

A Survey of Large Language Models for Autonomous Driving

Zhenjie Yang^{1*}, Xiaosong Jia^{1*}, Hongyang Li^{1,2}, Junchi Yan^{1†}

¹ Department of Computer Science and Engineering, Shanghai Jiao Tong University

² OpenDriveLab

{yangzhenjie, jiaxiaosong, hongyangli, yanjunchi}@sjtu.edu.cn

* Equal contributions, order decided by a coin toss †Correspondence author

Abstract

Autonomous driving technology, a catalyst for revolutionizing transportation and urban mobility, has the tend to transition from rule-based systems to data-driven strategies. Traditional module-based systems are constrained by cumulative errors among cascaded modules and inflexible pre-set rules. In contrast, end-to-end autonomous driving systems have the potential to avoid error accumulation due to their fully data-driven training process, although they often lack transparency due to their “black box” nature, complicating the validation and traceability of decisions. Recently, 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. By combining LLM with foundation vision models, it could open the door to open-world understanding, reasoning, and few-shot learning, which current autonomous driving systems are lacking. In this paper, we systematically review a research line about *Large Language Models for Autonomous Driving (LLM4AD)*. This study evaluates the current state of technological advancements, distinctly outlining the principal challenges and prospective directions for the field. For the convenience of researchers in academia and industry, we provide real-time updates on the latest advances in the field as well as relevant open-source resources via the designated link: <https://github.com/Thinklab-SJTU/Awesome-LLM4AD>.

1 Introduction

Autonomous driving is rapidly reshaping our understanding of transportation, heralding a new era of technological revolution. This transformation means not only the future of transportation but also a fundamental shift across various industries. In conventional autonomous driving systems, algorithms typically adopt the modular design [Liang *et al.*, 2020; Luo *et al.*, 2018; Sadat *et al.*, 2020], with separate components responsible for critical tasks such as perception [Li *et al.*, 2022c; Liu *et al.*, 2023d], prediction [Shi *et al.*, 2022;

Jia *et al.*, 2022a; Jia *et al.*, 2023b], and planning [Treiber *et al.*, 2000; Dauner *et al.*, 2023]. Specifically, the perception component handles object detection [Li *et al.*, 2022c; Liu *et al.*, 2023d], tracking [Zeng *et al.*, 2022], and sophisticated semantic segmentation tasks [Cheng *et al.*, 2022]. The prediction component analyzes the external environment [Jia *et al.*, 2021] and estimates the future states of the surrounding agents [Jia *et al.*, 2022b]. The planning component, often reliant on rule-based decision algorithms [Treiber *et al.*, 2000], determines the optimal and safest route to a predetermined destination. While the module-based approach provides reliability and enhanced security in a variety of scenarios, it also presents challenges. The decoupled design between system components may lead to key information loss during transitions and potentially redundant computation as well. Additionally, errors may accumulate within the system due to inconsistencies in optimization objectives among the modules, affecting the vehicle’s overall decision-making performance [Chen *et al.*, 2023a].

Rule-based decision systems, with their inherent limitations and scalability issues, are gradually giving way to data-driven methods. End-to-end autonomous driving solutions are increasingly becoming a consensus in the field [Wu *et al.*, 2022b; Chitta *et al.*, 2023; Chen and Krähenbühl, 2022; Jia *et al.*, 2023c; Jia *et al.*, 2023a; Hu *et al.*, 2023b]. By eliminating integration errors between multiple modules and reducing redundant computations, the end-to-end system enhances the expression of visual [Wu *et al.*, 2022a] and sensory information while ensuring greater efficiency. However, this approach also introduces the “black box” problem, meaning a lack of transparency in the decision-making process, complicating interpretation and validation.

