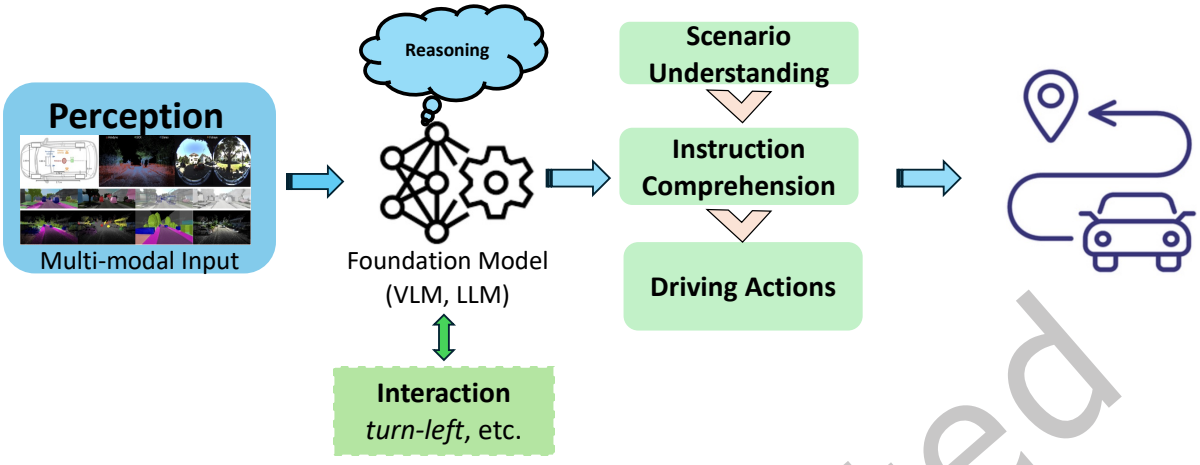

Fig. 2. A typical pipeline of enhancing autonomous driving leveraging FMs.

addressed by the community. Beyond a modular overview, Table 1 connects the three dimensions by showing how infrastructure-level decisions propagate through in-vehicle risks and ultimately manifest as failures in practice. Through this analysis, we aim to guide future research on building reliable, safe, and trustworthy FM-enabled ADSs.

The main contributions of this paper are summarized as follows:

- We provide a structured review of integrating FMs into autonomous driving, structured along three critical dimensions: the underlying infrastructure, the in-vehicle system, and practical real-world applications.
- For each dimension, we conduct an in-depth analysis to identify key challenges and highlight promising research opportunities, to actively pinpoint critical roadblocks and avenues for innovation.
- Based on our analysis, we formulate a forward-looking research roadmap with concrete short, mid, and long-term goals, designed to guide researchers and practitioners in prioritizing their efforts to build the next generation of safe, reliable, and trustworthy autonomous systems.

2 Background, Review Methodology, and Related Work

In this section, we first provide background on the current state of adopting foundation models in autonomous driving systems to orient the reader. We then describe the systematic review methodology used to select and synthesize the literature. Finally, we position our roadmap with respect to existing surveys and roadmaps, highlighting its scope and distinguishing characteristics.

2.1 Foundation Model in Autonomous Driving Systems

Existing integration of FMs into autonomous driving systems can be roughly categorized into *perception and scene understanding*, *navigation and planning*, *decision-making and control*, and *end-to-end autonomous driving*. Figure 2 shows a typical pipeline of leveraging FMs to enhance autonomous driving. In this section, we briefly discuss representative techniques and identify the challenges and opportunities. For more background and technique details, we refer readers to prior works [41, 64, 216, 245].

Table 1. End-to-End Dependency Analysis: Tracing Infrastructure Decisions to Real-World Practice Failures.

Infrastructure Decision *(Root Cause)*	**Vehicle Risk** *(System Effect)*	**Practice Failure** *(Final Consequence)*	**Ref.**
Biased Dataset Selection: Limited diversity in demographic attributes (e.g., age, sex, skin tone).	**Perception Blind Spots**: Inaccurate detection or classification of underrepresented groups; higher "Miss Rates" (MR).	**Systemic Safety Risks**: Higher accident rates for specific populations; ethical violations in real-world deployment.	Sec. 3.1 (Chal. I)
Unscrubbed PII in Training: Inclusion of sensitive data (faces, license plates) and lack of "Machine Unlearning" protocols.	**Data Memorization**: The FM inadvertently encodes and retains specific personal information within its parameters.	**Privacy Leaks**: Model regenerates private data during inference, causing GDPR violations or identity leaks.	Sec. 3.1 (Chal. I)
Data Collection Gaps: Absence of safety-critical edge cases (e.g., accident aftermath, extreme weather) in training data.	**Reasoning Defects**: Model inability to interpret or react appropriately to rare, high-risk scenarios (Long-tail distribution).	**Safety Validation Failure**: Catastrophic failure when the vehicle encounters unseen dangerous situations.	Sec. 3.1 (Chal. II)
Lack of Hardware Root-of-Trust: Deployment on commodity accelerators without TEEs.	**Execution Exposure**: Model weights, intermediate activations, and sensor streams are visible to the host OS.	**Model Theft & Spoofing**: Theft of proprietary model or injection of fake sensor data into the loop.	Sec. 3.2 (Chal. I)
Aggressive Model Optimization: High levels of pruning or quantization to meet on-board efficiency constraints.	**Security Vulnerabilities**: Lowered robustness thresholds; increased susceptibility to tailored hardware-level or bit-flip attacks.	**System Security Breach**: Unauthorized model weight manipulation or parameter leakage leading to loss of control.	Sec. 3.3 (Chal. II)
Untested FM Integration: Using FMs without robust multi-modal grounding or verification mechanisms.	**Hallucination**: Generation of "ghost objects," misinterpretation of traffic signs, or non-sensical path planning.	**Erratic Control**: "Phantom braking," sudden swerving, or dangerous maneuvers causing collisions.	Sec. 4 (Chal. I)
Weak Safety Alignment: Insufficient "Red Teaming" or guardrails during the instruction-tuning phase of LLM/VLM.	**Jailbreak Susceptibility**: System processes adversarial prompts (e.g., "Drive aggressively") without rejecting them.	**Reckless Driving**: Vehicle violates traffic rules or safety envelopes to satisfy user commands.	Sec. 4.1 (Chal. & Opp.)
Unverified Code Generation: Utilizing LLMs-generated code without careful inspection.	**Subtle Logic Defects**: Code is syntactically correct but contains semantic flaws or lacks context-aware safety logic.	**Runtime Software Failure**: Unexpected system crashes or unsafe execution paths during complex interactions.	Sec. 5 (Chal. II)

2.1.1 Perception and Scene Understanding. FMs enhance perception in autonomous driving by enabling context-aware environmental understanding [245]. VLMs such as LLaVA [119] and GPT-4V [147] support tasks like 3D open-vocabulary object detection [140, 150], language-guided retrieval [166, 209], and visual question answering (VQA) [34, 123, 142, 158]. Examples include OpenScene [155] for zero-shot 3D semantic segmentation and

NuScenes-QA [158] for VQA benchmarks. Additionally, DriveVLM [192] employs chain-of-thought reasoning for scene analysis, while DriveDreamer [205] predicts future states for proactive responses.

2.1.2 Decision-Making and Control. FMs improve decision-making and control by translating scene understanding into safe actions. LLMs in Drive as You Speak [39] and LanguageMPC [171] process complex data for real-time decisions. Hybrid systems like BEVGPT [202] and Driving with LLM [28] combine reasoning with traditional controls, while SurrealDriver [90] and Drive Like a Human [62] enhance robustness through safety and memory modules.

2.1.3 Navigation and Planning. FMs integrate natural language into navigation and planning by converting textual instructions into spatial representations. Systems like Talk to the Vehicle [180] and Ground then Navigate [82] generate waypoints and trajectories from multi-modal inputs. ALT-Pilot [145] enhances planning with language-augmented maps using CLIP [159], while GPT-Driver [132] and DriveVLM [192] support predictive planning and reasoning.

2.1.4 End-to-End Autonomous Driving. Recent advancements in FMs have enabled the development of unified models that integrate perception, reasoning, and control into a single differentiable framework. DriveGPT4 [226] processes sensor inputs and queries for control signals and explanations. ADAPT [88] maps video to actions and narratives, DriveMLM [44] integrates LLMs into closed-loop systems, and VLP [149] promotes generalization with context-aware frameworks.

2.2 Foundation Models for Autonomous Driving System Development

Besides the direct integration into autonomous driving systems, FMs have also been adopted in the development (i.e., testing) of autonomous driving systems.

2.2.1 Critical Scenario Generation and Comprehension. Due to its extraordinary capability in understanding the diverse driving environment and generating codes, FMs have been widely used in generating critical scenarios to test the autonomous driving systems [156, 178]. For instance, Tang *et al.* [188] propose a top-down fashion approach to generate diverse critical scenarios, which utilizes two LLMs to transform functional scenarios to formal scenarios and then searches with the logical scenarios for critical scenarios. Zhang *et al.* [241] propose ChatScene, a LLM-based agent that can generate domain-specific code from text instructions/descriptions, which can then be used to construct the desired scenario in simulators. While significant research effort has been made, existing approaches may still have problems in areas like long-tail scenario generation or multi-modal scenario fusion.

2.2.2 FM-Based Code Generation. As LLMs have been widely adopted in the daily code writing process, it is inevitable that autonomous driving developers use LLMs to generate code for production use. However, as pointed out by [87, 120], LLMs can generate erroneous code, or even worse, code that can pass the unit test but contain potential vulnerabilities. Nouri *et al.* [143] propose a prototype for automatic code generation and assessment using a designed pipeline of an LLM-based agent, simulation model, and a rule-based feedback generator. This pipeline can automatically evaluate the LLM-generated code and generate an assessment report as feedback to the LLM for modification or bug-fixing, which is a valuable attempt. But considering the ADS is a safety-critical software system, significant research effort still needs to be made.

2.3 Review Methodology

We conducted a structured literature review to identify representative studies on adopting FM in ADSs and to support the synthesis and roadmap proposed in this paper. Our objective is to provide a transparent and

reproducible account of how the literature was collected and organized, rather than to claim exhaustive coverage of all publications in this rapidly evolving area.

2.3.1 Scope and time span. We focus on the period from January 2020 to January 2026, which covers the emergence and rapid adoption of large-scale pre-trained (LLMs, VLMs, and world models) and their application to AD-related tasks. We considered works that address autonomous driving tasks, FM infrastructure, in-vehicle integration, real-world deployment, and engineering practices.

2.3.2 Database Search. We identified candidate papers through scholarly database search. In particular, we select *DBLP*[1] as our database, which is a popular bibliography database containing a comprehensive list of research venues in computer science. Furthermore, this initial search targets the titles of the papers, as the title often conveys the theme of a paper [187]. The search string is optimized in an iterative manner to cover as many related papers as possible. The final search string is shown as follows:

((“foundation model” OR “large language model” OR “vision language model” OR “large vision language model” OR “world model” OR “multimodal model” OR “FM” OR “LLM” OR “VLM” OR “MLLM”)
AND
(“automated vehicle” OR “automated driving” OR “autonomous car” OR “autonomous vehicle” OR “autonomous driving” OR “self-driving” OR “driver assistance system” OR “intelligent vehicle” OR “intelligent agent”))

The first group of terms (above “AND”) represents the identified synonyms of foundation models, containing the terms such as “large language model” and “foundation model”; The second group of terms (below “AND”) contains the synonyms of autonomous driving, containing the terms such as “autonomous driving” and “self-driving”. The terms in each group are connected with *OR* operator, while the two groups are connected using *AND* operators, which ensures that the results should contain the characteristics of both groups. Additionally, we observed that certain topics (e.g., hallucination, foundation model alignment) have a limited number of publications related to autonomous driving. For instance, the search string (“autonomous driving” AND “hallucination”) only returns two results. Thus, for specific topics related to FMs (e.g., “hallucination”), we substitute the second group in the search string with the corresponding *“<topic>”* to gain a more comprehensive understanding.

2.3.3 Abstract Analysis. To determine whether a paper is relevant to the scope of this paper, we manually screened the title and abstract of each candidate paper according to the inclusion (**IC#**) and exclusion (**EC#**) criteria below. When the relevance was unclear from the abstract, we further inspected the full text.
Inclusion Criteria:

- **IC1:** propose or evaluate FMs for autonomous driving tasks;
- **IC2:** introduce datasets, benchmarks, toolchains, or evaluation protocols that are used to study FM-based ADS capabilities;
- **IC3:** discuss deployment- and system-oriented issues for FM-based ADSs, such as latency, safety assurance.
- **IC4:** papers published between January 2020 and January 2026.

Exclusion Criteria:

- **EC1:** preprint papers or non-peer-reviewed papers (except for technical reports from leading companies or dataset/benchmarks);
- **EC2:** papers that are not related to ADS (e.g., general robotics, unmanned aerial vehicle);
- **EC3:** papers provide insufficient technical details or early results.

[1]https://dblp.org/

Specifically, we retain a small number of non-peer-reviewed arXiv papers when they serve as primary technical reports from leading industrial labs or companies and document influential FM systems, datasets, benchmarks, or deployment practices that are not yet available in peer-reviewed venues; we do not exclude survey or roadmap papers, which we instead record and compare in Section 2.4; for **EC2**, papers related to other intelligent systems, such as unmanned aerial vehicles, are excluded, since we focus on autonomous driving systems.

2.3.4 Limitations. Given the fast pace of FM research, the literature is continuously expanding. While we aimed for broad coverage, we do not claim exhaustiveness. Our focus is on representative and influential works that enable a structured synthesis and support actionable research directions for the community.

2.4 Comparison with Related Surveys and Roadmaps

We would like to clarify the difference between our roadmap and previous related surveys (or roadmaps) [178], [41], [227], [64], [245], [222], [206], [116], [217], [251], [233], [68], [196] [216] [66], [46], [169], [84], [60], [191], [52]. Many prior works primarily focus on discussing how ADS can be empowered by FMs, and how FMs may reshape the autonomous driving. Among these works, Zhou *et al.* [245] provide a comprehensive review of advances in VLMs for empowering autonomous driving tasks and summarize several challenges, and Cui *et al.* [41] presented a survey of the application of MLLMs in autonomous driving tasks. Li *et al.* [116] and Zhu *et al.* [251] both survey LLM-enabled autonomous driving, organizing applications across the autonomous driving tasks and highlighting key challenges for safe deployment (e.g., real-time constraints, hallucination), and Yang *et al.* [233] summarize the research landscape regarding the application of LLMs and VLMs in autonomous driving. Xie *et al.* [222] conduct an empirical study evaluating the readiness of VLMs for autonomous driving, highlighting their strengths in semantic scenario understanding while identifying critical weaknesses in spatial reasoning and long-horizon planning, while Dong *et al.* [52] survey end-to-end autonomous driving, tracing its evolution from classic imitation learning (IL)/reinforcement learning (RL) to FM-empowered approaches and summarizing key challenges and future directions. Gao *et al.* [64] survey foundation models for autonomous driving, proposing a modality-and-function taxonomy (LLMs, vision FMs, and multimodal FMs) and summarizing their roles in planning, perception, prediction, and simulation alongside key deployment limitations (e.g., hallucination and efficiency), Jiang *et al.* [84] survey vision-language-action (VLA) models for autonomous driving, unifying architectures, representative systems, datasets/benchmarks, and evaluation protocols, and outlining open challenges such as robustness, real-time efficiency, and formal verification, and Tian *et al.* [191] survey LLM/VLM integration for autonomous vehicles across modular and end-to-end pipelines, data generation, and evaluation resources, highlighting practical deployment issues such as real-time efficiency and ethical/regulatory concerns. Wu *et al.* [216] provide a comprehensive synthesis of the roles various FMs play in advancing autonomous driving safety by enhancing various autonomous driving tasks, and facilitating data augmentation to address long-tail distribution challenges, and Yan *et al.* [227] provide a systematic survey of over 250 papers to categorize the evolution of vision foundation models, detailing advancements in data curation, pre-training strategies, and downstream adaptation for autonomous vehicles. Wu *et al.* [217] investigate the integration of LLMs in multi-agent ADS, and Dai *et al.* [46] study the use of FMs for trajectory prediction in autonomous driving, while Sathyam *et al.* [169] examine their use for perception. Furthermore, Guan *et al.* [68], Tu *et al.* [196], and Feng *et al.* [60] conduct comprehensive reviews of the role of world models in autonomous driving tasks. Wang *et al.* [206] delivered a comprehensive synthesis of generative AI's role in autonomous driving, mapping foundational architectures like LLMs to frontier applications in multimodal data generation, simulation, and reasoning, while evaluating the technical and ethical challenges, Song *et al.* [178] surveys the application of generative AI in testing autonomous driving systems, and Gao *et al.* [66] survey how FMs enable driving scenario generation and scenario analysis for scenario-based testing, providing a unified taxonomy of methods, datasets/simulators, metrics, and open challenges.

