IEEE PES ISGT INNOVATIVE SMART GRID TECHNOLOGIES MIDDLE EAST

IEEE PES Power & Energy Society® | IEEE

Multimodal Large Language Model Framework for Safe and Interpretable Grid-Integrated EVs

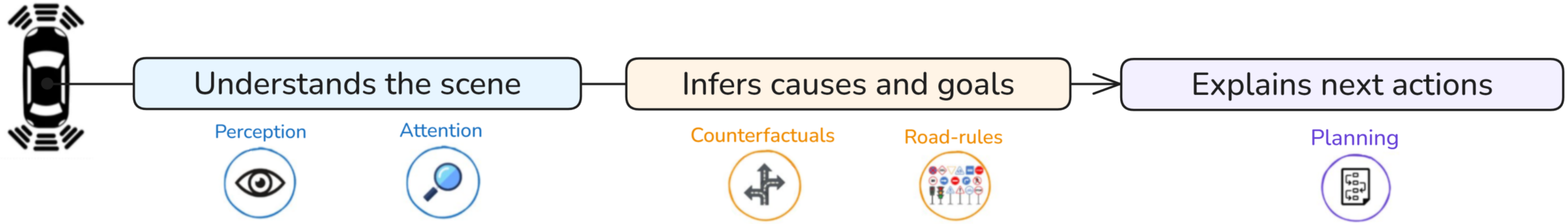
23-26th November 2025, Dubai, UAE

Jean D. Carvalho, Hugo T. Kenji, Ahmad M. Saber, Glaucia Melo,
Max Mauro Dias Santos, and Deepa Kundur

Summary

- Introduction
- Motivation and Challenges
- Research Goals and Contributions
- System Architecture Overview
  - Dataset and Experimental Setup
  - Visual Perception Layer
  - Structured Prompt Generation
  - LLM Configurations
- Validation Case Studies
- Results and Discussions
- Smart Grid and E-Mobility Implications
- Conclusions and Future Work

