Advanced Tool for Traffic Crash Analysis: An AI-Driven Multi-Agent Approach to Pre-Crash Reconstruction

Gerui Xu[a], Boyou Chen[a], Huizhong Guo[b], Dave LeBlanc[b], Ananna Ahmed[c], Zhaonan Sun[c], Shan Bao[a, b*]

[a]*Industrial and Manufacturing Systems Engineering Department, University of Michigan-Dearborn, 4901 Evergreen Rd, Dearborn, MI, 48128, USA*

[b]*University of Michigan Transportation Research Institute, 2901 Baxter Road, Ann Arbor, MI, 48109, USA*

[c]*Toyota Motor Engineering & Manufacturing North America, Inc., USA*

ABSTRACT

Traffic collision reconstruction traditionally relies on human expertise and, when performed properly, can be incredibly accurate. However, attempting to perform pre-crash reconstruction, i.e., reconstructing the driver and vehicle behaviors that preceded the actual crash, poses significantly more challenges. This study develops a multi-agent AI framework that reconstructs pre-crash scenarios and infers vehicle behaviors from fragmented collision data. We present a two-phase collaborative framework combining reconstruction and reasoning phases. The system processes 277 rear-end lead vehicle deceleration (LVD) collisions from the Crash Investigation Sampling System (2017-2022), integrating textual crash reports, structured tabular data, and visual scene diagrams. Phase I generates natural-language crash reconstructions from multimodal inputs. Phase II performs in-depth crash reasoning by combining these reconstructions with the temporal Event Data Recorder (EDR). This enables precise identification of striking and struck vehicles while isolating the EDR records most relevant to the collision moment, thereby revealing crucial pre-crash driving behaviors. For validation, we applied it to all LVD cases, focusing on a subset of 39 complex crashes where multiple EDR records per collision introduced ambiguity (e.g., due to missing or conflicting data). Ground truth was established via consensus between manual annotations (two independent researchers) and AI-assisted labeling, ensuring robustness.

The evaluation of the 39 LVD crash cases revealed our framework achieved perfect accuracy across all test cases, successfully identifying both the most relevant EDR event and correctly distinguishing striking versus struck vehicles, surpassing the 92% accuracy achieved by human researchers on the same challenging dataset. The system maintained robust performance even when processing incomplete data, including missing or erroneous EDR records and ambiguous scene diagrams. This study demonstrates superior AI capabilities in processing heterogeneous collision data, providing unprecedented precision in reconstructing impact dynamics and characterizing pre-crash behaviors.

Keywords: Pre-crash Analysis; Crash Reconstruction; Large Language Model (LLM); Crash Inference; Multi-Agent

1 INTRODUCTION

Road traffic crashes remain a critical public health and safety challenge worldwide. Every year, over a million people lose their lives in roadway accidents, with tens of millions more injured or disabled **[1]**. In the USA, the latest NHTSA report indicates that 40,901 people died in motor vehicle crashes in 2023 **[2]**. Furthermore, Michigan alone experienced over one million motor vehicle crashes from 2021 to 2024, with approximately 500,000 involving two-vehicle collisions. Hence, accurate reconstruction of pre-crash dynamics in these incidents is essential for developing effective prevention strategies and improving vehicle safety

* Corresponding author.
E-mail addresses: **geruixu@umich.edu** (Gerui Xu), **Boyou@umich.edu** (Boyou Chen), **shanbao@umich.edu** (Shan Bao)

systems.

However, traditional pre-crash reconstruction methods face substantial limitations when analyzing complex collision scenarios. Current approaches depend heavily on human experts to manually integrate fragmented evidence from multiple sources—witness accounts, physical evidence, vehicle damage, and Event Data Recorder (EDR) information [3-6]. This process becomes particularly challenging with incomplete data, ambiguous EDR records, and diverse multimodal information including crash narratives, vehicle dynamics, and scene diagrams [7]. The cognitive burden of processing such heterogeneous information may lead to analytical inconsistencies and reduced accuracy in reconstructing critical pre-crash dynamics. These limitations highlight the opportunity for sophisticated analytical frameworks capable of processing fragmented multimodal data to help reconstruct the critical pre-crash time window, where key driver behaviors and vehicle dynamics ultimately determine collision outcomes.

1.1 Pre-crash Reconstruction Challenges

Traditional traffic accident reconstruction methods rely heavily on human experts and researchers to analyze fragmented evidence to reconstruct the scene at the time of the accident, determining causal relationships and liability attribution [8, 9]. The core framework of accident reconstruction remains based on expert experience and rule-based methods, integrating and analyzing limited investigative evidence (such as scene investigation, road environment, eyewitness testimony, and other data), with experts ultimately interpreting, weighing, and making judgments [10, 11]. When performed properly, a qualified expert can, with shocking accuracy, reconstruct an accident. However, it is often of interest to understand the driver and vehicle behaviors that led to an accident, i.e. pre-crash reconstruction. Vehicle data, witness testimony, and evidence at the scene can all be extremely valuable inputs when performing a pre-crash reconstruction, in reality, this process faces the challenge of structurally complex and fragmented data, and traditional methods have limitations in multi-modal data fusion and subjective cognition.

From a data acquisition standpoint, pre-crash reconstruction typically begins with evidence collected post-incident by law enforcement or investigative personnel. In real-world scenarios, such evidence is often incomplete, inconsistent, and multi-modal in nature. Eyewitness statements, for instance, are susceptible to memory decay, constrained vantage points, and psychological stress, all of which degrade the accuracy and completeness of event descriptions [12]. Physical traces at the scene may be altered or lost due to environmental conditions or delayed response, and onboard data devices such as event data recorders (EDRs) may be damaged during impact, leading to partial or missing data [3, 6].

Another key limitation lies in the reliability and consistency of official crash reports, which frequently suffer from misreporting, underreporting, and internal contradictions—a global phenomenon well-documented in transportation safety research [13-15]. Studies have revealed widespread underestimation of critical factors such as pedestrian involvement, alcohol, and drug use. In certain U.S. states, alcohol-related elements are reportedly omitted in 30–50% of fatal or severe crashes [14, 16]. Even widely adopted datasets such as the CISS have been shown to exhibit systemic biases and sampling errors, potentially distorting estimates of crash severity and undermining data fidelity [17]. To compensate for these deficiencies, analysts often resort to manual interpretation of narrative records or cross-validation between data sources—processes that inevitably introduce further uncertainty and subjectivity [18]。

Although EDRs are regarded as reliable and objective sources of pre-crash information, interpreting EDR data presents substantial technical challenges [19]. A significant constraint is the devices' limited temporal scope; standard EDRs typically record only the final 5-10 seconds before a crash. This brief window is frequently

inadequate for documenting the complete chronology of evasive maneuvers or for determining the initial conditions that precipitated the event **[20]**. Another important issue is the significant variation in EDR recording methods across different manufacturers: Sequential Recording, Reverse Recording, and Mixed/Hybrid Recording. The same collision event may trigger multiple EDR recordings, and when multiple collision events have overlapping EDR data, it is difficult for researchers to determine the temporal relationship of events represented by the EDR recordings from the data alone.

Overall, although current pre-crash reconstruction frameworks attempt to integrate multi-source evidence, their reasoning processes remain heavily dependent on expert knowledge, sometimes under conditions of incomplete or uncertain data. Due to these potential limitations, there is an opportunity for intelligent, inference-capable systems that can support higher precision, efficiency, and objectivity in traffic pre-crash analysis.

1.2 AI Power of Large Language Models

The AI boom ignited by ChatGPT in 2022 has led to breakthrough advances in Large Language Models (LLMs) across natural language understanding and generation, multimodal information processing, and few-shot learning [21-23] With emerging capabilities in language comprehension, contextual learning, multimodal reasoning, and human-like decision-making, LLMs have evolved beyond text generation engines into versatile, general-purpose problem-solving frameworks. The technical evolution of LLMs is unfolding along two distinct trajectories： The first involves "General intelligence models" that integrate multimodal understanding capabilities and universal world knowledge (such as GPT-4, Claude-3.5, and Gemini) [24, 25]. These models acquire world knowledge foundations through large-scale pre-training and demonstrate superior performance in tasks requiring cross-domain knowledge integration and multimodal information understanding. The second path, emerging since late 2024, focuses on "Reasoning-centric models" (e.g., GPT-O1, Claude-Thinking, DeepSeek-R1), which exhibit superior performance in logical deduction, complex problem solving, and structured reasoning tasks [26-28]. These reasoning-oriented models have shown particular promise in complex scenarios such as traffic crash analysis, where integrating heterogeneous data and constructing coherent causal chains is critical—especially under conditions of incompleteness or contradiction.

Large language models have demonstrated unprecedented potential in traffic accident analysis. Research indicates that LLMs can effectively integrate heterogeneous multimodal data analysis. For example, AccidentGPT reconstructs accident process videos through multimodal inputs (audio, image, video, text, spatial and/or temporal tabular data), analyzes accident causes and liability determination, addressing the subjectivity and inefficiency issues of traditional manual analysis **[29]**. Furthermore, large language models have shown remarkable capabilities in traffic accident reporting analysis. For instance, an LLM-based analytical framework analyzed 500 traffic accident narrative texts, identifying previously unreported factors (such as alcohol-related incidents) in historical accident data with 96% accuracy in just 7 minutes, while human experts required an average of 25 hours **[18]** . More importantly, the TrafficRiskGPT system, which combines chain-of-thought with causal analysis, demonstrates that this approach outperforms traditional methods and mainstream LLMs in terms of collision prediction accuracy, reasoning speed, and overall risk evaluation metrics **[30]**.

1.2.1 LLM agent limitations

In recent years, AI agents have found widespread applications across various domains, including scientific research, healthcare, and engineering. While the definition of an AI agent remains broad and evolving [31-34], there is growing consensus that LLM-based agents—autonomous systems powered by large language models—are capable of performing complex tasks with minimal human intervention. These agents leverage the reasoning, planning, and generative capabilities of foundation models to act in dynamic and uncertain environments.

However, single-agent systems suffer from several inherent limitations.

First, the world knowledge boundaries and internal thinking patterns of single agents are constrained by the capabilities of the underlying LLM itself. Current research has shown that different types of large language models exhibit significant differences in performance when executing specific tasks [18, 35, 36].

Second, complex tasks often require step-by-step decomposition, and a single agent struggles to process multi-dimensional and multi-modal information streams simultaneously. Research indicates that single-agent systems tend to underperform across all subtasks when handling multi-modal data, particularly when the data is incomplete or contains contradictions [37, 38].

Third, the degree of specialization limits the ability of a single agent to achieve expert-level performance across multiple domains. For instance, a single agent often fails to comprehensively address the diverse subdomains involved in traffic crash analysis, especially in cross-disciplinary tasks spanning EDR data, visual evidence, and legal interpretation—leading to shallow analysis and reasoning bias [39, 40].

Furthermore, the inevitable hallucination problem of LLMs can lead to the propagation of errors and accumulation of risk. Mistakes made in early stages may compromise all subsequent steps, and a single agent typically lacks effective mechanisms for error detection and correction [41].

In summary, single-agent systems remain fundamentally constrained in knowledge coverage, task decomposition, domain specialization, and error robustness—highlighting the urgent need for more collaborative, modular AI architectures.

1.2.2 Multi-agent collaboration

To address the inherent limitations of single-agent systems, the research paradigm is shifting toward multi-agent collaboration. The core principle of multi-agent systems is "divide and conquer," whereby tasks are decomposed into subtasks (or systems are modularized), and specific agents are designed for each subtask [38, 39] . Furthermore, agents must be tailored to the requirements of subtasks based on the characteristics of LLMs [42] . For instance, tasks involving image recognition necessitate agents built on vision-based large models, whereas tasks requiring logical reasoning are best handled by models equipped with chain-of-thought capabilities.

