Review

Not peer-reviewed version

LLMs in the Loop: A Survey of Language-Driven Driver Monitoring and Assistance Systems

Vanchha Chandrayan * and Ignacio Alvarez

Posted Date: 11 March 2026

doi: 10.20944/preprints202603.0903.v1

Keywords: LLM reasoning; advanced driver assistance systems (adas); in-cabin sensing modalities; human-vehicle interaction

 doi:10.20944/preprints202603.0903.v1

Review

LLMs in the Loop: A Survey of Language-Driven Driver Monitoring and Assistance Systems

Vanchhha Chandrayan * and Ignacio Alvarez

Technische Hochschule Ingolstadt, Ingolstadt, Germany
* Correspondence: vanchha.chandrayan@thi.de

Abstract

In recent years we have seen Large Language Models (LLMs) demonstrating robust reasoning capabilities comparable to human performance. This makes them increasingly appealing for driver assistance, where adaptation to dynamic human context is essential. Yet, research in this area remains fragmented, often focusing on isolated applications, lacking utilization of LLM's full potential to deliver integrated, context-specific support and action. This survey synthesizes recent advancements in LLM-driven occupant monitoring systems, focusing on their capabilities for interpreting driver states and acting appropriately, enabling a new generation of intelligent driver assistance. We critically examine pioneering frameworks, benchmarks, and foundational datasets that employ techniques like reasoning chains, multimodality, and human-in-the-loop feedback to create personalized and safe driving experiences. We lay out the current trends, limitations, emerging patterns, in addition to a novel human-centered evaluation of the field, providing researchers with a roadmap towards transparent and trustworthy in-cabin systems, that bridge safety with driver experience.

Keywords: LLM reasoning; advanced driver assistance systems (adas); in-cabin sensing modalities; human-vehicle interaction

1. Introduction

Despite years of progress towards fully autonomous driving, most systems deployed today operate at intermediate levels of automation. At SAE level 2 and 3 vehicles can manage key vehicle controls such as lane centering or adaptive cruise, but they still depend on human oversight for safety and accountability [1]. Advanced Driver Assistance Systems (ADAS) and increasingly Driver Monitoring Systems (DMS) have become essential components of this technology roadmap, with regulatory mandates in regions like EU requiring in-cabin monitoring of driver states [1,2]. These systems promise to improve safety by detecting risky driver states and providing corrective support through alerts and warnings. However, the current DMS systems remain limited. They often reduce complex cognitive and affective states to simple labels, fail to generalize across driver populations and environments, and rarely offer explanations or dynamic feedback to the driver [3,4]. As a result, drivers may distrust or ignore them at the moments where assistance is most needed [5].

The emergence of LLMs offers new opportunities to rethink how assistance systems operate. They possess reasoning skills, where they apply complex correlations and causal analysis over the multimodal context, enabling more flexible interpretation and natural-language outputs. These capabilities set LLMs apart from conventional machine learning systems [1]. For instance, they can reason over multimodal inputs [6], adapt to novel situations with few or no examples [7], generate feedback in natural language [8], and retain memory across interactions [9]. In other safety critical domains, such as healthcare and aviation, LLMs are already being tested as human-in-the-loop decision aids [10–12]. For in-vehicle context, this means adoption of assistance systems that can integrate visual cues, physiological signals, and contextual information (e.g., time of day, traffic, weather conditions). These integrated signals can then be interpreted and reasoned by the system to provide adaptive interactions

 doi:10.20944/preprints202603.0903.v1

and explanations of their interpretations in natural language. Some initial prototypes have utilized LLMs as conversational agents [13,14] or as a mechanism to generate multimodal warnings for driver assistance [15]. These early experiments illustrate the potential of LLM-powered DMS to shift current rigid alert systems towards conversational adaptive assistant agents that act as "co-pilots".

Still, the path forward is far from straightforward. At the system's level, today's assistance technologies face challenges of accuracy, latency, efficiency, and reliability. Models that are too "heavy" (computationally and storage-wise) cannot be supported with real-time requirements on in-vehicle hardware [13]; interventions that are too aggressive may compromise driver autonomy; models that are too opaque may undermine driver's trust [16]. At the research level, studies focused on detecting and interpreting driver states remain fragmented as they operate on isolated features, modalities, and evaluation metrics [4]. Reasoning approaches are being proposed; from retrieval-augmented assistance [17] to reasoning chains for risk assessment [18], yet their incorporation to practical in-cabin interactions remain underexplored. Similarly, datasets remain isolated by state or task, limiting opportunities for an integrated benchmarking. Prior surveys and reviews of this state-of-the-art [1,7,19] provide broad overviews of LLMs in autonomous driving or multimodal systems, but they lack focus specifically on in-cabin, driver-in-the-loop applications, that connect sensing, reasoning, interaction, and evaluation under real-world constraints.

This paper addresses that gap by synthesizing the emerging literature on LLMs in driver monitoring and assistance as a system-level design space (Fig. 1). We focus specifically on LLM-driven systems that revolve around drivers, asking the following questions:

RQ1. How are LLMs used to detect and interpret driver states such as distraction, drowsiness, and emotion?

RQ2. How do LLMs reason over multimodal cues, context, and human intent to support decision making?

RQ3. How are in-cabin interactions and interventions enabled by LLMs?

RQ4. How do various datasets and benchmarks support or constrain research on LLM-based co-pilots?

The contributions of this paper are as follows. First, we propose a taxonomy of LLM roles in driver assistance i.e., as detectors, reasoners, and actors. Second, we provide design paradigms emerging across studies, from open-loop warnings to adaptive, conversational co-pilots and identify recurring design tensions. Third, we consolidate datasets across driver states, external context, and language-grounded tasks into a normalized map. Finally, we outline a human-centered roadmap for future intelligent driver assistance systems that integrate transparency, trustworthiness, and inclusivity by design. In doing so, this paper not only surveys a rapidly expanding field, but also argues for a shift i.e., from fragmented detectors towards integrated and adaptive co-pilots that reason *with* drivers, not just about them.

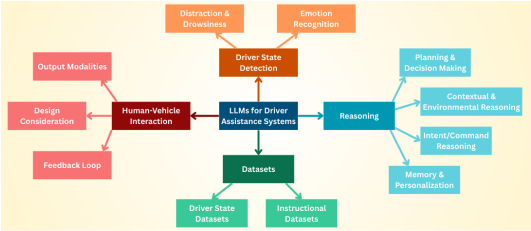

Figure 1. This survey paper focuses on the use of LLMs in the advancement of Driver Monitoring and Assistant Systems. Figure shows the scope of this paper.

2. Background

2.1. Traditional Driver Monitoring Systems

Driver monitoring has relied on three primary feature sources i.e., visual cues (e.g., eyelid closures, yawning, gaze metrics), physiological measures (e.g., electroencephalography (EEG), heart-rate variability (HRV) or respiration rate), and vehicle control behaviors such as steering variability

 doi:10.20944/preprints202603.0903.v1

or lane (Table 1) [3]. For instance, drowsiness is often measured through prolonged eye closure or reduced EEG indices [20]; distraction through gaze shifts, head-pose, or steering variability [18]; and emotion through facial expression analysis or variations in vocal prosody, and HRV [21].

In traditional DMS, these signals are processed through classical machine learning or early deep learning pipelines [4,22]. Facial and ocular features from camera feeds are extracted via Convolutional Neural Networks (CNNs) [23], while physiological indices are classified into driver states using models like Support Vector Machines (SVMs) and decision trees [11,12]. In some cases, temporal patterns are modeled through Recurrent Neural Networks (RNNs) [22]. Training of these systems relies on supervised learning with annotated datasets, often constrained to laboratory environments [24]. Upon detection of a safety critical driver state, hard-coded interventions were explored to mitigate risky situations (auditory warnings, dashboard icons, or seat vibration) [21]. These approaches provide early safety benefits but come with limitations including reliance on heavy annotation, rigid categorical states, limited robustness across drivers and contexts, and fixed feedback [4,8]. These limitations motivate more scalable, context-aware, and explainable approaches for next-generation in-cabin monitoring and assistance.

2.2. LLMs and Their Optimization

The development of LLMs has reshaped how Artificial Intelligence (AI) systems can understand and generate information [1,6]. Built on transformer architecture, LLMs learn patterns across massive text corpora by attending to the relationships between words and sequences (Fig. 2) [1]. This training allows them to go beyond simple word prediction. They can support multi-step inference, translate natural language instructions into structured outputs, and adapt to unfamiliar situations with minimal task-specific data [1,15]. Therefore, LLMs operate not as narrow classifiers but as flexible foundation models that can interpret, input, weigh context, and generate meaningful responses. These qualities make LLMs particularly appealing for domains where safety and system transparency are central, such as healthcare, aviation, and robotics [10–12]. In the vehicle domain, recent surveys note that LLMs have been tested for multiple roles e.g., supporting route planning and traffic predictions, as conversational assistants for navigation, or explaining outputs and actions of the system [1,6,7]. Such applications show that LLMs can support traditional ADAS by providing a reasoning layer that bridges technical output and driver-facing interpretation. Yet, the same surveys also highlight the limitations of using raw LLMs in real vehicles. Their large size makes them too slow and computational resource-hungry to run reliably on embedded automotive hardware. Their tendency to hallucinate and give nonsensical outputs can undermine system reliability and safety. And their outputs are not automatically aligned with driver preferences, regulatory constraints, and other task-specific requirements [6,12,13]. Therefore, LLMs must be optimized and adapted before they can serve as practical components of driver assistance systems.