Introduction



Understands the scene

Perception

Attention

Infers causes and goals

Counterfactuals

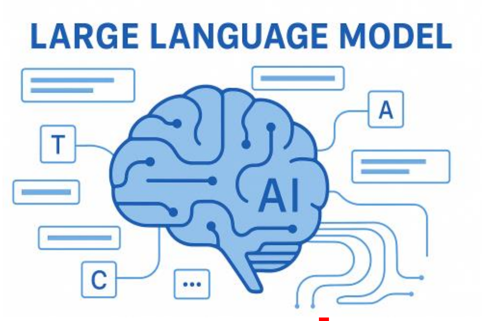
Road-rules

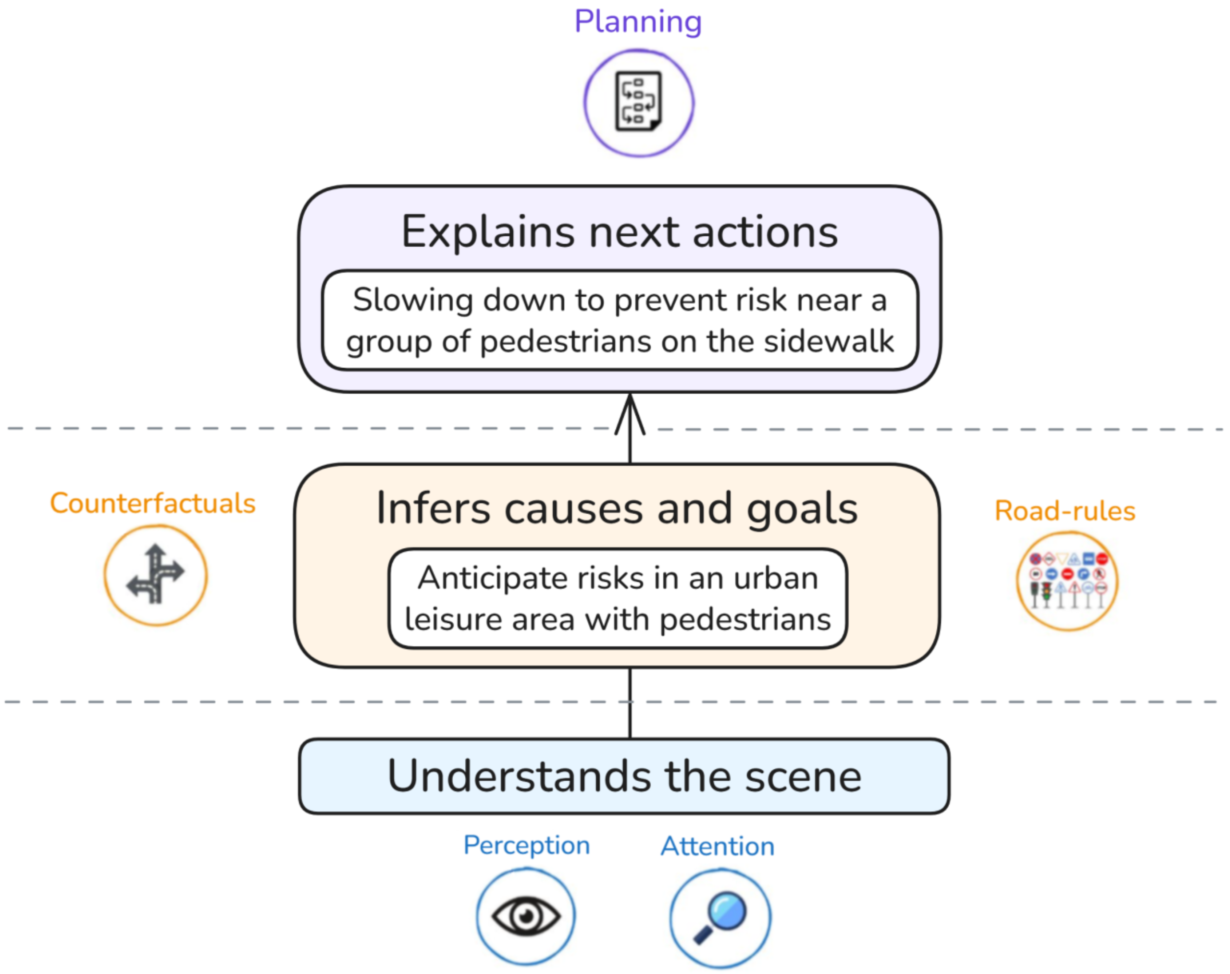
Explains next actions

Planning

Introduction

LLM





Explains next actions

Slowing down to prevent risk near a group of pedestrians on the sidewalk

Infers causes and goals

Anticipate risks in an urban leisure area with pedestrians

Understands the scene

ITS

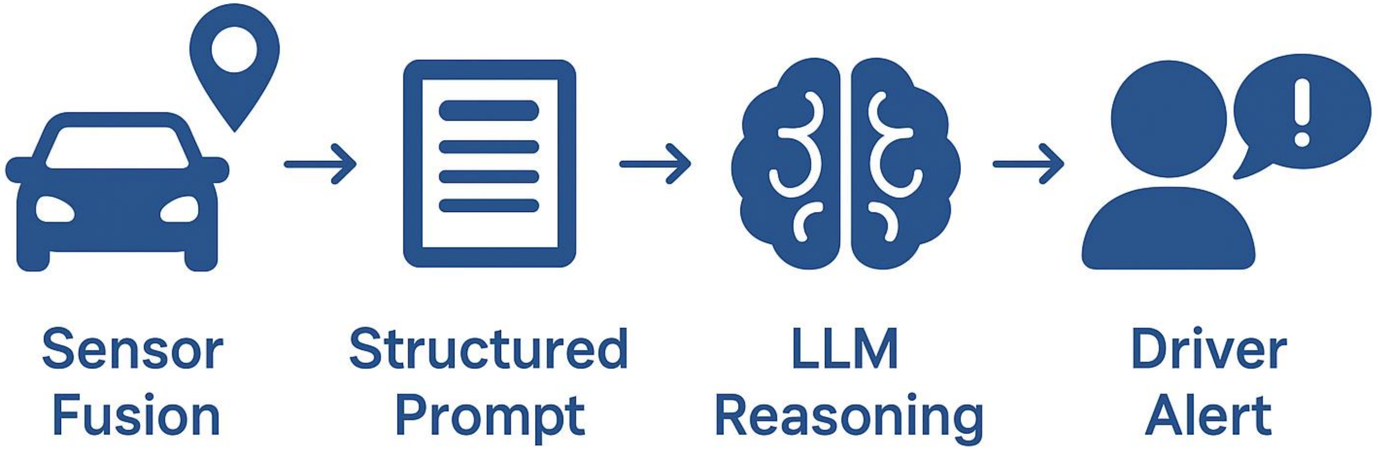
Intersection of LLM and ITS

Motivation and Challenges

- EVs are not just energy consumers — they are **mobile data and energy nodes**.
- Integration with smart grids → new opportunities for:
  - Energy optimization
  - Fleet coordination
  - Real-time traffic-energy planning
- **Challenge:** Lack of interpretable AI that translates multimodal data into human-understandable insights.