While some challenges discussed in this paper have already been proposed in the surveys listed above, for instance, hallucination by [116, 251] and real-time constraints by [116, 192], they generally only point out these challenges without in-depth investigation. Although Wu *et al.* [216] have conducted exploratory experiments using quantization techniques to showcase the possibility of deploying LLM-based ADS in practice, their experiments remain illustrative and do not consider the potential impact of quantizing LLMs on ADS tasks. In contrast, our work covers broader topics across the three dimensions of integrating FMs into autonomous driving systems and provides in-depth discussions. In particular, we approach these challenges through the lens of software engineering. For each challenge, we first discuss its current status and potential impact, then we summarize existing SE approaches to mitigate it and their limitations, and finally point out the opportunities.

2.5 Overview of the Research Landscape

To provide a structured overview of the research landscape, this section summarizes representative works identified through our systematic review, including their tasks, evaluation metrics, and key limitations. The tables also include representative background artifacts (e.g., widely used frameworks, industrial products) that are outside the primary DBLP review corpus but are included to situate the infrastructure and deployment landscape. We categorize the landscape into two primary dimensions: the underlying **Infrastructure** (Table 2), which covers the data, deployment toolchains, and hardware roots of trust required to support large-scale models, and the **In-Vehicle Application** (Table 3), which details how FMs are integrated into core ADS modules such as perception, planning, and control.

Table 2 highlights a critical shift in the field; while classical ADS research focused on modular performance, the integration of FMs introduces complex software engineering challenges such as GPU memory saturation during training and the need for TEE-based isolation for high-bandwidth sensor data. Simultaneously, Table 3 illustrates the transition toward end-to-end multi-modal models (e.g., DriveVLM [192] and DriveGPT4 [226]), which aim to bypass brittle perception-to-planning cascades but face new hurdles in real-time latency and ”hallucination” precision.

While the works presented in these tables are selected for their representative value in illustrating these SE challenges, they are by no means exhaustive. For readers seeking a more exhaustive list of foundation models and their specific architectural configurations in the broader AI domain, we recommend referring to recent comprehensive surveys such as those by Tian *et al.* [191] and Zhu *et al.* [251], which provide extensive taxonomies of individual model variants across the AD ecosystem.

3 FM4AD Infrastructure

The infrastructure is the cornerstone for integrating FMs into autonomous driving, encompassing the datasets, computational resources, and toolchains necessary for training, testing, and deployment of FMs.

3.1 High-quality Dataset for Autonomous Driving

High-quality datasets have played a critical role in advancing autonomous driving technology. Traditional datasets primarily focused on 2D annotations like bounding boxes and masks for RGB camera images [151]. With the emergence of FMs, datasets are evolving towards multi-modal integration, particularly incorporating language descriptions [245]. While this multi-modal approach promises to accelerate autonomous driving development, it introduces new challenges. Moreover, the massive data requirements for training FMs raise significant concerns about privacy protection and ethical/legal compliance [125]. Thus, we identify the following challenges:

Table 2. Infrastructure Landscape for Integrating Foundation Models into Autonomous Driving Systems

Category	Task	Representative Work (Product)	Key Metrics	Limitations
Dataset	Augmentation	Automated annotation approaches [12, 30], critical scenario generation [48, 200, 241], scenario reconstruction [188, 211], scenario transformation [17]	Caption accuracy, diversity, reconstruction consistency, multi-modal alignment	Sim2Real gap; hallucination in VLM annotated datasets; multi-modality alignment; diversity;
	Dataset Debiasing	Fairness testing [115] and evaluation [94, 95, 168], mitigation [10, 14, 92, 124]	Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD), Intersection over Union (IoU) Disparity, Detection Miss Rate (MR) (for pedestrian detection) [33, 115]	Primarily focus on perception module (e.g., pedestrian detection), limited investigation about its impact to the decision; integrating FMs may introduce inherited bias which needs further inspection.
	Privacy Preserving	Privacy-preserving machine learning [238, 247], federated learning [220, 229]	accuracy, communication overhead, client selection rate, data heterogeneity, differential privacy budgets	High communication overhead for FM models; FMs tend to memorize training data and can even infer personal attributes from anonymized data [181].
Development & Deployment	Development Framework	PyTorch [153], TensorFlow [5], JAX [20], PaddlePaddle [18]	Training time, accuracy, inference time, CPU/GPU utilization, memory usage, community support	Only offer basic distributed training functions, and exhibit critical limitations when scaling to FM-level workloads (e.g., GPU memory saturation, communication bottlenecks) [121, 237].
	Distributed Training	Ray [137], DeepSpeed [163], Megatron-LM [174], Colossal-AI [114]	Scaling efficiency, training throughput, memory efficiency, stability	High onboarding efforts for developers; limited research on bug comprehension [131, 237] and testing techniques [199].
	Efficient Deployment	Model pruning [79, 111], model distillation [70, 173, 244], and quantization [117, 197]	Inference latency, memory footprint, accuracy retention	Performance drop; limited investigation on real-world testing and its impact on robustness [15].
Computational Resources	Hardware	AVA-3510 (based on NVIDIA Quadro RTX 5000) [9], ADM-AL30 (based on NVIDIA RTX 4000 SFF Ada) [8], NVIDIA DRIVE AGX Thor [144], Tesla FSD Hardware (custom-designed SoC) [185]	TFLOPS, Memory Bandwidth, GPU Memory Size	Black-box proprietary designs are hard to verify; only a few products (e.g., AGX Thor) conform to safety standards like ISO 26262 ASIL-D [80].
	TEE for AD	CVShield [75], Mimer Trust [23]	Attestation speed, Enclave size, overhead	Current approaches are primarily limited to low-bandwidth sensor channels like GPS; limited evaluation for high-bandwidth sensors (LiDAR/Camera); Limited enclave size for FM-level model;

Challenge I: Dataset Cleaning and Curation. Dataset cleaning and curation serve as a critical foundation for developing FMs, ensuring data integrity, privacy protection, and efficient training. Key challenges include protecting privacy [24, 240, 248], and mitigating bias in training datasets [72, 115, 124]. The privacy challenge involves both preventing personal data from appearing in training datasets and ensuring it cannot be inferred from model outputs [99, 248]. Bias in training data can lead models to perpetuate systemic inequities and may introduce safety risks for unrepresented groups when deployed in real-world scenarios [124]. These challenges present several key opportunities for further research.

- **Opportunity: Bias Mitigation.** Dataset bias poses a critical challenge for autonomous driving. Recent studies [124] have revealed limited diversity in key demographic attributes (i.e., age, sex, and skin tone) within existing AD datasets. This lack of representation could lead to safety risks when deploying FMs trained on such datasets, particularly for underrepresented groups. While recent research has made progress in addressing these concerns [32, 72, 73, 115], these previous works mainly focus on the perception module (i.e., pedestrian detection), using fairness metrics such as Statistical Parity Differences (SPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD), and Miss Rate (MR). One open software engineering challenge lies in verifying how such biases propagate through the system pipeline. Future research could develop stratified, scenario-based testing suites where only demographic attributes vary, and evaluate whether downstream behaviors (e.g., safety margin, rule violations, intervention rate) exhibit systematic disparities, enabling fairness regression testing across releases. In addition, LLMs are known to exhibit bias and fairness issues [63]; integrating FMs into ADS reasoning/planning may introduce new bias pathways beyond perception, which calls for dedicated test suites and systematic investigation.
- **Opportunity: Privacy Preservation.** Another critical research opportunity lies in developing effective privacy-preserving algorithms for FMs in autonomous driving. According to Staab *et al.* [181], LLMs can even infer personal attributes from real-world data, even when the text is anonymized using commercial tools. Wang *et al.* [204] highlight that malicious attackers can infer and trace vehicle trajectories in accident warning systems, raising serious privacy concerns and introducing potential security risks. Established techniques such as differential privacy [38, 54, 190, 195], data cleaning [24, 91], and federated learning [182, 224, 240], have advanced the field, yet they consistently face challenges in balancing privacy preservation with data utility. The challenge is particularly acute with FMs, which tend to memorize training data extensively, potentially leading to privacy leakage even with the data used in fine-tuning processes [248]. There is an urgent need for novel techniques that can ensure robust privacy guarantees while maintaining the comprehensive nature of training datasets required for large FMs.
- **Opportunity: Machine Unlearning for FMs.** While curating training data can mitigate issues upfront, FMs already deployed may contain private information or learned biases. For these models, machine unlearning emerges as a critical capability [223]. The goal is to efficiently remove the influence of specific data points or concepts from a trained model without the need for a complete, costly retraining from scratch. This is essential for complying with data privacy regulations like the ”right to be forgotten” and for rectifying harmful biases discovered post-deployment [223]. For coding tasks, machine unlearning has proven its effectiveness [37]. The challenge of this opportunity lies in developing unlearning techniques that are not only effective at removing information but are also computationally efficient and maintain the model's overall performance on other tasks, which is particularly difficult for FMs in autonomous driving, since these models inherently need to deal with multiple tasks.

Challenge II: Augmenting Autonomous Driving Datasets. Despite significant investments in the development of autonomous driving datasets, current limitations in data quality and scale hinder their ability to comprehensively address the field's challenges [41]. Moreover, certain critical scenarios remain difficult or nearly impossible to capture in real-world data collection [179, 211]. These include high-risk situations such as accident

Table 3. FM in Vehicle Landscape

Task	Paper	Model	Dataset/Sim	Key Metrics	Limitations
Perception and Understanding	Semantic Anomaly Detection with LLMs [57]	GPT-3.5 [146]	CARLA [53]	TPR, FPR	Strong coupling with upstream perception; hallucination; poor spatial precision; temporal inconsistency;
	Zelda [166]	VIVA [165]	BDD-X [97]	MAP, APS	
	Talk2BEV [34]	BLIP-2 [110], MiniGPT-4 [249], Instruct-BLIP [45]	Talk2BEV-Bench [34]	MCQ Acc., IoU, Distance error	
Decision Making and Control	LanguageMPC [171]	GPT-3.5	IdSim [85]	Failure and collision cases, inefficiency, time efficiency, safety penalty	High on-device latency; struggles with long-horizon goals; high sensitivity to minor prompt disruptions; potential inheritance of bias from VLM/LLMs [251];
	Driving with LLMs [28]	LLaMA-7b [194]	Custom 2D Simulator	Mean Absolute Error (MAE) for perception and action; traffic light detection accuracy; GPT/Human Grading for QA	
	Drive Like a Human [62]	GPT-3.5 (with LLaMA-Adapter)	HighwayEnv [107]	Zero-shot pass rate, reasoning accuracy, and decision consistency	
Navigation and Planning	ALT-Pilot [145]	GPT-4, CLIP	CARLA + field test	Absolute Position Error (APE), Recall@K, Distance to Closest Landmark Region (DCLR), Distance to Converge, and Goal Reachability	Risk of generating unaligned/dangerous planning/routines; poor multi-turn interaction and instruction following;
	GPT-Driver [132]	GPT-3.5	nuScenes [22]	Avg. L2 Error (m) and Avg. Collision Rate	
	DriveVLM [192]	Qwen-VL [16]	SUP-AD [192], nuScenes	Displacement Error (DE), Collision Rate (CR), Scene Description Score, and Meta-action Score	
End-to-End Autonomous Driving	DriveGPT4 [226]	GPT-4	BDD-X	CIDEr, BLEU4, ROUGE-L, ChatGPT Score, RMSE, and Threshold Accuracies (A_τ)	High **on-device latency**; risk of executing unaligned/dangerous user commands; transparency and interpretability;
	DriveMLM [44]	LLaMA2-7B, EVA-CLIP	CARLA	Accuracy and F1-measure for decision prediction; BLEU-4, CIDEr, and METEOR for decision explanation; L2 distance, collision rate, and intersection violation rate for trajectory prediction	
	VLP [149]	CLIP [159]	nuScenes	Avg. L2 error, and Avg. collision rate	

aftermath and pedestrian-involved incidents. However, comprehensive testing of autonomous vehicles against these scenarios is crucial for safety validation. To overcome these challenges, researchers should increasingly explore methods for generating customized driving scenarios or automated data collection, enabling the effective simulation of these critical cases to augment existing datasets and enhance their utility.

- **Opportunity: Customizable Driving Scenario Generation.** Current autonomous driving systems are primarily trained and evaluated on datasets collected from daily driving scenarios or synthetic data [49, 211]. However, these datasets generally lack safety-critical scenarios that are crucial for robust system evaluation. Research in driving scenario generation has progressed along multiple directions, including data-driven approaches [101, 157, 186], adversarial generation methods [7, 48, 134, 200], and knowledge-based techniques [188, 211, 241]. Looking ahead, scenario generation algorithms need to address key challenges, such as maintaining consistency across multiple sensor modalities (e.g., LiDAR, camera images) and enhancing scenario complexity through interaction and collaboration between agents.

Challenge III: Dataset Licensing and Management. Dataset licensing and management pose a variety of challenges vital to ensuring the legal and ethical use of autonomous driving datasets. The massive amount of data required for training FMs heightens the risks of copyright breaches, licensing violations, and subsequent legal liabilities. Additionally, the terms of use for datasets released by leading autonomous driving companies vary widely, further complicating this task. The multimodal nature and diverse sources of autonomous driving datasets intensify these difficulties. Moreover, selecting/sampling the right training data is essential for producing capable FMs [21, 35, 212]. Recent studies [98, 214] have revealed the complex landscape of modern large dataset copyright and licensing, emphasizing the need for deeper exploration and development of innovative techniques. These challenges also open up opportunities for further research.

- **Opportunity: Dataset License Compliance.** The primary challenge of license management lies in the complexity and variety of licenses governing autonomous driving datasets [81, 225]. Unlike traditional datasets for LLMs, which primarily consist of publicly available data (crawled from the Internet) supplemented with proprietary datasets having usage restrictions, most autonomous driving datasets are released by leading autonomous driving companies with their own specific terms of use, necessitating careful review and understanding to ensure compliance [127, 198, 214]. As pointed out by Kim *et al.* [98], the scale of modern datasets renders manual compliance verification impractical, thereby requiring automated detection techniques. Promising research directions include the development of automated detection and audit systems for legal terms of use, providing developers with clear insights into the permissions and restrictions associated with each dataset.
- **Opportunity: Data Management Framework for FMs.** As FMs demonstrate performance improvements through data scaling and the significance of data becomes evident [35], effective data management becomes increasingly critical. While various tools and methods have been proposed to explore how to properly manage the training data, encompassing data deduplication [105], training data selection [113, 160, 201], sampling high-quality data [154, 212], and dataset license compliance [81, 98], there is still a lack of a unified framework and criteria. Although there have been some initial attempts in this area [148, 154, 207], systematic approaches to data management for FMs remain in their early stages. Given the massive scale and diverse sources of data required for training FMs, developing a comprehensive data management framework has become an urgent priority.

3.2 Computational Resources

Due to the computational-intensive nature of FMs, computational resources, including graphics processing units (GPUs), tensor processing units (TPUs), and other specialized AI accelerators, form the very foundation of the FM infrastructure. Building upon this hardware layer, distributed training frameworks and cloud computing enable

efficient resource utilization and management. However, the complexity of distributed, computation-intensive training, and reliable efficient deployment introduces unique challenges and opportunities in adopting FMs in autonomous driving. In this section, we mainly discuss the challenges and opportunities related to hardware, issues with the software layer (e.g., distributed training framework) are discussed in Section 3.3.

