The Review on Development of Large Language Models for Autonomous Vehicles

Sicheng Wang*

Jiangyang No. 2 School, Taizhou, Jiangsu, China

* 15850851970@139.com

Abstract

Autonomous driving systems rely heavily on visual inputs. The Large Language Model mainly trains this. A large language model is a deep-learning model trained with a lot of text data. It is one of the most important models for operating AI systems and their learning and training work. Autonomous vehicles are an emerging new operating system used for driving in recent decades, and large language models (LLMs) have demonstrated abilities including understanding context, logical reasoning, and generating answers. A natural thought is to utilize these abilities to empower autonomous driving. Recent advances in Visual Language Models (VLMs) drive a paradigm shift in autonomous driving research. The field is transitioning from training perception and policy models from scratch toward adopting a new "pre-training + fine-tuning" paradigm. Using a large language model on cars can help the AI system drive and provide car owners with a safer, faster, and more comfortable service. Thus, autonomous driving technology, a catalyst for revolutionizing transportation and urban mobility, tends to transition from rule-based systems to data-driven strategies.

Keywords

Large language model, autonomous driving, AI training systems, automation, humanless economy.

1. Introduction

After decades of development, the large language model has become one of the basic training models for AI systems. The AI system can judge and operate based on the training results by analyzing the massive text data. As one of the derivatizations of AI products, autonomous vehicles must use large language models to support their ability to explore the environment and make driving judgments. Large language models can help the development of autonomous cars and provide a better future of automation and a humanless economy. Autonomous vehicles face problems because they lack training and cannot judge emergencies. Then, a large language model can train the AI system and help it solve emergency changes. Autonomous vehicles must simultaneously perceive complex three-dimensional scenes, understand traffic context, and act safely in real time. As a more natural interaction modality, voice input began to be incorporated, bridging perception and interaction through voice-driven interfaces. This evolution-from static commands to dialogue-driven multi-step reasoning-reflects a broader trend: language is not just a tool for vehicle control but also the key to achieving interpretable and collaborative autonomy. Beyond employing memory and reflection mechanisms, large language models are now integrated with raw sensor data-such as camera feeds, Global Navigation Satellite Systems (GNSS), LiDAR, and millimeter-wave radar-to train AI systems. This multimodal approach enhances the model's environmental perception and decision-making robustness by grounding linguistic reasoning in real-world physical signals. Moreover, visual language models help autonomous cars drive, as VLMs bridge the gap between vision and language through aligned

textual semantics, enabling models to comprehend contextual relationships within complex visual scenes. More importantly, VLMs endow autonomous driving systems with cognitive capabilities that more closely resemble human reasoning, which is crucial for handling complex and dynamic traffic scenarios[1-10].

2. Development of Autonomous Vehicles and Large Language Models

Against current technological advancements, large language models represented by GPT-4 and GPT-4V have garnered significant attention due to their exceptional contextual understanding and in-context learning capabilities. Numerous approaches have emerged regarding applying large language models (LLMs) in the autonomous driving domain. The rich commonsense knowledge systems embedded in these models have already driven substantial breakthroughs in numerous downstream tasks. This raises a critical question: How can such large models empower the autonomous driving domain, particularly in playing a pivotal role in decision-making? The technological evolution of autonomous driving can be divided into four main paradigms: modular architecture, end-to-end learning, VLM4AD, and the recent wave of VLA4AD. Early attempts passively utilized language to enhance interpretability. The automatic driving technique has become increasingly important in driving systems, primarily using AI and ML algorithms to take dynamic accommodation in complex environments. In October 2011, the first real-environment test of autonomous vehicles occurred in the Mojave Desert in Western America. On 1 March 2012, the first law of autonomous vehicles took effect. In 2017, Alphabus was used to run in Futian, Shenzhen, China. This is the first time that autonomous cars have been used in business[11-14].