Simultaneously, the explainability of autonomous driving has become an important research focus [Jin *et al.*, 2023a]. Although smaller language models (like early versions of BERT [Devlin *et al.*, 2018] and GPT [Brown *et al.*, 2020]) employed in massive data collection from driving scenarios help address this issue, they often lack sufficient generalization capabilities to perform optimally. Recently, large language models [OpenAI, 2023; Touvron *et al.*, 2023] have demonstrated remarkable abilities in understanding context, generating answers, and handling complex tasks. They are also now integrated with multimodal models [Brohan *et al.*, 2023a; Liu *et al.*, 2023a; Driess *et al.*, 2023; Xu *et al.*, 2023;

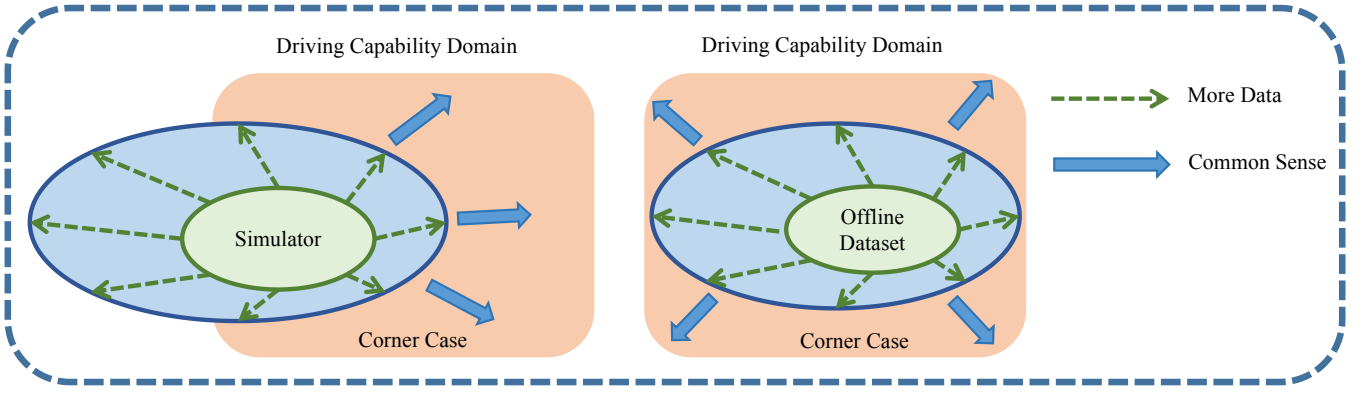


Figure 1: The limitation of current autonomous driving paradigm (green arrow) and where LLMs can potentially enhance autonomous driving ability (blue arrow).

Chen *et al.*, 2023b]. This integration achieves a unified feature space mapping for images, text, videos, point clouds, etc. Such consolidation significantly enhances the system’s generalization capabilities and equips them with the capacity to quickly adapt to new scenarios in a zero-shot or few-shot manner.

In this context, developing an interpretable and efficient end-to-end autonomous driving system has become a research hotspot [Chen *et al.*, 2023a]. Large language models, with their extensive knowledge base and exceptional generalization, could facilitate easier learning of complex driving behaviors. By leveraging the visual-language model (VLM)’s robust and comprehensive capabilities of open-world understanding and in-context learning [Bommasani *et al.*, 2021; Brohan *et al.*, 2023b; Liu *et al.*, 2023a; Driess *et al.*, 2023], it becomes possible to address the long-tail problem for perception networks, assist in decision-making, and provide intuitive explanations for these decisions.

This paper aims to provide a comprehensive overview of this rapidly emerging research field, analyze its basic principles, methods, and implementation processes, and introduce in detail regarding the application of LLMs for autonomous driving. Finally, we discuss related challenges and future research directions.

2 Motivation of LLM4AD

In today’s technological landscape, large language models such as GPT-4 and GPT-4V [OpenAI, 2023; Yang *et al.*, 2023b] are drawing attention with their superior contextual understanding and in-context learning capabilities. Their enriched common sense knowledge has facilitated significant advancements in many downstream tasks. We ask the question: *how do these large models assist in the domain of autonomous driving, especially in playing a critical role in the decision-making process?*

In Fig. 1, we give an intuitive demonstration of the limitation of current autonomous driving paradigm and where LLMs can potentially enhance autonomous driving ability. We summarize two primary aspects of driving skills. The orange circle represents the ideal level of driving competence, akin to that possessed by an experienced human driver.