Recent research has demonstrated the superior performance of multi-agent collaboration across diverse application domains. In autonomous driving scenarios, multi-agent systems that integrate reinforcement learning, large language models, and retrieval-augmented generation (RAG) mechanisms not only enhance human-like behavior and interpretability but also outperform alternative reinforcement learning approaches on critical metrics including merge success rates and collision rates [43]. In traffic flow prediction, xTP-LLM, which converts multimodal traffic data into natural language descriptions, exhibits significantly higher accuracy compared to deep learning baselines while providing intuitive and reliable explanations for traffic predictions [44] . Furthermore, regarding interpretability, researchers have proposed a closed-loop multi-agent collaborative system capable of automatically completing the entire process from model conception to code implementation, evaluation, and iterative optimization. This approach not only achieves substantial improvements in accuracy and robustness on classical models such as IDM, MOBIL, and LWR but also maintains model interpretability and demonstrates strong generalization capabilities across datasets [45].

1.3 Research Objectives and Contributions

To address the fundamental challenges in traffic crash analysis, this study develops an AI-driven multi-agent collaborative framework that revolutionizes how we reconstruct pre-crash scenarios and identify critical first collision events. The ability to accurately determine the first crash event and its corresponding EDR data is

crucial for understanding pre-crash driving behaviors that ultimately lead to collisions. The primary objective of this research is to construct and validate a two-stage multi-agent framework that overcomes the limitations of traditional human-centered pre-crash analysis. Unlike conventional approaches that struggle with fragmented multimodal data, our framework employs specialized AI agents working collaboratively to: (a) reconstruct comprehensive crash scenarios from disparate data sources, and (b) perform sophisticated reasoning to identify the precise first collision event and its associated pre-crash EDR records. The main contributions of this study include:

1) We present a novel multi-agent architecture where specialized agents collaborate sequentially. The Phase I agent synthesizes fragmented multimodal inputs (textual crash data, structured tabular information, and visual scene diagrams) into coherent crash scene reconstructions. The Phase II agent then performs temporal reasoning on these reconstructions combined with EDR data to analysis and inference the precrash collision.

2) Through comprehensive evaluation on 277 real-world LVD cases, our multi-agent framework achieved perfect accuracy in both simple cases and complex cases. Our framework shows superior performance even when processing incomplete data, conflicting information, or database labeling errors validates the robustness of the multi-agent collaborative approach.

3) We developed domain-specific reasoning anchors and structured prompts to ensure consistent analytical outcomes across different Large Language Model implementations. Testing with three reasoning models showed identical results across all 585 trials, confirming that the effectiveness of our framework derives from the prompt engineering design rather than specific LLM characteristics. This cross-model consistency establishes the reliability of AI-driven crash analysis systems.

This research focus on developing a multi-agent collaboration analytical framework based on LLMs. By successfully identifying first collision events and their corresponding pre-crash EDR data with unprecedented accuracy, our framework provides crucial insights into the driving behaviors and vehicle dynamics that precede crashes.

2 METHODS

2.1 Data Acquisition and Data Processing

2.1.1 Data Acquisition

The Crash Investigation Sampling System (CISS), maintained by the National Highway Traffic Safety Administration (NHTSA), provides a comprehensive and authoritative dataset of real-world motor vehicle crashes in the United States. For this study, CISS serves as the primary data source due to its detailed, multidimensional records of crash events, which are critical for reconstructing pre-crash scenarios using a multi-agent LLM framework. The dataset includes a wide range of crash-related information, including **textual crash data**, **structured tabular data** (e.g., vehicle dynamics, environmental conditions like weather and road surface, road configurations such as lane geometry and traffic control devices, and vehicle specifications like model, and damage panel), and **visual scene diagrams**, offering a robust foundation for multimodal data integration. Crucially, CISS incorporates EDR temporal data, which provide high-resolution time-series records of vehicle dynamics (e.g., speed, braking) in the seconds leading up to a collision. These EDR records capture detailed, time-stamped information on vehicle behavior, enabling precise analysis of pre-crash dynamics and supporting robust reconstruction of collision scenarios.

In this study, we analyzed 277 lead vehicle deceleration (LVD) collision cases from the CISS dataset (from 2017 to 2022), leveraging its rich and real-world data to validate the proposed multi-agent LLM framework for accurate and efficient pre-crash reconstruction.

2.1.2 Transforming Crash Events into Structured Textual Format

To support the integration and subsequent analysis by the multi-agent LLM, the diverse multimodal raw data were systematically transformed into coherent textual representations (shown in **Figure 1**).

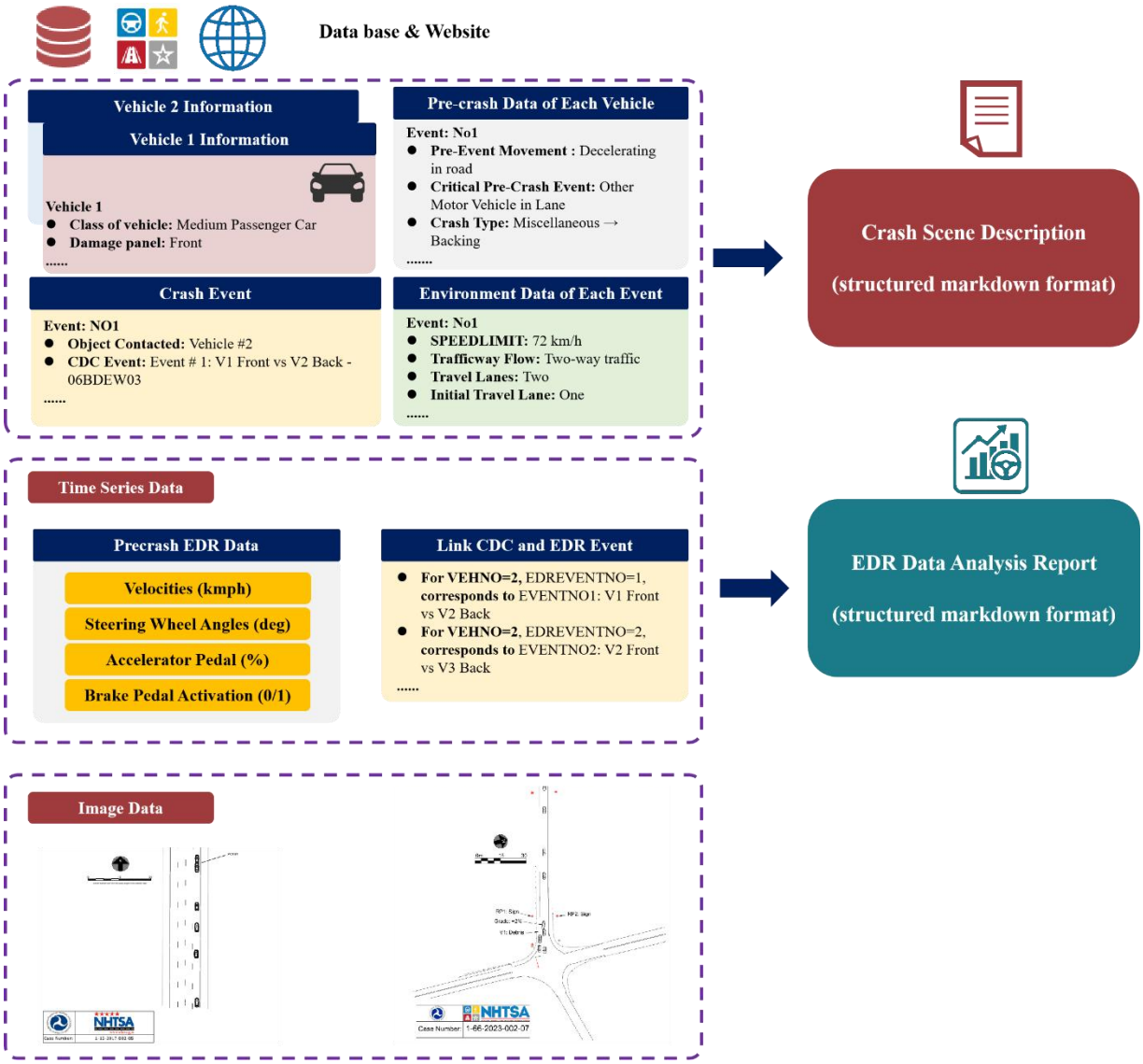

Figure 1. Extract textual crash data and structured data related to scene reconstruction from the CISS database.

First, crash data encompassing vehicle types, damage patterns, environmental attributes, and collision event details were structured into concise natural-language summaries. These summaries clearly identified each involved vehicle, described critical crash events, and contextualized environmental conditions relevant to the collision scenarios.

Secondly, temporal EDR data, capturing vehicle dynamics such as velocities, steering inputs, accelerator positions, and braking activities, were processed and reformatted into structured textual narratives. We specifically ensured accurate alignment between these EDR-derived records and the corresponding crash event sequences from CISS, facilitating subsequent precise crash scenario reconstructions and analyses (shown in **Figure 2**).

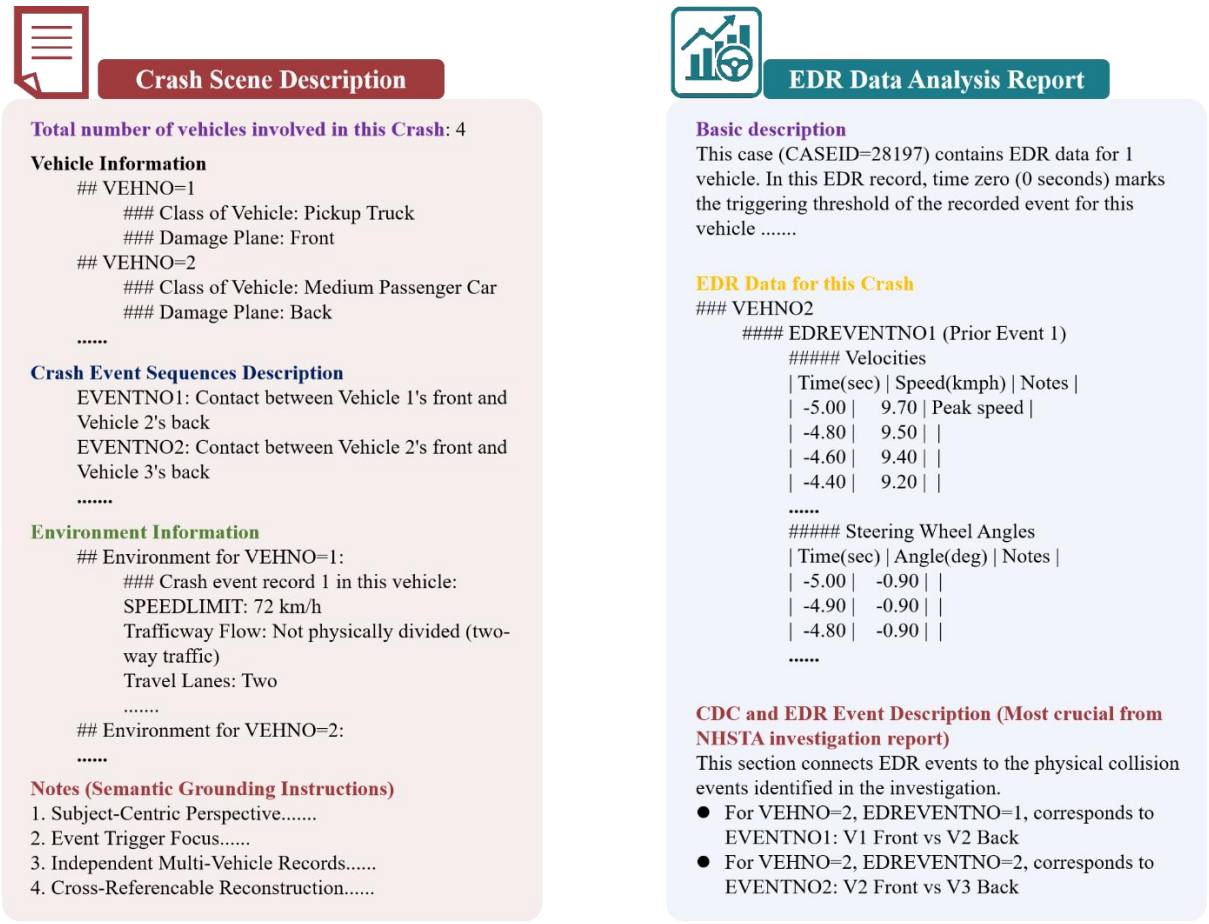

Figure 2. Transform the structured data into textual format.