Challenge I: Trustworthy Hardware Design. The computationally intensive nature of FMs necessitates a reliance on proprietary and specialized hardware, such as GPUs and other accelerators. This dependency creates significant challenges in ensuring security across the hardware stack, as the proprietary design of chips and firmware often results in a "black-box" environment where vulnerabilities can remain undetected [203, 230]. Furthermore, the parallel processing architectures and shared resources inherent to modern accelerators make them particularly susceptible to hardware-level attacks [86]. These attacks can lead to severe consequences, including the leakage of sensitive information like model parameters via side-channels [139] or even enabling arbitrary code execution [104]. Given the paramount importance of security in autonomous driving, these hardware vulnerabilities present a fundamental risk that necessitates urgent research and development. Beyond the general exposure of proprietary accelerators, FM-centric ADS deployments introduce additional constraints that exacerbate the hardware trust problem. First, FM inference (and especially VLM/world-model pipelines) typically requires large model footprints (weights, intermediate activations, and KV caches), which can exceed the secure memory/enclave capacity of many practical TEE designs and can trigger expensive enclave paging and frequent world switches. Second, ADS workloads ingest multi-channel, high-bandwidth sensor streams (e.g., camera and LiDAR), so naïvely confining the full perception–planning pipeline to a secure enclave can incur substantial I/O marshalling overhead and may violate real-time latency budgets. Existing TEE-for-AD prototypes are often evaluated on low-bandwidth channels (e.g., GPS) rather than on realistic high-throughput perception inputs [75], leaving the end-to-end feasibility for FM-scale workloads under-explored. Third, in today's deployment reality, widely used commodity accelerators may not expose confidential-computing capabilities suitable for isolating FM execution, while platforms that do provide stronger TEE-style protections can be substantially more costly and harder to provision at scale; this gap motivates practical designs that minimize the trusted computing base (TCB) and carefully trade off security guarantees against latency and integration overhead.

- **Opportunity: Security-by-Design Hardware Architectures.** A primary research opportunity lies in architecting hardware with intrinsic security guarantees, with a key approach being the development of Trusted Execution Environments (TEEs) for AI accelerators [26, 235]. A TEE [31] would leverage hardware to create an isolated enclave, protecting the confidentiality and integrity of an FM's parameters and its execution, even from a compromised high-level operating system. From a deployment perspective, widely used commodity accelerator stacks can provide limited support for enclave-style protection. Prior work reports that commonly used commercial GPUs in current AV/AD solutions/platforms (e.g., RTX A5000 in AVA-3510) do not provide TEE functionality, whereas confidential-computing-enabled accelerators (e.g., Hopper-class GPUs) can be substantially more costly [108]. This deployment gap motivates hybrid designs that minimize the trusted computing base (TCB) and confine only security-critical components to the trusted domain. Concretely, TEEs for FM-powered ADSs must satisfy architectural requirements such as (i) *low overhead and real-time predictability* (bounded enclave transitions and secure I/O costs), (ii) *scalability* to FM-scale working sets (or secure partitioning/streaming of model components), (iii) *accelerator-aware trust boundaries* that account for GPU/NPU driver stacks and DMA paths that often remain outside the enclave, and (iv) *multi-sensor support* for securely ingesting and attesting to high-throughput perception data. Significant research is therefore needed to design novel, low-overhead TEEs that can scale to serve FMs, establishing a verifiable hardware root of trust for critical AI computations in autonomous driving. To make the comparison

Table 4. TEE implementation categories, characteristics, and representative examples (adapted from Muñoz et al. [138]).

TEE Category	Description and Characteristics	Examples
1) Hardware-based & Privileged	**Description:** Rely on dedicated hardware isolation (e.g., ARM TrustZone) and run with high system privileges. **Characteristics:** (i) Access to broad system resources; (ii) typically uses a secure monitor; (iii) two domains: Secure World (SW) and Normal World (NW).	**Commercial:** Qualcomm QSEE; Trustonic t-base; Samsung TZ-RKP; Google Trusty. **Open/Academic:** Linaro OPTEE; Microsoft TLR; SafeG; Kinibi_M.
2) Hardware-based & Non-privileged	**Description:** Utilize hardware support (e.g., Intel SGX or AMD SEV) without granting the TEE instance total control over the system. **Characteristics:** (i) Supports multiple deployments (new instances can be added without extending the Trusted Computing Base); (ii) often referred to as *enclaves*.	**Commercial:** Intel SGX; AMD SEV; TrustICE. **Open/Academic:** Sanctum; SecureBlue; Haven; SCONE; Graphene-SGX.
3) Software-based & Privileged	**Description:** Software-enforced TEEs (without dedicated secure hardware elements) that run with high privileges to enforce isolation. **Characteristics:** (i) Often relies on hypervisor or kernel-level modifications; (ii) isolation is logical rather than physical.	**Open/Academic:** Nested Kernel; OpenTEE; MicroTEE; SoftTEE; TrustShadow; SKEE.
4) Software-based & Non-privileged	**Description:** Software-only implementations without elevated system-wide privileges. **Characteristics:** (i) Often rely on virtualization layers or application-level sandboxing; (ii) isolate specific processes/data without full system control.	**Open/Academic:** Overshadow; Virtual Ghost; InkTag; Flicker; TrustVisor; Multizone; Utango.

space explicit, as shown in Table 4, TEEs can be classified along two orthogonal axes—*software- vs. hardware-based* and *privileged vs. non-privileged* deployments—each implying different isolation granularity, attestation support, and performance costs [138]. A systematic evaluation of these design points for FM workloads (secure memory limits, sensor I/O throughput, and accelerator integration) remains an open problem and is essential for identifying feasible, automotive-grade TEE configurations.

- **Opportunity: Open and Verifiable Hardware Stacks.** To address the ”black-box” nature of proprietary hardware, there is an opportunity to develop and promote open-source, verifiable hardware designs or standards for FM acceleration. An open and transparent hardware stack would allow for community-driven security audits, reducing the risk of hidden backdoors or design flaws [203]. In addition to openness, an important opportunity is to standardize *measurement and attestation interfaces* (e.g., for firmware provenance, model binaries, and critical runtime configurations) so that safety cases can incorporate verifiable evidence about the hardware and software supply chain. Such interfaces would enable reproducible security evaluation across vendors and help bridge the gap between research prototypes and deployable, certifiable FM acceleration stacks.

Challenge II: Resource-Aware Engineering. Training and deploying FMs for autonomous driving tasks demand substantial computational resources, often necessitating large clusters of GPUs or TPUs [71, 161]. This

challenge is significantly exacerbated during in-vehicle deployment, where models must operate efficiently across diverse and resource-constrained hardware platforms. This is a fundamental infrastructure challenge for researchers and developers, opening up the following opportunities:

- **Opportunity: Distributed and Collaborative Training.** One key opportunity lies in the development of distributed and collaborative training frameworks [118, 137]. These frameworks could enable multiple smaller computing entities to pool their computational resources, meeting the demands of training large-scale FMs. This approach aims to overcome the barrier of high-cost infrastructure but also promote a more diverse and inclusive development ecosystem [161].
- **Opportunity: Resource-Aware Model Search for Foundation Models.** Given the diverse and resource-constrained hardware present in autonomous vehicles, resource-aware model search might be a potential direction for efficient deployment. This approach, often leveraging techniques such as neural architecture search (NAS) [58], aims to automatically discover specialized model architectures that optimally balance performance with certain constraints like latency, memory footprint [65].

3.3 Development and Deployment of FM4AD

Due to their formidable size and computational demands, developing and deploying FMs has posed significant new challenges for autonomous driving applications. In this section, we discuss the challenges and opportunities related to the development and deployment of FMs for autonomous driving.

Challenge I: Understanding the FM Development Toolchain. The development toolchain for Foundation Models presents unprecedented complexity compared to traditional deep learning frameworks. The enormous scale of these models significantly amplifies the intricacy of data pre-processing pipelines, distributed training systems, and model deployment workflows [203]. Compounding this issue is the rapid pace of innovation in the field, which hinders developers and researchers from maintaining a comprehensive understanding of the continuously evolving ecosystem of tools and libraries [203]. This complexity and opacity create significant opportunities to systematically analyze and improve how FMs are built and maintained.

- **Opportunity: Empirical Analysis for Toolchain Optimization.** A key research opportunity lies in the large-scale empirical analysis of the FM development toolchain. By systematically examining public repositories, development workflows, and the usage of popular libraries, researchers can identify common inefficiencies, performance bottlenecks, and resource-intensive anti-patterns [121, 141, 203]. The insights gained from such studies can lead to data-driven best practices and automated tools that help developers streamline complex processes, optimize resource consumption, and accelerate the overall development lifecycle.
- **Opportunity: Quality Assurance for the FM Development Toolchain.** While traditional deep learning frameworks benefit from a mature suite of quality assurance and testing techniques [208, 221, 242], these methods are often inadequate for the complex, distributed nature of the FM toolchain. Core components for large-scale training, such as Ray, introduce immense challenges related to network failures, asynchronous operations, and state management that conventional debuggers and testing methods cannot easily handle. Although initial research has begun to address these challenges [126, 199], these efforts are still in their early stages, leaving a critical opportunity to develop a new generation of QA tools and methodologies specifically for FM engineering. Research in this area could focus on scalable debuggers for distributed systems, fault injection frameworks to test resilience, and novel validation techniques for massive data-parallel pipelines, ultimately enhancing the reliability and reproducibility of the entire FM development process.

Challenge II: Efficient and Reliable FM Deployment in Vehicle. With growing concerns over data privacy and the strict low-latency requirements of autonomous driving, the on-board deployment of FMs in vehicles is becoming essential. A significant body of research has focused on making this feasible through model compression

techniques like pruning [130, 183], knowledge distillation [67, 96], and quantization [112, 117, 219], as well as inference optimizations such as parallel computation [175, 236] and KV cache management [103, 129]. However, while these techniques improve efficiency, they can inadvertently introduce new security and reliability vulnerabilities. For instance, researchers have demonstrated tailored attacks that specifically exploit the characteristics of quantized models [231, 243] and KV cache optimizations [177, 215]. This inherent tension between performance and security creates an urgent need for deployment strategies that are both highly efficient and fundamentally trustworthy. Furthermore, to navigate this tension and help practitioners better trade-off in practice, we provide the risk-utility guidelines (shown in Table 5).

Table 5. Risk-Utility assessment framework for FM deployment optimization techniques

Technique	Benefit	Introduced Risks	Mitigations
Quantization	Maximize speed and storage reduction; hardware compatibility; no re-training (generally);	Increases susceptibility to attacks targeting quantized models [56, 231, 243]; increases the chance of hallucination [109]	Safety patching [27]; safety-aware quantization-aware training;
Pruning	Reduction in computation and memory usage; less risk of hallucination [36]	Pruning-activated attacks [55]; privacy leakage [102, 172]	Security-aware calibration; model patching with repaired parameters [55]; iterative compensation [61]
Model distillation	Flexible student model architecture; high performance retention;	Exploitation of imperfections; teacher hacking [193] (transfer of unsafe behaviors, potential lack of generalization);	Prioritize online data generation (dynamic sampling) and high prompt diversity; limiting re-training epochs;
KV Cache Optimization	Enhances real-time responsiveness and throughput;	Potential verbatim input reconstruction and semantic exfiltration through side channel attack [215]; direct inversion/collision attack [128]	Obfuscation schemes like KV-Cloak [128]; execution isolation within TEEs to prevent leakage;

- **Opportunity: Empirical Analysis on the Impact of Optimization Techniques.** While specific vulnerabilities in optimized models have been identified [177, 215, 231, 243], there is currently a lack of large-scale empirical studies that systematically measure how these techniques impact a model's overall trustworthiness in autonomous driving tasks. This creates a vital research opportunity to conduct comprehensive analyses that quantify the effects of pruning, quantization, and other optimizations on key properties beyond performance. Such studies should evaluate the trade-offs concerning adversarial robustness, fairness, reliability on out-of-distribution data, and the model's propensity for hallucination. The findings would establish an evidence-based understanding for practitioners, leading to clear guidelines for safely applying optimization techniques in safety-critical systems like autonomous vehicles.
- **Opportunity: Robust-by-Design Optimization Techniques.** A key research opportunity lies in co-designing model optimization techniques with security and robustness as first-class objectives. Instead of

optimizing for performance alone, future research could focus on developing new pruning, distillation, or quantization algorithms that are inherently resistant to known attack vectors. For example, this could involve creating quantization schemes that provably maintain adversarial robustness or knowledge distillation processes that transfer security properties from a larger teacher model to a smaller student model. The goal is to create a new class of optimization methods where efficiency gains do not come at the cost of safety.

Challenge III: Model Maintenance and Lifecycle Management. The rapid evolution and high computational cost of FMs create a critical need for structured maintenance and lifecycle management within the ADS development process. A major engineering hurdle is the resource-intensive nature of model training; it is often computationally prohibitive to retrain large-scale models from scratch to adapt to new environmental shifts or specific task variations. Current practices frequently lack systematic frameworks for model versioning, discovery, and reuse. This results in significant redundant effort and inefficiency within the FM supply chain [203], as developers may lack the tools to identify and retrieve existing pre-trained models with similar capabilities that could be adapted through modular updates or lightweight fine-tuning rather than full retraining [164].

- **Opportunity: Model Recommendation and Reuse Frameworks.** To mitigate the maintenance overhead, there is a vital research opportunity in developing automated model recommendation engines for the FM-based ADS ecosystem. These frameworks could utilize metadata and gradient-based fingerprinting techniques, such as TENSORGUARD [218], to help developers discover similar FMs within a curated repository that best match a target domain's requirements. By treating FMs as software artifacts requiring provenance tracking, researchers can enable similarity detection and family classification independently of training data or specific model formats [218].

Challenge IV: Edge/Cloud Collaboration for FM Services. Deploying FMs in autonomous vehicles presents a fundamental trade-off between on-board (edge) and remote (cloud) computations. While optimization techniques like quantization can reduce the burden on edge devices, their limited computing power inherently caps model capability and complexity. Conversely, cloud infrastructure offers vast computational resources for complex reasoning but cannot meet the strict low-latency, high-reliability, and data privacy requirements essential for safety-critical driving decisions. This gap, where neither edge nor cloud alone provides a complete solution, creates a significant opportunity to design hybrid systems that strategically leverage the strengths of both environments.

- **Opportunity: Intelligent Task Orchestration.** A primary opportunity lies in creating intelligent frameworks that dynamically schedule and offload tasks across the edge-cloud continuum [69, 232]. Research in this area focuses on developing algorithms that can partition workloads in real-time: latency-critical functions like immediate hazard detection would remain on the edge, while computationally intensive, non-real-time tasks like complex scene interpretation or HD map updates could be sent to the cloud. The goal is to build an adaptive system that optimizes for performance, latency, and resource utilization based on current driving context and network conditions.
- **Opportunity: Robust and Asynchronous Communication.** The connection between a vehicle and the cloud is often intermittent and variable. A key opportunity is to design robust and asynchronous communication schemes that ensure system reliability despite unstable network connectivity [239]. This includes developing mechanisms that allow the edge model to operate autonomously when disconnected and then asynchronously sync its knowledge or receive updates from the cloud when a connection is re-established. Such schemes are vital for ensuring the vehicle remains safe and operational at all times while still benefiting from the power of the cloud.

4 Foundation Models in Vehicle

In this section, we examine how FMs enhance different modules of autonomous driving, summarizing techniques and methodological advances. Specifically, we mainly focus on how FMs can help achieve human-like driving

using LLMs, VLMs, and world model-based prediction. We also identify key challenges and research opportunities to guide future investigations in this rapidly evolving field.