There are two main methods to acquire such proficiency: one, through learning-based techniques within simulated environments; and two, by learning from offline data through similar methodologies. It’s important to note that due to discrepancies between simulations and the real world, these two domains are not fully the same, i.e. sim2real gap [Höfer *et al.*, 2021]. Concurrently, offline data serves as a subset of real-world data since it’s collected directly from actual surroundings. However, it is difficult to fully cover the distribution as well due to the notorious long-tailed nature [Jain *et al.*, 2021] of autonomous driving tasks.

The final goal of autonomous driving is to elevate driving abilities from a basic green stage to a more advanced blue level through extensive data collection and deep learning. However, the high cost associated with data gathering and annotation, along with the inherent differences between simulated and real-world environments, mean there’s still a gap before reaching the expert level of driving skills. In this scenario, if we can effectively utilize the innate common sense embedded within large language models, we might gradually narrow this gap. Intuitively, by adopting this approach, we could progressively enhance the capabilities of autonomous driving systems, bringing them closer to, or potentially reaching, the ideal expert level of driving proficiency. Through such technological integration and innovation, we anticipate significant improvements in the overall performance and safety of autonomous driving.

The application of large language models in the field of autonomous driving indeed covers a wide range of task types, combining depth and breadth with revolutionary potential. LLMs in autonomous driving pipelines is shown in the Fig. 2.

3 Application of LLM4AD

In the following sections, we divide existing works based on the perspective of applying LLMs: planning, perception, question answering, and generation. The corresponding taxonomy tree is shown in Fig. 3.

3.1 Planning

Large language models (LLMs) have achieved great success with their open-world cognitive and reasoning capa-

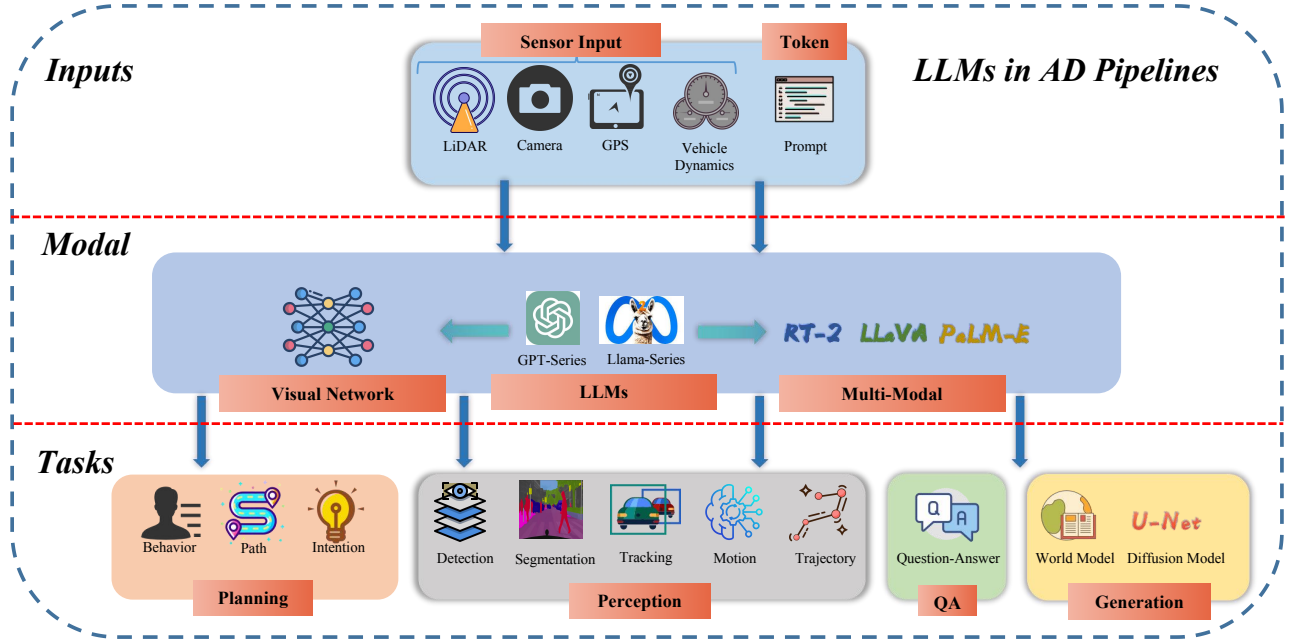


Figure 2: LLMs in Autonomous Driving Pipelines.