Finally, image data were preserved in the original format for direct input into vision-capable LLMs. It could be noted that the **Scene Diagrams** in the CISS database originate from on-site investigations conducted by NCSA (National Center for Statistics and Analysis) teams **[46]**. Investigators first perform precise field measurements at crash locations, then manually create collision scene diagrams using CAD software based on field survey records and interview information. The resulting **Scene Diagrams** are specialized PDF images that not only depict the relative spatial relationships among involved vehicles, road geometry, and environmental elements, but also provide scale information (such as reference object dimensions and distance annotations) along with embedded textual content and interfering elements (such as logos, tables, and annotations).

This multimodal-to-text transformation unified diverse data sources into coherent textual summaries, facilitating seamless integration within the subsequent scenario reconstruction and reasoning stages of our AI agent analytical framework.

2.2 AI-driven Multi-agent Analytical framework

The AI-driven multi-agent analytical framework proposed in this research operates through a structured two-phase workflow. The framework's architecture (shown in **Figure 3**), demonstrates how multiple specialized agents work in concert to transform raw, heterogeneous crash data into actionable insights about pre-crash vehicle behaviors and collision dynamics.

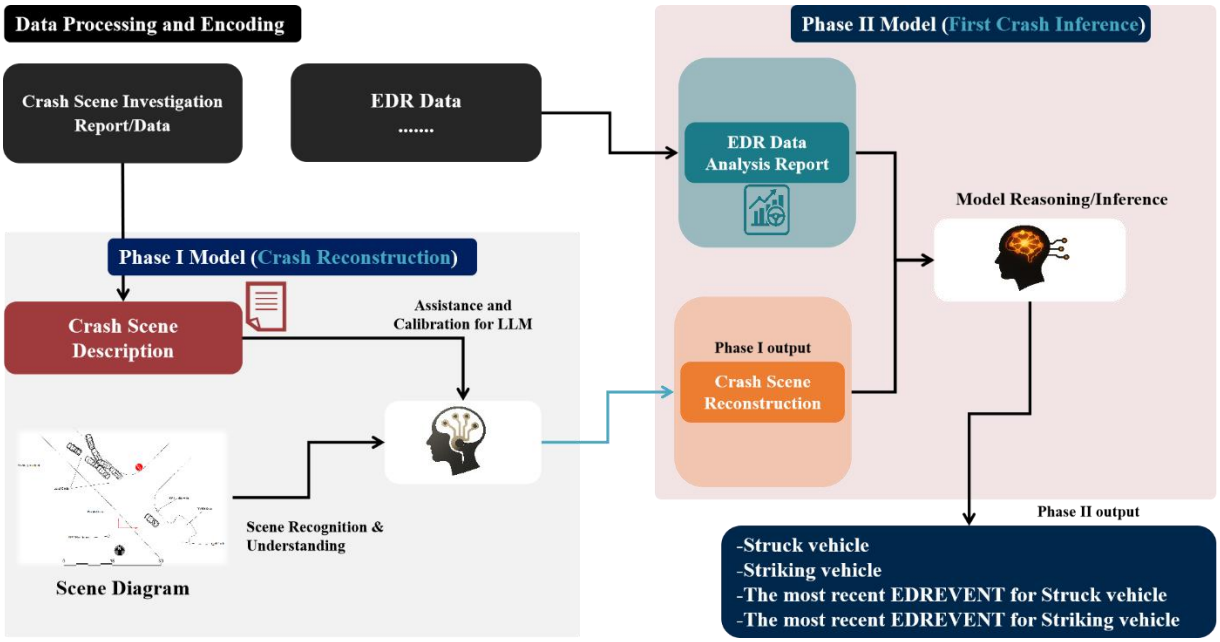

Figure 3. Overview of Proposed AI Agent Analytical framework.

Step1: The raw data from the database will be processed and encoded as described in **2.1,** transforming diverse crash-related information into structured natural language inputs. This process generates three key components: the **Crash Scene Description** containing detailed textual narratives of the collision event, the **EDR Data Analysis Report** presenting vehicle dynamics data in an interpretable format, and the **Scene Diagram** providing visual representation of the crash scenario. These processed outputs serve as standardized inputs for the subsequent agent-based analysis phases.

Step 2: The Crash Scene Description and Scene Diagram are fed into the Phase I Agent. This agent comprehensively understands the multi-dimensional crash data by processing all input data, capturing the essential elements of the crash scene through a predefined internal reasoning process, and leveraging multi-source data for assistance and calibration. The output is a structured, complete reconstruction of the crash scene.

Step 3: The output from the Phase I Agent, along with the **EDR Data Analysis Report**, is input into the Phase II Agent. This agent focuses on initial collision inference, a critical task for identifying the most relevant EDR events and distinguishing between the striking and struck vehicles. The Phase II Agent employs reasoning mechanisms to analyze the temporal patterns of EDR data within the reconstructed crash scene. By cross-referencing the reconstructed narrative with the time series of EDR events, the agent can infer the specific EDR record most relevant to the initial collision, even in complex scenarios involving multiple recorded events or ambiguous vehicle roles.

This AI analysis framework involves multi-agent collaboration, with each phase building upon and validating the previous one. The reconstruction from the first phase provides essential contextual grounding for the inference operations in the second phase, while the EDR analysis in the second phase reveals insights that deepen the understanding of the initial reconstruction. This iterative refinement process enables the system to handle complex scenarios with incomplete data, conflicting information, or multiple plausible interpretations. The final output includes precise identification of vehicle roles (striking and struck vehicles), determination of the most relevant EDR event for each involved vehicle, and an interpretable rationale for the inference.

2.3 Phase I Agent Design: Crash Reconstruction

2.3.1 Task definition

The core task of the first-phase agent is to understand and synthesize fragmented multimodal input data to generate structured natural language descriptions of collision scenario reconstructions. This agent needs to simultaneously understand two types of modal inputs: (1) Scene Diagram Input—accident scene diagrams manually drawn by NHTSA investigators (since CISS Scene Diagrams are in a special PDF format that contains both images and considerable interfering text, making it difficult for LLMs to directly understand these images); (2) Textual Input—structured Crash Scene Descriptions. Given that accident reconstruction is a knowledge-intensive task, this phase employs general-purpose large language models with world knowledge.

Regarding model selection, through visual image evaluation of multiple advanced large language models, including GPT-4, GPT-4O, Grok3, Deepseek-v3, Claude, and Gemini Pro [25, 47, 48], we found that only Claude 3.7 could preliminarily understand images containing substantial interfering text without requiring additional prompt optimization. Other models tended to degrade the image understanding task to simple text extraction, failing to truly comprehend the spatial relationships of accident scenes. Therefore, this study selected Claude 3.7 as the core model for the Phase I agent.

2.3.2 Prompt construction

Numerous studies have shown that carefully designed prompts can greatly enhance an AI's performance and depth of understanding [49-51]. To ensure accurate scene reconstruction, we develop a prompt-engineering framework tailored to the CISS database that is more generalizable, more intelligent, and more structured (Phase I Agent structure is shown in **Figure 3**).

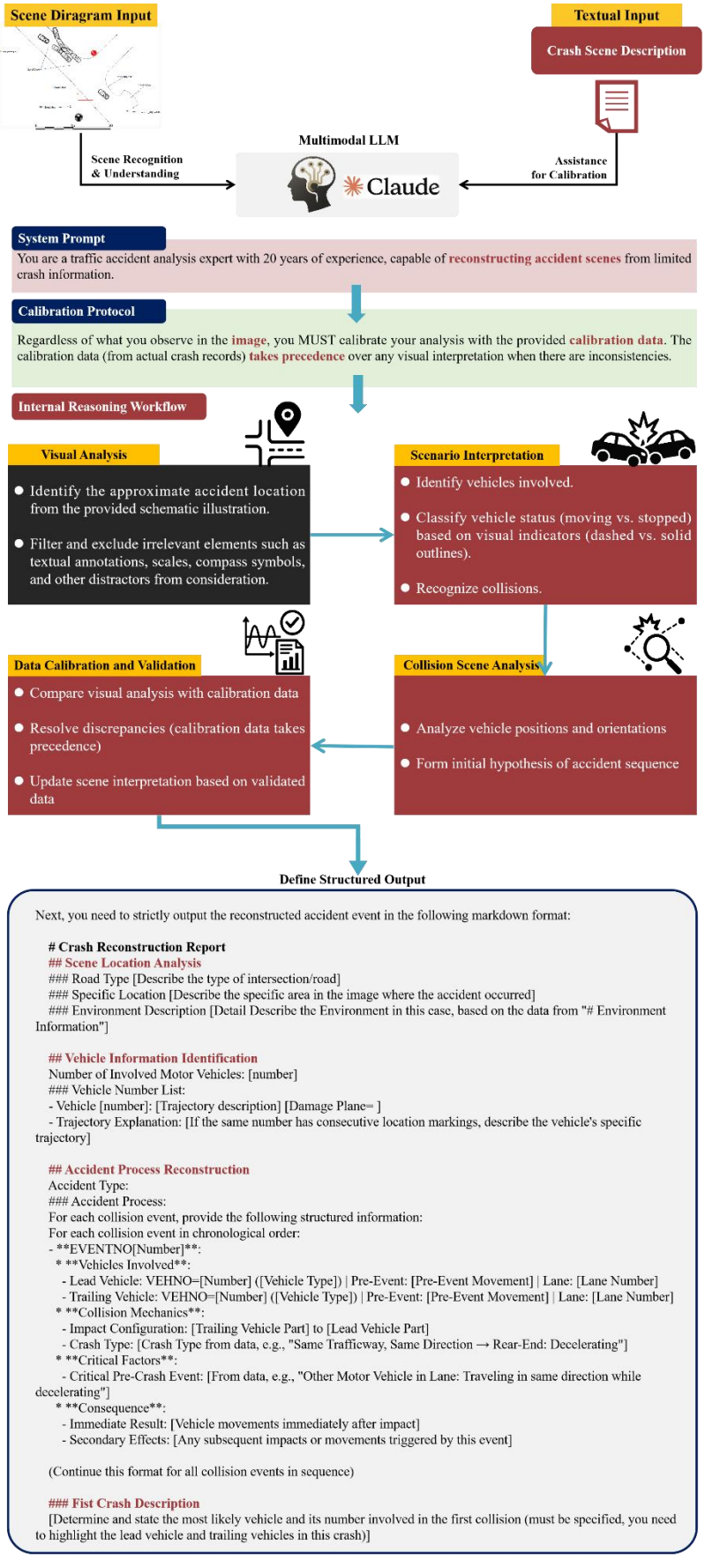

Figure 4. Prompt design for accident-reconstruction agent.

2.3.2.1 System prompt

In this study, the Phase I agent is framed as a traffic-accident analysis expert with 20 years of experience, endowed with the ability to reconstruct crash scenes from limited information. This "role prompting" strategy helps activate the model's domain knowledge and professional analytical framework.

2.3.2.2 Calibration protocol

We clearly define the priority order in which the agent processes different data types. To minimize hallucinations during text and image interpretation and to ensure consistency in understanding the crash scene, image-based inferences must be calibrated against the structured data—in other words, scene understanding is achieved through bidirectional calibration between visual and textual inputs.

2.3.2.3 Internal reasoning workflow

To ensure precise processing of crash scenarios and enhance the robustness of analysis results, this study designs an Internal Reasoning Workflow that explicitly defines the logical pathways for the model when processing multimodal input data. This workflow comprises four core steps:

Step1. Visual Analysis: The agent first conducts preliminary visual analysis of the Scene Diagram. By identifying and filtering out irrelevant interference information (such as symbols, arrows, scales, directional markers, and accident-unrelated text in the diagram), it focuses on the accident location, vehicle numbers, vehicle motion states (dashed lines for motion, solid lines for stationary), and potential collision points.

Step2. Scenario Interpretation: Following visual analysis, the agent further interprets the scene by identifying specific vehicles involved in the image, determining each vehicle's motion state, and explicitly marking vehicles involved in collisions along with their spatial relationships.