Optimization strategies fall broadly into two categories i.e., making models efficient to run in vehicles, and aligning their behavior with the expectations of human users [6]. These approaches are highlighted here as they recur across the frameworks we review, providing technical foundation for how LLM-based driver assistance systems are being optimized. In an automotive setting, these strategies are closely tied to deployment constraints such as on-device compute, memory, latency, connectivity, and privacy. Following are some techniques widely explored for improving efficiency of LLMs (Fig. 2). Knowledge distillation compresses a larger "teacher" model into small "student" ones that retains performance but operate faster with fewer resources [19,25]. Quantization reduces the precision of stored LLM parameters, thereby shrinking memory utilization and increasing inference speed. Both distillation and quantization approaches enable integration (embedding) of LLM-based systems on vehicle hardware with reduced latency [22,26]. Parameter-efficient fine-tuning (PEFT) techniques like Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) refine only small portions of the model's weights, making it possible to personalize models for individual automotive tasks or even individual user support without retraining the full system [18,27] or losing model knowledge/performance.

 doi:10.20944/preprints202603.0903.v1

Alignment strategies focus on ensuring that the model's outputs are reliable, transparent, and human-centered (Fig. 2) [1]. Chain-of-Thought (CoT) prompting encourages models to make their reasoning steps explicit, improving the interpretability of decisions such as why a particular action is unsafe [6,9,15]. Retrieval-Augmented Generation (RAG) grounds LLM outputs in stored knowledge bases such as vehicle manuals, regulations, or real-time weather and traffic databases, thus reducing hallucinations and providing verifiable justifications [8,17]. Reinforcement Learning with Human Feedback (RLHF) tunes model responses to reflect human expectations, ensuring that decisions are not only correct but acceptable to the users [28]. By combining these efficiency and alignment techniques, studies have explored and transformed LLMs from general purpose models into domain-specific assistant systems that are capable of reasoning in real time, grounding outputs in domain knowledge, and adapting to the needs of their users [1,26].

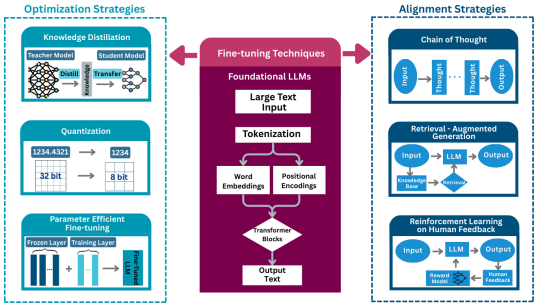

Figure 2. Foundational LLM architecture with fine-tuning and optimization strategies highlighted in the reviewed literature.

2.3. Vision-Language Models (VLMs) and Multimodal LLMs (MLLMs)

While optimization techniques adapt LLMs for embedded and safety-critical use, another important direction has been extending reasoning capabilities of these models beyond text. VLMs (such as the CLIP model), jointly learn embeddings (numerical representations of patterns and relationships) from paired images and text. This allows them to connect visual content with natural language descriptions, enabling zero-shot classification (the ability to perform successfully on new visual data from natural language descriptions) and cross-modal retrieval (e.g. searching in text and returning visual data) [4,29,30]. An exploration of VLMs in autonomous driving reports their use in traffic sign recognition, pedestrian detection, and scene captioning, where they provide flexible perception without the need for exhaustive training [4,23,31]. Extending beyond two modalities, MLLMs incorporate diverse streams such as video, audio, telemetry, and sensor data into a unified reasoning framework [1]. Their applications are explored in traffic accident prediction, contextual risk assessment, and integrated perception systems, demonstrating how multimodality improves robustness in dynamic driving environments [1,21].

The trajectory from traditional monitoring pipelines to LLMs, and their extension into multimodal systems illustrates the rapid transformation of driver assistance technologies. What was once limited to rigid detection and fixed interventions is now evolving into flexible systems capable of reasoning, adaptation, and interaction. Motivated by this rapid emergence of VLM/MLLM-enabled in-cabin sensing and assistance, we conduct this survey to consolidate scattered evidence into a coherent taxonomy and benchmarking-oriented view, clarifying how language models are being integrated into driver monitoring and assistance pipelines and what system-level trade-offs recur across studies. To support transparency and reproducibility of this survey, the next section details our search strategy, screening procedure, and inclusion criteria used to identify and select the studies reviewed in this paper.

 doi:10.20944/preprints202603.0903.v1

3. Methodology

This work presents a narrative review of recent research on language-driven driver monitoring and in-cabin driver assistance systems, with a focus on approaches that integrate LLMs and related VLMs/MLLMs into sensing, reasoning, and interaction pipelines. The scope of the survey covers English-language publications from 2020 onward, reflecting the period in which transformer-based and language-driven models began to meaningfully appear in automotive driver monitoring and assistance contexts. To capture both established and emerging work in this rapidly evolving area, literature searches were conducted using two complementary scholarly databases: Scopus, to identify indexed peer-reviewed publications, and Google Scholar, to broaden coverage to recent conference papers and early-access or preprint work. Searches were organized around thematic keyword classes spanning (i) LLMs and generative AI, (ii) human state interpretation (e.g., distraction, drowsiness, emotion, workload), (iii) reasoning and decision support, (iv) intervention and interaction strategies, (v) multimodal sensing systems, (vi) personalization and adaptation, (vii) explainable AI, and (viii) automotive application contexts such as driver monitoring systems and ADAS.

Across both databases, 146 candidate papers were initially identified (109 from Scopus and 37 from Google Scholar). Dataset papers frequently used for evaluation were identified through backward citation analysis of included studies and retained if they played an important role in the model development. After deduplication, 12 records were removed. The remaining papers were screened based on title, abstract, and when necessary full text, using inclusion criteria that required explicit use of LLMs, VLMs, or closely related language-driven transformer models within in-cabin driver monitoring or driver assistance scenarios. Studies relying solely on classical machine learning without language models, or applying LLMs outside the in-cabin driver assistance scope, were excluded. Following this screening, 89 papers were excluded, resulting in a final set of 45 studies included in the survey (Fig. 3). For each included work, key attributes were extracted and synthesized, including the driver state(s) addressed, model modality, the functional role of language models within the system (e.g., detection, reasoning, interaction), datasets used, and deployment or optimization considerations. Studies employing language models across multiple stages of the assistance loop were analyzed from each relevant perspective and may therefore appear in multiple sections of the review, reflecting the different roles language models play within integrated driver assistance pipelines. The next sections build on this foundation by reviewing how LLMs are being applied across four dimensions: driver state detection, reasoning over driver context, interaction and intervention, and the datasets that enable these developments (Fig. 1).

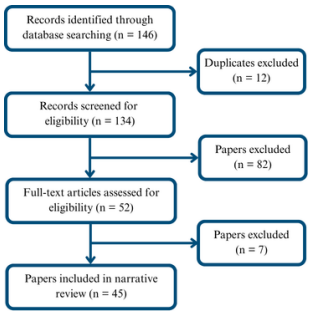

Figure 3. Flowchart of Literature Review Process.

4. Driver State Detection with LLMs

Although traditional approaches based on handcrafted features and conventional classifiers have significant limitations, they continue to play an important role in today's systems; particularly for initial feature extraction and for generating supervisory labels. As outlined in the Background section, these pipelines provide the raw inputs upon which newer architectures build. What distinguishes recent research is the exploration of LLMs and VLMs as complementary components: rather than

 doi:10.20944/preprints202603.0903.v1

replacing traditional detectors, they extend them by taking on the more complex tasks of reasoning and interpretation [15,20,22].

4.1. In-cabin Sensing Modalities

LLM-based driver assistance systems remain fundamentally grounded in the sensing pipelines that provide access to the driver's internal states and behaviors. As a result, the choice of sensing modality continues to shape driver states that can be detected, their interpretation, and whether such systems can be deployed in real-time, on-device constraints. Before reviewing state-specific LLM-based detection approaches, we therefore summarize the in-cabin sensing modalities referred or employed across surveyed literature and highlight their system level trade-offs. Table 1 synthesizes how different vision-based, physiological, and vehicle telemetry sensors support driver state inference, alongside their key failure modes, privacy implications, and feasibility when integrated into LLM-driven pipelines.

 doi:10.20944/preprints202603.0903.v1

Table 1. In-cabin sensing modalities for LLM/VLM/MLLM-based driver monitoring and assistance, including supported driver states and deployment trade-offs.

Sensors	Used for	Key limitations / failure modes	Edge feasibility	Privacy risk
Vision / Audio				
FIR (thermal) [20]	Drowsiness (yawn, head droop, head pose); distraction	Glasses block eyes; temperature variance; low spatial detail	Med–high efficiency	High
NIR / IR [23][3]	Facial emotion; drowsiness; distraction; robust low-light monitoring	Occlusion (hair, glasses); pose variation; specular reflections	Efficient perception; temporal complexity	Med–High
RGB [3][29][30][18]	Distraction / secondary tasks; facial emotion; drowsiness; body/pose behavior	Low-light sensitivity; occlusion; motion blur; false detections	Low–Med; time-consuming inference	High
Depth / RGB-D [32][21]	Posture / geometry cues; object detection; silent interaction	Noisy depth in real time; reflective surfaces; limited range / placement constraints	Low; 3D processing overhead	Med–High
Audio [27][14][33][34][35]	Emotion; intent / command; fatigue	Background noise; cognitive interference; latency; speech variability	Med–high with hybrid edge-cloud frameworks	Med–High
Physiological				
EEG [36][21]	Drowsiness; cognitive state; stress; workload; emotion	Intrusive; noisy; high inter-subject variability	Low; high-dimensional data	High
ECG / HRV [16][21][36][24][32][3]	Arousal; drowsiness; workload	Motion artefacts; signal noise; delayed HRV response	Med; temporal complexity	Med–High
EDA / GSR [37][28][24][38]	Stress; arousal; frustration	Contact sensitivity; drift; influenced by hydration/skin properties/placement	Wearables feasible; ambiguous LLM interpretation	Med
Eye tracking [39][3][4][14][36]	Attention; situational awareness; drowsiness; workload	Lighting variation; occlusion; calibration sensitivity	Med–low efficiency	Med
Vehicle telemetry				
Steering wheel (angle/torque/variability) [14][23][13][28][21]	Fatigue; drowsiness; distraction	Reactive (not proactive); indirect correlation; limited personalization	High efficiency	Low
Pedal interaction (brake/throttle) [23][33][40]	Attention level; situational awareness; arousal proxies; driving performance	Context-dependent; confounded by driving style/traffic; one-size-fits-all assumptions	High efficiency	Low