bilities [Radford *et al.*, 2018; Radford *et al.*, 2019; Brown *et al.*, 2020; Ouyang *et al.*, 2022; OpenAI, 2023]. These capabilities could provide a transparent explanation of the autonomous driving decision-making process, significantly enhancing system reliability and user trust in the technology [Deruyttere *et al.*, 2019; Kim *et al.*, 2019a; Atakishiyev *et al.*, 2023; Jin *et al.*, 2023a; Malla *et al.*, 2023]. Within this domain, based on whether tuning the LLM, related research can be categorized into two main types: fine-tuning pre-trained models and prompt engineering.

In the application of fine-tuning pre-trained models, MTD-GPT [Liu *et al.*, 2023b] translates multi-task decision-making problems into sequence modeling problems. Through training on a mixed multi-task dataset, it addresses various decision-making tasks at unsignaled intersections. Although this approach outperforms the performance of single-task decision-making RL models, the used scenes are limited to unsignaled intersections, which might be enough to demonstrate the complexity of the real world application. Driving with LLMs [Chen *et al.*, 2023b] designs an architecture that fuses vectorized inputs into LLMs with a two-stage pretraining and fine-tuning method. Due to the limitation of vectorized representations, their method are only tested in the simulation. DriveGPT4 [Xu *et al.*, 2023] presents a multimodal LLM based on Valley [Luo *et al.*, 2023a] and develops a visual instruction tuning dataset for interpretable autonomous driving. Beside predicting a vehicle’s basic control signals, it also responds in real-time, explaining why the action was taken. It outperforms baseline models in a variety of QA tasks while the experiments about planning is simple.

In the prompt engineering perspective, some methods tried to tap into the deep reasoning potential of the LLMs through clever prompt design. DiLu [Wen *et al.*, 2023] designs a

framework of LLMs as agents to solve closed-loop driving tasks. This method introduces a memory module to record experience, to leverage LLMs to facilitate reasoning and reflection processes. DiLu exhibits strong generalization capabilities compared with SOTA RL-based methods. However, the reasoning and reflection processes require multiple rounds of question-answering, and its inference time cannot be ignored. Similarly, Receive Reason and React [Cui *et al.*, 2023b] and Drive as You Speak [Cui *et al.*, 2023a] integrate the language and reasoning capabilities of LLMs into autonomous vehicles. In addition to memory and reflection processes, these methods introduce additional raw sensor information such as camera, GNSS, lidar, and radar. However, the inference speed is unsolved as well. Furthermore, SurrealDriver [Jin *et al.*, 2023b] divides the memory module into short-term memory, long-term guidelines, and safety criteria. Meanwhile, it interviews 24 drivers and uses their detailed descriptions of driving behaviors as chain-of-thought prompts to develop a ‘coach agent’ module. However, there is a lack of comparison with traditional algorithms to prove that large language models indeed bring performance improvements. LanguageMPC [Sha *et al.*, 2023] also designs a chain-of-thought framework for LLMs in driving scenarios and it integrates with low-level controllers by guided parameter matrix adaptation. Although its performance exceeds MPC and RL-based methods in the simplified simulator environments, it lacks validation in complex environments. TrafficGPT [Zhang *et al.*, 2023b] is a fusion of ChatGPT and traffic foundation models which can tackle complex traffic-related problems and provide insightful suggestions. It leverages multimodal data as a data source, offering comprehensive support for various traffic-related tasks. Talk2BEV [Dewangan *et al.*, 2023] introduces a large vision-language model (LVLM) interface