Step3. Data Calibration and Validation: The model performs bidirectional calibration between preliminary scene interpretation results and structured text (EDR records, crash scene descriptions). When inconsistencies arise between visual analysis results and textual records, the model takes textual records as the baseline to correct and validate the crash scene reconstruction.

Step4. Collision Scene Analysis: After calibration, the agent conducts in-depth analysis of vehicle positions, orientations, and relative dynamics to formulate preliminary crash sequence hypotheses. This stage integrates calibrated information from both visual and textual modalities, ensuring reasoning accuracy and logical consistency.

This internal reasoning workflow effectively enhances the model's comprehension capability for complex crash scenarios while reducing the probability of analytical errors and reasoning hallucinations.

2.3.2.4 Structured output

To ensure clarity and consistency in Phase I agent outputs, this study designs a structured output format. This structured output is presented in markdown format, which facilitates LLM text comprehension, and is distinctly organized into three core sections:

A) Scene Location Analysis: Includes road type description, specific accident location, and environmental characteristics. This information is completed by the agent through integration of visual analysis and calibration data, ensuring spatial positioning accuracy of the scene.

B) Vehicle Information Identification: The agent lists, in a clear and standardized manner, the number of all involved vehicles, their identification numbers, corresponding motion trajectory descriptions, and vehicle damage plane information, providing detailed explanations of each vehicle's trajectory and dynamic changes.

C) Accident Process Reconstruction: The agent presents structured descriptions of collision events in detail, including accident type and accident process (event chronology, involved vehicles, collision mechanics, critical factors, and consequences), with particular emphasis on precise identification of the first crash event and clear differentiation between primary and secondary vehicles.

The structured output design enables the LLM to generate standardized, clear, and easily comprehensible crash scene reconstruction reports while providing complete collision context for Phase II agent's deep reasoning tasks. Through this comprehensive context, the Phase II agent can directly understand natural language to further extract key temporal features, complete first crash determination, and other high-level analyses. This achieves a closed-loop multi-agent collaborative workflow, enhancing the reliability and consistency of agent pre-crash reasoning and decision-making capabilities.

2.4 Phase II Agent Design: First Crash Inference

The Phase II agent represents the critical reasoning component of our multi-agent framework, designed to perform sophisticated temporal analysis and causal inference on reconstructed crash scenarios. While the Phase I agent focuses on scene understanding and reconstruction, the Phase II agent advances the analysis by identifying the precise moment of initial impact and determining vehicle roles within complex collision sequences. This agent must navigate the inherent challenges of asynchronous data streams, mis-ordered EDR records, and ambiguous vehicle interactions to extract meaningful insights about pre-crash driving behaviors.

The core challenge addressed by this agent lies in the temporal complexity of real-world crashes. EDR systems record multiple events that may or may not correspond to the actual collision moment, creating a need for intelligent filtering and correlation. Furthermore, the agent must reconcile potential discrepancies between the narrative reconstruction from Phase I and the quantitative EDR data, while maintaining logical consistency in its reasoning process. To address these challenges, we employ advanced reasoning models with demonstrated capabilities in complex logical inference, including **DeepSeek-R1, Grok 3-mini and Gemini-2.5Pro**, selected for their deep thinking and chain-of-thought reasoning abilities [26-28].

2.4.1 Prompt structure

The prompt architecture for the Phase II agent follows a hierarchical design that systematically guides the LLM through the deeper reasoning task (shown in **Figure 5**).

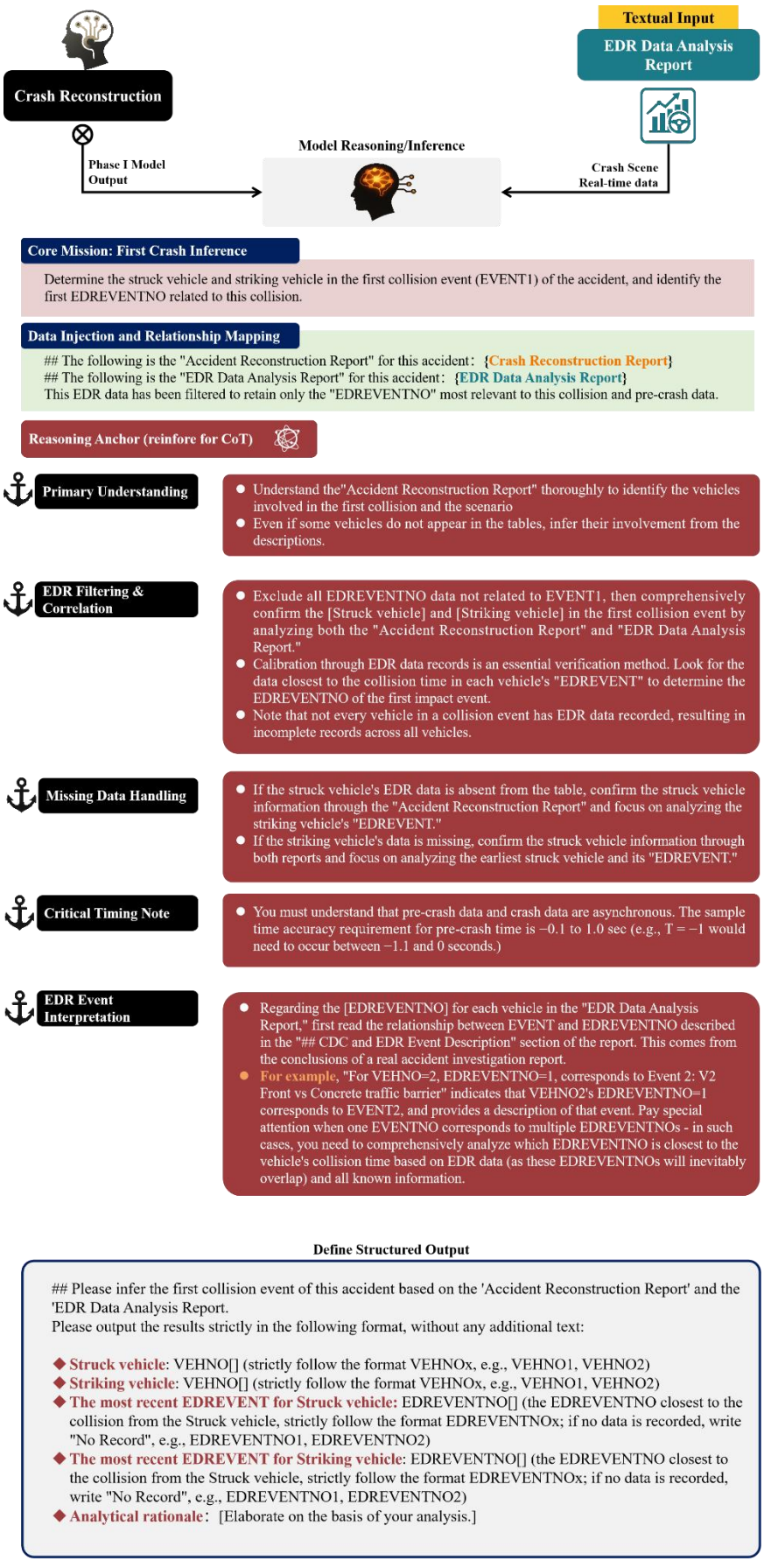

Figure 5. Prompt design for crash inference agent.

- **Core Mission:** A statement instructing the LLM to locate first crash event.
- **Data & Relationship:** Defined the data and relationship.
- **Reasoning Anchor:** Anchors that design for guiding reasoning process.
- **Structured Output:** A predefined format for preventing free-form responses.

This “mission–data–procedure–output” layout exposes the intended reasoning path, ensuring different models follow an identical logic chain while producing uniform, machine-readable outputs for automated evaluation.

2.4.2 Reasoning anchor

In multi-path reasoning scenarios such as crash analysis, LLMs may explore various logical pathways that could lead to different conclusions in different experiment trial. Reasoning anchors serve as calibration points that guide the model toward correct reasoning trajectories while preventing drift into erroneous inference patterns.

Our framework implements five primary reasoning anchors, each addressing a specific aspect of the inference challenge.

The first anchor, **Primary Understanding**, establishes the foundational context by ensuring the model thoroughly comprehends the accident reconstruction before attempting EDR correlation. This prevents premature conclusions based on incomplete scene understanding.

EDR Filtering & Correlation provides explicit criteria for excluding irrelevant data, particularly important given that EDR systems may record multiple events unrelated to the primary collision. The anchor incorporates two key mechanisms to ensure data fidelity: first, it filters out events labeled "Related, unknown event" to discard probable sensor anomalies; and second, it corrects for timing discrepancies caused by triggering thresholds, which can misalign recorded events with the true moment of impact.

Missing Data Handling addresses the reality that not all vehicles in a collision have complete EDR records. By providing explicit protocols for these scenarios, the anchor prevents the model from making unfounded assumptions or failing to reach conclusions due to data gaps.

Critical Timing calibrates the model's temporal reasoning by explicitly stating the asynchronous relationship between pre-crash and crash data, along with precise timing tolerances. The anchor's critical function is to define time zero in EDR records as the data trigger point, distinct from the actual collision moment. This addresses the inherent uncertainties of triggering thresholds and sampling frequency delays, thereby preventing a misalignment between recorded events and the true crash timeline.

EDREVENTNO Interpretation guides the model through the complex mapping between physical events and their digital representations in EDR systems. This anchor is particularly crucial when multiple EDREVENTNO records correspond to a single physical event, requiring the model to analyze temporal proximity and data patterns to identify the most relevant record. By explicitly instructing the model to consider overlapping EDREVENTNOs and comprehensively analyze timing data, this anchor ensures consistent and accurate identification of first crash events.

The effectiveness of these reasoning anchors is demonstrated through their impact on model consistency and accuracy. By constraining the reasoning space while still allowing for flexible analysis, the anchors enable different models to converge on similar conclusions when presented with the same crash scenario. This convergence is critical for establishing the reliability and trustworthiness of AI-driven crash analysis systems in real-world applications.

2.5 Performance Evaluation of LLM

The evaluation dataset included 277 LVD cases, each with ground truth results from the CISS dataset. Of these, 238 cases were simple EDR cases, defined by a one-to-one correspondence between crash events and EDR records, where each vehicle's crash event matched a single, unique EDR record. These cases provided the benchmark for evaluating the LLM performance in crash event reconstruction and first crash inference under standard conditions.

Additionally, 39 cases were classified as "Complicated EDR" cases, defined by a many-to-one relationship where a single crash event was linked to multiple EDR records from at least one vehicle. Some records also contained data logging errors, adding further complexity. These edge cases were included to evaluate the LLM's robustness in handling ambiguous or erroneous data, specifically testing its ability to discern relevant information and perform reliable inference under non-ideal conditions.

Two experienced researchers independently analyzed each of the 39 cases with a structured textual format, which underwent the data processing described above, to establish a human judgment baseline. To address potential human judgment errors, both researchers compared their results with a third-party model using identical prompts, and inconsistent cases were manually verified. The researchers then synthesized the final consensus results as the ground truth for each case, ensuring reliability and minimizing subjective bias.

Once the ground truth was established, the target LLM was formally evaluated by comparing its outputs against the validated benchmark results.

3 RESULTS

3.1 Overall Performance

3.1.1 AI-driven framework performance

The comprehensive evaluation of the AI-driven multi-agent analytical framework was conducted across all 277 LVD cases from the CISS dataset. A total of **4,155** experimental trials were conducted, with each of the 277 cases evaluated five times across three reasoning LLMs (277 × 5 × 3), comprising 1,190 trials for the 238 simple EDR cases and 195 trials for the 39 complicate EDR cases per model. The evaluation criterion was defined—for each experimental trial, our framework was required to inference four critical outputs: (1) the striking vehicle, (2) the struck vehicle, (3) the most relevant EDR event for the striking vehicle, and (4) the most relevant EDR event for the struck vehicle. A single error in any of these four determinations resulted in the entire trial being classified as a failure.

The AI-driven framework achieved perfect classification accuracy across all experimental trials. As shown in **Figure 6**, the confusion matrices demonstrate that the system correctly identified all 1,190 instances in simple EDR cases and all 195 instances in complicate EDR cases, with no false positives or false negatives recorded. This resulted in precision, recall, and F1 scores of 1.00 for both case categories.