Autonomous driving approaches based on large language models (LLMs) primarily follow two technical pathways: direct control generation (identifying textual descriptions of objects to output driving operation commands directly) and scene interpretation (feeding driving scenario images into LLMs to generate environmental state descriptions). The large language model's rapid development showed its strong ability to deal with natural language, information generation, and decision support. The developments of OpenAI and DeepSeek-R1 have demonstrated that large language models can help AI systems analyze text data and develop their logical reasoning ability[15-17].

Another significant development for autonomous driving is the question-answering system. Question-answering systems are critical technologies with significant application value in intelligent transportation, driver assistance, and autonomous driving domains. Their implementations primarily manifest in two paradigms: traditional QA mechanisms and more refined visual question-answering (VQA) approaches. Applying large language models (LLMs) to autonomous driving is limited by their ability to comprehend and adhere to traffic regulations. While traditional methods rely on complex rule-based systems to achieve regulatory compliance, LLMs enable two innovative paradigms: explicitly incorporating regulations through prompt engineering (in-context learning) or directly training the models on legal frameworks[18-20].

3. Limitations of Autonomous Vehicles

Despite their advantages, autonomous vehicles continue to face serious limitations, with traffic accidents serving as reminders of unresolved challenges. One core issue is latency: frame-by-frame generation of long textual outputs requires visual encoders to process thousands of tokens per image, introducing delays in decision-making. At the same time, general-purpose visual encoders expend computational resources on irrelevant details, undermining efficiency. These factors contribute to delayed responses in predicting emergent behaviors and reduced reasoning efficiency in complex interactions-failures that can directly cause accidents. For

example, when vehicles misinterpret lighting conditions, occlusions, or missing lane markings, they may fail to recognize obstacles, incorrectly assume a clear path, or neglect to trigger emergency braking in time. Such failures illustrate the inability of current models to handle edge cases and long-tail scenarios.

Another major limitation is domain shift. Variations in weather, lighting, road quality, and object appearance mean that models trained on general-purpose data or small, biased driving datasets cannot reliably generalize to real-world driving. This mismatch between training data and real deployment conditions often leads to poor performance in unusual but safety-critical situations. Moreover, VLM-based systems remain largely passive: they can describe or reason about a scene but are weakly coupled to low-level control, sometimes producing hazardous outputs or misinterpreting natural language instructions.

Efforts to improve robustness through simulation-based training also face barriers, as synthetic platforms cannot yet replicate the full visual realism and behavioral unpredictability of the real world, leaving a significant "simulation-to-reality gap." Standardization further compounds the problem: there is currently no unified benchmark for evaluating large language models in driving tasks, making it difficult to compare them to traditional approaches. Finally, while some progress has been made in trajectory prediction and interaction understanding, current systems still depend heavily on first-person annotated datasets, and they lack effective methods for handling abstract, information-dense bird's-eye-view traffic scenarios[21-24].

4. Large Language Models-Trained AI System

Recent breakthroughs in autonomous driving research have highlighted how large language models (LLMs) can enhance complex scenario handling, multimodal fusion, and reasoning ability, leading to the development of several integrated paradigms. One representative approach is the Bird's-Eye View reasoning system, exemplified by PlanAgent, which combines BEV representation extraction with lane map textualization. This architecture employs a hierarchical chain-of-thought process-progressing from scene understanding to motion instruction generation and ultimately to planning code writing-demonstrating superior closed-loop performance in the demanding nuPlan benchmark. Beyond single-vehicle reasoning, researchers have also proposed multi-agent collaboration frameworks such as AGENTSCODRIVER, which integrates a reasoning engine, cognitive memory, reinforced reflection, and a communication protocol. This design enables coordinated decision-making among multiple vehicles in complex traffic scenarios and supports lifelong learning, allowing capabilities to evolve over time. Another important advance is the vision–language fusion paradigm, as seen in DriveVLM's dual-mode architecture, which features a central system for VLM-enhanced scene understanding and planning, complemented by an auxiliary system that coordinates traditional three-dimensional perception and planning. This approach achieves breakthroughs in dynamic scene spatial reasoning and significantly improves computational efficiency.