4.2. Distraction and Drowsiness

Distraction and drowsiness have been among the earliest driver states explored using LLMs and VLMs, largely motivated by the limitations of traditional supervised pipelines that rely on heavily annotated data and predefined state categories. Recent work leverages multimodal language-conditioned models to interpret complex driving scenes by jointly processing visual inputs and linguistic context, enabling a more holistic representation of driver behavior and surrounding interactions [4,22]. Hasan et al. introduced DriveCLIP, which adapts CLIP-based vision-language representations to recognize distracted driving actions in naturalistic images and videos [29]. Their framework compares frame-based and video-based variants, showing that temporal modeling substantially improves the recognition of distracting actions. VideoCLIP achieved high Top-1 accuracy on benchmark distrac-

 doi:10.20944/preprints202603.0903.v1

tion datasets. Complementing this, Girbacia et al. evaluated multiple open-source VLMs, including PaliGemma, for phone-based distraction detection and showed that limited fine-tuning can yield competitive performance exceeding 95% accuracy [30]. Moving beyond action recognition, Zhang et al. proposed the Distracted Driving Language Model (DDLM), which integrates whole-body pose estimation with a VLM to classify 22 distraction categories and extend recognition toward risk-oriented interpretation [18]. By incorporating reasoning chains, DDLM produces explainable assessments of driver behavior and demonstrates strong zero- and few-shot performance relative to standard baselines. Collectively, these studies illustrate how LLM/VLM-based systems are being used not only for categorical distraction detection but also for temporally aware and explanation-oriented safety assessment.

Parallel efforts have focused on improving the robustness and reliability of distraction detection. Hu et al. proposed the Human-Centric Context and Self-Uncertainty driven MLLM (HSUM), a training-free framework that introduces self-uncertainty estimation and contextual grounding through scene graphs of faces, hands, objects, and environment-specific prompts [22]. In related work, Hu et al. introduced Trustworthy Driver State Perception (TDSP), which combines visual and language features with evidence-based learning to provide calibrated confidence estimates alongside distraction predictions [38]. Both HSUM and TDSP achieve competitive or superior results on public benchmarks such as AIDE and 3MDAD for distraction recognition, with TDSP outperforming DriveCLIP [29]. These studies emphasize uncertainty estimation and calibration as critical system-level requirements for deployment-ready driver monitoring.

A broader survey of VLMs for driver monitoring confirms the strong performance of these models across heterogeneous distraction datasets, while also underscoring persistent challenges related to generalization and evaluation design [4]. Across the literature, in-cabin RGB video remains the dominant sensing modality for distraction detection. At the same time, recent work has begun exploring complementary sensing and multimodal fusion strategies to improve robustness. Knapik et al. demonstrated the use of far-infrared thermal imaging to capture distraction-related cues such as head movements and yawning under low-light conditions, positioning FIR sensors as a viable complement to conventional RGB cameras [20]. Other approaches combine visual cues with language-based contextual priors to produce more calibrated and trustworthy predictions [38]. Overall, the distraction literature suggests that video-based perception forms the backbone of current systems, while contextual reasoning and uncertainty modeling are increasingly treated as first-class design objectives.

Drowsiness detection represents another critical safety concern, as fatigue contributes to a substantial proportion of serious road accidents worldwide [33]. Compared to distraction, explicit drowsiness detection using LLMs and VLMs remains relatively underexplored, though early work indicates that similar modeling principles may extend to fatigue-related states. Canas et al. evaluated the open-source VLM Idefics2 on a driver monitoring dataset using zero- and one-shot prompting to detect yawning [4]. While the approach showed some potential for generalization, performance was inconsistent and highly sensitive to prompt design, highlighting the difficulty of modeling subtle and ambiguous states such as drowsiness. Tavakkoli et al. proposed a conceptual framework for multimodal bio-signal fusion that leverages LLM-based contextual reasoning to support holistic fatigue recognition, though the framework was not empirically evaluated [21]. Knapik et al. further demonstrated that far-infrared imaging can robustly capture yawning, head drooping, and eye closure in low-light environments [20]. Several models originally designed for distraction detection, including DDLM [18], HSUM [22], and TDSP [38], also incorporate modules for detecting eye closure or eyelid drooping, underscoring the significant overlap between distraction and drowsiness cues. This overlap suggests opportunities for shared perception components, while also highlighting ambiguity when individual cues may correspond to multiple driver states.

Beyond detection, a smaller body of work explores how LLM-based assistants might intervene once drowsiness is detected, emphasizing the role of adaptive feedback and closed-loop assistance

 doi:10.20944/preprints202603.0903.v1

[33,36]. Although primarily focused on intervention, these studies reinforce the importance of reliable and timely detection as a prerequisite for effective driver assistance.

Despite ongoing progress, several challenges remain in achieving robust distraction and drowsiness detection. Real-time monitoring of psychophysiological states is affected by lighting variation, sensor noise, and inter-individual differences [20]. Annotation of distraction and drowsiness remains costly and subjective due to the lack of standardized definitions and ground truth. In addition, privacy and security considerations arise when processing sensitive in-cabin data, alongside technical constraints related to computational load and reliance on cloud-based APIs [18,21]. Across the reviewed literature, these limitations consistently cluster around sensing robustness, labeling subjectivity and dataset comparability, and deployment constraints such as compute, privacy, and connectivity.

4.3. Emotion Recognition

Emotion detection plays a critical role in both driving safety and driver acceptance of assistance, shaping vigilance, and compliance with the guidance. Other alarming states, such as driver distraction, could often stem from unsafe emotional states such as anger or anxiety, leading to traffic accidents [23]. However, unlike distraction or drowsiness, emotion is more subjective and context-specific. Therefore, effective detection should integrate multiple cues and reasoning about how they relate in a situation. As the automated driving systems advance towards SAE level 3 or higher, integrating emotion aware capabilities into driver assistance systems becomes a necessity to also build trust and ensure a personalized user experience [23,41].

In practice, automotive systems infer emotions from complementary modalities. Non-invasive vision remains mostly applied, using facial expressions and audio cues that can capture prosodic shifts linked to arousal or valence [3,23]. Other modalities also include physiological signals (e.g., HRV, skin conductance, EEG) and behavioral measures (e.g., vehicle behaviors, posture) [3,21,23].

LLM/VLM based detectors have begun to exploit this rich supervision [21]. At the level of facial emotion detection, a study by Li et al [23] integrated an InternVL vision transformer (an open-source VLM) with Qwen-2 (an LLM), reporting 100% performance on KMU-FED (camera facing driver dataset) and 74.36% on FER2013 (facial emotion recognition dataset). These results outperformed strong conventional CNN baselines. Other limited modality frameworks like Talk2Drive [42] use LLMs to translate natural verbal commands, including emotional cues, into executable vehicle controls, continuously adapting to individual preferences. However, challenges like occlusion, lighting variations and lack of environmental context, further emphasize the need for multimodal systems to comprehensively capture the emotional cues with the contextual information. MLLMs can improve emotional recognition and reasoning, even for subtle changes often missed by traditional models [27]. Models like Emotion-LLaMA employ instruction tuning and a combination of encoders to process diverse modalities and align features for in-depth emotional analysis [27]. According to the conceptual framework introduced by Tavakkolli et al. [21], LLMs also offer contextual understanding, allowing them to interpret nuances like sarcasm and implicit meanings, and perform sentiment analysis to assess emotional tone from speech. Training-free pipelines such as HSUM [22] and TDSP [38] also suggest that context- and ambiguity-aware systems can enhance reliability in detecting emotions, even without task-specific training. A recent step toward integrated state monitoring is VLM-DM, proposed by Chi et al [32]. Their approach shows that combining parameter-efficient LoRA-tuned VLMs with traditional vision encoders can unify distraction, drowsiness, and emotion recognition in a single model with competitive accuracy.

Despite these advancements, several significant challenges remain in developing robust emotion detection systems. One of the main limitations is the limited availability of comprehensive, high-quality datasets, including individual and cultural diversity [32]. The black-box nature of many models also raises concerns about their reliability and capability to accurately understand emotional states [24].

In sum, recent systems strive to detect driver states more holistically, by conditioning on human-centric context, quantifying uncertainty, and integrating multimodal reasoning.

 doi:10.20944/preprints202603.0903.v1

5. LLM Reasoning Over Driver Context

To synthesize how LLMs are used beyond perception-level processing, we organize prior work along four reasoning dimensions: planning and decision making, environmental or contextual reasoning, command or intent reasoning, and memory and personalization. These dimensions are intended as analytical lenses rather than mutually exclusive categories, capturing distinct functional roles that language models can play within a driver monitoring and assistance pipeline. Table 2 provides a cross-sectional summary of the reviewed studies, mapping these reasoning dimensions against different system focuses to illustrate how LLM capabilities are distributed across the end-to-end pipeline.

Table 2. Overview of current research in LLM-based driver assistance systems, categorized by reasoning dimensions (planning, contextual, intent, and memory) and system focus.