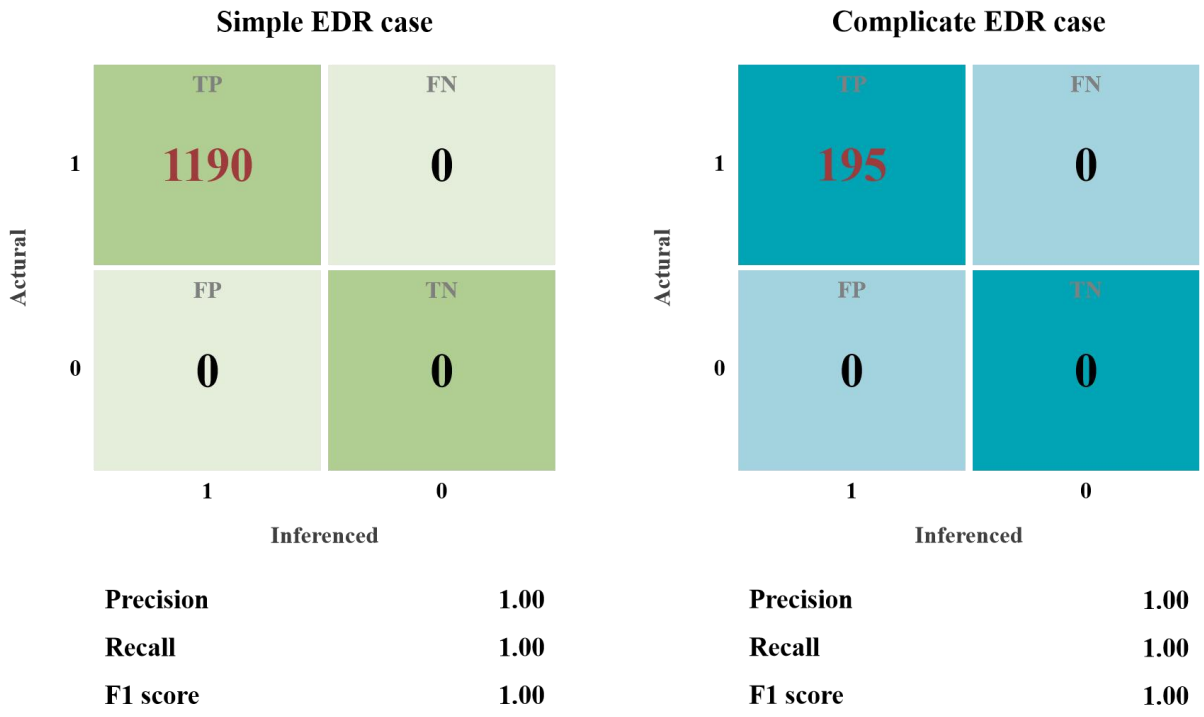

Figure 6. Evaluation results of overall performance.

3.1.2 Human vs AI analytical performance

A comparative evaluation showed that the research analysts achieved differential performance based on case complexity. The evaluation revealed a clear performance distinction between simple and complex EDR scenarios. In the 238 simple cases, which feature a clear one-to-one relationship between events and records, both the AI system and the two research analysts achieved 100% accuracy. This confirms that such straightforward cases are unambiguous for both automated and human review.

Performance on the 39 complex EDR cases—characterized by many-to-one relationships and occasional data errors—provides a more nuanced insight. The two individuals performing the human analysis in this study were research analysts, who are experts in crash data analysis, but not certified crash reconstruction experts. In these challenging cases, their combined accuracy was 92.31% (72 correct determinations out of 78, as shown in **Figure 7**). The performance of the research analysts should therefore be viewed as a benchmark for *non-specialist* interpretation. The significant finding is not that AI outperformed humans in an absolute sense, but that it demonstrated robust performance in complex scenarios where individuals without specialized expertise are prone to error. This highlights the AI's potential utility as a reliable tool for initial analysis and for flagging cases that contain the kind of ambiguities that may challenge non-experts.

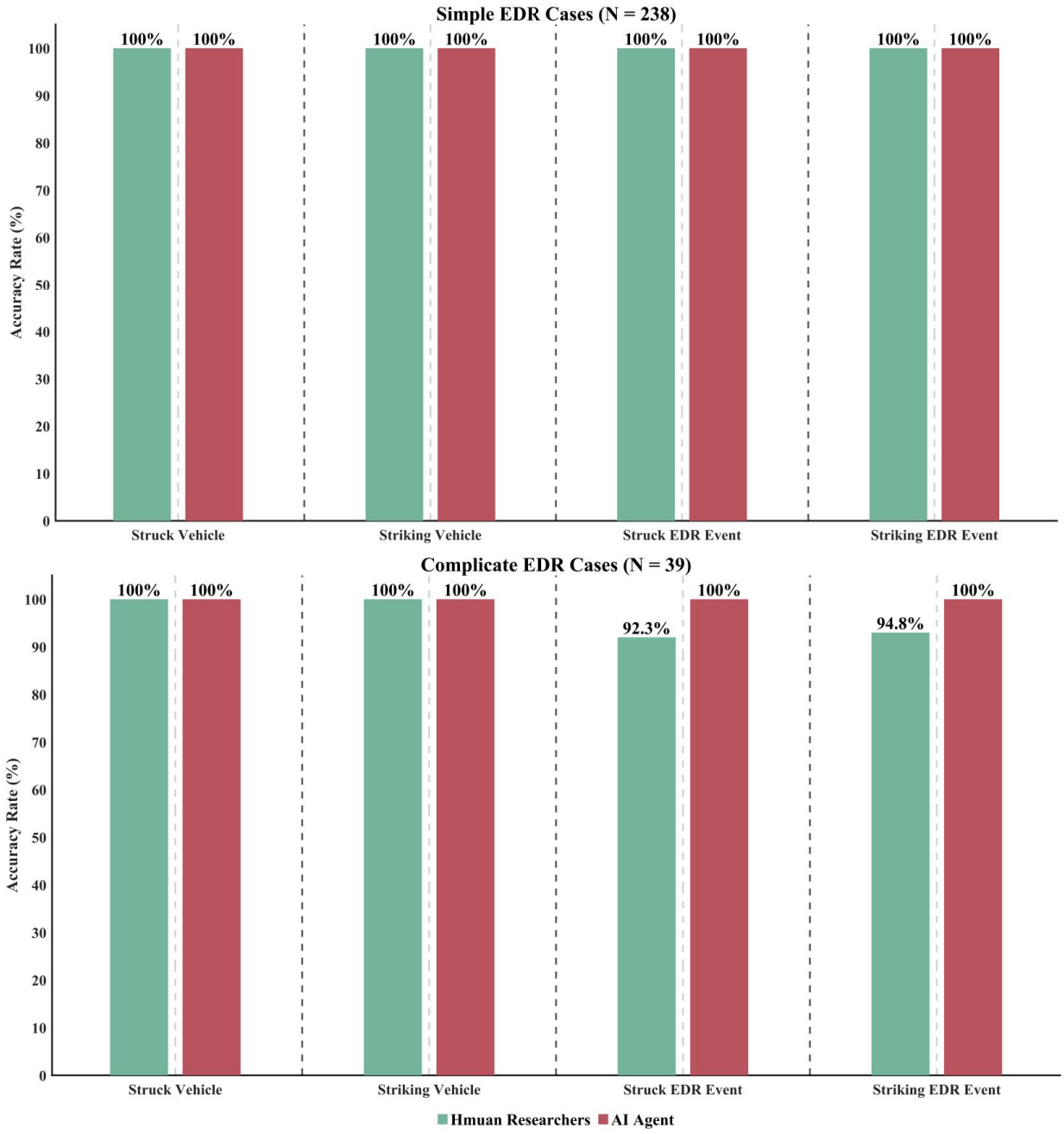

Figure 7. Accuracy comparison between research analysts and AI agents for vehicle and EDR event detection in simple (N=238) and complicated (N=39) EDR cases.

The evaluation revealed a consistent performance in vehicle role identification, with both the AI and research analysts achieving 100% accuracy. The critical differentiator was the task of determining the initial collision EDR event within complex scenarios. Here, the inherent ambiguities of multi-record cases and data logging errors led to errors in human analysis. The AI framework, however, maintained perfect accuracy across all five trials for each of the 39 complicated cases, demonstrating its resilience to the specific complexities that impeded human analysts.

3.2 Case Study on Complicate EDR Cases

To illustrate this AI-driven framework performance on extreme cases and demonstrate the complexities that challenge human analysts, we present a detailed analysis of a multi-vehicle collision case that exemplifies the

analytical difficulties inherent in extreme EDR scenarios.

3.2.1 Case study

Case ID = 32548 involved a three-vehicle collision that occurred on a straight, two-lane undivided roadway with a 56 km/h speed limit. The crash sequence began when Vehicle 2, a small passenger car, was decelerating in the travel lane. Vehicle 3, traveling in the same direction behind Vehicle 2, failed to maintain adequate following distance and struck Vehicle 2 from behind, resulting in a typical rear-end collision (Crash event 1). The impact from this initial collision appears to have altered Vehicle 2's position or trajectory, subsequently leading to a secondary crash. In this second event (Crash event 2), Vehicle 1, which was also a small passenger car, collided front-to-front with the already-impacted Vehicle 2. The scene diagram indicates that all three vehicles sustained front-end damage, with the crash sequence progressing from the initial rear-end impact to the subsequent frontal collision.

From the perspective of EDR data analysis, this case exemplified complications arising from multiple EDR records associated with a single crash event and mislabeling within the database. Vehicle 1, involved in only one collision, had missing EDR data. Vehicle 3 had a single collision event with only one corresponding EDR record. Vehicle 2, however, involved in both crash events, presented a notably complex EDR scenario with six separate records.

The initial collision event, occurring exclusively between Vehicle 2 and Vehicle 3, clearly positioned Vehicle 2 as the struck vehicle and Vehicle 3 as the striking vehicle. For Vehicle 3, with only one available EDR event (EDREVENTNO1), the system easily identified this as the relevant pre-crash data. Vehicle 2's EDR situation was significantly more complicated. According to the database, EDREVENTNO1, EDREVENTNO3, and EDREVENTNO4 related to the second collision event, while EDREVENTNO2 and EDREVENTNO5 corresponded to the initial collision event, with EDREVENTNO6 marked “Event not related to this crash”.

3.2.2 Comparative performance

Analysis of Vehicle 2’s EDR records revealed temporal overlaps in pre-crash data, expected due to the short time interval between Crash Event 1 and Crash Event 2. Specifically, the pre-crash data for EDREVENTNO1, 3, 4 (labeled for Crash Event 2), and EDREVENTNO2 (labeled for Crash Event 1) were identical across their recorded timeframes. Additionally, EDREVENTNO5 (labeled for Crash Event 1) shared partial temporal overlap with these records, with data from -3.4 to -0.4 seconds aligning with segments of EDREVENTNO1, 3, 4, and EDREVENTNO2 from -4.6 to -2.1 seconds. Hence, the CISS database incorrectly labeled EDREVENTNO2 as corresponding to Crash Event 1, as its pre-crash data matched those associated with Crash Event 2.