Challenge I: Hallucination. While FMs (i.e., LLMs and VLMs) have achieved significant advancements in autonomous driving, hallucination remains a critical challenge for their safe real-world deployment. Hallucination refers to the generation of outputs that are factually incorrect, inconsistent, or nonsensical, a phenomenon to which FMs are particularly prone [25, 184]. Following the formal taxonomy established in recent research [25], these failures can be categorized into four core characteristics: (1) **Compliance**, where the generation violates non-negotiable hard constraints (e.g., executing an illegal maneuver); (2) **Desirability**, where the output fails to meet soft constraints measured by optimization objectives like fuel efficiency or passenger comfort; (3) **Relevancy**, where the model introduces extraneous or off-topic details that do not belong to the driving task; and (4) **Plausibility**, which assesses the syntactic soundness of the output. Crucially, a hallucination may appear highly plausible and believable to a human critic while still being non-compliant or undesirable, making it a severe risk for safe deployment. Table 6 summarizes the representative hallucination metrics across various task settings, adapted from [25]. In the context of autonomous driving, a hallucinated object detection, such as mistakenly identifying a non-existent pedestrian, could trigger a severe safety incident like an abrupt stop or a potential collision. Although substantial research has addressed hallucination in general-purpose LLMs [13, 76, 77, 176, 234, 246] and in FMs for autonomous driving [51, 59], the underlying triggers and effective detection methods remain largely unclear, creating vital opportunities for further research.

- **Opportunity: Hallucination Detection and Mitigation.** While the underlying mechanism of hallucination is currently unclear, the first opportunity lies in the detection and mitigation of hallucination. To combat hallucination, research efforts have been made by both the AI and SE communities. For instance, Yang *et al.* [228] utilise metamorphic relations to detect hallucinations in LLMs. Future research could focus on developing detection and mitigation strategies that reduce hallucinations without hurting the model's performance.
- **Opportunity: Multi-modal Grounding and Verification.** One promising direction to tackle hallucination is to leverage the inherently multi-modal nature of the perception data. Research may focus on developing methods to ground an FM's generated output (e.g., textual caption) against raw sensor information from cameras, LiDAR, or radar [25]. Developing robust grounding techniques would anchor the model's outputs to physical reality, which can significantly reduce the likelihood of factually incorrect or unverified statements.

Table 6. Summary of Hallucination Metrics across Task Settings (Adapted from [25])

Task Setting	Compliance Metrics	Desirability Metrics	Relevancy Metrics	Plausibility Metrics
Question-Answering	Accuracy, FactScore, BERTScore, Contradictions	Calibration Error, Succinctness, Prudence	Cross Encoder, Perplexity	ROUGE, Semantic Preciseness
Image Captioning	CHAIR, POPE, CLIP Score, METEOR	CIDER Human Alignment, Detailedness	CHAIR, POPE	ROUGE, SPICE, Perplexity
Planning	Feasibility, Action Probabilities, Plan/Action Accuracy	Success Rate, Clarification Rate, Unsafe Action Rate, Calibration Error, Estimated Payoff	Action Probabilities	Non-compliance Rate, Preciseness

Challenge II: Multi-modality Adaptation. While Foundation Models (FMs), particularly Large Language and Multi-modal Large Language Models, have demonstrated remarkable reasoning capabilities in autonomous

driving, their practical application is often hindered by a critical vulnerability. Most current approaches rely heavily on the processed outputs of upstream perception modules, treating them as absolute ground truth [64]. This tight coupling means the entire system is brittle; even minor perception inaccuracies, such as a slight error in an object's heading estimation, can cascade and lead to catastrophic failures in the downstream decision-making process [133]. This dependency on imperfect perception highlights a critical need for robust adaptation methods, opening up several research opportunities.

- **Opportunity: Benchmarking Robustness in Multi-Modal Fusion.** As research moves towards end-to-end models, there is a pressing need for standardized benchmarks to rigorously evaluate the quality and robustness of their sensor fusion capabilities. Current evaluations often focus on overall task performance, making it difficult to isolate how well a model handles sensor noise, failure, or conflicting information. This creates a significant opportunity to develop novel evaluation protocols and challenging datasets designed specifically to probe the fusion process. The creation of these benchmarks is critical for systematically comparing different architectures and guiding the development of more reliable and truly robust multimodal systems.
- **Opportunity: End-to-End Multi-Modal Fusion.** A promising research direction is to move beyond cascaded pipelines and develop integrated, end-to-end FMs. The goal is to create architectures that can directly ingest and fuse raw, heterogeneous sensor data (e.g., camera pixels, LiDAR point clouds, radar signals). By learning directly from raw data, the model can develop its own robust internal representations, bypassing the vulnerabilities of a brittle intermediate perception layer. This approach allows the FM to learn complex cross-modal correlations, enabling it to rely on one modality to compensate for noise or errors in another.

Challenge & Opportunity: Domain-Specific Foundation Models for Autonomous Driving. While the open-source landscape for code-centric large language models (LLMs) has thrived with examples like Magicoder [210] and CodeLlama [167] setting benchmarks, most FMs for autonomous driving remain proprietary, such as GAIA-1 [74]. This restricts academic and independent researchers from advancing innovation in a field where safety and robustness are critical. The core problem is the absence of a pre-trained, powerful foundation model designed for the specific tasks of autonomous driving, such as integrating multi-modal data (e.g., cameras, LiDAR) and handling complex decision-making in dynamic environments. An open-source, domain-specific foundation model is urgently needed to bridge this gap. By providing a robust starting point for tasks like perception, planning, and control, this model would empower researchers to address real-world driving challenges efficiently.

4.1 FM-Enabled Intelligent User Experience

Beyond enhancing core autonomous driving capabilities, FMs are revolutionizing user interaction and experience in autonomous vehicles. This subsection explores two key aspects: intelligent user interfaces with personalization, and enhanced surrounding awareness capabilities.

Challenge & Opportunity: Intelligent User Interface and Personalization. While FMs enable more intelligent and personalized user experiences in autonomous vehicles, several challenges need to be addressed. MLLMs like GPT-4V can interpret natural language instructions to control vehicles according to user preferences. For example, Cui *et al.* demonstrated that LLM-based planners can respond to personalized commands such as "drive aggressively," adjusting vehicle behavior across different speeds and risk levels [40]. However, this flexibility raises significant safety concerns. As shown in [42], LLMs may interpret and execute potentially dangerous commands like "drive as fast as you can." Although research has explored methods to ensure compliance with traffic rules and safety requirements [42, 232], the vulnerability to jailbreak attacks remains a concern, particularly given the proliferation of LLM exploitation techniques [89, 162, 250, 252]. Additionally, balancing real-time responsiveness with user privacy presents another significant challenge, as discussed in Section 3.3.

Challenge & Opportunity: FM-Enabled Surrounding Awareness. FMs could enhance surrounding awareness by providing users with real-time, interpretable insights about the vehicle's environment. For instance,

DriveGPT4 [226] integrates this awareness into the driving loop, offering passengers explanations for vehicle actions (e.g., "veers left to avoid collision"). This awareness extends to both safety and convenience features, such as alerting users to nearby hazards or points of interest, enhancing the overall experience [50, 135]. However, ensuring the accuracy and reliability of FM-generated insights remains challenging, as hallucinations or misinterpretations could mislead users. Additionally, presenting complex information requires careful UI design to maintain user-friendliness. Potential opportunities include developing robust multi-modal grounding techniques to reduce errors through cross-validation of visual and textual data, and creating intuitive visualization methods such as augmented reality overlays to effectively convey FM insights. These advancements could transform vehicles into intelligent companions that enhance both safety and user engagement.

5 Foundation Model Application in Practice

This section explores the practical deployment of FMs in autonomous driving. We distinguish between modular integration and full adoption of FMs, showcasing their role in enhancing vehicle capabilities. Besides, we also articulate their application in automating the development of autonomous driving systems, along with the potential challenges and opportunities.

Several initiatives employ FMs as specialized components within autonomous driving systems. Xiaomi SU7 integrates a VLM via OTA update to enhance scene interpretation and safety alerts [4]. Li Auto combines a VLM with an end-to-end framework in its OTA-updated smart driving system, improving scene recognition and maneuver accuracy [2]. TIER IV utilizes an LLM to enable vehicles to reason and communicate, enhancing human-vehicle interaction [3]. Similarly, Bosch researchers apply natural language processing to predict traffic behaviors, boosting situational awareness [93]. These cases demonstrate FMs augmenting specific functions like perception and communication.

Meanwhile, full adoption leverages FMs as the core of autonomous driving systems. Cui et al. deploy an LLM in Talk2Drive for end-to-end control, personalizing driving through language and vision inputs [42]. Their subsequent work fully integrates a VLM onboard for motion control, unifying perception and decision-making [43]. Waymo's MotionLM uses FMs to transform multi-agent motion prediction into a language task, streamlining dynamic interactions [170]. These efforts highlight FMs driving comprehensive, adaptive autonomy.

Challenge I: Foundation Model Alignment. As FMs become increasingly integrated into autonomous driving systems, their potential societal risks demand careful consideration. The undesired behaviors exhibited by FMs, such as hallucination, raise particular concerns in safety-critical domains like autonomous driving where they directly impact public safety. AI alignment has emerged as a potential solution, aiming to ensure AI systems behave in accordance with human intentions and values [106]. Despite its critical importance for the safe deployment of FMs in autonomous driving, research in this area remains limited [6, 78, 100, 213]. The complexity of foundation model systems, encompassing fairness, privacy, and security concerns, urgently calls for more attention and investigation into alignment strategies. Current alignment research can basically be divided into two key components: forward alignment and backward alignment [83]. Below are potential opportunities:

- **Opportunity: Enhancing Feedback Mechanisms (Forward Alignment).** Forward alignment, which focuses on proactively shaping model behavior during training, presents a significant opportunity for improving FMs in autonomous driving. By incorporating human-value feedback during the training process, developers can construct more robust systems where FMs not only continuously learn but also maintain alignment with human intentions and safety requirements [203].
- **Opportunity: Safety Benchmarks and Evaluation for Assurance (Backward Alignment).** Datasets and benchmarks are crucial for safety evaluation, serving as fundamental tools for ensuring AI alignment. A key opportunity lies in developing comprehensive metrics and benchmarks for FMs to better evaluate their safety performance and ability to minimize accidents during task execution [83]. Unlike traditional deep

learning models, FMs can leverage general knowledge rather than actual cues to achieve unexpectedly high scores on existing metrics [222]. This limitation highlights the need for more comprehensive benchmarks and metrics that can accurately assess both FM capabilities and potential deviations from intended behaviors [11].

- **Opportunity: Online Methods for Safety Alignment.** While safety benchmarks can partially ensure alignment, online methods are still needed to guarantee safety during task execution in autonomous vehicles. However, due to the complexity of FMs and the requirement for low-latency execution, designing such methods remains challenging. As highlighted by OpenAI, alignment must embrace uncertainty while maintaining rigorous measurement [1]. This limitation highlights the need for designing suitable online methods for FMs in autonomous driving systems.

Challenge II: Ethical and Human-Like Autonomous Driving. The ultimate objective of integrating FMs into autonomous driving is to enable ethical and human-like driving behavior. However, as pointed out by Tian *et al.* [191], existing regulatory frameworks for autonomous vehicles (e.g., ISO 26262) primarily address functional safety and cybersecurity for conventional rule-based systems, and do not fully capture the complexities introduced by foundation models; thus, these regulations may need to be revisited and updated. Moreover, transparency and interpretability in ADS have been increasingly emphasised due to their critical roles in regulatory approval and accountability. The integration of FMs may offer new opportunities for interpretability—for example, in speech-to-speech systems, researchers can inspect intermediate textual representations produced by LLMs to facilitate error analysis [152]. Nevertheless, further advances are still required to systematically embed explainability into FM-driven ADS pipelines. Finally, since FMs are often adopted for reasoning and decision-making, it is crucial to test them under morally sensitive scenarios and to characterise how risks are distributed among different road users. Recent work has begun to operationalise such moral testing via simulation-based methods (e.g., metamorphic testing) to expose potentially problematic decision patterns [189]. These efforts are essential for developing ethical and human-like autonomous driving.

- **Opportunity: Moral Testing and Assurance for FM-based ADS.** As FMs are increasingly adopted for high-level reasoning and decision-making in ADS, a key opportunity is to establish systematic *moral testing* and assurance pipelines that complement functional safety validation. Unlike conventional safety requirements, moral expectations are often value-laden, context-dependent, and may not admit a single ground-truth label, making classic test oracles difficult to define. Recent work has started to operationalise moral evaluation via simulation-based *metamorphic testing*, where moral meta-principles are encoded as relational properties across paired scenarios to expose inconsistent or discriminatory decision patterns [189]. Building on this direction, future research could develop scalable moral test suites and benchmarks, integrate scenario generation and coverage metrics tailored to morally sensitive situations.

Challenge III: Ensuring Correctness of FM-Generated Code. With the widespread application of FM-based coding assistance (i.e., GitHub Co-pilot), autonomous driving developers will inevitably use FM-generated code during the development of the autonomous driving system. As pointed out by Chen *et al.* [29], LLM could misunderstand the code description and generate code without syntactic mistakes, but still defective. Furthermore, code generated by LLMs often lacks description or context and is more difficult for human developers to understand and maintain [122], and may introduce critical issues into the codebase if not carefully verified [136].

- **Opportunity: Automated Verification for FM-generated Code.** With the growing coding capabilities of LLMs, the use of LM-generated code is becoming inevitable. However, errors in such code can be more difficult for human developers to detect. Therefore, in safety-critical domains like autonomous driving, there is a pressing need to develop automated verification techniques for FM-generated code. Nouri *et al.* [143] proposed an iterative loop to automatically generate, verify, and refine LLM-generated code in the context of autonomous driving. Yet, due to issues such as LLMs' sensitivity to minor text disruptions in code descriptions or prompts, more robust automated verification techniques are still required.

6 Future Research Roadmap

The development and deployment of Foundation Models for Autonomous Driving (FM4AD) is a monumental task that requires a coordinated, multi-stage research effort. To guide the community, we propose a roadmap divided into three phases: short-term foundational work, mid-term system integration, and a long-term vision for achieving verifiable trustworthiness at scale.

6.1 Short-Term Goal

The immediate focus must be on establishing the fundamental building blocks required for robust and scalable research and development of FMs for autonomous driving. This involves addressing the most pressing challenges in data, development tools, and deployment efficiency.

- **Establishing Data Best Practices:** The quality and management of data are paramount. The community's short-term goals should be to develop **unified data management frameworks** to handle the massive scale of autonomous driving datasets and to create automated tools for **dataset license compliance** to ensure legal and ethical usage. Simultaneously, advancing techniques for **bias mitigation** and **privacy preservation** during data cleaning and curation is critical to building a fair and secure foundation. Success in this phase will be measured by the development of automated fairness testing suites that achieve a high correlation between perception-level bias metrics (e.g., AOD) and decision-level safety violations.
- **Improving the Development Toolchain:** To manage the unprecedented complexity of the FM toolchain, the immediate priority is to conduct **large-scale empirical analyses** to identify and eliminate inefficiencies in current workflows. A parallel effort is required to develop a new generation of **Quality Assurance** and testing techniques specifically designed for the distributed systems (e.g., Ray) that underpin FM training, thereby improving the reliability and reproducibility of the development process.
- **Enabling Efficient and Safe Deployment:** Before FMs can be widely integrated, their efficiency and safety on resource-constrained vehicle hardware must be understood and improved. A crucial short-term task is to perform a comprehensive **empirical analysis of the impact of optimization techniques** like quantization and pruning on model trustworthiness. This knowledge will directly inform the development of **robust-by-design optimization** methods, where security and reliability are co-designed with performance. Furthermore, success in this area will be measured by the ability of optimized models to maintain an inference latency within the 500ms range, a critical threshold for ensuring real-time responsiveness in complex driving reasoning tasks [191].

6.2 Mid-term Goal

With a solid foundation in place, the research focus can shift to the complex challenges of integrating FMs into the vehicle as a cohesive and reliable system. This phase emphasizes system-level reliability and architecture.