	Planning and Decision Making	**Environmental or Contextual Reasoning**	**Command or Intent Reasoning**	**Memory and Personalization**
Driver State Detection and Interpretation	[18]: VLM, reasoning chain framework (7 steps); [20]: LLM (thermal fusion) + zero-shot	[22]: MLLMs with Human-Centric Context Generator (HCG) for scene graphs; [38]: VLM+LLM for contextual interactions; evidential fusion for uncertainty; [4][30]: VLM for Driver Monitoring (Idefics2, PaliGemma) + prompt engineering; [27]: Emotion-LLaMA for aligning emotional tone; [32]: LoRA-optimized multi-state detection	–	–
Interaction with Humans	[15][8]: LLM with CoT prompting; [16]: LLM (GPT-4) with prompt engineering	[17]: RAG to retrieve from external knowledge; [16]: LLM feedback by contextual relevance; [33][14]: ChatGPT-4 voice assistant; [27]: MLLM for affect-aware reasoning	[9][43][13][37]: LLM interprets natural language commands; [35][39]: LLM + CoT for command interpretation; [17]: LLM with RAG for queries	[15][8]: memory module (individualization profile) + RAG; [13]: memory module (working/procedural/semantic); [28]: human data into preferences; [14]: ChatGPT for adaptive dialogue
Interaction with Vehicles	[9][43][44][13]: LLM generates action policies for vehicles; [34]: LLM (Llama3) plans & refines vehicle maneuvers	[9][43]: LLMs process contextual data with CoT prompting; [34]: LLM estimates driving styles	[44][13]: LLM generates driving policy code via intent reasoning; [34]: human instructions as prompts	[34]: interaction memory database + memory partition module
Closed-Loop Systems	[42]: GPT-4 translates natural commands into controls; [36]: EEG + LLM dialogue agent; [21] (Conceptual): LLMs for decision support (emotion/fatigue recognition)	[42]: GPT-4 for context and emotional state; [36]: affect-aware dialogue interaction; [21]: LLMs with bio-signal + context fusion	[42]: GPT-4 interprets direct vs. indirect intentions	[42]: memory module for past interactions; [36]: LoRA fine-tuned LLMs on driver dialogue, training on driver's dialogue data to embody user's personality and emotional traits; [21]: personalized ADAS through empathetic interactions

5.1. Planning and Decision Making

A central capability of LLMs in the driving domain is their ability to act as high-level planners and decision makers that move beyond rigid rules and statistical controllers. Unlike traditional pipelines, that fail to adapt to driver's intent, preferences, or state changes. LLMs bring the capacity to generate, refine, and justify plans that are responsive to both environment and the human in the loop.

In one approach, LLMs are being leveraged as central intelligence for translating complex human instructions and intentions into executable driving actions, making it the decision-making brain of the vehicle. Multiple frameworks such as "Drive As You Speak" [9], LaMPilot [44], and "Receive, Reason, and React" [43] position LLMs as the core decision maker. Verbal commands from the driver, alongside sensory inputs are interpreted to output language-model programs (LMPs) that can be executed by classical planners. This results in action plans that respect both safety constraints and human intent.

 doi:10.20944/preprints202603.0903.v1

Experiments such as "ChatGPT as Vehicle Co-Pilot" [13] also applied this idea by embedding a general LLM into the control loop. This demonstrates that even without extensive fine-tuning, models can adjust trajectories or select appropriate controllers that align with the driver's preferences expressed in natural language. These frameworks explicitly emphasize LLM's role in strategic planning, positioning interpretation and planning as two sides of the same reasoning process [44].

Beyond direct command translations, a second cluster of work emphasizes adaptive planning that incorporates driver context and interaction history. The "Actor-Reasoner" framework introduced by Fang et al. [34], couples a lightweight module that proposes vehicle maneuvers with an LLM-based reasoner that uses CoT prompting to refine these actions in line with human driving styles. Their dual system approach combined quantitative scenario descriptions with qualitative experience to enhance the adaptability of AV-decision making. Furthermore, LLMs can guide core planning functions using human physiological and behavioral signals. For instance, a framework by Song et al. [26] integrates human behavior and cognitive data (eye tracking and EEG signals), to guide the autonomous driving model's planning. In parallel, LLM-enhanced RLHF approach directly integrates LLMs to model human preferences to bias decision policies towards more safe, driver-preferred outcomes [28]. These systems do not simply execute plans but learn how drivers expect decisions to be made, reducing gap between algorithmic efficiency and human trust.

These developments collectively highlight that by embedding reasoning mechanisms that can interpret natural language commands, adapt to different driving styles and integrate human behavioral signals, LLMs enable planning pipelines that make interpretable, reliable, and situationally-appropriate decisions. While most frameworks emphasize motion planning, some also extend to decision making for the design of assistance strategies, thereby determining not only *what* the vehicle does but also *how* it communicates with the driver [15,16]. These approaches are revisited in section 5. However, the effectiveness of these planning systems depend on how well they interpret the broader driving scene and contextual cues surrounding the driver. The following section examines these environmental and contextual reasoning mechanisms, which provide situational awareness that support the planning decisions.

5.2. Environmental and Contextual Reasoning

Environmental and contextual reasoning refers to how systems go beyond classifying isolated signals to form a narrative understanding of the driver's state and surroundings [4].

Several recent frameworks demonstrate this shift by integrating information about the traffic environment into LLM reasoning. These systems combine external sensor data, localization, and traffic rules with LLM planning to evaluate road conditions, vehicle distances, and lane availability in real-time [17,39,44]. Similarly, multimodal warning systems adapt their guidance not only to the driver profile but also to situational hazards such as sudden cut-ins or collisions [8]. Collectively, these approaches highlight that assistance quality improves when environmental cues are considered alongside driver state.

In parallel, contextual reasoning also deepens how driver cues themselves are interpreted. As noted earlier in Section 3, systems are increasingly moving beyond a single signal to combine multiple streams (such as body pose, hand movements, voice, and affect) to infer context in richer ways. Approaches like HSUM and TDSP show how adding structured context (scene graphs, descriptive expansions, calibrated confidence) makes detection more interpretable and cautious [22,38]. Another conceptual framework extends this by recommending fusion of vision, speech, and text to align emotional tone with semantic meaning [21]. These works demonstrate that human context is shifting from reactive classification toward proactive interpretation, enabling systems that can explain why a driver might act a certain way, not just what state they are in.

Overall, environmental and contextual reasoning broadens the scope of driver assistance from monitoring isolated signals to reasoning about scenes and interactions. However, this rich understanding comes at the cost of greater system complexity, higher computational demands, and new risks of false positives if contextual signals are ambiguous or culturally variable. These challenges

 doi:10.20944/preprints202603.0903.v1

underscore the need for another layer of reasoning i.e., combining information about what drivers do and what surrounds them with what they mean in context when they issue commands or exhibit critical behaviors.

5.3. Command and Intent Reasoning

Making decisions and planning actions also rely on a more fundamental capability possessed by LLMs, that is, correctly interpreting what the driver actually means, from explicit directiveness to abstract requests, and even through their emotional tone. Command or intent reasoning addresses this ambiguity by enabling LLMs to infer meaning beyond literal input and ensuring that the assistance systems align with the driver's underlying goals.

Some studies have framed LLMs as interpreters that translate natural language instructions into structured inputs for the vehicle. Yang et al. [35] demonstrated how LLMs can parse commands into tasks across perception, localization, and control modules, thereby bridging free-form driver input with vehicle functions. Similarly, LaMPilot frames spontaneous user instructions as input for LLMs to generate LMPs [44]. These works highlight the potential of LLMs as intent decoders, but they come with limitations. Benchmarks show that while structured phrasing is handled well by LLMs, ambiguous instructions such as "find a faster way" remain difficult to parse reliably [44]. Such limitations underscore the need for models that can interpret both syntax and human/situational factors simultaneously.

Frameworks like Talk2Drive [42] evaluate LLM based interpretation of verbal input in field experiments. They demonstrate that the systems must adapt to personal variation in phrasing and the indirectness of commands/requests. By learning from repeated interactions, the model has shown to improve its accuracy to create LMPs even for the indirect expressions like *"I am really in a hurry now"*. Their system has shown to significantly reduce takeover rate in diverse driving scenarios, indicating enhanced human trust. The Intelligent Driving Assistant System (IDAS) framework expands on this by using RAG to ground driver queries in external knowledge bases like vehicle manuals or other regulatory documents [17]. This reasoning moves beyond "what action to perform" towards "why this request is made", and it also enables systems to contextualize instructions according to safety criteria or operational restrictions.

Intent is rarely expressed in a neutral form as emotional states, stress levels, and other affective components greatly alter the meaning of command. Context-aware frameworks are increasingly exploring how such factors shape interpretation of intent [14]. As explored by Tavakkoli et al. [21], LLMs function as a "cognitive bridge", integrating emotion and fatigue recognition with intent reasoning. Their framework shows that same command (e.g., "slow down") carries different urgency depending on whether it is uttered under stress or casually. Emotion-LLaMA also provides a blueprint for multimodal affect-aware intent reasoning. They propose combining audio, visual, and text encoders to align emotional tone with semantic meaning [27]. These concepts illustrate that intent cannot be separated from *how* is it expressed, hence integrating affective modulations increases personalization and reliability. Supporting these concepts, a preliminary study introduced an in-vehicle conversational agent, CARA, that uses multi-turn dialogues to turn its responses more empathetic and tested it against pre-determined responses[14]. Their approach significantly enhanced perceived competence and trust among drivers. However, these studies also show the risk of reducing robustness when emotional cues are subtle, ambiguous, or culturally varied [14,24,37].

Together, these studies highlight the importance of reasoning over driver commands but also highlight multiple trade offs. Structured parsing of intent provides precision but struggles with ambiguity, field experiments adapt to variation but rely on limited data, and emotion/affect-aware systems capture the nuances of ambiguous commands but can limit robustness. However, intent is not interpreted in isolation. Humans tend to repeat, refine, or contradict their own instructions over time. To remain effective, another layer of reasoning, utilizing memory and personalization mechanisms has been explored among the researchers.

 doi:10.20944/preprints202603.0903.v1

5.4. Memory and Personalization

As intent reasoning interprets what the driver means in the moment, assistance systems remain limited if they restart the inference process with every interaction [13]. Without continuity, interactions might become repetitive, mechanical, and poorly aligned with driver expectations [13,42]. Memory and personalization mechanisms are being explored to address this limitation. Memory enables the system to accumulate and organize knowledge across interaction and personalization ensures that the system adapts its responses to the same driver over time. Collectively, they move driver assistance from reactive support to adaptive collaboration.