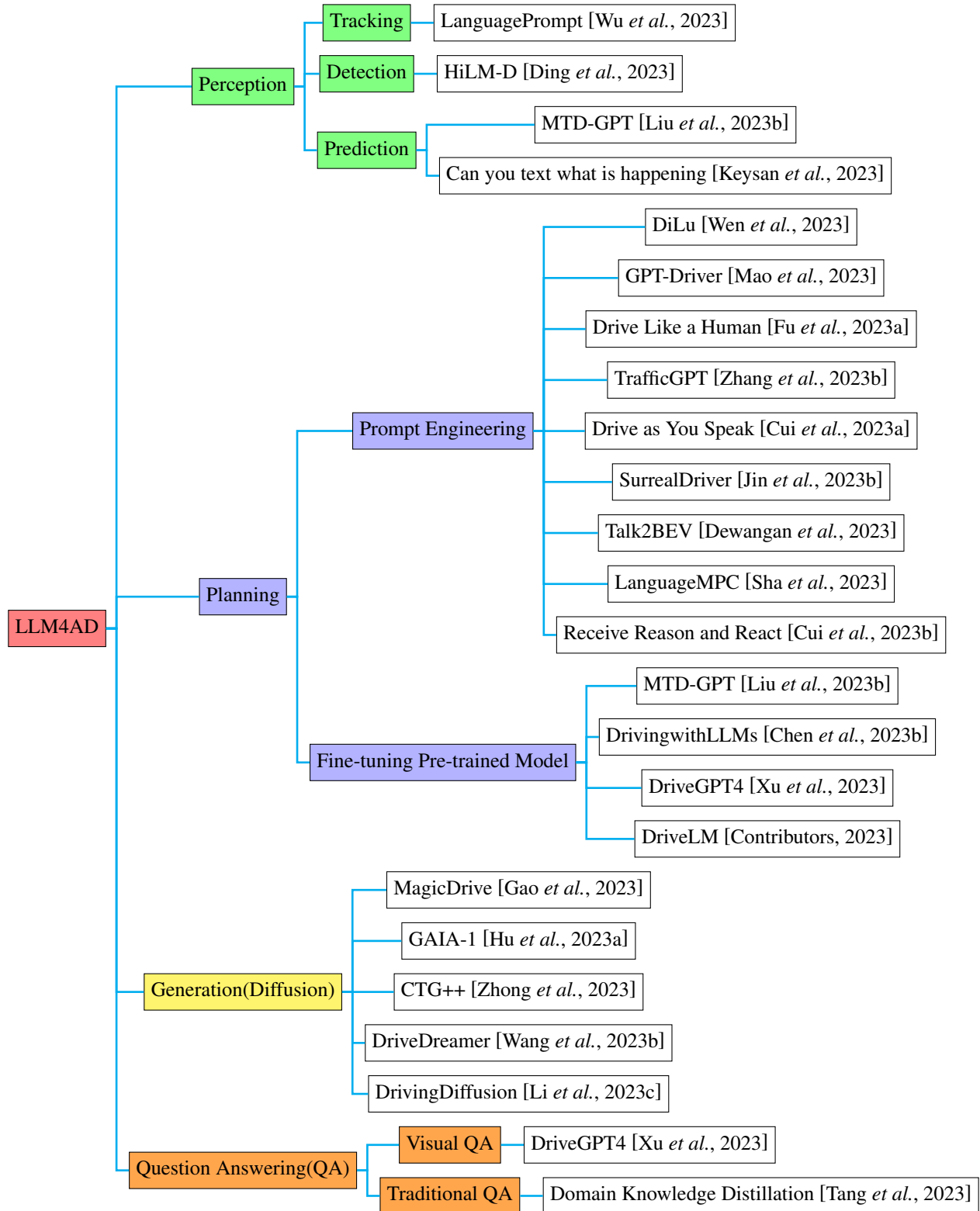


Figure 3: Large Language Models for Autonomous Driving Research Tree

for bird’s-eye view (BEV) maps in autonomous driving contexts. It does not require any training or fine-tuning, only relying on pre-trained image-language models. In addition, it presents a benchmark for evaluating subsequent work in LVLMs for AD applications. GPT-Driver [Mao *et al.*, 2023] transforms the motion planning task into a language modeling problem. It exceeds the UniAD [Hu *et al.*, 2023b] in the L2 metric. Nevertheless, since it uses past speed and acceleration information, there is concern about unfair comparison with UniAD. Additionally, L2 only reflects the fitting degree of the driving route and might not reflect the driving performance [Dauner *et al.*, 2023]. DriveLM [Contributors, 2023] uses a trajectory tokenizer to process ego-trajectory signals to texts, making them belong to the same domain space. Such a tokenizer can be applied to any general vision language models. Moreover, they utilize a graph-structure inference with multiple QA pairs in logical order, thus improving the final planning performance.

Metric:

MTD-GPT [Liu *et al.*, 2023b] uses single-subtask success rates as the metric in simulation and it exceeds RL expert. DriveGPT4 [Xu *et al.*, 2023] uses root mean squared error (RMSE) and threshold accuracies for evaluation. In vehicle action description, justification, and full sentences, it uses BLEU-4 [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], CIDER [Vedantam *et al.*, 2015] and chatgpt score [Fu *et al.*, 2023b]. Driving with LLMs [Chen *et al.*, 2023b] uses the Mean Absolute Error (MAE) for the predictions of the number of cars and pedestrians, normalized acceleration, and brake pressure. Additionally, it measures the accuracy of traffic light detection as well as the mean absolute distance error in meters for traffic light distance prediction. Besides perception-related metrics, it also uses GPT-3.5 to grade their model’s answers which is a recently emerging technique - grading natural language responses [Fu *et al.