Retrieval-augmented models such as RAG-Driver extend these advances by introducing interpretable end-to-end control, causal reasoning for driving behaviors, and zero-shot generalization to new environments. Similarly, training-free adaptation mechanisms, exemplified by LLaDA, introduce a human-assisted driving mode alongside self-adaptive strategy optimization, enabling systems to adjust seamlessly to novel traffic conditions without additional training. Finally, unified vision–language–planning models such as VLP integrate dual learning paradigms, including ALP for BEV generation aligned with real maps and SLP for aligning ego-vehicle query features with textual planning features. These LLM-driven paradigms demonstrate how natural language reasoning, multimodal representation, and

adaptive control can be synergistically combined to address the central challenges of autonomous driving[25-28].

5. Vision–Language–Action Models for Autonomous Driving

Recent advances in autonomous driving have shown how vision–language–action (VLA) models reshape the field by integrating perception, reasoning, and control into unified architectures. Systems such as DME-Drive demonstrate the potential of this approach by combining vision–language decision models with planning-oriented perception mechanisms and high-precision planning trained on the HBD dataset. These innovations mark a decisive shift from modular pipelines toward integrated intelligent driving systems capable of addressing critical challenges of generalization, interpretability, and adaptability to real-world environments. VLA paradigms unify camera streams, natural language instructions, and low-level control policies into a single framework, although they continue to struggle with edge cases, such as generating feasible trajectories when encountering disabled vehicles at intersections.

Key research has focused on resolving these limitations through three primary directions: enhanced spatial understanding, latency reduction, and mitigation of hallucinations. Object-centric tokenization methods such as TOKEN and WKAD improve scene representation, while models like BEVDriver combine BEV features with language queries to enable three-dimensional spatial reasoning and multimodal trajectory prediction. Other approaches, including See2DriveX and MPDrive, model spatial relationships through relational graphs or visual prompting to strengthen spatial reasoning. To reduce system latency, dual-architecture designs employ vision–language models as intermediate modules that provide auxiliary feedback to end-to-end systems, while knowledge distillation methods transfer VLM capabilities offline into lighter traditional networks. Mitigating hallucinations has become equally important: in-context learning methods such as Dilu use memory banks to preserve critical driving knowledge, ReasonPlan generates explicit step-by-step rationales to increase interpretability, and agent-style models like AgentDriver enhance robustness with tool-augmented chain-of-thought prompting.

These research directions are already influencing industry practice. Companies such as Li Auto have adopted VLA frameworks like MindVLA, which integrate three-dimensional spatial intelligence, natural language reasoning, and control into unified policies that enable real-time decision-making while maintaining strict safety constraints. Training strategies for VLA4AD systems typically follow a four-stage curriculum of pre-training, alignment, targeted enhancement, and compression. Yet challenges remain, as language reasoning enhances contextual understanding and introduces new failure modes: LLMs may hallucinate unsafe maneuvers or misinterpret colloquial instructions such as “step on the gas.” Maintaining robustness under sensor degradation from rain, snow, or glare further complicates deployment. Logic-based safety overrides such as SafeAuto are early attempts to ensure constraint compliance, but formal verification frameworks and socially compliant driving strategies remain pressing open problems[29].

Recent representative systems demonstrate the breadth of this paradigm by anchoring free-form instructions to the ego-vehicle’s visual context, generating corresponding trajectories, and outputting chain-of-thought explanations through models such as DriveCoT and CoT-VLA. These systems go beyond direct control token prediction by integrating diffusion-based planning heads, hierarchical reasoning planners, and hybrid discrete–continuous control strategies. Collectively, such advances signal a transition from perception-centric pipelines toward action-aware, interpretable, and instruction-following multimodal agents. Inspired by broader progress in artificial intelligence, the convergence of vision, language, and action has

emerged as a central trend in autonomous driving, driven by three practical requirements of real-world deployment: robust reasoning in rare and long-tail scenarios, noise-tolerant control under dynamic and partially occluded conditions, and the capacity to interpret spontaneous high-level language instructions such as “overtake the truck.”