- **Key Message:** "We need safe, interpretable, and context-aware intelligence inside EVs — bridging raw sensor data and driver comprehension."

Research Goals and Contributions

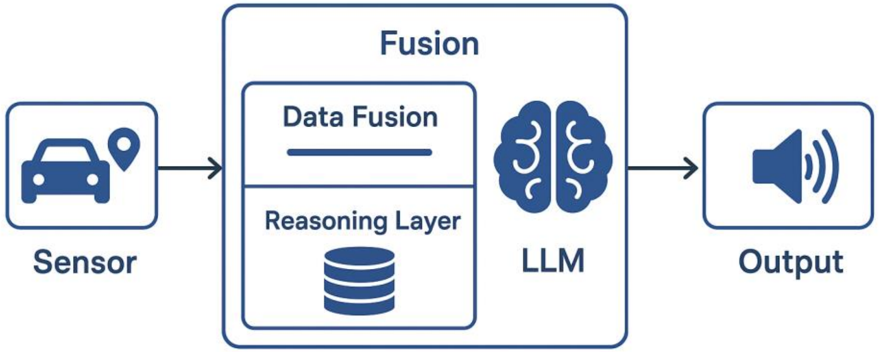
- **Multimodal LLM-based framework** integrating vision, telemetry, and context for interpretable alerts.
- **Structured Prompt Engineering** to transform heterogeneous sensor data into textual reasoning input.
- **Validation with real-world vehicle data**, demonstrating human-aligned, natural-language alerts.



System Architecture Overview

Introduce modular layers:

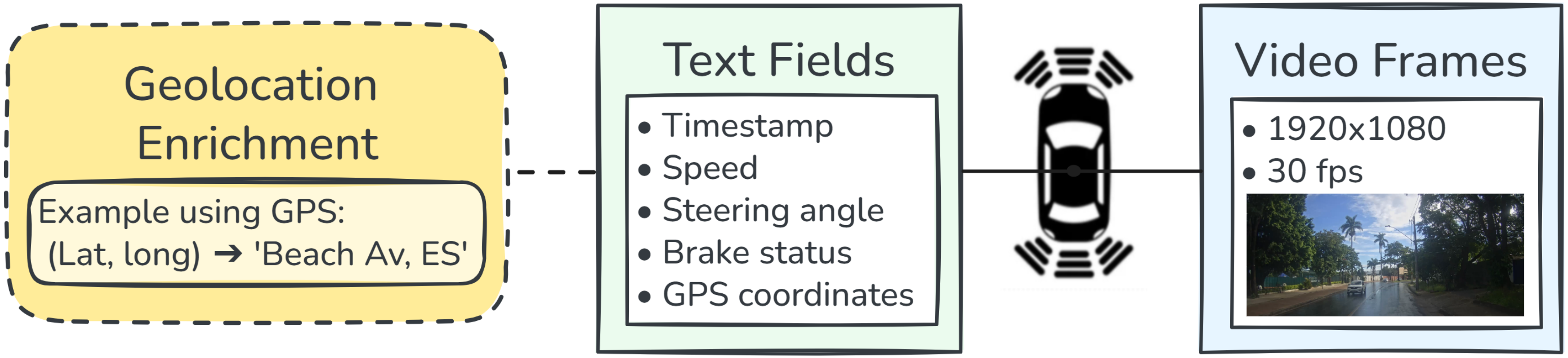
- **Data Acquisition** – Camera, GPS, CAN
- **Preprocessing** – YOLOv8 + Semantic Segmentation
- **Prompt Generation** – Structured textual fusion
- **LLM Reasoning** – GPT-5, Gemini, DeepSeek, GPT-Vision
- **Output** – Natural-language driver alerts and grid-aware data



Key Message: "This pipeline transforms multimodal signals into interpretable reasoning in near real-time."