*, 2023b; Wang *et al.*, 2023a; Liu *et al.*, 2023c]. DiLu [Wen *et al.*, 2023] uses Success Steps in simulation as a metric to evaluate generalization and transformation abilities. SurrealDriver [Jin *et al.*, 2023b] evaluates agents based on two main dimensions: safe driving ability and humanness. Safe driving capabilities are assessed through collision rates, while human likeness is assessed through user experiments with 24 adult participants (age 29.3 ± 4.9 years, male = 17 years) who are legal to drive. LanguageMPC [Sha *et al.*, 2023] customizes some metrics: failure/collision cases, the efficiency of traffic flow, time cost by ego vehicle, and the safety of the ego vehicle’s driving behavior. Similarly, Talk2BEV [Dewangan *et al.*, 2023] measures their methods from the perspective of spatial reasoning, instance attribute, instance counting, and visual reasoning. GPT-Driver [Mao *et al.*, 2023] and DriveLM [Contributors, 2023] contains two metrics: L2 error (in meters) and collision rate (in percentage). The average L2 error is calculated by measuring the distance of each waypoint in the planned trajectory and the offline recorded human driver trajectory. It reflects the fitting of the planned trajectory to the human driving trajectory. The collision rate is calculated by placing an ego vehicle box at each waypoint of the planned trajectory and then checking for collisions with the ground truth bounding boxes of other objects. It reflects

the safety of the planned trajectory.

At present, LLM4AD regarding the planning task lacks a unified metric and cannot uniformly evaluate the pros and cons between each method and traditional counterparts.