A prominent architectural pattern for achieving this involves integrating a dedicated memory module as a core component of the agent's framework. This design moves beyond the short context window of a single query to establish an evolving knowledge base [13]. For instance, the LLM-MW system [15] initializes a detailed "individualization profile" for each driver, which is stored in its memory to guide the generation of multimodal warnings. This profile is built and continuously updated by retriving domain-specific knowledge relevant to the driver's demographic, physical, and experiential characteristics. Other frameworks have further exemplified this approach by designing their system around the memory component. For example, the LLM-PDA framework [8] also constructs personalization profiles that are continuously updated through a dedicated memory system that utilizes MultiQuery-RAG for efficient retrieval. Talk2Drive framework [42,43] similarly incorporates a memory component to log past interactions and user feedback. Thus, enabling the system to adapt to the driver's phrasing and preferences over repeated interactions. Another framework that uses ChatGPT-based driver assistant [13], extends this approach by structuring memory into working, semantic, and episodic layers. This approach offers a more human-like model of remembering past events to support decision making. The Actor-Reasoner framework [34] also integrates an interaction memory database, however in a lighter form. Their memory module retrieves relevant experiences to inform the system to reason about new driving scenarios. These works demonstrate that explicit memory architectures are important to transform assistance into a continuous learning process.

Building on this architectural foundation, the functional application of personalization aims to create a more human-centric and intuitive interaction between the driver and vehicle. Multiple systems achieve this by adapting their behavior to driver's immediate cognitive, emotional, and preferential states. The RLHF framework has been used to align vehicle decision policies with human comfort and safety preference [28]. This approach represents the population-level personalization where the system's behavior is tuned by aggregated feedback. At a more individual level, the D-Twins framework [36] embodies a user's unique personality and emotional traits by training them on their specific dialogue data, hence enhancing emotional resonance with the user. More lightweight adaptations of personalization are also explored in approaches such as the fine-tuning and quantization framework by Song et al. [26]. Their approach identifies drivers through facial recognition and retrieves stored personalized guidance while remaining resource efficient. Similarly, other studies [3,17,21] also note that psychophysiological, preferential, and environmental cues can also trigger personalization component and are essential for trust and acceptance of autonomous systems.

Ultimately the goal of memory and personalization components is to enable systems that continuously learn and align their decision making with the user preferences. However, they also introduce new challenges like protecting the driver data, ensuring retrieval efficiency for real-time adaptation, and avoiding cultural biases in personalization aspects. Despite these limitations, such approaches are not just optional enhancements, but essential components of LLM-based driver assistance to build more intuitive and trustworthy systems.

6. Interaction and Intervention

So far, we demonstrated how LLMs reason to interpret driver states and contextual cues, but their value to the user lies in how assistance is delivered to them. Interaction and intervention form the outward-facing dimension of these systems, where plans and predictions are translated into feedback,

 doi:10.20944/preprints202603.0903.v1

guidance, or a dialogue that drivers can perceive and act upon. The emphasis of this section is more about practicality; which outputs and design systems actually work in the cabin.

6.1. Output Modalities and Design Considerations

Driver assistance systems rely on three primary channels for feedback; visual, auditory, and haptic [40]. Visual cues are particularly rich sources of information, ranging from dashboard icons, texts, to projections on HUDs, but they depend on the driver's gaze [21]. While auditory (ranging from beeps to spoken commands) and haptic (like steering wheel or seat vibrations) modalities prove effective in capturing attention regardless of visual focus [21,40]. Early systems often relied on a single channel, but research has consistently shown that multimodal feedback outperforms unimodal approaches in reaction times, accuracy, and satisfaction [15].

Recent LLM systems have started to build on this insight by introducing dynamic modality selection. The LLM based multimodal warning system and LLM-PDA framework [8,15] integrates a planning module that selects what to warn as well as the most effective modality to relay the warning based on contextual factors and driver profiles. Beyond warning and alerts, conversational modalities have also emerged to foster closed-loop interaction mechanisms. Systems like voice-based chatbot, for fatigue mitigation [33], or D-Twins, for boredom mitigation, use a dialogue agent to engage drivers and extend feedback from alerts to ongoing, adaptive interactions. These examples illustrate a shift from static warnings towards more flexible, multimodal interaction strategies where LLMs decide how information is conveyed to maximize salience and minimize disturbance.

Effective interaction design extends beyond selecting modalities, it requires adhering to human-centric principles to ensure interactions are helpful rather than harmful. A primary goal of such interactive systems is managing the driver's cognitive load, and other critical states, as poorly designed or mistimed feedback can become a distraction, inducing stress or even panic [8,21,40]. Following this principle, Xiang et al. proposed an LLM-based persuasion tool that strategically delivers "humanized" persuasive advice, assessing real-time road risks and driver attention to provide load-aware interactions [39]. Other human-centered approaches emphasize that drivers are more likely to accept interventions if they are personalized, empathetic, and transparent [24,41,43]. This was further demonstrated by an empathetic conversational agent proposed by Huang et al., that adapted its interactions based on the induced driver emotional states [37]. Their evaluations against predetermined baselines showed that affective systems led to higher compliance, but conflicting emotions also significantly increased safety critical scenarios. Researchers are also moving towards design choices that address accessibility issues. For instance, shape coding instead of color coding was proposed for the visual warnings for drivers with color-vision deficiency. In the same study, more visual than auditory warnings were provided to foreign drivers who do not understand the local language [15].

These insights highlight that the effectiveness of LLM-based interventions depends on the careful choice of modalities and adherence to human-centric design principles. These principles also help in building driver trust, a prerequisite for the acceptance and adoption of advanced driving technologies.

6.2. Open- vs closed-loop

When considering how driver assistance is delivered, another important factor is the system's underlying adaptivity, ranging from static, predefined responses to real-time, context- and driver state-aware interventions. We divide LLM-based assistance systems into three paradigms based on their interaction loop i.e., open-loop, hybrid, and closed-loop systems. This spectrum reflects a progression towards increasingly sophisticated and human-centric interactions, where the system's capacity to perceive, reason, and react to the dynamic user and environmental states dictates the depth of human-AI collaboration.

Open-loop systems deliver interventions based on predefined rules, without adapting to the driver's real time state or environmental feedback. In this paradigm, driver is largely a passive recipient of information, and the systems responses are fixed. While foundational and low-weight due to their computational simplicity, these systems can sometimes fall short in complex or ambiguous

 doi:10.20944/preprints202603.0903.v1

scenarios where adaptation is critical. For example, studies that targeted fatigue mitigation or engaging drivers using multimodal reminders, mainly used pre-generated outputs as their interactions. These systems, although effective in their isolated tasks, would have less impact under complex and dynamic real world changes due to their lack of feedback integration [33,40].

Hybrid systems advance beyond open-loop models by integrating LLM reasoning with elements of feedback and action, though their adaptation is typically partial or discrete rather than continuous. Here, the LLMs function as intelligent intermediary, performing high level, human-like reasoning, which then informs a subsequent action or interaction mechanism. There are multiple approaches through which such hybrid interactions are achieved. Command interpretation and action generation systems, where LLMs interpret diverse human intentions, translate them into actionable vehicle controls or interaction strategies [18,42,43]. Another such hybrid system augments LLM reasoning with external knowledge bases or iterative refinement processes to provide context-based aids or personalized responses. Examples of such systems include the IDAS framework that utilizes RAG to ground driver queries in vehicle manuals [17], the actor-reasoner framework and RLHF loop that performs iterative feedback refinement to bias decision policies towards safer, driver-preferred outcomes [28,34].

Closed-loop systems push further towards continuous adaptation, with interventions shaped by real-time monitoring of driver states and contextual cues [39]. D-twins for boredom framework exemplifies it by detecting boredom in real time through physiological measures and re-engaging driver with adaptive dialogue [36]. Another study that caters to stress mitigation of drivers, altered the length of interaction depending on the real time monitoring of driver stress [16]. A cross-domain study in aviation also demonstrated the importance of adaptive modification of modality and content of interaction depending upon pilot's mental workload, attention, and memory [10]. These systems embody the vision of a cognitive co-pilot, where human states and preferences are continuously integrated in the loop (Fig. 4). However, they introduce even higher computational overhead, and privacy of sensitive, continuously observed data.

Finally, the effectiveness of driver assistance also depends on whether the driver can understand and trust the system's reasoning. Current studies already explore different ways to achieve this through interaction design. LLMs enable assistance systems to articulate their decision-making process in natural language, moving beyond "black-box" operations to foster trust between the technology and its users [16,43]. Furthermore, as discussed above, adaptive and personalized transparency through closed-loop interactions is also crucial for enhancing trust and overall driving experience. For example, short and concise feedback is less stressful in manual driving whereas longer, more detailed explanations can enhance feelings of safety in AV by increasing the understanding of system's decision making [16]. The concept of bidirectional transparency, where the system not only explains its actions but also shows that it understands the user inputs and intent, also significantly increases user trust, especially in complex and safety-critical scenarios [45]. Collectively, these interactive explainability mechanisms are important for ensuring that LLM-driven assistance is not just functional, but also comprehensible, accepted, and genuinely trusted by its users.

7. Datasets and Benchmarks

Bridging the gap between perceptive assistance and intuitive interaction, the development of LLM-driven in-cabin systems critically relies on comprehensive and diverse datasets. These datasets provide the ground truth for training models that can understand, reason, and adapt to the complexities of human behavior and dynamic driving environments. This section reviews the key datasets and benchmarks that underpin the advancements of LLM based driver monitoring and interaction (Table 3).

The first category comprises datasets for driver state detection. Widely used distraction corpora such as StateFarm [46], SynDD1 [47], DMD [48], NTHU-DDD [49] have enabled recognition of secondary tasks and are still benchmarks for vision-language approach [29][32]. More advanced datasets,

 doi:10.20944/preprints202603.0903.v1

like AIDE [3] and 3MDAD [50], extend this scope to include multi-view and multimodal streams for distraction, drowsiness, and emotion [38]. Specialized datasets expand the modality space, e.g., thermal datasets (TFW [51], SF-TL54 [52]) for robust monitoring under low-lighting conditions [20]. For emotion detection, some studies use general datasets like KMU-FED and FER2013 or AIDE for five basic emotions in driving context [23,27]. However, for more sophisticated emotional reasoning, datasets like MERR and CA-MER [53] offer extensive fine-grained multimodal annotations (visual, audio, text) [23]. These datasets illustrate progress but also highlight fragmentation as each focuses on a particular driver state, underscoring the need to build integrated resources that span multiple states and affect in naturalistic conditions.