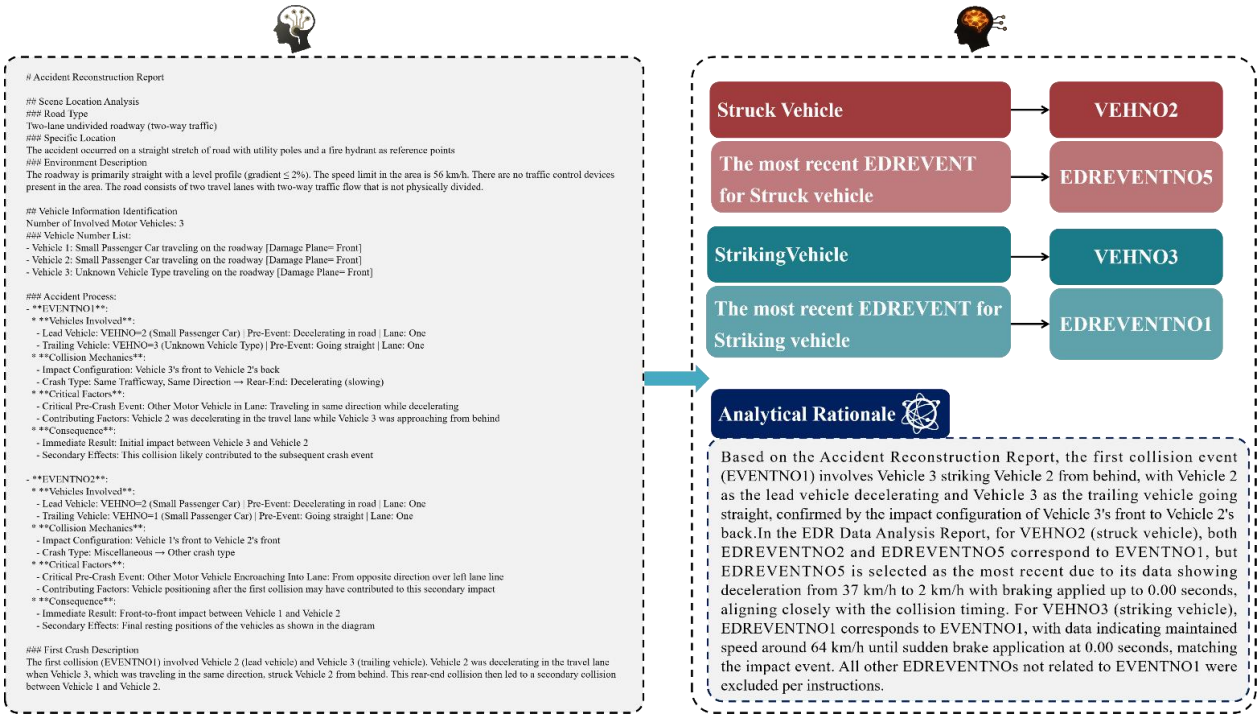

Figure 8. Phase I and Phase II agent outputs for case 32548, showing (left) condensed accident reconstruction and (right) automated identification of struck/striking vehicles and their nearest pre-crash EDR records.

The framework analytical process proceeded through two distinct phases (shown in **Figure 8**). In the first phase, the system identified Vehicle 2 as the struck vehicle and Vehicle 3 as the striking vehicle, with initial selections of EDREVENTNO5 for Vehicle 2 and EDREVENTNO1 for Vehicle 3. The second phase involved systematic cross-validation between accident reconstruction reports and EDR temporal data patterns. The AI system determined that EDREVENTNO5 contained pre-crash data showing vehicle deceleration from 37 km/h to 2 km/h with brake application timing that corresponded most closely to the first collision event. For Vehicle 3, EDREVENTNO1 showed maintained speed around 64 km/h until sudden brake application at the moment of impact.

The AI-driven framework final determined EDREVENTNO5 as the most relevant EDR record for Vehicle 2's involvement in the first collision event and EDREVENTNO1 for Vehicle 3. This selection was based on temporal proximity to the first collision and consistency between the recorded vehicle dynamics and the collision sequence described in the accident reconstruction report. In contrast, all human analysts incorrectly selected EDREVENTNO2 as the EDR record for Vehicle 2 in first crash, misleading by the erroneous CISS database label.

The complicate EDR cases demonstrate the AI framework capability to maintain analytical consistency in extreme scenarios with complex data structures and database labeling inconsistencies. The results highlight the framework robustness in large-scale data analysis when confronted with the CISS's erroneous or misreported information.

3.3 Cross-model Consistency of Inference

To evaluate the robustness and reliability of the Phase II agent's reasoning capabilities, we assessed the consistency of inference results across three reasoning models: DeepSeek-R1, Grok3-Mini, and Gemini-2.5-Pro. Each model was independently applied to all 39 complicate EDR cases, with five experimental trials conducted per case per model, resulting in a total of 585 individual inference experiments (39 cases × 3 models × 5 trials).

The cross-model evaluation results demonstrated remarkable consistency, with all three reasoning models achieving perfect accuracy (100%) across all experimental trials. Each model correctly processed all 195 individual trials (39 cases × 5 trials), producing identical outputs for every inference task. The inter-model agreement rate was 100%, with no discrepancies observed between models when analyzing the same crash scenarios (shown on **Figure 9**). This consistency demonstrates the robustness of the reasoning anchor framework and prompt design in guiding reliable analytical pathways across different LLM architectures, validating that the observed performance reflects the effectiveness of the structured reasoning framework rather than model-specific characteristics.

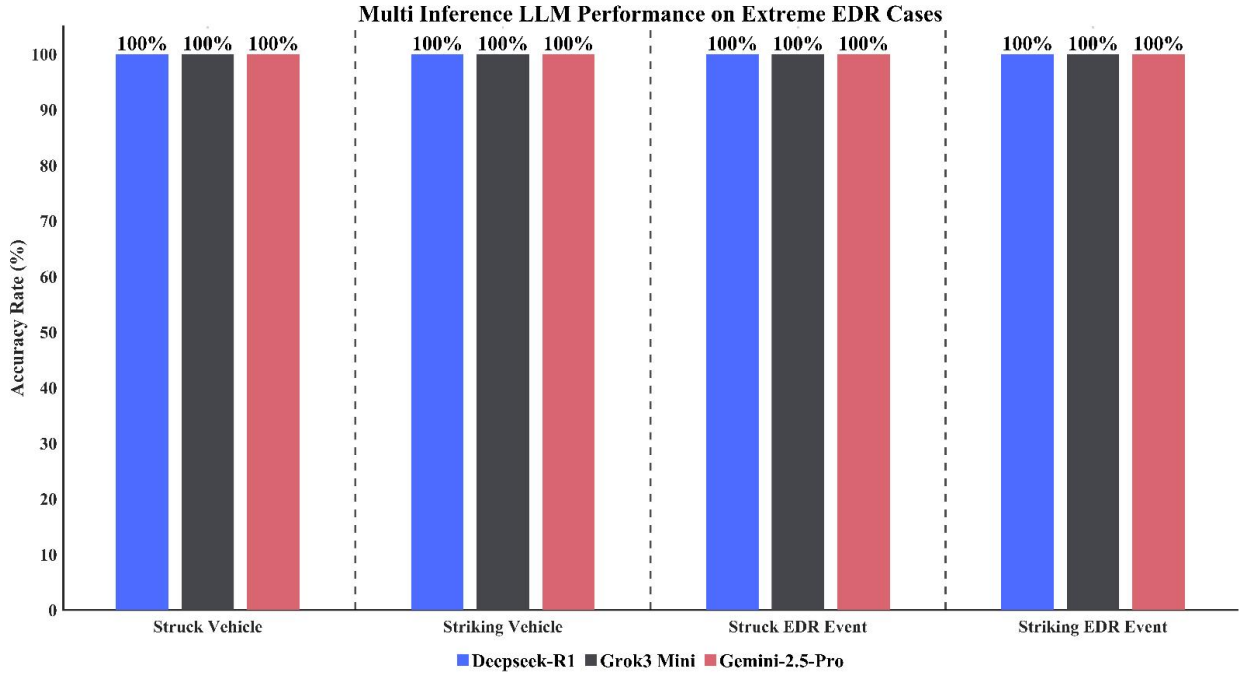

Figure 9. Evaluation on LLM inference performance.

3.4 Time Use for AI Agent Analytical Framework

Figure 10 shown the time consumption characteristics of different AI agent configurations within the analytical framework. During phase I, the Claude agent demonstrated a consistent execution time of approximately 15.10 seconds across all configurations. For phase II, notable differences emerged among the three models employed:

a) Claude + Deepseek-R1 incurred the highest total runtime, averaging 44.18 seconds (15.10s + 29.07s), reflecting Deepseek-R1's relatively higher computational cost and broader inference time distribution. As seen in the violin plot, Deepseek-R1 exhibited substantial variability, with response times ranging from 20s to 65s.

b) Claude + Gemini-Pro achieved a moderate total runtime of 32.97 seconds, supported by a more stable inference distribution centered around 18s.

c) Claude + Grok-Mini demonstrated the lowest total runtime at 22.71 seconds, benefitting from its second-phase agent's short response time (mean = 7.60s). This efficiency is likely attributable to Grok-Mini's design as a lightweight inference model, optimized for fast reasoning under constrained computational budgets.

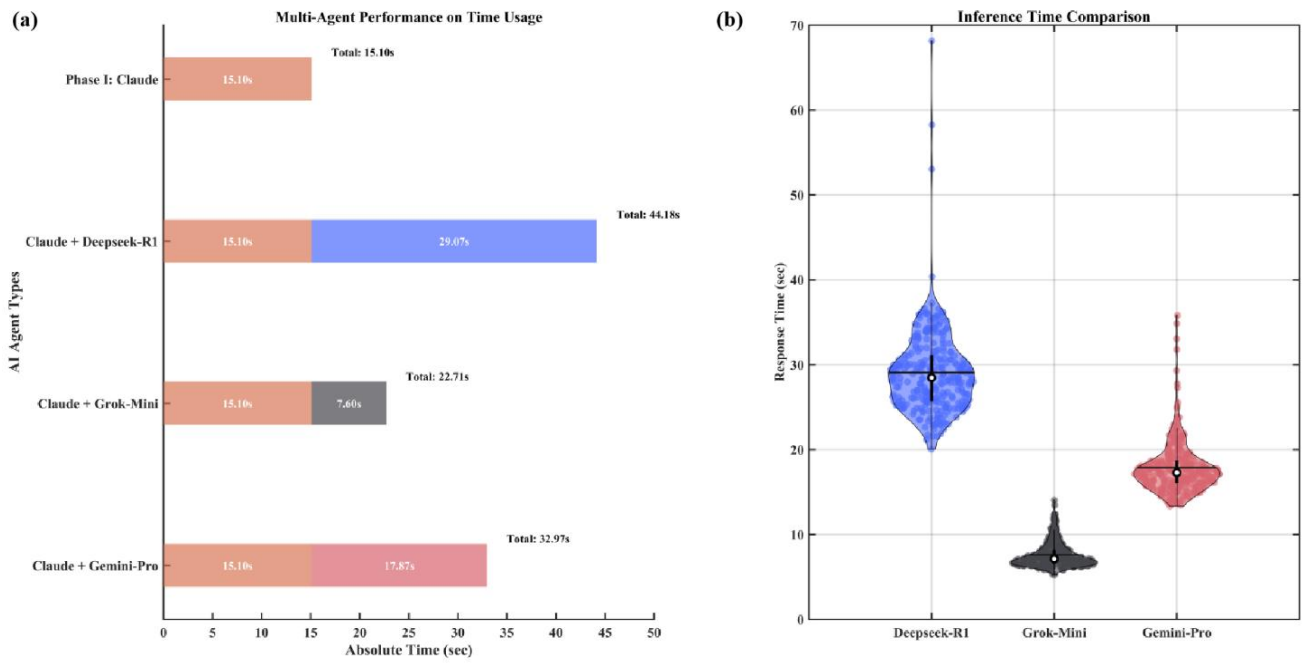

Figure 10. Comparison of computing time across multi-agent configurations. (a) absolute execution time for Phase I and Phase II agents. () distribution of inference times for Phase II models.

For comparison, human analysts required substantially longer durations to complete the same set of analytical tasks. In the 39 extreme LVD cases, two independent researchers spent **4 hours and 5 minutes** and **4 hours and 20 minutes** respectively, resulting in an average per-case processing time of **6.47 minutes**. This indicates that even the slowest AI configuration (Claude + Deepseek-R1) operated at a speed more than five times faster than human experts, while the fastest configuration (Claude + Grok-Mini) reduced analytical time by a factor of over seventeen.

These results highlight the significant time-efficiency advantage of the AI-driven multi-agent framework. By dramatically reducing analysis time without compromising result fidelity, the system demonstrates strong potential for scalable deployment in real-world traffic crash analysis tasks, particularly under time-sensitive or high-throughput conditions.

4 DISCUSSION

The discussion process of this study: first, the analytical strengths of our multi-agent collaborative system in complex collision analysis are thoroughly explored; second, we examine the efficacy of reasoning anchors in enhancing consistency across inference models; and third, we discuss the system's limitations and future directions for extending the framework’s applicability.