- **Enhancing System Reliability and Adaptation:** The mid-term priority is to address the core reliability issues of FMs in dynamic environments, namely **hallucination** and **multi-modality adaptation**. This involves developing **end-to-end multi-modal fusion** models that are less vulnerable to upstream perception errors and using **multi-modal grounding** to verify model outputs against raw sensor data. To measure progress, creating **standardized benchmarks for robustness in multi-modal fusion** is essential.
- **Designing Hybrid System Architectures:** An integrated FM-powered vehicle will not operate in isolation. A key mid-term goal is to architect effective **edge-cloud collaboration** systems. Research must deliver frameworks for **intelligent task orchestration** and **robust, asynchronous communication** schemes that can dynamically balance the low-latency needs of the vehicle (edge) with the immense computational power of the cloud.

- **Fostering a Domain-Specific Ecosystem:** To accelerate innovation and democratize research, the community must address the lack of powerful, open-source models. A central mid-term objective should be the development of an **open-source, domain-specific foundation model for autonomous driving** that can serve as a strong baseline for academic and independent researchers.

6.3 Long-term Goal

The ultimate vision is the widespread deployment of autonomous vehicles powered by FMs that are not just capable but are verifiably safe, trustworthy, and aligned with human values. This requires tackling the most fundamental challenges from the hardware to the AI's core objectives.

- **Establishing a Hardware Root of Trust:** Trustworthiness must begin at the silicon level. The long-term grand challenge is to overcome the risks of "black-box" hardware by developing **security-by-design hardware architectures**. This includes creating **Trusted Execution Environments (TEEs)** tailored for AI accelerators and promoting **open and verifiable hardware stacks** that allow for community-driven security audits, providing a transparent and secure foundation for all software.
- **Achieving Foundation Model Alignment:** Ensuring that FMs behave in accordance with human intentions is perhaps the most critical long-term goal. This demands a sustained research program in **AI alignment**. The focus should be on developing robust methods for both **forward alignment**, such as enhancing human-feedback mechanisms during training, and **backward alignment**, by creating comprehensive **safety benchmarks** to evaluate model behavior. A key frontier is the design of effective **online methods for safety alignment** that can provide real-time assurance in a deployed vehicle.
- **Guaranteeing Code and System Correctness:** As FMs increasingly contribute to writing the software that runs autonomous vehicles, verifying its correctness is a long-term imperative. The research community must work towards powerful **automated verification techniques** that can formally prove the safety and reliability of **FM-generated code**, ensuring that this powerful tool enhances, rather than compromises, the safety of autonomous systems.

7 Conclusion

In this paper, we presented a structured literature review and an initial roadmap for integrating foundation models into autonomous driving systems. We organized the discussion around three dimensions: FM infrastructure, in-vehicle integration, and practical deployment. Across these dimensions, we synthesized the state of the art, identified key challenges, and highlighted promising research opportunities, which we further distilled into short-, mid-, and long-term research goals. Although substantial technical and engineering challenges remain, FMs offer significant potential to advance the capabilities of ADSs. We hope this paper serves as a useful reference for future research and helps guide the development of safer, more reliable, and more trustworthy autonomous driving systems.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China: No. 52408039, JST CRONOS Grant (No. JPMJCS24K8) and JSPS KAKENHI Grant (No.JP24K02920). This research/project was also supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 2 (Proposal ID: T2EP20223-0043; Project ID: MOE-000613-00). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

[1] [n. d.]. How we think about safety and alignment. https://openai.com/safety/how-we-think-about-safety-alignment/

[2] [n. d.]. Li Auto rolls out 'end-to-end + VLM' smart driving system via OTA. https://autonews.gasgoo.com/icv/70034940.html

[3] [n. d.]. TIER IV introduces LLM for autonomous driving: Enabling cars that think and talk. https://tier4.jp/en/media/detail/?sys_id=7EB3ywJsqIdelR0ozhqYzg

[4] [n. d.]. Xiaomi SU7 rolls out OTA update with VLM integration. https://www.metal.com/en/newscontent/103104153

[5] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[6] Giulio Antonio Abbo, Serena Marchesi, Agnieszka Wykowska, and Tony Belpaeme. 2024. Social Value Alignment in Large Language Models. In *Value Engineering in Artificial Intelligence*, Nardine Osman and Luc Steels (Eds.). Springer Nature Switzerland, Cham, 83–97.

[7] Yasasa Abeysirigoonawardena, Florian Shkurti, and Gregory Dudek. 2019. Generating Adversarial Driving Scenarios in High-Fidelity Simulators. In *2019 International Conference on Robotics and Automation (ICRA)*. 8271–8277. doi:10.1109/ICRA.2019.8793740

[8] ADLINK Technology. 2026. ADM-AL30: Autonomous Driving AI Decision Making ECU. https://www.adlinktech.com/Products/Automotive-Computing/Autonomous-Driving/ADM-AL30. Accessed: 2026-01-22.

[9] ADLINK Technology. 2026. AVA-3510: Autonomous Driving Solutions. https://www.adlinktech.com/products/automotive-computing/autonomous-driving/ava-3510. Accessed: 2026-01-22.

[10] Alexander Amini, Wilko Schwarting, Guy Rosman, Brandon Araki, Sertac Karaman, and Daniela Rus. 2018. Variational Autoencoder for End-to-End Control of Autonomous Driving with Novelty Detection and Training De-biasing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid, Spain). IEEE Press, 568–575. doi:10.1109/IROS.2018.8594386

[11] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565 [cs.AI] https://arxiv.org/abs/1606.06565

[12] Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. 2025. CoVLA: Comprehensive Vision-Language-Action Dataset for Autonomous Driving. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. 1933–1943.

[13] Gabriel Y. Arteaga, Thomas B. Schön, and Nicolas Pielawski. 2024. Hallucination Detection in LLMs: Fast and Memory-Efficient Finetuned Models. In *Northern Lights Deep Learning Conference 2025*. https://openreview.net/forum?id=8T8QkDsuO9

[14] Sai Siddhartha Chary Aylapuram, Veeraraju Elluru, and Shivang Agarwal. 2025. Bias-Aware Machine Unlearning: Towards Fairer Vision Models via Controllable Forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2571–2579.

[15] Sudharshana B, Nandhini V, and AkilaGandhi G S ME. 2025. A Comprehensive Review of Llm Neural Network Enhancements for Advanced Driving Assistance Systems Through Quantization. In *2025 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*. 1–7. doi:10.1109/ASSIC64892.2025.11158258

[16] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966* (2023).

[17] Luciano Baresi, Davide Yi Xian Hu, Muhammad Irfan Mas'udi, and Giovanni Quattrocchi. 2025. DILLEMA: Diffusion and Large Language Models for Multi-Modal Augmentation. In *2025 IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest)*. 29–36. doi:10.1109/DeepTest66595.2025.00010

[18] Ran Bi, Tongtong Xu, Mingxue Xu, and Enhong Chen. 2022. PaddlePaddle: A Production-Oriented Deep Learning Platform Facilitating the Competency of Enterprises. In *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*. 92–99. doi:10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00046

[19] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray

Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG] https://arxiv.org/abs/2108.07258

[20] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. *JAX: composable transformations of Python+NumPy programs*. http://github.com/jax-ml/jax

[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[22] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 11618–11628. doi:10.1109/CVPR42600.2020.01164

[23] Simin Cai, Fredrik Ålund, Bengt Gunne, and Richard Hayton. 2022. Mimer Trust: Efficient and Secure Data Processing for Trusted Execution Environment in Automotive Systems. In *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*. 1–8. doi:10.1109/ETFA52439.2022.9921649

[24] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[25] Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. 2025. Hallucination Detection in Foundation Models for Decision-Making: A Flexible Definition and Review of the State of the Art. *ACM Comput. Surv.* (Feb. 2025). doi:10.1145/3716846 Just Accepted.

[26] Daihang Chen, Yonghui Liu, Mingyi Zhou, Yanjie Zhao, Haoyu Wang, Shuai Wang, Xiao Chen, Tegawendé F. Bissyandé, Jacques Klein, and Li Li. 2024. LLM for Mobile: An Initial Roadmap. *ACM Trans. Softw. Eng. Methodol.* (Dec. 2024). doi:10.1145/3708528 Just Accepted.

[27] Kejia Chen, Jiawen Zhang, Jiacong Hu, Yu Wang, Jian Lou, Zunlei Feng, and Mingli Song. 2025. Assessing Safety Risks and Quantization-aware Safety Patching for Quantized Large Language Models. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=jywq7qJLt5

[28] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 14093–14100. doi:10.1109/ICRA57147.2024.10611018

[29] QiHong Chen, Jiachen Yu, Jiawei Li, Jiecheng Deng, Justin Tian Jin Chen, and Iftekhar Ahmed. 2025. A Deep Dive Into Large Language Model Code Generation Mistakes: What and Why? arXiv:2411.01414 [cs.SE] https://arxiv.org/abs/2411.01414

[30] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13320–13331.

[31] Yu Chen, Fang Luo, Tong Li, Tao Xiang, Zheli Liu, and Jin Li. 2020. A training-integrity privacy-preserving federated learning scheme with trusted execution environment. *Information Sciences* 522 (2020), 69–79. doi:10.1016/j.ins.2020.02.037

[32] Zhenpeng Chen, Xinyue Li, Jie M. Zhang, Federica Sarro, and Yang Liu. 2025. Diversity Drives Fairness: Ensemble of Higher Order Mutants for Intersectional Fairness of Machine Learning Software. In *Proceedings of the 47th IEEE/ACM International Conference on Software Engineering, ICSE 2025*.

[33] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Trans. Softw. Eng. Methodol.* 32, 4, Article 106 (May 2023), 30 pages. doi:10.1145/3583561

[34] Tushar Choudhary, Vikrant Dewangan, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K. Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K. Madhava Krishna. 2024. Talk2BEV: Language-enhanced Bird's-eye View Maps for Autonomous Driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 16345–16352. doi:10.1109/ICRA57147.2024.10611485

[35] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski,

Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113. http://jmlr.org/papers/v24/22-1144.html

[36] George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating Hallucinations in Pruned Large Language Models for Abstractive Summarization. *Transactions of the Association for Computational Linguistics* 12 (2024), 1163–1181. doi:10.1162/tacl_a_00695

[37] Zhaoyang Chu, Yao Wan, Zhikun Zhang, Di Wang, Zhou Yang, Hongyu Zhang, Pan Zhou, Xuanhua Shi, Hai Jin, and David Lo. 2025. Scrub It Out! Erasing Sensitive Memorization in Code Language Models via Machine Unlearning. *arXiv preprint arXiv:2509.13755* (2025).

[38] Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Daogao Liu, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. 2024. Mind the Privacy Unit! User-Level Differential Privacy for Language Model Fine-Tuning. In *First Conference on Language Modeling*. https://openreview.net/forum?id=Jd0bCD12DS

[39] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Drive as You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE Computer Society, Los Alamitos, CA, USA, 902–909. doi:10.1109/WACVW60836.2024.00101

[40] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Receive, Reason, and React: Drive as You Say, With Large Language Models in Autonomous Vehicles. *IEEE Intelligent Transportation Systems Magazine* 16, 4 (2024), 81–94. doi:10.1109/MITS.2024.3381793

[41] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2024. A Survey on Multimodal Large Language Models for Autonomous Driving . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE Computer Society, Los Alamitos, CA, USA, 958–979. doi:10.1109/WACVW60836.2024.00106

[42] Can Cui, Zichong Yang, Yupeng Zhou, Yunsheng Ma, Juanwu Lu, Lingxi Li, Yaobin Chen, Jitesh Panchal, and Ziran Wang. 2024. Personalized Autonomous Driving with Large Language Models: Field Experiments. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. 20–27. doi:10.1109/ITSC58415.2024.10919978

[43] Can Cui, Zichong Yang, Yupeng Zhou, Juntong Peng, Sung-Yeon Park, Cong Zhang, Yunsheng Ma, Xu Cao, Wenqian Ye, Yiheng Feng, Jitesh Panchal, Lingxi Li, Yaobin Chen, and Ziran Wang. 2024. On-Board Vision-Language Models for Personalized Autonomous Vehicle Motion Control: System Design and Real-World Validation. arXiv:2411.11913 [cs.AI] https://arxiv.org/abs/2411.11913

[44] Erfei Cui, Wenhai Wang, Zhiqi Li, Jiangwei Xie, Haoming Zou, Hanming Deng, Gen Luo, Lewei Lu, Xizhou Zhu, and Jifeng Dai. 2025. DriveMLM: aligning multi-modal large language models with behavioral planning states for autonomous driving. *Visual Intelligence* 3, 1 (26 Nov 2025), 22. doi:10.1007/s44267-025-00095-w

[45] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=vvoWPYqZJA

[46] Wei Dai, Shengen Wu, Wei Wu, Zhenhao Wang, Sisuo Lyu, Haicheng Liao, Limin Yu, Weiping Ding, Runwei Guan, and Yutao Yue. 2025. Large Foundation Models for Trajectory Prediction in Autonomous Driving: A Comprehensive Survey. arXiv:2509.10570 [cs.RO] https://arxiv.org/abs/2509.10570

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[48] Wenhao Ding, Baiming Chen, Minjun Xu, and Ding Zhao. 2020. Learning to Collide: An Adaptive Safety-Critical Scenarios Generating Method. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2243–2250. doi:10.1109/IROS45743.2020.9340696

[49] Wenhao Ding, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. 2023. A Survey on Safety-Critical Driving Scenario Generation—A Methodological Perspective. *IEEE Transactions on Intelligent Transportation Systems* 24, 7 (2023), 6971–6988. doi:10.1109/TITS.2023.3259322

[50] Veronika Domova, Rebecca Maria Currano, and David Sirkin. 2024. Comfort in Automated Driving: A Literature Survey and a High-Level Integrative Framework. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 98 (Sept. 2024), 23 pages. doi:10.1145/3678583

[51] Malsha Ashani Mahawatta Dona, Beatriz Cabrero-Daniel, Yinan Yu, and Christian Berger. 2024. LLMs Can Check Their Own Results to Mitigate Hallucinations in Traffic Understanding Tasks. In *Testing Software and Systems: 36th IFIP WG 6.1 International Conference, ICTSS 2024, London, UK, October 30 – November 1, 2024, Proceedings* (London, United Kingdom). Springer-Verlag, Berlin, Heidelberg,

114–130. doi:10.1007/978-3-031-80889-0_8

[52] Wei Dong, Sikai Lu, Xinhe Chen, Shunyao Zhang, Qingchao Liu, Ze Liu, Long Chen, Hai Wang, and Yingfeng Cai. 2025. End-to-end Autonomous Driving: From Classic Paradigm to Large Model Empowerment - A Comprehensive Survey. *IEEE Internet of Things Journal* (2025), 1–1. doi:10.1109/JIOT.2025.3635092

[53] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 78)*, Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (Eds.). PMLR, 1–16. https://proceedings.mlr.press/v78/dosovitskiy17a.html

[54] Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. 2023. Flocks of Stochastic Parrots: Differentially Private Prompt Learning for Large Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 76852–76871. https://proceedings.neurips.cc/paper_files/paper/2023/file/f26119b4ffe38c24d97e4c49d334b99e-Paper-Conference.pdf

[55] Kazuki Egashira, Robin Staab, Thibaud Gloaguen, Mark Vero, and Martin Vechev. 2026. Fewer Weights, More Problems: A Practical Attack on LLM Pruning. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=YRwe9fP7j5

[56] Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. 2024. Exploiting LLM Quantization. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 41709–41732. doi:10.52202/079017-1319

[57] Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa A. D. Nesnas, and Marco Pavone. 2023. Semantic anomaly detection with large language models. *Autonomous Robots* 47, 8 (01 Dec 2023), 1035–1055. doi:10.1007/s10514-023-10132-6

[58] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: a survey. *J. Mach. Learn. Res.* 20, 1 (Jan. 2019), 1997–2017.

[59] Jiaqi Fan, Jianhua Wu, Hongqing Chu, Quanbo Ge, and Bingzhao Gao. 2026. Hallucination Elimination and Text Annotation Framework for Large Vision-Language Models in Traffic Scenarios. *IEEE Transactions on Intelligent Transportation Systems* 27, 1 (2026), 358–374. doi:10.1109/TITS.2025.3625700

[60] Tuo Feng, Wenguan Wang, and Yi Yang. 2025. A Survey of World Models for Autonomous Driving. arXiv:2501.11260 [cs.RO] https://arxiv.org/abs/2501.11260

[61] Elias Frantar and Dan Alistarh. 2023. SparseGPT: massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 414, 15 pages.