A second category consists of instruction datasets and benchmarks, which enable LLM systems to evolve into cognitive co-pilots by understanding natural language commands. Language-augmented datasets like Talk2Car [9], nuScenes-QA [54], and DriveLM [55] enrich existing autonomous driving data with natural language commands and visual question answering, facilitating the interpretation of human intent [9,44]. BDD-X further contributes with textual explanations for vehicle's self driving actions, enhancing transparency [6]. To evaluate end-to-end instruction following, benchmarks like LaMPilot-Bench and UCU Dataset quantitatively assess LLM's ability to reason and classify system requirements based on driver's demands, from safety instructions to comfort requests [44][35]. These datasets are crucial for developing LLM-driven systems that can monitor driver states, infer and interpret, and translate human intent to personalized and contextually appropriate actions. Yet, they remain limited in scale and diversity, reinforcing the need for integrated benchmarks for cohesive evaluation of intelligent driver assistance systems.

Finally, a notable gap lies in datasets that integrate external driving context with in-cabin states. While general autonomous driving datasets like KITTI [56], nuScenes [57], and BDD100K [58] dominate perception research in tasks such as 3D object detection and motion planning, the driver assistance systems rarely employ them directly. It has been widely demonstrated that external scene factors such as traffic density, weather, or lighting conditions strongly influence driver states including stress, cognitive load, and distraction [8,17,44]. Therefore, bridging these datasets with in-cabin corpora would allow copilots to align driver state recognition with environmental triggers, enabling more timely, contextually grounded, and empathetic interventions.

 doi:10.20944/preprints202603.0903.v1

Table 3. Datasets used in reviewed studies, with their focus and modalities.

Type	Focus	Datasets	Modalities	Utilization in Reviewed Studies
Driver State	Distraction	StateFarm	RGB	[30] [29]
		SynDD1, DMD	RGB	[29]
		SAM-DD	RGB	[29][32]
	Drowsiness	NTHU-DDD	RGB	[32]
	Emotion	KMU-FED	RGB	[32][23]
		FER2013	RGB	[23]
		MERR, CA-MER	Multimodal (Audio/Visual/Text)	[24]
	Multi-state	AIDE (Emotion, Drowsiness, Distraction)	RGB (internal & external view)	[22][38]
		3MDAD (Distraction, Drowsiness)	RGB	[22][38]
		TFW / SF-TL54 (Distraction, Drowsiness)	Far-infrared spectrum	[20]
Instruction	Command grounding	UCU	Text	[35]
		Talk2Car	RGB + LiDAR + RADAR + GPS + Text (commands)	[9]
	Reasoning / QA	DriveLM, nuScenes-QA	RGB + LiDAR + RADAR + GPS + textual Q/A	[17]
	Program / Policy synthesis	LaMPilot-Bench	Text (commands) + simulation environment	[44]

8. Discussion

This review highlights how the integration of LLMs into driver assistance has moved from fragmented state detectors towards reasoning-driven, adaptive, and interactive copilots. Yet the opportunities and tensions that emerge extend beyond technical improvements; they redefine the interaction paradigm between humans and vehicles. In this discussion, we analyze these tensions across four threads: first, the paradigm shift from traditional ADAS to cognitive co-pilots; second, critical design tradeoffs in human-centered systems; third, rethinking trust in driver assistance; and finally, broader implications for research and ethics.

8.1. From ADAS to Cognitive Co-Pilots

Driver monitoring and assistance systems have traditionally been built around modular ADAS pipelines that perceive signals and interpret them into predefined categories. While this architecture has yielded some progress, our review shows it remains fragmented. For instance, only a handful of studies integrate detection and reasoning about context; and interventions are mostly delivered without considering how drivers will make sense of them.

However, studies show that LLMs and MLLMs are capable of enabling a more integrated and human-centric approach. Systems are emerging that utilize LLMs as the central decision making brains, enabling a loop from perception to feedback adaptation (Fig. 4). This involves integrating detection, reasoning over human and environmental context, decision-making, human-machine interaction, and continuous learning into unified frameworks. So far, at state interpretation level, works like VLM-DM [32], DriveCLIP [29] and DDLM [18] demonstrate that VLMs can generalize multiple states across datasets with minimal fine-tuning. At reasoning level, LLMs and VLMs are being introduced to deliver contextual prompts, reasoning chains, and uncertainty calibrations to connect raw signals with explanations and risk assessments [18,22,38]. At the action level, frameworks increasingly embed reasoning into interventions, parsing vague or effective commands or tailoring warnings to driver

 doi:10.20944/preprints202603.0903.v1

profiles, with some pipelines also closing the loop by providing human feedback to enhance system's learning [16,36,42,43].

Despite this progress, most systems fall short of integrating all four dimensions within a single deployable pipeline. Many excel at interaction design but lack robust contextual reasoning, while others demonstrate advanced reasoning without tight coupling to real-time perception. These limitations are primarily driven by computational complexity and latency constraints. Due to their scale, LLMs demand substantial compute resources and often incur nontrivial inference delays. Multimodal architectures further increase memory and processing requirements, particularly when continuous learning or memory components are included, limiting feasibility for embedded in-cabin deployment. Addressing these constraints is therefore central to advancing integrated driver assistance systems.

A promising direction lies in adopting a "train heavy, deploy light" paradigm, in which resource-intensive multimodal reasoning is leveraged during development and distilled into compact models suitable for in-vehicle execution (Fig. 5). This can be achieved through techniques such as knowledge distillation, quantization, and edge deployment. For example, quantized LLMs such as PHI-3 have demonstrated improved inference efficiency in pilot assistance systems [10], while smaller VLMs like Idefics2 are increasingly favored for their computational efficiency [4]. Memory optimization is also critical for reducing system overhead and supporting long-term adaptation, with approaches including selective retention, driver profiling, and lightweight memory modules [13,15]. However, open challenges remain in designing efficient retention strategies and privacy-preserving mechanisms that enable continuous learning without excessive storage or data exposure. Moving beyond siloed state detection, future systems must also integrate multiple safety-critical driver states and affective factors within unified reasoning and action pipelines. This requires leveraging LLMs to interpret rich multimodal inputs, contextualize them with external factors, and translate system inferences into timely, deployable responses [21].

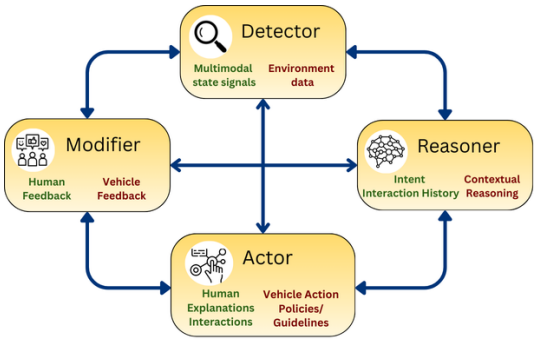

Figure 4. Conceptual closed-loop framework for LLM-based driver assistance.

8.2. Design Tensions in Human-Centric Driver Assistance

The development of LLM-based driver assistance systems is not only a technical challenge but also a design one. These tensions highlight that effective systems cannot be optimized on a single axis i.e., personalization, transparency, or efficiency, without carefully considering their trade-offs. Below, we discuss three recurring tensions and outline potential directions to address them.

8.2.1. Personalization vs. Generalization

Several systems emphasize personalization by creating driver profiles, memory modules, or behavior and cognitive models that adapt interactions to individual characteristics and preferences [8,36,42]. While these approaches show promise in improving user acceptance and reducing false positives and unnecessary interactions, they risk overfitting idiosyncrasies that may not translate across various driver profiles, cultures, or driving contexts. In contrast, training-free or evidential-fusion models demonstrate greater robustness across multiple datasets but remain limited in catering to individual needs [22,38]. This tension highlights the lack of evaluation frameworks that test both personalization and generalization simultaneously. Moving forward, frameworks that employ shadow

 doi:10.20944/preprints202603.0903.v1

learning and contextual personalization, where systems learn and adapt for certain drives and then deploy, can be studied (Fig. 5). Such systems could maintain general safety actions and let phrasing, timing, and modality according to individual preferences. Additionally, validation studies could utilize split-by-driver or split-by-demographics protocols to report what transfers and what does not [8,43].

8.2.2. Transparency vs. Cognitive Load

Explainability of system's planning and decisions is increasingly viewed as an important factor to foster driver trust. However, studies also highlight that more explanation does not always equate to better outcomes. Specifically, in high-risk or stressful situations, shorter feedback is considered more efficient and non-intrusive as compared to detailed explanations [16,33]. On the other hand, autonomous systems with higher agency, when provide detailed descriptions, lead to fewer takeovers and higher satisfaction [16]. The key challenge is not the absence of explainability but the lack of workload-informed mechanisms for when and how to provide it [39]. A potential solution is to reconceptualize co-pilots as independent agents that not only track driver's state but also maintains awareness of the vehicle's operational design domain (ODD) (Fig. 5) [17,44]. By combining the two streams of information, these systems can mediate between the driver and ADAS, delivering explanations that are both contextually grounded and do not risk information overload. This framing could support more adaptive explanation strategies, such as tailoring modality, verbosity, and timing, based on real-time workload assessments. Explanations could also be provided in multiple stages i.e., brief instruction or feedback during high workload situations and more detailed explanations on demand.