6. Conclusion

The ultimate goal of autonomous driving is to elevate driving capabilities from foundational to expert-level through large-scale data collection and deep learning. Nowadays, autonomous vehicles have become a part of our real life. The automatic car has 6 levels, and the recent technique is the L3 type, which ensures that cars can drive by their own system; however, it still needs drivers’ help. With the training of a large language model, the AI system will have a stronger ability to judge the emergency, and the higher-level AI cars will come to reality. At that time, we will have a safer and more comfortable driving experience when we take autonomous vehicles. In summary, these output modalities demonstrate the evolutionary objectives of VLA4AD systems: achieving not just driving functionality, but accomplishing it in a robust, interpretable, and context-aware manner.

Fundamentally, achieving these goals requires advances in large-scale multimodal learning, formal verification, communication standards, and human-AI interaction. Success will yield a versatile "driving brain" capable of rapid adaptation, safety auditing, and seamless integration into the global transportation ecosystem. It is imperative to note that significant challenges remain in enabling LLMs to comprehend complex scenario information. Substantial exploration space still exists for integrating rich contextual information with motion prediction models. Current research from various countries worldwide demonstrates this field's remarkable development potential.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022. 3

[2] Florent Altche and Arnaud de La Fortelle. An LSTM network for highway trajectory prediction. In 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pages 353–359. IEEE, 2017. 2

[3] Florent Bartoccioni, Elias Ramzi, Victor Besnier, Shashanka Venkataramanan, Tuan-Hung Vu, Yihong Xu, Loick Chambon, Spyros Gidaris, Serkan Odabas, David Hurych, et al. Vavim and vavam: Autonomous driving through video generative modeling. arXiv preprint arXiv:2502.15672, 2025. 5, 8

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024. 4

[5] Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. Fine-grained affective processing capabilities emerging from large language models. In the 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8. IEEE, 2023. 4

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020. 5

[7] Holger Caesar, Alex Bankiti, Orien Lang, et al. nuScenes: A multimodal dataset for autonomous driving. In CVPR, 2020. 2, 7, 9

[8] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11621–11631, 2020. 4

[9] Dian Chen and Philipp Krahenbuhl. Learning from all vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17222–17231, 2022. 2

[10] Dianwei Chen, Zifan Zhang, Yuchen Liu, and Xianfeng Terry Yang. Insight: Enhancing autonomous driving safety through vision-language models on context-aware hazard detection and edge case evaluation. arXiv e-prints, pages arXiv–2502, 2025. 4

[11] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as I can, not as I say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691, 2022.

[12] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99:101805, 2023.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[14] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving. arXiv preprint arXiv:2310.01957, 2023.

[15] J. Mao, J., Ye, Y., Qian, M., Pavone, M., and Y. Wang, "A language agent for autonomous driving," arXiv preprint arXiv:2311.10813, 2023.

[16] Y. Jin, X. Shen, H. Peng, X. Liu, J. Qin, J. Li, J. Xie, P. Gao, G. Zhou, and J. Gong, "Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model," arXiv preprint arXiv:2309.13193, 2023.

[17] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," arXiv preprint arXiv:2312.14150, 2023.

[18] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, and D. Rus, "Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models," arXiv preprint arXiv:2310.17642, 2023.

[19] D. Wu, W. Han, T. Wang, Y. Liu, X. Zhang, and J. Shen, "Language prompt for autonomous driving," arXiv preprint arXiv:2309.04379, 2023.

[20] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario." arXiv preprint arXiv:2305.14836, 2023.

[21] M. Nie, R., Peng, C., Wang, X., Cai, J., Han, H., Xu, H., and L. Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving." arXiv preprint arXiv:2312.03661, 2023.

[22] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15 793–15 803.

[23] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving, 2023.

[24] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13668–13677, 2024.

[25] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint,

Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

[26] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles Qi, Yin Zhou, Zoey Yang, Aurelien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large-scale interactive motion forecasting for autonomous driving: The Waymo open motion dataset, 2021.

[27] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic Nuscene: A large-scale benchmark for lidar panoptic segmentation and tracking. arXiv preprint arXiv:2109.03805, 2021.

[28] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models, 2023.

[29] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire, 2023.