3.2 Perception

Large language models have demonstrated their unique value and strong capabilities in “perception” tasks [Radford *et al.*, 2021; Li *et al.*, 2022b; Li *et al.*, 2023a; Li *et al.*, 2023b; Li *et al.*, 2022a]. Especially in environments where data is relatively scarce, these models can rely on their few-shot learning characteristics to achieve fast and accurate learning and reasoning [P *et al.*, 2023; Lin *et al.*, 2023]. This learning ability is of significance in the perception stage of the autonomous driving system, which greatly improves the system’s adaptability and generalization capabilities in changing and complex driving environments. PromptTrack [Wu *et al.*, 2023] fuses cross-modal features in a prompt reasoning branch to predict 3D objects. It uses language prompts as semantic cues and combines LLMs with 3D detection tasks and tracking tasks. Although it achieves better performance compared to other methods, the advantages of LLMs do not directly affect the tracking task. Rather, the tracking task serves as a query to assist LLMs in performing 3D detection tasks. HiLM-D [Ding *et al.*, 2023] incorporates high-resolution information into multimodal large language models for the Risk Object Localization and Intention and Suggestion Prediction (ROLISP) task. It combines LLMs with 2D detection tasks and obtains better performance in detection tasks and QA tasks compared to other multi-modal large models such as Video-LLaMa [Zhang *et al.*, 2023a], ePALM [Shukor *et al.*, 2023]. It is worth noting to point out one potential limitation of the dataset: each video contains only one risk object, which might not capture the complexity of real-world scenarios. [Keysan *et al.*, 2023] integrates pre-trained language models as text-based input encoders for the autonomous driving trajectory prediction task. Joint encoders (image and text) over both modalities perform better than using a single encoder in isolation. While the joint model significantly improves the baseline, its performance has not reached the state-of-the-art level yet [Deo *et al.*, 2021; Gilles *et al.*, 2021].

Metric:

PromptTrack [Wu *et al.*, 2023] uses the Average Multiple Object Tracking Precision (AMOTA) metric [Bernardin and Stiefelhagen, 2008], the Average MultiObject Tracking Precision (AMOTP) [Bashar *et al.*, 2022] and Identity Switches (IDS) [Huang *et al.*, 2023] metrics. HiLM-D [Ding *et al.*, 2023] uses the BLEU-4 [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], CIDER [Vedantam *et al.*, 2015] and SPICE [Anderson *et al.*, 2016], IoU [Rezatofighi *et al.*, 2019] as metrics to compare with the state-of-the-art. [Keysan *et al.*, 2023] uses the standard evaluation metrics that are provided in the nuScenes-devkit [Caesar *et al.*, 2019; Fong *et al.*, 2021]: minimum Average Displacement Error (minADEk), Final Displacement Error (minFDEk), and the miss rate over 2 meters.

3.3 Question Answering

Question-Answering is an important task that has a wide range of applications in intelligent transportation, assisted driving, and autonomous vehicles [Xu *et al.*, 2021a; Xu *et al.*, 2021b]. It mainly reflects through different question and answer paradigms, including traditional QA mechanism [Tang *et al.*, 2023] and more detailed visual QA methods [Xu *et al.*, 2023]. [Tang *et al.*, 2023] constructs the domain knowledge ontology by “chatting” with ChatGPT. It develops a web-based assistant to enable manual supervision and early intervention at runtime and it guarantees the quality of fully automated distillation results. This question-and-answer system enhances the interactivity of the vehicle, transforms the traditional one-way human-machine interface into an interactive communication experience, and might be able to cultivate the user’s sense of participation and control. These sophisticated models [Tang *et al.*, 2023; Xu *et al.*, 2023], equipped with the ability to parse, understand, and generate human-like responses, are pivotal in real-time information processing and provision. They design comprehensive questions related to the scene, including but not limited to vehicle states, navigation assistance, and understanding of traffic situations.

Metric:

In terms of QA tasks, NLP’s metric is often used. In DriveFPT4 [Xu *et al.*, 2023], it uses BLEU-4 [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], CIDER [Vedantam *et al.*, 2015] and chatgpt score [Fu *et al.*, 2023b].

3.4 Generation

In the realm of “generation” task, large language models leverage their advanced knowledge-base and generative capabilities to create realistic driving videos or intricate driving scenarios under specific environmental factors [Khachatryan *et al.*, 2023; Luo *et al.*, 2023b]. This approach offers revolutionary solutions to the challenges of data collection and labeling for autonomous driving, also constructing a safe and easily controllable setting for testing and validating the decision boundaries of autonomous driving systems. Moreover, by simulating a variety of driving situations and emergency conditions, the generated content becomes a crucial resource for refining and enriching the emergency response strategies of autonomous driving systems.