Dataset and Experimental Setup

- Instrumented vehicle: Renault Captur with Camera, GPS, and CAN interface.
- Stored on **CarCará** platform for synchronized acquisition.
- Collected in urban driving conditions (Ponta Grossa & São Paulo).
- Modalities:



Geolocation Enrichment

Example using GPS:
(Lat, long) ➔ 'Beach Av, ES'

Text Fields

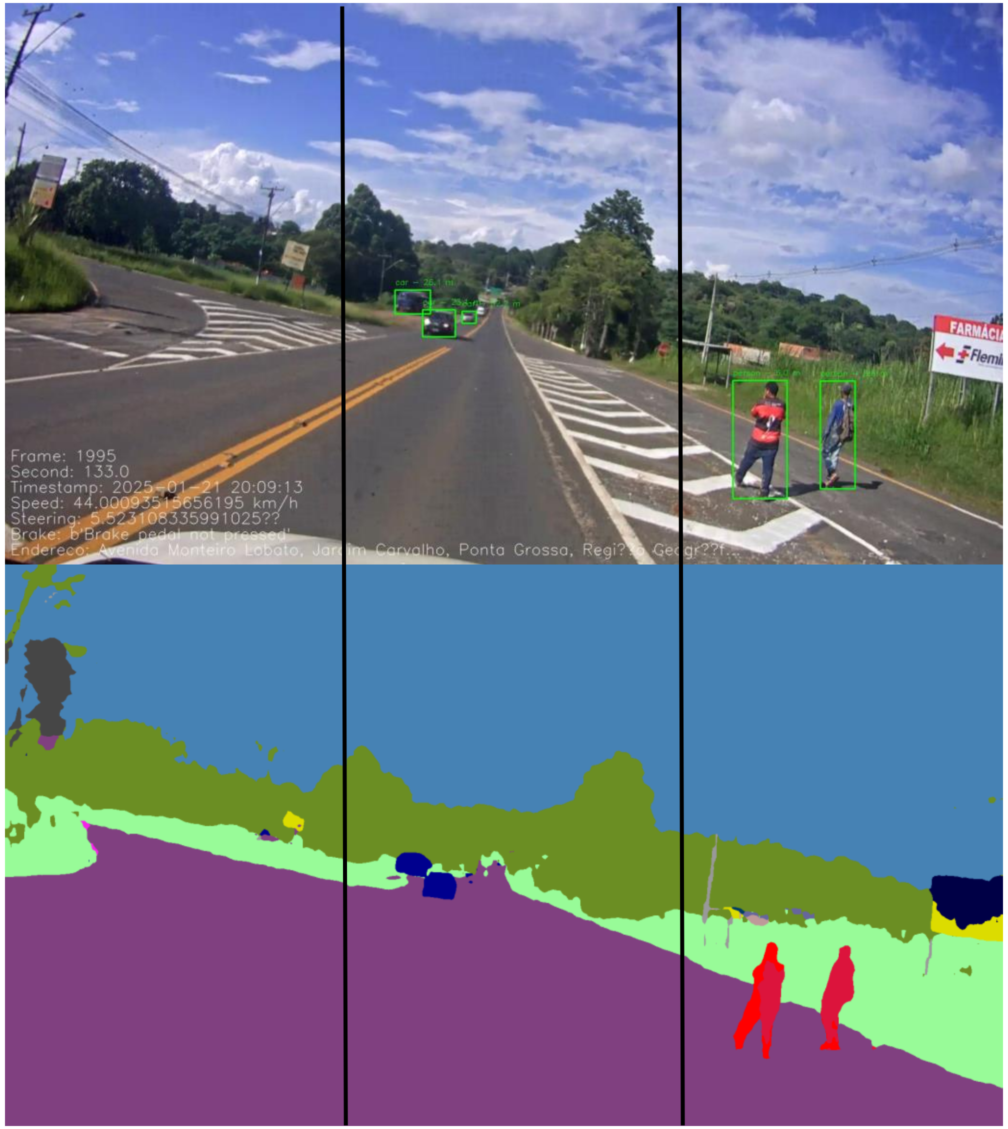
- Timestamp
- Speed
- Steering angle
- Brake status
- GPS coordinates