8.2.3. Efficiency vs. Reliability

As previously discussed, computational costs and latency issues have motivated researchers to develop more efficient and lightweight systems that rely on quantization or knowledge distillation [10]. Although these systems improve on technical integration challenges, they can compromise the subtlety and reliability of state detection and action strategies. For instance, VLM-DM [32] demonstrates that LoRA can unify multiple state interpretation with competitive accuracy, but still struggles to reliably reason over overlapping features (reaching behind or dropping down). To mitigate this tradeoff, reliability must be treated as a primary criterion, with efficiency serving as a constraint rather than goal. Approaches such as "train heavy, deploy light" offer a way forward (Fig. 5). Reliability can also be enhanced by integrating fallback mechanisms, such as escalating uncertain or ambiguous cases to slower but more robust models [34]. Adding modality redundancy, like using thermal vision under low lighting conditions or incorporating various state cues from multiple modalities, can also mitigate failures in specific cases [20,21].

8.3. Rethinking Trust

Trust in ADAS has traditionally been tied to the predictability and reliability of the system, if the system works smoothly, without interruptions, drivers could trust it. However, with autonomous systems working on higher agency and LLMs supporting with higher reasoning tasks, this conception is no longer sufficient. Trust now hinges on whether the systems can articulate their reasoning, account for uncertainty, and demonstrate self-awareness in their outputs. Frameworks that quantify uncertainty or provide stepwise reasoning explanations have begun to shift trust from binary correctness toward calibrated confidence [18,22,38]. At the same time, state-aware explanations offer a way to manage cognitive load and overwhelming interactions [10,37,39]. Beyond transparency, trust is also shaped by whether the drivers feel the system "knows" them and adapts to their characteristics. Personalization strategies show promise in strengthening the bond between drivers and vehicles, however with critical caveats. As previously discussed, highly personalized systems raise tensions in scalability, privacy, and cultural diversity, since over-standardized empathy can feel manipulative, without personalizing on the individual level [13,36,42]. Accessibility further complicates this picture: multi-modal warnings

 doi:10.20944/preprints202603.0903.v1

may foster trust in general, but if visual cues rely solely on color, or auditory responses overlook sensory overload, trust will erode for users with different cognitive and perceptual needs. Few studies have adopted approaches to cater to users with color-vision deficiency or balancing between modalities to achieve inclusivity [8,15]. However, there remains a significant gap in addressing accessibility requirements for a broader spectrum of users, including those with age-related changes, sensory and motor impairments, neuro-diverse conditions (e.g., ADHD, autism), and psychological or affective challenges such as heightened anxiety or stress.

Future works in intelligent driver assistance systems must reframe trust as a design goal that integrates adaptive explainability, context personalization, and accessibility by default. Promising directions include: maintaining driver profiles by shadow learning and continuous feedback integration, explain-on-demand mechanisms that balance trust with overwhelming interactions, and adaptive monitoring of safety critical and pathological states, in order to deliver human-like responses and interventions. Equally, the traditional reliability-dependent trust must be safeguarded by constantly improving vehicle actions and incorporating fail-safe mechanisms to mimic human driving behaviors. As ultimately, the credibility of cognitive co-pilots will rest on both perfect predictability and decision making, and how transparent, empathetic, and inclusive they behave with human drivers.

Figure 5 summarizes our design recommendations for future research on LLM-based driver assistance. We propose conceptualizing the cognitive co-pilot as an agent distinct from the ADAS, mediating between vehicle systems and the human driver. The co-pilot integrates both heavy (deep, fallback) and lightweight (fast, low-load) subsystems, enabling adaptive reasoning, context-sensitive explanations, and reliable actions. By separating this reasoning layer from ADAS control, the architecture supports human-centric collaboration while balancing efficiency, transparency, and trust.

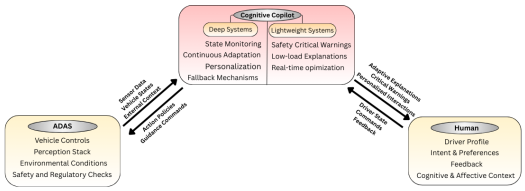

Figure 5. Design Recommendations for Future Driver Assistance Systems.

8.4. Datasets, Methods, and Ethical Considerations

Another hurdle in the progress towards cognitive co-pilots is the infrastructure that supports their development and deployment. Existing datasets remain fragmented, with some focusing on task specific driver states, other on traffic scenes, and a very small subset on language-grounded instructions. With only a few benchmarks integrating multimodality sensing with human-centric instructions, this lack of cohesion hinders comparability and evaluation across diverse datasets and frameworks. Future work must prioritize integrated benchmarks that combine multiple driver state signals with contextual reasoning and natural-language annotations. At the same time, LLMs can support data augmentation, for example by generating synthetic corner cases or complex scenarios that reflect real driving conditions. This is an important future direction for HCI, as the human-partnership capabilities of new LLMs call for benchmarks that are human-centric by design.

Methodologically, most works continue to rely on dataset-bound metrics for evaluation with limited ecological testing. As shown in the reviewed studies, models perform strongly on benchmarks but falter in real-life or simulated testing with human participants. There is a growing need for hybrid methodologies where scalability of automated benchmarks can be combined with the depth of human-centered user studies. Evaluations should extend beyond accuracy checks to include results central to HCI, such as explanation quality, acceptance, and trust.

Finally, the integration of driver's sensitive data e.g., faces, voices, and other physiological signals raises critical ethical concerns. Ensuring privacy-preserving machine learning, for instance, on-device inference and federated learning, are essential for such systems. Additionally, systems should also account for transparent data collection practices with explicit user agency in deciding what is shared.

 doi:10.20944/preprints202603.0903.v1

9. Conclusions

This survey reviewed how LLMs are being integrated into driver monitoring and assistance, moving the field from rigid detectors toward reasoning-driven and interactive systems. We synthesized progress in driver state detection, reasoning over driver context, interaction design, and supporting datasets, while also highlighting persistent design tensions around personalization, transparency, efficiency, and trust. Our contribution lies in framing these developments through an HCI lens, providing design recommendations (Fig. 5), and emphasizing the need for research infrastructure that reflects human partnership rather than narrow classification. In reviewing this literature, a key challenge was that many of the systems remain prototypes or proofs of concept, with findings often constrained to limited datasets or simulation. Looking forward, we identify several directions for future work, including the development of multimodal and multi-agent systems that can close the loop of human–vehicle interaction, integrated benchmarks that combine driver state with contextual and natural language annotations, and more real-world evaluations to assess the usability of such systems. Overall, this survey argues for a closer integration between sensing technologies, language-based reasoning, and human-centered evaluation to support the development of transparent, trustworthy, and deployable driver assistance systems.

Author Contributions: Conceptualization, I.A. and V.C.; methodology, V.C.; validation, V.C. and I.A.; resources, V.C. and I.A.; data curation, V.C.; writing—original draft preparation, V.C.; writing—review and editing, I.A.; visualization, V.C.; supervision, I.A.; project administration, I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, version 5.2) for the refinement of language. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Li, J.; Li, J.; Yang, G.; Yang, L.; Chi, H.; Yang, L. Applications of large language models and multimodal large models in autonomous driving: A comprehensive review **2025**.
2. Commission, E.; Directorate-General for Internal Market, Industry, E.; SMEs.; TRL.; Seidl, M.; Edwards, M.; Hynd, D.; McCarthy, M.; Livadeas, A.; Carroll, J.; et al. *General Safety Regulation – Technical study to assess and develop performance requirements and test protocols for various measures implementing the new General Safety Regulation, for accident avoidance and vehicle occupant, pedestrian and cyclist protection in case of collisions – Final report*; Publications Office, 2021. https://doi.org/doi/10.2873/499942.
3. Yang, D.; Huang, S.; Xu, Z.; Li, Z.; Wang, S.; Li, M.; Wang, Y.; Liu, Y.; Yang, K.; Chen, Z.; et al. Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20459–20470.
4. Cañas, P.N.; Nieto, M.; Otaegui, O.; Rodríguez, I. Exploration of VLMs for Driver Monitoring Systems Applications. *arXiv preprint arXiv:2503.12281* **2025**.
5. Rumpf, S. Sleepy driver puts Tesla on autopilot, leading police on high-speed chase in Germany, 2023.
6. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.D.; et al. A survey on multimodal large language models for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2024, pp. 958–979.
7. Yan, Y.; Liao, Y.; Xu, G.; Yao, R.; Fan, H.; Sun, J.; Wang, X.; Sprinkle, J.; An, Z.; Ma, M.; et al. Large language models for traffic and transportation research: Methodologies, state of the art, and future opportunities. *arXiv preprint arXiv:2503.21330* **2025**.