[62] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. 2024. Drive Like a Human: Rethinking Autonomous Driving with Large Language Models . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE Computer Society, Los Alamitos, CA, USA, 910–919. doi:10.1109/WACVW60836.2024.00102

[63] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (Sept. 2024), 1097–1179. doi:10.1162/coli_a_00524

[64] Haoxiang Gao, Zhongruo Wang, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. 2025. A Survey for Foundation Models in Autonomous Driving. In *2025 6th International Conference on Computer Vision and Data Mining (ICCVDM)*. 63–71. doi:10.1109/ICCVDM66874.2025.11290083

[65] Jianhua Gao, Zeming Liu, Yizhuo Wang, and Weixing Ji. 2025. RaNAS: Resource-Aware Neural Architecture Search for Edge Computing. *ACM Trans. Archit. Code Optim.* 22, 1, Article 18 (March 2025), 18 pages. doi:10.1145/3703353

[66] Yuan Gao, Mattia Piccinini, Yuchen Zhang, Dingrui Wang, Korbinian Moller, Roberto Brusnicki, Baha Zarrouki, Alessio Gambi, Jan Frederik Totz, Kai Storms, Steven Peters, Andrea Stocco, Bassam Alrifaee, Marco Pavone, and Johannes Betz. 2025. Foundation Models in Autonomous Driving: A Survey on Scenario Generation and Scenario Analysis. arXiv:2506.11526 [cs.RO] https://arxiv.org/abs/2506.11526

[67] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge Distillation of Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=5h0qf7IBZZ

[68] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Guohui Zhang, and Chengzhong Xu. 2024. World Models for Autonomous Driving: An Initial Survey. *IEEE Transactions on Intelligent Vehicles* (2024), 1–17. doi:10.1109/TIV.2024.3398357

[69] Zixu Hao, Huiqiang Jiang, Shiqi Jiang, Ju Ren, and Ting Cao. 2024. Hybrid SLM and LLM for Edge-Cloud Collaborative Inference. In *Proceedings of the Workshop on Edge and Mobile Foundation Models* (Minato-ku, Tokyo, Japan) *(EdgeFM '24)*. Association for Computing Machinery, New York, NY, USA, 36–41. doi:10.1145/3662006.3662067

[70] Deepti Hegde, Rajeev Yasarla, Hong Cai, Shizhong Han, Apratim Bhattacharyya, Shweta Mahajan, Litian Liu, Risheek Garrepalli, Vishal M. Patel, and Fatih Porikli. 2025. Distilling Multi-Modal Large Language Models for Autonomous Driving . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 27575–27585. doi:10.1109/CVPR52734.2025.02568

[71] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc,

Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=iBBcRUlOAPR

[72] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM J. Responsib. Comput.* 1, 2, Article 11 (June 2024), 52 pages. doi:10.1145/3631326

[73] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2024. Search-based Automatic Repair for Fairness and Accuracy in Decision-making Software. *Empir. Softw. Eng.* 29, 1 (2024), 36. doi:10.1007/S10664-023-10419-3

[74] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. 2023. GAIA-1: A Generative World Model for Autonomous Driving. arXiv:2309.17080 [cs.CV] https://arxiv.org/abs/2309.17080

[75] Shengtuo Hu, Qi Alfred Chen, Jiwon Joung, Can Carlak, Yiheng Feng, Z. Morley Mao, and Henry X. Liu. 2020. CVShield: Guarding Sensor Data in Connected Vehicle with Trusted Execution Environment. In *Proceedings of the Second ACM Workshop on Automotive and Aerial Vehicle Security* (New Orleans, LA, USA) *(AutoSec '20)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3375706.3380552

[76] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[77] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13418–13427.

[78] IEEE. 2025. Standard for Human Intentions and Artificial Intelligence Alignment in Autonomous Driving Agent. IEEE P3474, Draft Standard. https://standards.ieee.org/ieee/3474/11639/ Accessed: Feb. 21, 2025.

[79] Fatih Ilhan, Gong Su, Selim Furkan Tekin, Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. Resource-Efficient Transformer Pruning for Finetuning of Large Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16206–16215.

[80] ISO. 2018. *Road vehicles — Functional safety — Part 9: Automotive Safety Integrity Level (ASIL)-oriented and safety-oriented analyses*. Technical Report ISO 26262-9:2018. International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/68391.html Accessed: 2026-01-22.

[81] Mahmoud Jahanshahi and Audris Mockus. 2025. Cracks in The Stack: Hidden Vulnerabilities and Licensing Risks in LLM Pre-Training Datasets. arXiv:2501.02628 [cs.SE] https://arxiv.org/abs/2501.02628

[82] Kanishk Jain, Varun Chhangani, Amogh Tiwari, K. Madhava Krishna, and Vineet Gandhi. 2023. Ground then Navigate: Language-guided Navigation in Dynamic Scenes. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 4113–4120. doi:10.1109/ICRA48891.2023.10160614

[83] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Hua Xu, Aidan O'Gara, Kwan Ng, Brian Tse, Jie Fu, Stephen Mcaleer, Yanfeng Wang, Mingchuan Yang, Yunhuai Liu, Yizhou Wang, Song-Chun Zhu, Yike Guo, Yaodong Yang, and Wen Gao. 2025. AI Alignment: A Contemporary Survey. *ACM Comput. Surv.* 58, 5, Article 132 (Nov. 2025), 38 pages. doi:10.1145/3770749

[84] Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, Hao Ye, Zihao Sheng, Xin Zhao, Tuopu Wen, Zheng Fu, Sikai Chen, Kun Jiang, Diange Yang, Seongjin Choi, and Lijun Sun. 2025. A Survey on Vision-Language-Action Models for Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 4524–4536.

[85] Yuxuan Jiang, Guojian Zhan, Zhiqian Lan, Chang Liu, Bo Cheng, and Shengbo Eben Li. 2024. A Reinforcement Learning Benchmark for Autonomous Driving in General Urban Scenarios. *IEEE Transactions on Intelligent Transportation Systems* 25, 5 (2024), 4335–4345. doi:10.1109/TITS.2023.3329823

[86] Zhen Hang Jiang, Yunsi Fei, and David Kaeli. 2017. A Novel Side-Channel Timing Attack on GPUs. In *Proceedings of the Great Lakes Symposium on VLSI 2017* (Banff, Alberta, Canada) *(GLSVLSI '17)*. Association for Computing Machinery, New York, NY, USA, 167–172. doi:10.1145/3060403.3060462

[87] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=VTF8yNQM66

[88] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. 2023. ADAPT: Action-aware Driving Caption Transformer. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 7554–7561. doi:10.1109/ICRA48891.2023.10160326

[89] Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. 2024. JailbreakZoo: Survey, Landscapes, and Horizons in Jailbreaking Large Language and Vision-Language Models. arXiv:2407.01599 [cs.CL] https://arxiv.org/abs/2407.01599

[90] Ye Jin, Ruoxuan Yang, Zhijie Yi, Xiaoxi Shen, Huiling Peng, Xiaoan Liu, Jingli Qin, Jiayang Li, Jintao Xie, Peizhong Gao, Guyue Zhou, and Jiangtao Gong. 2024. SurrealDriver: Designing LLM-powered Generative Driver Agent Framework based on Human Drivers'

Driving-thinking Data. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 966–971. doi:10.1109/IROS58592.2024.10802229

[91] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating Training Data Mitigates Privacy Risks in Language Models. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 10697–10707. https://proceedings.mlr.press/v162/kandpal22a.html

[92] Dewant Katare, David Solans Noguero, Souneil Park, Nicolas Kourtellis, Marijn Janssen, and Aaron Yi Ding. 2025. Analyzing and Mitigating Bias for Vulnerable Road Users by Addressing Class Imbalance in Datasets. *IEEE Open Journal of Intelligent Transportation Systems* 6 (2025), 590–604. doi:10.1109/OJITS.2025.3564558

[93] Ali Keysan, Andreas Look, Eitan Kosman, Gonca Gürsun, Jörg Wagner, Yu Yao, and Barbara Rakitsch. 2023. Can you text what is happening? Integrating pre-trained language encoders into trajectory prediction models for autonomous driving. arXiv:2309.05282 [cs.CV]

[94] Mohammad Khoshkdahan, Arman Akbari, Arash Akbari, and Xuan Zhang. 2025. Beyond Overall Accuracy: Pose- and Occlusion-driven Fairness Analysis in Pedestrian Detection for Autonomous Driving. arXiv:2509.26166 [cs.CV] https://arxiv.org/abs/2509.26166

[95] Mohammad Khoshkdahan, Nicholas Kjär, and Fabian B. Flohr. 2025. FAIR-PED: Fairness Evaluation in Pedestrian Detection Using CLIP. In *2025 IEEE Intelligent Vehicles Symposium (IV)*. 1504–1509. doi:10.1109/IV64158.2025.11097691

[96] Gyeongman Kim, Doohyuk Jang, and Eunho Yang. 2024. PromptKD: Distilling Student-Friendly Knowledge for Generative Language Models via Prompt Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 6266–6282. doi:10.18653/v1/2024.findings-emnlp.364

[97] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 577–593.

[98] Jaekyeom Kim, Sungryull Sohn, Gerrard Jeongwon Jo, Jihoon Choi, Kyunghoon Bae, Hwayoung Lee, Yongmin Park, and Honglak Lee. 2024. Do Not Trust Licenses You See—Dataset Compliance Requires Massive-Scale AI-Powered Lifecycle Tracing. https://lgresearch.ai/data/upload/LG_AI_Research_Data_compliance_arxiv_EST.pdf

[99] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing Privacy Leakage in Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=QkLpGxUboF

[100] Xiangrui Kong, Thomas Braunl, Marco Fahmi, and Yue Wang. 2024. A Superalignment Framework in Autonomous Driving with Large Language Models. In *2024 IEEE Intelligent Vehicles Symposium (IV)*. 1715–1720. doi:10.1109/IV55156.2024.10588403

[101] Friedrich Kruber, Jonas Wurst, Eduardo Sánchez Morales, Samarjit Chakraborty, and Michael Botsch. 2019. Unsupervised and Supervised Learning with the Random Forest Algorithm for Traffic Scenario Clustering and Classification. In *2019 IEEE Intelligent Vehicles Symposium (IV)*. 2463–2470. doi:10.1109/IVS.2019.8813994

[102] Wenxin Kuang, Qizhuang Liang, Peng Sun, Wei Fu, Qiao Hu, and Yupeng Hu. 2025. Unveiling the Pruning Risks on Privacy Vulnerabilities of Deep Neural Networks. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49660.2025.10889113

[103] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles* (Koblenz, Germany) *(SOSP '23)*. Association for Computing Machinery, New York, NY, USA, 611–626. doi:10.1145/3600006.3613165

[104] Jaewon Lee, Yonghae Kim, Jiashen Cao, Euna Kim, Jaekyu Lee, and Hyesoon Kim. 2022. Securing GPU via region-based bounds checking. In *Proceedings of the 49th Annual International Symposium on Computer Architecture* (New York, New York) *(ISCA '22)*. Association for Computing Machinery, New York, NY, USA, 27–41. doi:10.1145/3470496.3527420

[105] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 8424–8445. doi:10.18653/v1/2022.acl-long.577

[106] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. arXiv:1811.07871 [cs.LG] https://arxiv.org/abs/1811.07871

[107] Edouard Leurent. 2018. *An Environment for Autonomous Driving Decision-Making*. https://github.com/eleurent/highway-env Release date: 2018-05-01.

[108] Ding Li, Ziqi Zhang, Mengyu Yao, Yifeng Cai, Yao Guo, and Xiangqun Chen. 2025. TEESlice: Protecting Sensitive Neural Network Models in Trusted Execution Environments when Attackers Have Pre-Trained Models. *ACM Trans. Softw. Eng. Methodol.* 34, 6, Article 166 (July 2025), 49 pages. doi:10.1145/3707453

[109] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 10879–10899. doi:10.18653/v1/2024.acl-long.586

[110] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 814, 13 pages.

[111] Jiaxi Li, Lu Yin, and Xilu Wang. 2025. OWLed: Outlier-weighed Layerwise Pruning for Efficient Autonomous Driving Framework. In *2025 International Joint Conference on Neural Networks (IJCNN)*. 1–8. doi:10.1109/IJCNN64981.2025.11227233

[112] Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Junxian Guo, Xiuyu Li, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. 2025. SVDQuant: Absorbing Outliers by Low-Rank Component for 4-Bit Diffusion Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=vWR3KuiQur

[113] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 7602–7635. doi:10.18653/v1/2024.naacl-long.421

[114] Shenggui Li, Hongxin Liu, Zhengda Bian, Jiarui Fang, Haichen Huang, Yuliang Liu, Boxiang Wang, and Yang You. 2023. Colossal-AI: A Unified Deep Learning System For Large-Scale Parallel Training. In *Proceedings of the 52nd International Conference on Parallel Processing* (Salt Lake City, UT, USA) *(ICPP '23)*. Association for Computing Machinery, New York, NY, USA, 766–775. doi:10.1145/3605573.3605613

[115] Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. 2024. Bias Behind the Wheel: Fairness Testing of Autonomous Driving Systems. *ACM Trans. Softw. Eng. Methodol.* (Nov. 2024). doi:10.1145/3702989 Just Accepted.

[116] Yun Li, Kai Katsumata, Ehsan Javanmardi, and Manabu Tsukada. 2024. Large Language Models for Human-Like Autonomous Driving: A Survey. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. 439–446. doi:10.1109/ITSC58415.2024.10919629

[117] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. In *Proceedings of Machine Learning and Systems*, P. Gibbons, G. Pekhimenko, and C. De Sa (Eds.), Vol. 6. 87–100. https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf

[118] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. 2018. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkhQHMW0W

[119] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=w0H2xGHlkw

[120] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 943, 15 pages.

[121] Xuanzhe Liu, Diandian Gu, Zhenpeng Chen, Jinfeng Wen, Zili Zhang, Yun Ma, Haoyu Wang, and Xin Jin. 2023. Rise of Distributed Deep Learning Training in the Big Model Era: From a Software Engineering Perspective. *ACM Trans. Softw. Eng. Methodol.* 32, 6, Article 156 (Sept. 2023), 26 pages. doi:10.1145/3597204

[122] Yue Liu, Thanh Le-Cong, Ratnadira Widyasari, Chakkrit Tantithamthavorn, Li Li, Xuan-Bach D. Le, and David Lo. 2024. Refining ChatGPT-Generated Code: Characterizing and Mitigating Code Quality Issues. *ACM Trans. Softw. Eng. Methodol.* 33, 5, Article 116 (June 2024), 26 pages. doi:10.1145/3643674

[123] Yang Liu, Ying Tan, Jingzhou Luo, and Weixing Chen. 2024. VCD: Visual Causality Discovery for Cross-Modal Question Reasoning. In *Pattern Recognition and Computer Vision*, Qingshan Liu, Hanzi Wang, Zhanyu Ma, Weishi Zheng, Hongbin Zha, Xilin Chen, Liang Wang, and Rongrong Ji (Eds.). Springer Nature Singapore, Singapore, 309–322.