4.1 Analytical Performance of AI in Complex Collision Scenarios

The results of this study highlight the significant advantages of a multi-agent collaborative analysis framework in vehicle accident investigation, particularly its performance in analyzing multimodal traffic data. In experiments conducted on 277 LVD cases from the CISS dataset, the AI system demonstrated exceptional performance: 1) achieving a 100% success rate in determining the first collision event, identifying the impacting vehicle, the impacted vehicle, and the EDR driving data closest to the collision, and 2) exhibiting high consistency, with results remaining consistent across five experiments using the same reasoning model for a given case, as well as across different reasoning models.

In simple EDR cases, characterized by unique one-to-one correspondence between crash events and EDR

records, both the AI framework and human analysts achieved consistent performance, attaining 100% accuracy across all cases. The clear data mapping minimized the need for complex reasoning, enabling AI to perform as effectively as human analysts in routine inference tasks, regardless of whether collisions involved simple two-vehicle accidents or more complex multi-vehicle collision chains. This finding aligns with Farber's research, which demonstrated that AI can achieve accuracy comparable to human experts in criminal cases with clear evidence chains **[52]**.

However, in the complicate EDR cases, the presence of multiple EDR records per crash event, temporal overlaps, and data errors (such as labeling errors in the CISS database) significantly increased task complexity. Among these 39 extreme cases, human errors occurred exclusively in the critical task of identifying the EDR event most relevant to the initial collision, as human analysts were apparently misled by hidden false positives or contradictory information in the database while attempting to untangle the causal logic of collision processes amid vast amounts of information [**53**, **54**] . In contrast, the AI framework maintained 100% accuracy across all 585 experiments (195 × 3 trials) on these cases, correctly identifying striking and struck vehicles and inferring the most relevant pre-crash driving EDR data. A notable example is Case ID 32548, where the AI system accurately selected EDREVENTNO5 for Vehicle 2's initial collision, analyzed the temporal relationships in overlapping data, and corrected the mislabeled EDREVENTNO2 in the database—an error that misled two human analysts. This ability to reconstruct logically consistent collision sequences and precisely infer causal chains underscores the framework's remarkable judgment in handling complex temporal and causal relationships.

The robustness of the AI system stems from its multi-agent collaborative analysis architecture, which integrates Phase I reconstruction with Phase II reasoning. In Phase I agent synthesize fragmented multimodal inputs (crash scene descriptions, scene graphs, and structured data) through the internal reasoning workflow, calibrating visual and textual information to generate coherent crash scene reconstructions. This foundation enables Phase II agent to leverage advanced reasoning models for complex temporal and causal inference, guided by reasoning anchors that ensure logical consistency. Furthermore, our AI analytical framework demonstrated superior efficiency, processing each case in an average time of less than one minute, significantly outperforming human experts' average processing time of 6.474 minutes. Recent research supports the efficiency of LLMs in addressing underreporting phenomena in traffic accident data, achieving 96% accuracy in processing 500 traffic accident reports in just 7 minutes **[18]**.

Fundamentally, the AI-driven multi-agent framework redefines the paradigm of collision analysis, demonstrating AI-Agent's substantial advantages in assisting human decision-making and even dominating analytical tasks in scenarios requiring high precision with incomplete data. This advancement marks a significant leap toward more reliable, efficient, and scalable real-world collision reconstruction solutions, with profound implications for improving the accuracy and impact of accident investigations.

4.2 Reasoning Anchors for Consistent Multi-Model Inference

The exceptional consistency of our AI-driven multi-agent framework across multiple large language models and repeated experimental trials underscores the crucial role of **Reasoning Anchors** in ensuring reliable and unified inference outcomes, particularly in complex collision analysis tasks. The evaluation of 39 complicated EDR cases, involving 585 independent inference experiments (39 cases × 3 models × 5 trials) across three reasoning LLMs, revealed perfect inter-model agreement, with each model achieving 100% accuracy (195/195 correct inferences per model, with precision, recall, and F1 scores of 1.00). This remarkable consistency is not merely a reflection of individual model capabilities but a direct outcome of the structured

reasoning framework, anchored by meticulously designed reasoning anchors that guide models along a unified logical pathway, mitigating inference drift and ensuring robust performance across diverse architectures.

Chain-of-thought (CoT) reasoning has shown significant promise in improving LLM performance on complex reasoning tasks by encouraging models to generate intermediate reasoning steps [28, 55]. Unstructured CoT allows models to explore multiple reasoning paths, which can lead to inconsistent conclusions when dealing with fragmented or contradictory information typical in crash scenarios. Our reasoning anchors address this limitation by implementing structured guidelines that constrain reasoning within domain-appropriate boundaries while maintaining analytical flexibility. Reasoning anchors serve as structured guides comprising Primary Understanding, EDR Filtering & Correlation, Missing Data Handling, Critical Timing, and EDREVENTNO Interpretation. These anchors address specific challenges in crash analysis, such as temporal overlaps, incorrect EDR records, and unclear vehicle interactions. For instance, the EDREVENTNO Interpretation anchor guided all three models to consistently select EDREVENTNO5 for Vehicle 2's initial collision in Case ID 32548, correcting database errors that misled human experts.

The power of reasoning anchors lies in their ability to standardize inference processes across heterogeneous LLM architectures, ensuring that models with distinct internal mechanisms converge on consistent outputs. Different reasoning models exhibit significant variations in their inference approaches: DeepSeek-R1, trained through pure reinforcement learning, excels at deep chain-of-thought reasoning but exhibits higher hallucination rates; Grok3-Mini's smaller parameter size enables high operational efficiency but limits its reasoning chain capabilities (**https://x.ai/news/grok-3**)； Gemini 2.5 Pro possesses superior long-context understanding and multimodal reasoning capabilities, yet its reasoning mode may lead to overthinking in simple tasks [27, 56]. Traditional chain-of-thought approaches in LLM applications often suffer from uncontrollable stochastic behaviors stemming from architectural variations and reasoning chain inconsistencies, leading to output variability that compromises analytical reliability. In contrast, reasoning anchors impose a rigorous inference framework that aligns chain-of-thought model outputs with task-specific logical requirements. This structured approach transforms the reasoning process into a reproducible, deterministic workflow, thereby significantly enhancing the reliability of AI-driven crash analysis in real-world applications.

Furthermore, our results demonstrate to a certain extent that combining reasoning anchors with domain-knowledge-based prompt engineering to guide LLMs can also reduce hallucination risks, which represents a common challenge for LLM-based systems when processing fragmented, inconsistence or incomplete data. For example, the ***Missing Data Handling anchor*** equips the system to make informed inferences in the absence of complete EDR records, as shown in **3.2.1 Case study**, where Vehicle 1's missing EDR data did not impede the accurate identification of collision dynamics. By embedding traffic-specific analytical protocols—such as temporal calibration and cross-modal validation—into the reasoning framework, the system ensures that LLMs operate within a knowledge-rich context, enhancing their robustness and reducing errors that might arise from generic reasoning approaches.

In summary, reasoning anchors provide a robust methodological framework for building more reliable and reproducible LLM-based analysis systems. By explicitly addressing the challenges of analytical transparency and consistency, this prompt design technique paves the way for trustworthy AI integration in specialized, high-stakes vertical domains.

4.3 Limitation and Future Works

Although our AI-driven analytical framework demonstrated exceptional accuracy and robustness in the lead vehicle deceleration (LVD) collision scenarios tested in this study, its validation has not yet been extended

to other types of vehicle collision scenarios. As a result, the current performance conclusions are specific to LVD conditions and cannot be directly generalized to broader contexts, such as side impacts, head-on collisions, or multi-vehicle pile-ups, each presenting distinct analytical complexities and data characteristics. It is important to contextualize the performance claims of our AI framework. While we have demonstrated a 96% accuracy rate achieved in a remarkably short timeframe, this metric inherently acknowledges that AI-based models are not infallible. Consequently, in scenarios demanding absolute certainty, sole reliance on AI may be insufficient or inappropriate. Furthermore, the comparison between AI and human analyst performance must be interpreted with precision. The human benchmark in this study was established using research analysts, and their performance (against which the AI was measured) is specific to that context. We explicitly acknowledge that a certified crash reconstructionist with deep expertise in vehicle data systems would likely achieve near-perfect accuracy, potentially matching or exceeding the AI's performance. Conversely, less experienced analysts would likely make more errors. The purpose of our comparison was not to claim general superiority over all human experts, but to provide a concrete reference point demonstrating that AI can deliver value by performing at a high level on a complex analytical task, offering a combination of speed and consistency that can augment human expertise.

Future research could therefore focus on extending the applicability of our analytical framework to diverse crash scenarios to assess its generalizability and adaptability comprehensively. Additionally, further real-world road tests and validations in live, real-time traffic conditions are essential. Such practical evaluations will enable more accurate assessments of the framework's performance under dynamically changing environments and unexpected circumstances, ultimately contributing to robust, scalable, and broadly applicable AI-driven collision analysis tools.

5 CONCLUSION

The primary contribution of this study lies in the development of a novel two-stage multi-agent framework for pre-crash reconstruction, leveraging advanced AI to enhance the accuracy and efficiency of traffic collision analysis. The following conclusions can be drawn:

(1) The framework's ability to integrate fragmented multimodal data (textual, tabular, and visual) into comprehensive crash reconstructions reveals a dynamic analytical process, with first-stage reconstructions significantly influencing subsequent reasoning outcomes.

(2) The system's precise identification of striking and struck vehicles and inference the most relevant EDR records to precrash, demonstrates coherent analytical patterns that support robust crash reasoning, offering insights for designing AI-driven crash analysis tools.

(3) The framework achieves superior performance (100% accuracy in complex LVD cases) compared to traditional human expert methods (92% accuracy), highlighting the potential of AI to overcome challenges posed by incomplete or ambiguous data.

Future research will focus on expanding the framework's applicability across diverse crash scenarios to enhance its robustness and generalizability. Additionally, optimizing the integration of multimodal data and addressing computational demands will guide the development of AI-driven tools that contribute to advancements in traffic crash analysis.

REFERENCES

1. WHO. 2023. *Global Status Report on Road Safety 2023.* World Health Organization (Geneva: World Health Organization). **https://www.who.int/publications/i/item/9789240086517**.