 doi:10.20944/preprints202603.0903.v1

8. Xu, Z.; Chen, T.; Huang, Z.; Xing, Y.; Chen, S. Personalizing driver agent using large language models for driving safety and smarter human–machine interactions. *IEEE intelligent transportation Systems magazine* **2025**.
9. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 902–909.
10. Wen, S.; Middleton, M.; Ping, S.; Chawla, N.N.; Wu, G.; Feest, B.S.; Nadri, C.; Liu, Y.; Kaber, D.; Zahabi, M.; et al. AdaptiveCoPilot: Design and Testing of a NeuroAdaptive LLM Cockpit Guidance System in both Novice and Expert Pilots. In Proceedings of the 2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR). IEEE, 2025, pp. 656–666.
11. Gao, Y.; Yue, L.; Sun, J.; Shan, X.; Liu, Y.; Wu, X. WorkloadGPT: A Large Language Model Approach to Real-Time Detection of Pilot Workload. *Applied Sciences* **2024**, *14*, 8274.
12. Wahab, O.; Adda, M. Comprehensive Literature Review on Large Language Models and Smart Monitoring Devices for Stress Management. *Procedia Computer Science* **2025**, *257*, 166–173.
13. Wang, S.; Zhu, Y.; Li, Z.; Wang, Y.; Li, L.; He, Z. ChatGPT as your vehicle co-pilot: An initial attempt. *IEEE Transactions on Intelligent Vehicles* **2023**, *8*, 4706–4721.
14. Bond, Y.L.; Choe, M.; Hasan, B.K.; Siddiqui, A.; Jeon, M. ChatGPT on the Road: Leveraging Large Language Model-Powered In-vehicle Conversational Agents for Safer and More Enjoyable Driving Experience. *arXiv preprint arXiv:2508.08101* **2025**.
15. Xu, Z.; Chen, T.; Chen, S. A LLM-based Multimodal Warning System for Driver Assistance. In Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2024, pp. 1527–1532.
16. Markelius, A.; Lou, Y.; Galazka, M.; Lundgren, S.; Zemblys, R.; Lind, H.; Lowe, R. Investigating the Mitigation of Stress in Autonomous and Non-autonomous Vehicles Using LLM Feedback. In Proceedings of the International Conference on Human-Computer Interaction. Springer, 2025, pp. 108–127.
17. Hernandez-Salinas, B.; Terven, J.; ChaveZ-Urbiola, E.; Cordova-Esparza, D.M.; Romero-Gonzalez, J.A.; Arguelles, A.; Cervantes, I. Idas: Intelligent driving assistance system using rag. *IEEE Open Journal of Vehicular Technology* **2024**.
18. Zhang, K.; Wang, S.; Jia, N.; Zhao, L.; Han, C.; Li, L. Integrating visual large language model and reasoning chain for driver behavior analysis and risk assessment. *Accident Analysis & Prevention* **2024**, *198*, 107497.
19. Wu, Y.; Li, D.; Chen, Y.; Jiang, R.; Zou, H.P.; Fang, L.; Wang, Z.; Yu, P.S. Multi-agent autonomous driving systems with large language models: A survey of recent advances. *arXiv preprint arXiv:2502.16804* **2025**.
20. Knapik, M.; Cyganek, B.; Balon, T. Multimodal driver condition monitoring system operating in the far-infrared spectrum. *Electronics* **2024**, *13*, 3502.
21. Tavakkoli, V.; Mohsenzadegan, K.; Kyamakya, K. Leveraging Context-Aware Emotion and Fatigue Recognition Through Large Language Models for Enhanced Advanced Driver Assistance Systems (ADAS). In *Recent Advances in Machine Learning Techniques and Sensor Applications for Human Emotion, Activity Recognition and Support*; Springer, 2024; pp. 49–85.
22. Hu, C.; Li, X. Human-Centric Context and Self-Uncertainty-Driven Multi-Modal Large Language Model for Training-Free Vision-Based Driver State Recognition. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
23. Li, B.; Luo, M.; Zhang, D. Facial Emotion Detection Research Based on an Improved Multi-modal LLM. In Proceedings of the 2024 3rd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR). IEEE, 2024, pp. 61–66.
24. Chen, X.; Wang, X.; Fang, C.; Fang, L.; Gong, W.; Liu, C.; Wang, S.J. Emotion-aware Design in Automobiles: Embracing Technology Advancements to Enhance Human-vehicle Interaction. In Proceedings of the Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–18.
25. Taveekitworachai, P.; Suntichaikul, P.; Nukoolkit, C.; Thawonmas, R. Speed up! Cost-effective large language model for ADAS via knowledge distillation. In Proceedings of the 2024 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2024, pp. 1933–1938.
26. Song, G.; Lim, J.; Jeong, C.; Kang, C.M. Enhancing Inference Performance of a Personalized Driver Assistance System through LLM Fine-Tuning and Quantization. In Proceedings of the 2025 International Conference on Electronics, Information, and Communication (ICEIC). IEEE, 2025, pp. 1–4.

 doi:10.20944/preprints202603.0903.v1

27. Cheng, Z.; Cheng, Z.Q.; He, J.Y.; Wang, K.; Lin, Y.; Lian, Z.; Peng, X.; Hauptmann, A. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems* **2024**, *37*, 110805–110853.
28. Sun, Y.; Salami Pargoo, N.; Jin, P.; Ortiz, J. Optimizing autonomous driving for safety: A human-centric approach with LLM-enhanced RLHF. In Proceedings of the Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2024, pp. 76–80.
29. Hasan, M.Z.; Chen, J.; Wang, J.; Rahman, M.S.; Joshi, A.; Velipasalar, S.; Hegde, C.; Sharma, A.; Sarkar, S. Vision-language models can identify distracted driver behavior from naturalistic videos. *IEEE Transactions on Intelligent Transportation Systems* **2024**, *25*, 11602–11616.
30. Girbacia, F.; Voinea, G.D.; Danu, M.D.; Buzdugan, I.D.; Duguleana, M. Detection of Phone Distraction While Driving Using Open Visual-Language Models. In Proceedings of the International Congress of Automotive and Transport Engineering. Springer, 2024, pp. 281–286.
31. Dona, M.A.M.; Cabrero-Daniel, B.; Yu, Y.; Berger, C. Evaluating and Enhancing Trustworthiness of LLMs in Perception Tasks. In Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2024, pp. 431–438.
32. Chi, H.; Yang, H.; Yang, L.; Lv, C. VLM-DM: Visual Language Models for Multitask Domain Adaptation in Driver Monitoring. In Proceedings of the 2025 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2025, pp. 1280–1285.
33. Huang, S.; Zhao, X.; Wei, D.; Song, X.; Sun, Y. Chatbot and fatigued driver: Exploring the use of LLM-based voice assistants for driving fatigue. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–8.
34. Fang, S.; Liu, J.; Xu, C.; Lv, C.; Hang, P.; Sun, J. Interact, instruct to improve: A llm-driven parallel actor-reasoner framework for enhancing autonomous vehicle interactions. *arXiv preprint arXiv:2503.00502* **2025**.
35. Yang, Y.; Zhang, Q.; Li, C.; Marta, D.S.; Batool, N.; Folkesson, J. Human-centric autonomous systems with llms for user command reasoning. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 988–994.
36. Lo, I.C.; Rau, P.L.P. D-Twins: Your Digital Twin Designed for Real-Time Boredom Intervention. In Proceedings of the Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–15.
37. Huang, B.; Lv, J.; Qiang, L. Influencing driving safety by matching AI assistant's verbal emotions to driver: A randomized controlled trial on performance, attention, and emotion. *Computers in Human Behavior* **2025**, *169*, 108667.
38. Hu, C.; Li, X.; Pang, J. Trustworthy Driver State Perception via Contextual Interaction-Driven Evidential Vision-Language Fusion in Vehicular Cyber-Physical Systems. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
39. Xiang, W.; Li, M.; Yan, J.; Zheng, M.; Zhu, H.; Jiang, M.; Sun, L. Driver Assistant: Persuading Drivers to Adjust Secondary Tasks Using Large Language Models. *arXiv preprint arXiv:2508.05238* **2025**.
40. Zou, Z.; Khan, A.; Lwin, M.; Alnajjar, F.; Mubin, O. Investigating the impacts of auditory and visual feedback in advanced driver assistance systems: a pilot study on driver behavior and emotional response. *Frontiers in Computer Science* **2025**, *6*, 1499165.
41. Choe, M.; Bosch, E.; Dong, J.; Alvarez, I.; Oehl, M.; Jallais, C.; Alsaid, A.; Nadri, C.; Jeon, M. Emotion GaRage Vol. IV: Creating empathic in-vehicle interfaces with generative AIs for automated vehicle contexts. In Proceedings of the Adjunct Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2023, pp. 234–236.
42. Cui, C.; Yang, Z.; Zhou, Y.; Ma, Y.; Lu, J.; Li, L.; Chen, Y.; Panchal, J.; Wang, Z. Personalized autonomous driving with large language models: Field experiments. In Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2024, pp. 20–27.
43. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE Intelligent Transportation Systems Magazine* **2024**, *16*, 81–94.
44. Ma, Y.; Cui, C.; Cao, X.; Ye, W.; Liu, P.; Lu, J.; Abdelraouf, A.; Gupta, R.; Han, K.; Bera, A.; et al. Lampilot: An open benchmark dataset for autonomous driving with language model programs. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 15141–15151.
45. Krömker, H. *HCI in Mobility, Transport, and Automotive Systems*; Springer, 2021.
46.

 doi:10.20944/preprints202603.0903.v1

47. Rahman, M.S.; Venkatachalapathy, A.; Sharma, A.; Wang, J.; Gursoy, S.V.; Anastasiu, D.; Wang, S. Synthetic distracted driving (syndd1) dataset for analyzing distracted behaviors and various gaze zones of a driver. *Data in brief* **2023**, *46*, 108793.
48. Ortega, J.D.; Kose, N.; Cañas, P.; Chao, M.A.; Unnervik, A.; Nieto, M.; Otaegui, O.; Salgado, L. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 387–405.
49. Liu, W.; Qian, J.; Yao, Z.; Jiao, X.; Pan, J. Convolutional two-stream network using multi-facial feature fusion for driver fatigue detection. *Future Internet* **2019**, *11*, 115.
50. Jegham, I.; Khalifa, A.B.; Alouani, I.; Mahjoub, M.A. A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD. *Signal Processing: Image Communication* **2020**, *88*, 115960.
51. Kuzdeuov, A.; Aubakirova, D.; Koishigarina, D.; Varol, H.A. TFW: Annotated thermal faces in the wild dataset. *IEEE Transactions on Information Forensics and Security* **2022**, *17*, 2084–2094.
52. Kuzdeuov, A.; Koishigarina, D.; Aubakirova, D.; Abushakimova, S.; Varol, H.A. Sf-tl54: A thermal facial landmark dataset with visual pairs. In Proceedings of the 2022 IEEE/SICE International Symposium on System Integration (SII). IEEE, 2022, pp. 748–753.
53. Han, Z.; Zhu, B.; Xu, Y.; Song, P.; Yang, X. Benchmarking and Bridging Emotion Conflicts for Multimodal Emotion Reasoning. *arXiv preprint arXiv:2508.01181* **2025**.
54. Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; Jiang, Y.G. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 4542–4550.
55. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; Li, H. Drivelm: Driving with graph visual question answering. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 256–274.
56. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *The international journal of robotics research* **2013**, *32*, 1231–1237.
57. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11621–11631.
58. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2636–2645.