[124] David Fernández Llorca, Pedro Frau, Ignacio Parra, Rubén Izquierdo, and Emilia Gómez. 2024. Attribute annotation and bias evaluation in visual datasets for autonomous driving. *J. Big Data* 11, 1 (2024), 137. doi:10.1186/S40537-024-00976-9

[125] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt D. Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2024. A large-scale audit of dataset licensing and attribution in AI. *Nat. Mac. Intell.* 6, 8 (2024), 975–987. doi:10.1038/S42256-024-00878-8

[126] Yunchi Lu, Youshan Miao, Cheng Tan, Peng Huang, Yi Zhu, Xian Zhang, and Fan Yang. 2025. TrainVerify: Equivalence-Based Verification for Distributed LLM Training. In *Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles* (Lotte Hotel World, Seoul, Republic of Korea) *(SOSP '25)*. Association for Computing Machinery, New York, NY, USA, 237–253. doi:10.1145/3731569.3764850

[127] Nicola Lucchi. 2024. ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. *European Journal of Risk Regulation* 15, 3 (2024), 602–624. doi:10.1017/err.2023.59

[128] Zhifan Luo, Shuo Shao, Su Zhang, Lijing Zhou, Yuke Hu, Chenxu Zhao, Zhihao Liu, and Zhan Qin. 2025. Shadow in the Cache: Unveiling and Mitigating Privacy Risks of KV-cache in LLM Inference. arXiv:2508.09442 [cs.CR] https://arxiv.org/abs/2508.09442

[129] Shi Luohe, Hongyi Zhang, Yao Yao, Zuchao Li, and hai zhao. 2024. Keep the Cost Down: A Review on Methods to Optimize LLM's KV-Cache Consumption. In *First Conference on Language Modeling*. https://openreview.net/forum?id=8tKjqqMM5z

[130] Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. LLM-Pruner: On the Structural Pruning of Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=J8Ajf9WfXP

[131] Xiaoxue Ma, Wanwei Zhan, Jiale Chen, Yishu Li, Jacky Keung, and Federica Sarro. 2025. A Comprehensive Study of Bugs in Modern Distributed Deep Learning Systems. arXiv:2512.20345 [cs.SE] https://arxiv.org/abs/2512.20345

[132] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. 2023. GPT-Driver: Learning to Drive with GPT. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*. https://openreview.net/forum?id=Pvjk9lxLJK

[133] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. 2024. A Language Agent for Autonomous Driving. In *First Conference on Language Modeling*. https://openreview.net/forum?id=UPE6WYE8vg

[134] Yuewen Mei, Tong Nie, Jian Sun, and Ye Tian. 2025. LLM-Attacker: Enhancing Closed-Loop Adversarial Scenario Generation for Autonomous Driving With Large Language Models. *IEEE Transactions on Intelligent Transportation Systems* 26, 10 (2025), 15068–15076. doi:10.1109/TITS.2025.3578383

[135] Dave Miller, Annabel Sun, and Wendy Ju. 2014. Situation awareness with different levels of automation. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 688–693. doi:10.1109/SMC.2014.6973989

[136] Alfred Santa Molison, Marcia Moraes, Glaucia Melo, Fabio Santos, and Fabio Assuncao. 2025. Is LLM-Generated Code More Maintainable & Reliable Than Human-Written Code? . In *2025 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE Computer Society, Los Alamitos, CA, USA, 151–162. doi:10.1109/ESEM64174.2025.00036

[137] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. 2018. Ray: a distributed framework for emerging AI applications. In *Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation* (Carlsbad, CA, USA) *(OSDI'18)*. USENIX Association, USA, 561–577.

[138] Antonio Muñoz, Ruben Ríos, Rodrigo Román, and Javier López. 2023. A survey on the (in)security of trusted execution environments. *Computers & Security* 129 (2023), 103180. doi:10.1016/j.cose.2023.103180

[139] Hoda Naghibijouybari, Ajaya Neupane, Zhiyun Qian, and Nael Abu-Ghazaleh. 2021. Side Channel Attacks on GPUs. *IEEE Transactions on Dependable and Secure Computing* 18, 4 (2021), 1950–1961. doi:10.1109/TDSC.2019.2944624

[140] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. 2023. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8602–8612.

[141] Giang Nguyen, Stefan Dlugolinsky, Martin Bobák, Viet Tran, Álvaro López García, Ignacio Heredia, Peter Malík, and Ladislav Hluch? 2019. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artif. Intell. Rev.* 52, 1 (June 2019), 77–124. doi:10.1007/s10462-018-09679-z

[142] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. 2025. Reason2Drive: Towards Interpretable and Chain-Based Reasoning for Autonomous Driving. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 292–308.

[143] Ali Nouri, Johan Andersson, Kailash De Jesus Hornig, Zhennan Fei, Emil Knabe, Hakan Sivencrona, Beatriz Cabrero-Daniel, and Christian Berger. 2025. On Simulation-Guided LLM-based Code Generation for Safe Autonomous Driving Software. In *Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering (EASE '25)*. Association for Computing Machinery, New York, NY, USA, 1097–1106. doi:10.1145/3756681.3756987

[144] NVIDIA Developer. 2026. DRIVE AGX Thor Developer Kit. https://developer.nvidia.com/drive/agx#section-thor-specifications. Accessed: 2026-01-22.

[145] Mohammad Omama, Pranav Inani, Pranjal Paul, Sarat Chandra Yellapragada, Krishna Murthy Jatavallabhula, Sandeep Chinchali, and Madhava Krishna. 2023. ALT-Pilot: Autonomous navigation with Language augmented Topometric maps. arXiv:2310.02324 [cs.RO] https://arxiv.org/abs/2310.02324

[146] OpenAI. 2023. ChatGPT 3.5. https://platform.openai.com/docs/models/gpt-3-5. Accessed: 2024-07-09.

[147] OpenAI. 2023. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf

[148] Malte Ostendorff, Pedro Ortiz Suarez, Lucas Fonseca Lage, and Georg Rehm. 2024. LLM-Datasets: An Open Framework for Pretraining Datasets of Large Language Models. In *First Conference on Language Modeling*. https://openreview.net/forum?id=5RdIMlGLXL

[149] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. 2024. VLP: Vision Language Planning for Autonomous Driving . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 14760–14769. doi:10.1109/CVPR52733.2024.01398

[150] Chenbin Pan, Burhaneddin Yaman, Senem Velipasalar, and Liu Ren. 2024. CLIP-BEVFormer: Enhancing Multi-View Image-Based BEV Detector with Ground Truth Flow . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 15216–15225. doi:10.1109/CVPR52733.2024.01441

[151] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Spatial as deep: spatial CNN for traffic scene understanding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Article 891, 8 pages.

[152] Yu Pan, Xiongfei Wu, Yuguang Yang, Jixun Yao, Lei Ma, and Jianjun Zhao. 2026. S2ST-Omni: Hierarchical Language-Aware SpeechLLM Adaptation for Multilingual Speech-to-Speech Translation. arXiv:2506.11160 [eess.AS] https://arxiv.org/abs/2506.11160

[153] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

[154] Ru Peng, Kexin Yang, Yawen Zeng, Junyang Lin, Dayiheng Liu, and Junbo Zhao. 2025. DataMan: Data Manager for Pre-training Large Language Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=eNbA8Fqir4

[155] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. 2023. OpenScene: 3D Scene Understanding with Open Vocabularies . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 815–824. doi:10.1109/CVPR52729.2023.00085

[156] Nenad Petrovic, Krzysztof Lebioda, Vahid Zolfaghari, André Schamschurko, Sven Kirchner, Nils Purschke, Fengjunjie Pan, and Alois Knoll. 2024. LLM-Driven Testing for Autonomous Driving Scenarios. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. 173–178. doi:10.1109/FLLM63129.2024.10852505

[157] Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nicholas Roy. 2023. Scenario Diffusion: Controllable Driving Scenario Generation With Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=99MHSB98yZ

[158] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. NuScenes-QA: A Multi-Modal Visual Question Answering Benchmark for Autonomous Driving Scenario. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 5 (Mar. 2024), 4542–4550. doi:10.1609/aaai.v38i5.28253

[159] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[160] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. arXiv:2112.11446 [cs.CL] https://arxiv.org/abs/2112.11446

[161] Dezhi Ran, Mengzhou Wu, Wei Yang, and Tao Xie. 2025. Foundation Model Engineering: Engineering Foundation Models Just as Engineering Software. *ACM Trans. Softw. Eng. Methodol.* 34, 5, Article 143 (May 2025), 18 pages. doi:10.1145/3719005

[162] Abhinav Sukumar Rao, Atharva Roshan Naik, Sachin Vashistha, Somak Aditya, and Monojit Choudhury. 2024. Tricking LLMs into Disobedience: Formalizing, Analyzing, and Detecting Jailbreaks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 16802–16830. https://aclanthology.org/2024.lrec-main.1462/

[163] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3505–3506. doi:10.1145/3394486.3406703

[164] Xiaoning Ren, Yun Lin, Yinxing Xue, Ruofan Liu, Jun Sun, Zhiyong Feng, and Jin Song Dong. 2023. DeepArc: Modularizing Neural Networks for the Model Maintenance. In *Proceedings of the 45th International Conference on Software Engineering* (Melbourne, Victoria, Australia) *(ICSE '23)*. IEEE Press, 1008–1019. doi:10.1109/ICSE48619.2023.00092

[165] Francisco Romero, Johann Hauswald, Aditi Partap, Daniel Kang, Matei Zaharia, and Christos Kozyrakis. 2022. Optimizing Video Analytics with Declarative Model Relationships. *Proc. VLDB Endow.* 16, 3 (Nov. 2022), 447–460. doi:10.14778/3570690.3570695

[166] Francisco Romero, Caleb Winston, Johann Hauswald, Matei Zaharia, and Christos Kozyrakis. 2023. Zelda: Video Analytics using Vision-Language Models. arXiv:2305.03785 [cs.DB] https://arxiv.org/abs/2305.03785

[167] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code Llama: Open Foundation Models for Code. arXiv:2308.12950 [cs.CL] https://arxiv.org/abs/2308.12950

[168] Jimin Ryu and Yong Ik Yoon. 2024. Designing Middleware for Ethical Decision-Making(EDM) in Autonomous Driving: Bias Detection Algorithms(BDA) for Enhanced Pedestrian Situation Awareness. In *Workshop Proceedings of the 53rd International Conference on Parallel Processing* (Gotland, Sweden) *(ICPP Workshops '24)*. Association for Computing Machinery, New York, NY, USA, 106–107. doi:10.1145/3677333.3678268

[169] Rajendramayavan Sathyam and Yueqi Li. 2025. Foundation Models for Autonomous Driving Perception: A Survey Through Core Capabilities. *IEEE Open Journal of Vehicular Technology* 6 (2025), 2554–2582. doi:10.1109/OJVT.2025.3604823

[170] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. 2023. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8579–8590.

[171] Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. 2023. LanguageMPC: Large Language Models as Decision Makers for Autonomous Driving. arXiv:2310.03026 [cs.RO] https://arxiv.org/abs/2310.03026

[172] Jing Shang, Jian Wang, Kailun Wang, Nan Jiang, and Jiqiang Liu. 2026. MSG: Stealing data from pruned neural networks via malicious sparsity guidance. *Neural Networks* 193 (2026), 108036. doi:10.1016/j.neunet.2025.108036

[173] Ankit Kumar Shaw, Kun Jiang, Tuopu Wen, Chandan Kumar Sah, Yining Shi, Mengmeng Yang, Diange Yang, and Xiaoli Lian. 2025. CleanMAP: Distilling Multimodal LLMs for Confidence-Driven Crowdsourced HD Map Updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3837–3846.

[174] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv preprint arXiv:1909.08053* (2019).

[175] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv:1909.08053 [cs.CL] https://arxiv.org/abs/1909.08053

[176] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. 2024. LUNA: A Model-Based Universal Analysis Framework for Large Language Models. *IEEE Transactions on Software Engineering* 50, 7 (2024), 1921–1948. doi:10.1109/TSE.2024.3411928

[177] Linke Song, Zixuan Pang, Wenhao Wang, Zihao Wang, XiaoFeng Wang, Hongbo Chen, Wei Song, Yier Jin, Dan Meng, and Rui Hou. 2025. The Early Bird Catches the Leak: Unveiling Timing Side Channels in LLM Serving Systems. *IEEE Transactions on Information Forensics and Security* 20 (2025), 11431–11446. doi:10.1109/TIFS.2025.3622954

[178] Qunying Song, He Ye, Mark Harman, and Federica Sarro. 2025. Generative AI for Testing of Autonomous Driving Systems: A Survey. arXiv:2508.19882 [cs.SE] https://arxiv.org/abs/2508.19882

[179] Zhihang Song, Zimin He, Xingyu Li, Qiming Ma, Ruibo Ming, Zhiqi Mao, Huaxin Pei, Lihui Peng, Jianming Hu, Danya Yao, and Yi Zhang. 2024. Synthetic Datasets for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Vehicles* 9, 1 (2024), 1847–1864. doi:10.1109/TIV.2023.3331024

[180] N. N. Sriram, Tirth Maniar, Jayaganesh Kalyanasundaram, Vineet Gandhi, Brojeshwar Bhowmick, and K Madhava Krishna. 2019. Talk to the Vehicle: Language Conditioned Autonomous Navigation of Self Driving Cars. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5284–5290. doi:10.1109/IROS40897.2019.8967929

[181] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=kmn0BhQk7p

[182] Jingwei Sun, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yudong Liu, Zhixu Du, Yiran Chen, and Holger R. Roth. 2024. FedBPT: Efficient Federated Black-box Prompt Tuning for Large Language Models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. https://openreview.net/forum?id=AoYhtJ4A90

[183] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2024. A Simple and Effective Pruning Approach for Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=PxoFut3dWW

[184] Shiliang Sun, Zhilin Lin, and Xuhan Wu. 2025. Hallucinations of large multimodal models: Problem and countermeasures. *Information Fusion* 118 (2025), 102970. doi:10.1016/j.inffus.2025.102970

[185] Emil Talpes, Debjit Das Sarma, Ganesh Venkataramanan, Peter Bannon, Bill McGee, Benjamin Floering, Ankit Jalote, Christopher Hsiong, Sahil Arora, Atchyuth Gorti, and Gagandeep S. Sachdev. 2020. Compute Solution for Tesla's Full Self-Driving Computer. *IEEE Micro* 40, 2 (2020), 25–35. doi:10.1109/MM.2020.2975764

[186] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. 2021. SceneGen: Learning To Generate Realistic Traffic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 892–901.

[187] Shuncheng Tang, Zhenya Zhang, Yi Zhang, Jixiang Zhou, Yan Guo, Shuang Liu, Shengjian Guo, Yan-Fu Li, Lei Ma, Yinxing Xue, and Yang Liu. 2023. A Survey on Automated Driving System Testing: Landscapes and Trends. *ACM Trans. Softw. Eng. Methodol.* 32, 5, Article 124 (July 2023), 62 pages. doi:10.1145/3579642

[188] Shuncheng Tang, Zhenya Zhang, Jixiang Zhou, Lei Lei, Yuan Zhou, and Yinxing Xue. 2024. LeGEND: A Top-Down Approach to Scenario Generation of Autonomous Driving Systems Assisted by Large Language Models. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Sacramento, CA, USA) *(ASE '24)*. Association for Computing Machinery, New York, NY, USA, 1497–1508. doi:10.1145/3691620.3695520

[189] Wenbing Tang, Mingfei Cheng, Yuan Zhou, and Yang Liu. 2025. Moral Testing of Autonomous Driving Systems . In *2025 IEEE/ACM 1st International Workshop on Software Engineering for Autonomous Driving Systems (SE4ADS)*. IEEE Computer Society, Los Alamitos, CA, USA, 26–30. doi:10.1109/SE4ADS66461.2025.00011

[190] Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2024. Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=oZtt0pRnOl

[191] Hanlin Tian, Kethan Reddy, Yuxiang Feng, Mohammed Quddus, Yiannis Demiris, and Panagiotis Angeloudis. 2026. Large (Vision) Language Models for Autonomous Vehicles: Current Trends and Future Directions. *IEEE Transactions on Intelligent Transportation Systems* 27, 1 (2026), 187–210. doi:10.1109/TITS.2025.3628969

[192] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. 2024. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. In *8th Annual Conference on Robot Learning*. https://openreview.net/forum?id=928V4Umlys

[193] Daniil Tiapkin, Daniele Calandriello, Johan Ferret, Sarah Perrin, Nino Vieillard, Alexandre Rame, and Mathieu Blondel. 2025. On Teacher Hacking in Language Model Distillation. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=qxSFIigPug

[194] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971

[195] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. 2024. Position: Considerations for Differentially Private Learning with Large-Scale Public Pretraining. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=ncjhi4qAPV

[196] Sifan Tu, Xin Zhou, Dingkang Liang, Xingyu Jiang, Yumeng Zhang, Xiaofan Li, and Xiang Bai. 2025. The Role of World Models in Shaping Autonomous Driving: A Comprehensive Survey. arXiv:2502.10498 [cs.CV] https://arxiv.org/abs/2502.10498

[197] Ishparsh Uprety and Xinghui Zhao. 2025. Edge-Deployable LLMs for Autonomous Vehicle Intelligence. In *Proceedings of the Tenth ACM/IEEE Symposium on Edge Computing* (the Hilton Arlington National Landing, Arlington, VA, USA) *(SEC '25)*. Association for Computing Machinery, New York, NY, USA, Article 73, 7 pages. doi:10.1145/3769102.3774639

[198] Christopher Vendome, Mario Linares-Vásquez, Gabriele Bavota, Massimiliano Di Penta, Daniel German, and Denys Poshyvanyk. 2017. Machine Learning-Based Detection of Open Source License Exceptions. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*. 118–129. doi:10.1109/ICSE.2017.19

[199] Jiannan Wang, Hung Viet Pham, Qi Li, Lin Tan, Yu Guo, Adnan Aziz, and Erik Meijer. 2025. D3: Differential Testing of Distributed Deep Learning With Model Generation. *IEEE Transactions on Software Engineering* 51, 1 (2025), 38–52. doi:10.1109/TSE.2024.3461657

[200] Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Casas, Mengye Ren, and Raquel Urtasun. 2021. AdvSim: Generating Safety-Critical Scenarios for Self-Driving Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9909–9918.