2. NHTSA. 2025. *Early Estimates of Motor Vehicle Traffic Fatalities and Fatality Rate by Sub-Categories in 2024.* U.S. Department of Transportation (1200 New Jersey Avenue SE, Washington, DC 20590). **https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813729**.
3. Scanlon, J. M., K. D. Kusano, and H. C. Gabler, *Analysis of Driver Evasive Maneuvering Prior to Intersection Crashes Using Event Data Recorders.* Traffic Inj Prev, (2015), *16 Suppl 2*, S182-9. **https://doi.org/10.1080/15389588.2015.1066500**
4. Kolla, E., V. Adamova, and P. Vertal, *Simulation-based reconstruction of traffic incidents from moving vehicle mono-camera.* Sci Justice, (2022), *62*(1), 94-109. **https://doi.org/10.1016/j.scijus.2021.11.001**
5. Doecke, S. D., M. R. J. Baldock, C. N. Kloeden, and J. K. Dutschke, *Impact speed and the risk of serious injury in vehicle crashes.* Accid Anal Prev, (2020), *144*, 105629. **https://doi.org/10.1016/j.aap.2020.105629**
6. Santos, Kenny, Nuno M. Silva, João P. Dias, and Conceição Amado, *A methodology for crash investigation of motorcycle-cars collisions combining accident reconstruction, finite elements, and experimental tests.* Engineering Failure Analysis, (2023), *152*. **https://doi.org/10.1016/j.engfailanal.2023.107505**
7. Shen, Z., W. Ji, S. Yu, G. Cheng, Q. Yuan, Z. Han, H. Liu, and T. Yang, *Mapping the knowledge of traffic collision Reconstruction: A scientometric analysis in CiteSpace, VOSviewer, and SciMAT.* Sci Justice, (2023), *63*(1), 19-37. **https://doi.org/10.1016/j.scijus.2022.10.005**
8. Vida, Gábor and Árpád Török, *Expected effects of accident data recording technology evolution on the identification of accident causes and liability.* European Transport Research Review, (2023), *15*(1), 17. **https://doi.org/10.1186/s12544-023-00591-4**
9. Evtiukov, Sergei, Elena Kurakina, Valery Lukinskiy, and Aleksey Ushakov, *Methods of Accident Reconstruction and Investigation Given the Parameters of Vehicle Condition and Road Environment.* Transportation Research Procedia, (2017), *20*, 185-192. **https://doi.org/https://doi.org/10.1016/j.trpro.2017.01.049**
10. Davis, Gary A., *Crash reconstruction and crash modification factors.* Accident Analysis & Prevention, (2014), *62*, 294-302. **https://doi.org/10.1016/j.aap.2013.09.027**
11. Deng, X., Z. Du, H. Feng, S. Wang, H. Luo, and Y. Liu, *Investigation on the Modeling and Reconstruction of Head Injury Accident Using ABAQUS/Explicit.* Bioengineering (Basel), (2022), *9*(12). **https://doi.org/10.3390/bioengineering9120723**
12. Maulina, Dewi, Diandra Yasmine Irwanda, Thahira Hanum Sekarmewangi, Komang Meydiana Hutama Putri, and Henry Otgaar, *How accurate are memories of traffic accidents? Increased false memory levels among motorcyclists when confronted with accident-related word lists.* Transportation Research Part F: Traffic Psychology and Behaviour, (2021), *80*, 275-294. **https://doi.org/https://doi.org/10.1016/j.trf.2021.04.015**
13. Chung, Younshik and IlJoon Chang, *How accurate is accident data in road safety research? An application of vehicle black box data regarding pedestrian-to-taxi accidents in Korea.* Accident Analysis & Prevention, (2015), *84*, 1-8. **https://doi.org/10.1016/j.aap.2015.08.001**
14. Miller, Timothy, *The Hippies and American Values*. 2011: The University of Tennessee Press.
15. Janstrup, Kira H., Kaplan Sigal, Hels Tove, Lauritsen Jens, and Carlo G. and Prato, *Understanding traffic crash under-reporting: Linking police and medical records to individual and crash*

characteristics. Traffic Injury Prevention, (2016), *17*(6), 580-584. **https://doi.org/10.1080/15389588.2015.1128533**

16. Bhatti, Junaid A. and Louis-Rachid and Salmi, *Challenges in Evaluating the Decade of Action for Road Safety in Developing Countries: A Survey of Traffic Fatality Reporting Capacity in the Eastern Mediterranean Region.* Traffic Injury Prevention, (2012), *13*(4), 422-426. **https://doi.org/10.1080/15389588.2012.655431**
17. Viano, David C., *NHTSA crash investigation sampling system (CISS) is unreliable, inaccurate and does not estimate serious injury in motor vehicle crashes.* Traffic Injury Prevention, (2025), *26*(1), 61-75. **https://doi.org/10.1080/15389588.2025.2451572**
18. Arteaga, Cristian and JeeWoong Park, *A large language model framework to uncover underreporting in traffic crashes.* Journal of Safety Research, (2025), *92*, 1-13. **https://doi.org/10.1016/j.jsr.2024.11.009**
19. Ruth, Richard, Charles King, Andrew Rich, and Hamed Sadrnia. *Accuracy of 2016-2022 EDRs in IIHS Crash Tests*. in *WCX SAE World Congress Experience*. 2024.
20. Adeel, Muhammad, King , Meredith, Usman , Sheikh M., and Asad J. and Khattak, *Advancing crash investigation with connected and automated vehicle data: Insights from a survey of law enforcement.* Journal of Transportation Safety & Security, (2025), *0*(0), 1-30. **https://doi.org/10.1080/19439962.2025.2462803**
21. OpenAI, *GPT-4 Technical Report.* (2024). **https://doi.org/10.48550/arXiv.2303.08774**
22. Chiriatti, Massimo, Marianna Ganapini, Enrico Panai, Mario Ubiali, and Giuseppe Riva, *The case for human–AI interaction as system 0 thinking.* Nature Human Behaviour, (2024), *8*(10), 1829-1830. **https://doi.org/10.1038/s41562-024-01995-5**
23. Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, *Language Models are Few-Shot Learners.* (2020). **https://doi.org/10.48550/arXiv.2005.14165**
24. Gemini Team, Google, *Gemini: A Family of Highly Capable Multimodal Models.* (2025). **https://doi.org/10.48550/arXiv.2312.11805**
25. Nie, Tong, Jian Sun, and Wei Ma, *Exploring the roles of large language models in reshaping transportation systems: A survey, framework, and roadmap.* Artificial Intelligence for Transportation, (2025), *1*, 100003. **https://doi.org/10.1016/j.ait.2025.100003**
26. Chen, Yanda, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez, *Reasoning Models Don't Always Say What They Think.* Alignment Science Team, Anthropic, (2025).
27. DeepSeek-AI, *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* (2025). **https://doi.org/10.48550/arXiv.2501.12948**
28. Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou, *Chain-of-thought prompting elicits reasoning in large language models.* Advances in neural information processing systems, (2022), *35*, 24824-24837.

29. Wu, Kebin, Wenbin Li, and Xiaofei Xiao, *AccidentGPT: Large Multi-Modal Foundation Model for Traffic Accident Analysis.* (2024). **https://doi.org/10.48550/arXiv.2401.03040**
30. Zhong, Wuchang, Jinglin Huang, Maoqiang Wu, Weinan Luo, and Rong Yu, *Large language model based system with causal inference and Chain-of-Thoughts reasoning for traffic scene risk assessment.* Knowledge-Based Systems, (2025), *319*, 113630. **https://doi.org/10.1016/j.knosys.2025.113630**
31. Schluntz, Erik and Barry Zhang. 2024. *Building Effective Agents.* Anthropic (Anthropic). **https://www.anthropic.com/engineering/building-effective-agents**.
32. Julia Wiesinger, Patrick Marlow, Vladimir Vuskovic. 2025. *Agents White Paper.* Google (Google). **https://www.kaggle.com/whitepaper-agents**.
33. Xi, Zhiheng, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui, *The rise and potential of large language model based agents: a survey.* Science China Information Sciences, (2025), *68*(2). **https://doi.org/10.1007/s11432-024-4222-0**
34. Lu, Yikang, Alberto Aleta, Chunpeng Du, Lei Shi, and Yamir Moreno, *LLMs and generative agent-based models for complex systems research.* Physics of Life Reviews, (2024), *51*, 283-293. **https://doi.org/10.1016/j.plrev.2024.10.013**
35. Vrdoljak, Josip, Zvonimir Boban, Ivan Males, Roko Skrabic, Marko Kumric, Anna Ottosen, Alexander Clemencau, Josko Bozic, and Sebastian Völker, *Evaluating large language and large reasoning models as decision support tools in emergency internal medicine.* Computers in Biology and Medicine, (2025), *192*, 110351. **https://doi.org/https://doi.org/10.1016/j.compbiomed.2025.110351**
36. Wu, Chaoyi, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie, *Towards evaluating and building versatile large language models for medicine.* npj Digital Medicine, (2025), *8*(1), 58. **https://doi.org/10.1038/s41746-024-01390-4**
37. Ashqar, Huthaifa I., Ahmed Jaber, Taqwa I. Alhadidi, and Mohammed Elhenawy, *Advancing Object Detection in Transportation with Multimodal Large Language Models (MLLMs): A Comprehensive Review and Empirical Testing.* Computation, (2025), *13*(6), 133.
38. Li, Xinyi, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang, *A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges.* Vicinagearth, (2024), *1*(1), 9. **https://doi.org/10.1007/s44336-024-00009-2**
39. Gao, Mingyan, Yanzi Li, Banruo Liu, Yifan Yu, Phillip Wang, Ching-Yu Lin, and Fan Lai, *Single-agent or Multi-agent Systems? Why Not Both?* (2025). **https://doi.org/10.48550/arXiv.2505.18286**
40. Kitajima, Sou, Shimono Keisuke, Tajima Jun, Antona-Makoshi Jacobo, and Nobuyuki and Uchida, *Multi-agent traffic simulations to estimate the impact of automated technologies on safety.* Traffic Injury Prevention, (2019), *20*(sup1), S58-S64. **https://doi.org/10.1080/15389588.2019.1625335**
41. Chen, Junzhou and Sidi Lu. *An Advanced Driving Agent with the Multimodal Large Language Model for Autonomous Vehicles*. in *2024 IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)*. 2024. IEEE.
42. Haji, Fatemeh, Mazal Bethany, Maryam Tabar, Jason Chiang, Anthony Rios, and Peyman Najafirad, *Improving LLM Reasoning with Multi-Agent Tree-of-Thought Validator Agent.* (2024). **https://doi.org/10.48550/arXiv.2409.11527**

43. Zhang, Miao, Zhenlong Fang, Tianyi Wang, Shuai Lu, Xueqian Wang, and Tianyu Shi, *CCMA: A framework for cascading cooperative multi-agent in autonomous driving merging using Large Language Models.* Expert Systems with Applications, (2025), *282*, 127717. **https://doi.org/10.1016/j.eswa.2025.127717**
44. Guo, Xusen, Qiming Zhang, Junyue Jiang, Mingxing Peng, Meixin Zhu, and Hao Frank Yang, *Towards explainable traffic flow prediction with large language models.* Communications in Transportation Research, (2024), *4*, 100150. **https://doi.org/10.1016/j.commtr.2024.100150**
45. Guo, Xusen, Xinxin Yang, Mingxing Peng, Hongliang Lu, Meixin Zhu, and Hai Yang, *Automating Traffic Model Enhancement with Ai Research Agent.* (2024). **https://doi.org/10.2139/ssrn.4995746**
46. NHTSA. 2025. *NHTSA Field Crash Investigation 2023 Coding and Editing Manual.* National Highway Traffic Safety Administration (Washington, DC, United States). **https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813614**.
47. Sandmann, Sarah, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese, *Benchmark evaluation of DeepSeek large language models in clinical decision-making.* Nature Medicine, (2025). **https://doi.org/10.1038/s41591-025-03727-2**
48. Yin, Shukang, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen, *A survey on multimodal large language models.* National Science Review, (2024), *11*(12). **https://doi.org/10.1093/nsr/nwae403**
49. Knoth, Nils, Antonia Tolzin, Andreas Janson, and Jan Marco Leimeister, *AI literacy and its implications for prompt engineering strategies.* Computers and Education: Artificial Intelligence, (2024), *6*, 100225. **https://doi.org/https://doi.org/10.1016/j.caeai.2024.100225**
50. Chen, Banghao, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu, *Unleashing the potential of prompt engineering for large language models.* Patterns, (2025), *6*(6), 101260. **https://doi.org/https://doi.org/10.1016/j.patter.2025.101260**
51. Garg, Ashish, K. Nisumba Soodhani, and Ramkumar Rajendran, *Enhancing data analysis and programming skills through structured prompt training: The impact of generative AI in engineering education.* Computers and Education: Artificial Intelligence, (2025), *8*, 100380. **https://doi.org/https://doi.org/10.1016/j.caeai.2025.100380**
52. Farber, Shai, *AI as a decision support tool in forensic image analysis: A pilot study on integrating large language models into crime scene investigation workflows.* Journal of Forensic Sciences, (2025), *70*(3), 932-943. **https://doi.org/10.1111/1556-4029.70035**
53. Behimehr, Sara and Hamid R. Jamali, *Relations between Cognitive Biases and Some Concepts of Information Behavior.* Data and Information Management, (2020), *4*(2), 109-118. **https://doi.org/10.2478/dim-2020-0007**
54. Acciarini, Chiara, Federica Brunetta, and Paolo Boccardelli, *Cognitive biases and decision-making strategies in times of change: a systematic literature review.* Management Decision, (2020), *59*(3), 638-652. **https://doi.org/10.1108/md-07-2019-1006**
55. Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa, *Large Language Models are Zero-Shot Reasoners.* (2023). **https://doi.org/10.48550/arXiv.2205.11916**
56. Gemini Team, Google. 2025. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.* Google DeepMind, Google

(Google DeepMind, Google). **https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf**.