The common generative models include the Variational Auto-Encoder(VAE) [Kingma and Welling, 2022], Generative Adversarial Network(GAN) [Goodfellow *et al.*, 2014], Normalizing Flow(Flow)[Rezende and Mohamed, 2016], and Denoising Diffusion Probabilistic Model(Diffusion)[Ho *et al.*, 2020]. With diffusion models have recently achieved great success in text-to-image [Ronneberger *et al.*, 2015; Rombach *et al.*, 2021; Ramesh *et al.*, 2022], some research has begun to study using diffusion models to generate autonomous driving images or videos. DriveDreamer [Wang *et al.*, 2023b] is a world model derived from real-world driving scenarios. It uses text, initial image, HDmap, and 3Dbox as input, then generates high-quality driving videos and reasonable driving policies. Similarly, Driving Diffusion [Li *et al.*, 2023c] adopts a 3D layout as a control signal to generate realistic multi-view videos. GAIA-1 [Hu *et al.*, 2023a] leverages video, text, and action inputs to generate traffic scenarios, environmental elements, and potential risks. In these methods, text encoder both adopt CLIP [Radford *et al.*, 2021] which has a better alignment between image and text. In addition to generating autonomous driving videos, traffic scenes can also be generated. CTG++ [Zhong *et al.*, 2023] is a scene-level diffusion model that can generate realistic and controllable traffic. It leverages LLMs for translating a user query into a differentiable loss function and use a diffusion model to transform the loss function into realistic, query compliant trajectories. MagicDrive [Gao *et al.*, 2023] generates highly realistic images, exploiting geometric information from 3D annotations by independently encoding road maps, object boxes, and camera parameters for precise, geometry-guided synthesis. This approach effectively solves the challenge of multi-camera view consistency. Although it achieves better performance in terms of generation fidelity compared to BEVGen [Swerdlow *et al.*, 2023] and BEVControl [Yang *et al.*, 2023a], it also faces huge challenges in some complex scenes, such as night views and unseen weather conditions.

These methods explore the customized authentic generations of autonomous driving data. Although these diffusion-based models achieved good results on video and image-generated metrics, it is still unclear whether they could really be used in closed-loop to really boost the performance of the autonomous driving system.

Metric:

DriveDreamer [Wang *et al.*, 2023b] and DrivingDiffusion [Li *et al.*, 2023c] use the frame-wise Frechet Inception Distance (FID) [Parmar *et al.*, 2022] to evaluate the quality of generated images and the Frechet Video Distance (FVD) [Unterthiner *et al.*, 2019] for video quality evaluation. DrivingDiffusion also uses average intersection crossing (mIoU) [Rezatofighi *et al.*, 2019] scores for drivable areas and NDS [Yin *et al.*, 2021] for all the object classes by comparing the predicted layout with the ground-truth BEV layout. CTG++ [Zhong *et al.*, 2023] following [Xu *et al.*, 2022; Zhong *et al.*, 2022], uses the failure rate, Wasserstein distance between normalized histograms of driving profiles, realism deviation (real), and scene-level realism metric (rel real) as metrics. MagicDrive [Gao *et al.*, 2023] utilizes segmentation metrics such as Road mIoU and Vehicle mIoU [Taran *et al.*, 2018], as well as 3D object detection metrics like mAP [Henderson and Ferrari, 2017] and NDS [Yin *et al.*, 2021].

4 Datasets in LLM4AD

Traditional datasets such as nuScenes dataset [Caesar *et al.*, 2019; Fong *et al.*, 2021] lack action description, detailed caption, and question-answering pairs which are used to interact with LLMs. The BDD-x [Kim *et al.*, 2018], Rank2Tell [Sachdeva *et al.*, 2023], DriveLM [Contributors, 2023], DRAMA [Malla *et al.*, 2023], NuPrompt [Wu *et al.*, 2023] and NuScenes-QA [Qian *et al.*, 2023] datasets represent key developments in LLM4AD research, each bringing unique contributions to understanding agent behaviors and urban traffic dynamics through extensive, diverse, and