Video Frames

- 1920x1080
- 30 fps

Visual Perception Layer

- **YOLOv8**: Detects dynamic objects (pedestrians, vehicles, traffic signs).
- **Cityscapes Segmentation**: Identifies static elements (roads, sidewalks, vegetation).
- **Distance Estimation**: Derived from bounding box size and pinhole model.
- **Spatial Division:** Splits frame into three zones (Left, center and right) for position-based reasoning.



```
Frame: 1995
Second: 133.0
Timestamp: 2025-01-21 20:09:13
Speed: 44.00093515656195 km/h
Steering: 5.523108335991025??
Brake: b'Brake pedal not pressed'
Endereco: Avenida Monteiro Lobato, Jardim Carvalho, Ponta Grossa, Regi??o Geogr??f...
```

Structured Prompt Generation

- Each scene → structured textual prompt:
  - Instruction
  - Vehicle (speed, brake, steering)
  - Location (address)
  - Scene (object + segmentation results)
- Example:

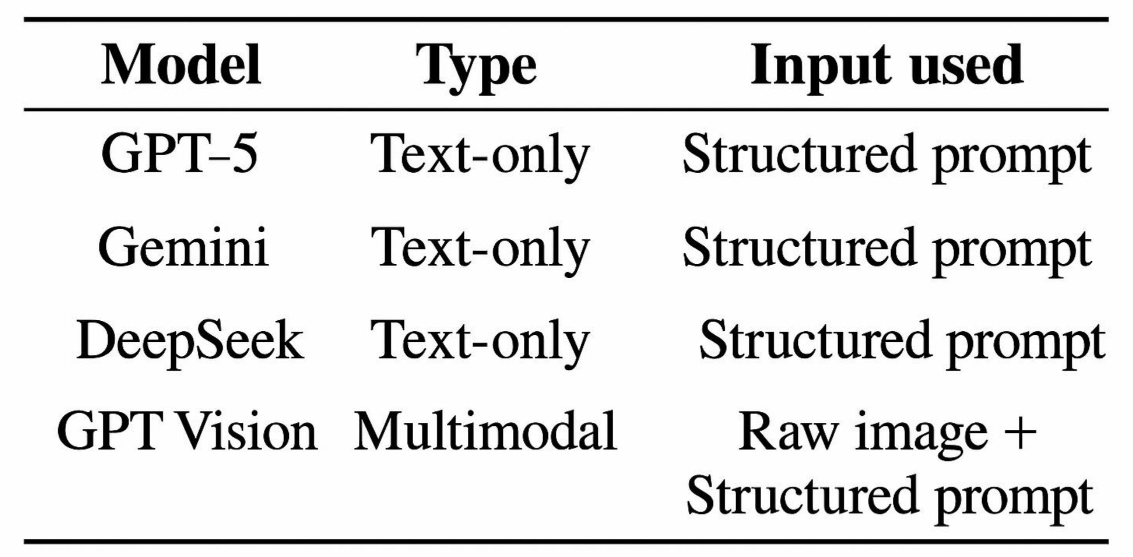
Instruction: Analyze the scene, send an alert to the driver if necessary quickly
Vehicle: Speed = 40 km/h; Brake = not pressed
Location: Main St, NY
Scene: person (5.9m right), car (23 m center left)
Sidewalk: False (right)

Benefits:

- Converts complex multimodal inputs → interpretable format
- Enables consistency, robustness, and explainability

LLM Configurations

- **Text-only**: GPT-5, Gemini, DeepSeek (structured prompts)
- **Multimodal**: GPT-Vision (raw image + text)
- **Tradeoff**:
  - Text-only → faster inference (~1 s)
  - Multimodal → richer context (~3–4 s)