[201] Jiachen T. Wang, Tong Wu, Dawn Song, Prateek Mittal, and Ruoxi Jia. 2024. GREATS: Online Selection of High-Quality Data for LLM Training in Every Iteration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=232VcN8tSx

[202] Pengqin Wang, Meixin Zhu, Xinhu Zheng, Hongliang Lu, Hui Zhong, Xianda Chen, Shaojie Shen, Xuesong Wang, Yinhai Wang, and Fei-Yue Wang. 2024. BEVGPT: Generative Pre-trained Foundation Model for Autonomous Driving Prediction, Decision-Making, and Planning. *IEEE Transactions on Intelligent Vehicles* (2024), 1–13. doi:10.1109/TIV.2024.3449278

[203] Shenao Wang, Yanjie Zhao, Xinyi Hou, and Haoyu Wang. 2024. Large Language Model Supply Chain: A Research Agenda. *ACM Trans. Softw. Eng. Methodol.* (Dec. 2024). doi:10.1145/3708531 Just Accepted.

[204] Xiaonan Wang and Jiajia Xu. 2026. Secure traffic accident warning system with privacy support for autonomous driving. *Reliability Engineering & System Safety* 266 (2026), 111803. doi:10.1016/j.ress.2025.111803

[205] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. 2025. DriveDreamer: Towards Real-World-Drive World Models for Autonomous Driving. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 55–72.

[206] Yuping Wang, Shuo Xing, Cui Can, Renjie Li, Hongyuan Hua, Kexin Tian, Zhaobin Mo, Xiangbo Gao, Keshu Wu, Sulong Zhou, Hengxu You, Juntong Peng, Junge Zhang, Zehao Wang, Rui Song, Mingxuan Yan, Walter Zimmer, Xingcheng Zhou, Peiran Li, Zhaohan Lu, Chia-Ju Chen, Yue Huang, Ryan A. Rossi, Lichao Sun, Hongkai Yu, Zhiwen Fan, Frank Hao Yang, Yuhao Kang, Ross Greer, Chenxi Liu, Eun Hak Lee, Xuan Di, Xinyue Ye, Liu Ren, Alois Knoll, Xiaopeng Li, Shuiwang Ji, Masayoshi Tomizuka, Marco Pavone, Tianbao Yang, Jing Du, Ming-Hsuan Yang, Hua Wei, Ziran Wang, Yang Zhou, Jiachen Li, and Zhengzhong Tu. 2025. Generative AI for Autonomous Driving: Frontiers and Opportunities. arXiv:2505.08854 [cs.CV] https://arxiv.org/abs/2505.08854

[207] Zige Wang, Wanjun Zhong, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Data Management For Training Large Language Models: A Survey. arXiv:2312.01700 [cs.CL] https://arxiv.org/abs/2312.01700

[208] Anjiang Wei, Yinlin Deng, Chenyuan Yang, and Lingming Zhang. 2022. Free lunch for testing: fuzzing deep-learning libraries from open source. In *Proceedings of the 44th International Conference on Software Engineering* (Pittsburgh, Pennsylvania) *(ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 995–1007. doi:10.1145/3510003.3510041

[209] Dafeng Wei, Tian Gao, Zhengyu Jia, Changwei Cai, Chengkai Hou, Peng Jia, Fu Liu, Kun Zhan, Jingchen Fan, Yixing Zhao, and Yang Wang. 2024. BEV-CLIP: Multi-modal BEV Retrieval Methodology for Complex Scene in Autonomous Driving. *CoRR* abs/2401.01065 (2024). doi:10.48550/ARXIV.2401.01065 arXiv:2401.01065

[210] Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. 2024. Magicoder: Empowering Code Generation with OSS-Instruct. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 52632–52657. https://proceedings.mlr.press/v235/wei24h.html

[211] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. 2024. Editable Scene Simulation for Autonomous Driving via Collaborative LLM-Agents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 15077–15087. doi:10.1109/CVPR52733.2024.01428

[212] Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. QuRating: selecting high-quality data for training language models. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2171, 57 pages.

[213] Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2024. Fundamental limitations of alignment in large language models. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2176, 34 pages.

[214] Thomas Wolter, Ann Barcomb, Dirk Riehle, and Nikolay Harutyunyan. 2023. Open Source License Inconsistencies on GitHub. *ACM Trans. Softw. Eng. Methodol.* 32, 5, Article 110 (July 2023), 23 pages. doi:10.1145/3571852

[215] Guanlong Wu, Zheng Zhang, Yao Zhang, Weili Wang, Jianyu Niu, Ye Wu, and Yinqian Zhang. 2025. I Know What You Asked: Prompt Leakage via KV-Cache Sharing in Multi-Tenant LLM Serving. In *32nd Annual Network and Distributed System Security Symposium, NDSS 2025, San Diego, California, USA, February 2025*. The Internet Society. https://www.ndss-symposium.org/ndss-paper/i-know-what-you-asked-prompt-leakage-via-kv-cache-sharing-in-multi-tenant-llm-serving/

[216] Jianhua Wu, Bingzhao Gao, Jincheng Gao, Jianhao Yu, Hongqing Chu, Qiankun Yu, Xun Gong, Yi Chang, H. Eric Tseng, Hong Chen, and Jie Chen. 2024. Prospective Role of Foundation Models in Advancing Autonomous Vehicles. *Research* 7 (2024), 0399. doi:10.34133/research.0399 arXiv:https://spj.science.org/doi/pdf/10.34133/research.0399

[217] Yaozu Wu, Dongyuan Li, Yankai Chen, Renhe Jiang, Henry Peng Zou, Wei-Chieh Huang, Yangning Li, Liancheng Fang, Zhen Wang, and Philip S. Yu. 2025. Multi-Agent Autonomous Driving Systems with Large Language Models: A Survey of Recent Advances, Resources, and Future Directions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 12756–12773. doi:10.18653/v1/2025.findings-emnlp.683

[218] Zehao Wu, Yanjie Zhao, and Haoyu Wang. 2025. Gradient-Based Model Fingerprinting for LLM Similarity Detection and Family Classification. arXiv:2506.01631 [cs.LG] https://arxiv.org/abs/2506.01631

[219] Haocheng Xi, Han Cai, Ligeng Zhu, Yao Lu, Kurt Keutzer, Jianfei Chen, and Song Han. 2025. COAT: Compressing Optimizer states and Activations for Memory-Efficient FP8 Training. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=XfKSDgqIRj

[220] Tianao Xiang, Yuanguo Bi, Mingjian Zhi, and Lin Cai. 2025. FLAD: Federated-Trained Large Language Models for Autonomous Driving. *IEEE Network* (2025), 1–7. doi:10.1109/MNET.2025.3634409

[221] Dongwei Xiao, Zhibo LIU, Yuanyuan Yuan, Qi Pang, and Shuai Wang. 2022. Metamorphic Testing of Deep Learning Compilers. *Proc. ACM Meas. Anal. Comput. Syst.* 6, 1, Article 15 (Feb. 2022), 28 pages. doi:10.1145/3508035

[222] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. 2025. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data and Metric Perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6585–6597.

[223] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. 2023. Machine Unlearning: A Survey. *ACM Comput. Surv.* 56, 1, Article 9 (Aug. 2023), 36 pages. doi:10.1145/3603620

[224] Mengwei Xu, Dongqi Cai, Yaozong Wu, Xiang Li, and Shangguang Wang. 2024. FwdLLM: Efficient Federated Finetuning of Large Language Models with Perturbed Inferences. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*. USENIX Association, Santa Clara, CA, 579–596. https://www.usenix.org/conference/atc24/presentation/xu-mengwei

[225] Weiwei Xu, Kai Gao, Hao He, and Minghui Zhou. 2025. LiCoEval: Evaluating LLMs on License Compliance in Code Generation. In *Proceedings of the 47th IEEE/ACM International Conference on Software Engineering, ICSE 2025*.

[226] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. 2024. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. *IEEE Robotics and Automation Letters* 9, 10 (2024), 8186–8193. doi:10.1109/LRA.2024.3440097

[227] Xu Yan, Haiming Zhang, Yingjie Cai, Jingming Guo, Weichao Qiu, Bin Gao, Kaiqiang Zhou, Yue Zhao, Huan Jin, Jiantao Gao, Zhen Li, Lihui Jiang, Wei Zhang, Hongbo Zhang, Dengxin Dai, and Bingbing Liu. 2024. Forging Vision Foundation Models for Autonomous Driving: Challenges, Methodologies, and Opportunities. arXiv:2401.08045 [cs.CV]

[228] Borui Yang, Md Afif Al Mamun, Jie M. Zhang, and Gias Uddin. 2025. Hallucination Detection in Large Language Models with Metamorphic Relations. *Proc. ACM Softw. Eng.* 2, FSE, Article FSE020 (June 2025), 21 pages. doi:10.1145/3715735

[229] Haomiao Yang, Dongyun Xue, Tianyi Wang, Jin Kwak, and Hyunsung Kim. 2025. FedGPL: Gradient Priority-based Federated Learning Enhancement for In-Vehicle Language Models. *ACM Trans. Auton. Adapt. Syst.* (Feb. 2025). doi:10.1145/3717835 Just Accepted.

[230] Kaiyuan Yang, Matthew Hicks, Qing Dong, Todd Austin, and Dennis Sylvester. 2016. A2: Analog Malicious Hardware. In *2016 IEEE Symposium on Security and Privacy (SP)*. 18–37. doi:10.1109/SP.2016.10

[231] Yulong Yang, Chenhao Lin, Qian Li, Zhengyu Zhao, Haoran Fan, Dawei Zhou, Nannan Wang, Tongliang Liu, and Chao Shen. 2024. Quantization Aware Attack: Enhancing Transferable Adversarial Attacks by Model Quantization. *Trans. Info. For. Sec.* 19 (Jan. 2024), 3265–3278. doi:10.1109/TIFS.2024.3360891

[232] Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. 2024. Human-Centric Autonomous Systems With LLMs for User Command Reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 988–994.

[233] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. 2024. LLM4Drive: A Survey of Large Language Models for Autonomous Driving. In *NeurIPS 2024 Workshop on Open-World Agents*. https://openreview.net/forum?id=ehojTglbMj

[234] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2024. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. arXiv:2310.01469 [cs.CL] https://arxiv.org/abs/2310.01469

[235] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing* 4, 2 (2024), 100211. doi:10.1016/j.hcc.2024.100211

[236] Chengye Yu, Tianyu Wang, Zili Shao, Linjie Zhu, Xu Zhou, and Song Jiang. 2024. TwinPilots: A New Computing Paradigm for GPU-CPU Parallel LLM Inference. In *Proceedings of the 17th ACM International Systems and Storage Conference* (Virtual, Israel) *(SYSTOR '24)*. Association for Computing Machinery, New York, NY, USA, 91–103. doi:10.1145/3688351.3689164

[237] Xiao Yu, Haoxuan Chen, Feifei Niu, Xing Hu, Jacky Wai Keung, and Xin Xia. 2025. Towards Understanding Bugs in Distributed Training and Inference Frameworks for Large Language Models. arXiv:2506.10426 [cs.SE] https://arxiv.org/abs/2506.10426

[238] Chenkai Zeng, Debiao He, Qi Feng, Xiaolin Yang, and Qingcai Luo. 2025. EPAuto: Efficient Privacy-Preserving Machine Learning on AI-powered Autonomous Driving Systems using Multi-Party Computation. *ACM Trans. Auton. Adapt. Syst.* (Feb. 2025). doi:10.1145/3718743 Just Accepted.

[239] Dayong Zhang, Hao Men, and Zhaoxin Zhang. 2024. Assessing the stability of collaboration networks: A structural cohesion analysis perspective. *Journal of Informetrics* 18, 1 (2024), 101490. doi:10.1016/j.joi.2024.101490

[240] Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. 2024. Towards Building The Federatedgpt: Federated Instruction Tuning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6915–6919. doi:10.1109/ICASSP48485.2024.10447454

[241] Jiawei Zhang, Chejian Xu, and Bo Li. 2024. ChatScene: Knowledge-Enabled Safety-Critical Scenario Generation for Autonomous Vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 15459–15469. doi:10.1109/CVPR52733.2024.01464

[242] Xiaoyu Zhang, Weipeng Jiang, Chao Shen, Qi Li, Qian Wang, Chenhao Lin, and Xiaohong Guan. 2025. Deep Learning Library Testing: Definition, Methods and Challenges. *ACM Comput. Surv.* 57, 7, Article 187 (March 2025), 37 pages. doi:10.1145/3716497

[243] Yedi Zhang, Lei Huang, Pengfei Gao, Fu Song, Jun Sun, and Jin Song Dong. 2025. Verification of Bit-Flip Attacks against Quantized Neural Networks. *Proc. ACM Program. Lang.* 9, OOPSLA1, Article 115 (April 2025), 31 pages. doi:10.1145/3720471

[244] Li Zhong, Ahmed Ghazal, Jun-Jun Wan, Frederik Zilly, Patrick Mackens, Joachim Vollrath, and Bogdan Coseriu. 2025. Clip4Retrofit: Enabling Real-Time Image Labeling on Edge Devices via Cross-Architecture CLIP Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3868–3876.

[245] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. 2024. Vision Language Models in Autonomous Driving: A Survey and Outlook. *IEEE Transactions on Intelligent Vehicles* (2024), 1–20. doi:10.1109/TIV.2024.3402136

[246] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=oZDJKTlOUe

[247] Zhaodong Zhou and Jun Chen. 2025. Privacy-Preserving Personalized Autonomous Vehicle Lane Change Using Inverse Reinforcement Learning. *IEEE Transactions on Vehicular Technology* (2025), 1–13. doi:10.1109/TVT.2025.3617416

[248] Derui Zhu, Dingfan Chen, Xiongfei Wu, Jiahui Geng, Zhuo Li, Jens Grossklags, and Lei Ma. 2024. PrivAuditor: Benchmarking Data Protection Vulnerabilities in LLM Adaptation Techniques. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=VpkfxuVXwx

[249] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=1tZbq88f27

[250] Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. 2024. AutoDAN: Interpretable Gradient-Based Adversarial Attacks on Large Language Models. In *First Conference on Language Modeling*. https://openreview.net/forum?id=INivcBeIDK

[251] Yuxuan Zhu, Shiyi Wang, Wenqing Zhong, Nianchen Shen, Yunqi Li, Siqi Wang, Zhiheng Li, Cathy Wu, Zhengbing He, and Li Li. 2025. A Survey on Large Language Model-Powered Autonomous Driving. *Engineering* (2025). doi:10.1016/j.eng.2025.07.038

[252] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043 [cs.CL]