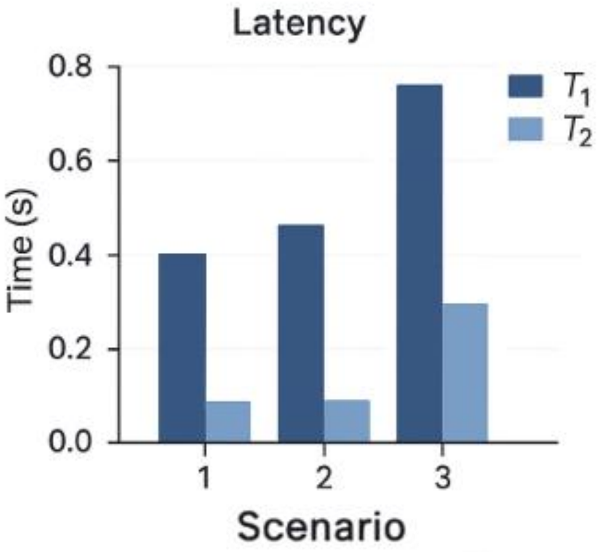
Model	Type	Input used
GPT-5	Text-only	Structured prompt
Gemini	Text-only	Structured prompt
DeepSeek	Text-only	Structured prompt
GPT Vision	Multimodal	Raw image + Structured prompt

Validation Case Studies

- **Scenario 1**: Pedestrians ahead, no sidewalk → "Brake, pedestrians may enter lane."
- **Scenario 2**: Bus left, car right, close proximity → "Avoid lateral collision."
- **Scenario 3**: Multiple elements on wide avenue → "Reduce speed; prepare to stop."

All alerts matched expert human assessments.

Scenario	T_1	T_2
1	0.40	0.01
2	0.43	0.01
3	0.72	0.03

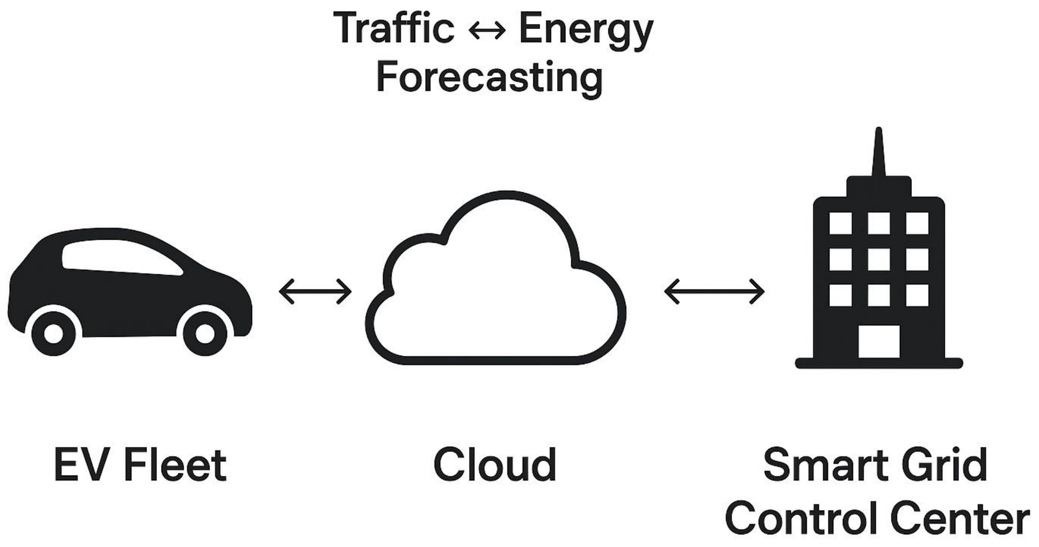


Results and Discussions

- Table VI summary: GPT outputs vs human evaluation → 100% match on risk detection.
- LLM generated timely, context-rich alerts.
- Demonstrated:
  - **Human-aligned reasoning**
  - **Low latency**
  - **Interpretability and scalability**

Smart Grid and E-Mobility Implications

- Structured textual outputs can feed smart grid decision systems:
  - Fleet coordination
  - Load forecasting
  - Traffic-aware energy planning
- EVs act as **intelligent, distributed sensing agents**.



Conclusions and Future Work

Summary:

- LLM-based multimodal fusion enables human-interpretable alerts.
- Framework validated with real-world data.
- Scalable and deployable on edge devices.

Future Directions:

- More scenarios (weather, nighttime)
- Fleet-level integration
- Edge deployment for real-time safety

IEEE PES Power & Energy Society® | IEEE

IEEE PES ISGT INNOVATIVE SMART GRID TECHNOLOGIES MIDDLE EAST

“That's all Folks!”