Autonomous Driving Assistant

Carlos Illán Aldariz
Intelligent Systems Engineering
UIE
A Coruña, Spain

***Abstract*—Autonomous driving technologies increasingly rely on the integration of multiple fields such as Natural Language Processing (NLP), Machine Learning, Intelligent Systems, and Computer Vision. This project presents a modular framework for an intelligent driving assistant composed of three main components. First, a conversational chatbot based on a Large Language Model (LLM) evaluates vehicle conditions and road environments, offering adaptive recommendations. Second, an audio processing module extracts mentioned cities from spoken input, computes the optimal driving route, and generates a visual representation of the path. Third, deep learning models are trained for pedestrian and traffic sign detection to enhance environmental awareness, addressing key challenges in road safety. Together, these components demonstrate the potential of combining language understanding, route optimization, and visual perception for advanced driver assistance systems.**

***Index Terms*—Autonomous Driving, Driving System, NLP, Natural Language Processing, Computer Vision, Route Optimization, Intelligent Systems, Deep Learning.**

I. Introduction

Autonomous driving has emerged as one of the most challenging and promising fields in modern artificial intelligence research. Building robust systems capable of perceiving their environment, interpreting multimodal data, making intelligent decisions, and interacting naturally with human operators demands the convergence of multiple advanced technologies, especially Natural Language Processing (NLP), Machine Learning, Intelligent Systems, and Computer Vision.

This project proposes a modular approach to enhance autonomous driving assistance through the integration of these disciplines. The system is composed of three interconnected modules. First, a conversational chatbot powered by a Large Language Model (LLM) engages the driver or operator in a natural dialogue to assess the vehicle's condition and environmental factors, providing real-time, adaptive recommendations. This leverages the latest advances in NLP and dialogue systems to promote safer driving behavior.

Second, a speech-driven route optimization tool processes audio inputs to identify mentioned cities, extracts structured information via information extraction techniques and computes the optimal route between the origin and destination using graph-based algorithms. Additionally, a visual map is generated to support intuitive understanding of the computed route. This component showcases the combination of speech processing, named entity recognition (NER), and pathfinding algorithms.

Third, a set of deep learning models specialized in pedestrian and traffic sign detection are developed and evaluated. Using supervised learning methods and computer vision architectures, these models provide crucial real-time perception capabilities, allowing the system to recognize key elements in the driving environment and thus improve decision-making. Although traffic light detection was initially considered, technical limitations necessitated focusing on traffic signs and pedestrian detection.

The interdisciplinary nature of this project highlights the synergies between language understanding, sensory data processing, and intelligent decision-making, offering a comprehensive framework for future advancements in intelligent and autonomous vehicles.

II. State of the Art

The development of autonomous driving systems increasingly requires the integration of multiple cutting-edge technologies across language processing, speech recognition, decision-making, and computer vision. In this section, we review the main technological pillars that support the proposed system.

A. Conversational Agents in Driving Assistance

Conversational agents based on Large Language Models (LLMs) have shown significant potential in enhancing human-vehicle interaction. Traditional dialogue systems often relied on rule-based structures or limited context windows, but recent advancements, such as the Mistral model, have enabled more sophisticated natural language understanding and generation. However, due to memory and context window constraints, LLMs like Mistral require careful input structuring to maintain coherence. In this project, the system gathers predefined answers before querying the model in a single prompt, mitigating the limitations of context retention and ensuring consistent feedback generation. Similar approaches are becoming increasingly common in constrained-resource environments where lightweight LLM deployment is necessary.

B. Speech-to-Text and Named Entity Recognition for Route Planning

Speech recognition technologies have evolved considerably with the advent of models such as Whisper, which provides robust and high-fidelity transcription even under noisy conditions. Transforming spoken language into structured information is critical for intelligent route planning systems. Named Entity Recognition (NER) models, widely available through platforms like Hugging Face, allow the automatic extraction of geographic entities such as cities from transcribed audio. Once the cities are identified, pathfinding algorithms are employed to compute optimal routes. Greedy Best-First Search (GreedyBFS), while not guaranteeing globally optimal solutions like Dijkstra's algorithm, offers a computationally efficient strategy well-suited for real-time or resource-constrained applications.

C. Object Detection for Driving Safety

Visual perception is a cornerstone of autonomous driving, with object detection models playing a key role in identifying environmental elements such as pedestrians and traffic signs. The YOLO (You Only Look Once) family, particularly YOLOv8, represents a state-of-the-art approach, offering a balance between accuracy and inference speed suitable for real-time applications. In this project, customized models were trained on curated datasets from Roboflow, focusing on pedestrian and traffic sign detection after traffic light detection proved unfeasible. Fine-tuning YOLOv8 on domain-specific data enables the system to adapt to varied real-world conditions while maintaining lightweight deployment suitable for local execution.

D. Local Deployment and System Integration

A critical design choice of the project is the exclusive use of local resources for both model inference and user interaction, avoiding reliance on external APIs or cloud services. Python serves as the backbone of the system, integrating libraries such as Streamlit for user interfaces, YOLO frameworks for detection, and Whisper for speech processing. This local-first approach enhances system autonomy, reduces latency, and ensures greater privacy and data control — essential attributes for intelligent driving assistance in practical applications.

III. Experimentation

To assess the feasibility and integration of multimodal AI technologies for autonomous driving assistance, the project was divided into three independent experimental modules: a dialogue-based assistant, a speech-driven routing system, and visual detection models. Each module was implemented and tested in isolation, then integrated within a unified local Python-based environment.

A. Conversational Assistant with LLMs

The chatbot module was developed using the Mistral language model, deployed locally ensuring system autonomy and data privacy. Given the model's context window limitations, a sequential approach was employed: the system first conducted a predefined series of questions to gather information about the vehicle's status and road conditions, and subsequently aggregated the user's responses into a single prompt submitted to the LLM. The resulting feedback was generated based on the model's internal knowledge and served as guidance for safe driving behavior.

B. Audio-Based Route Extraction and Optimization

The second module involved the processing of audio inputs to dynamically generate optimal driving routes. Whisper was utilized to transcribe audio files mentioning an origin and a destination city. Subsequently, Named Entity Recognition (NER) was applied using a model sourced from Hugging Face to extract the cities from the text.

To account for potential transcription errors or slight variations in city names produced by the NER module, a Levenshtein distance-based string matching algorithm was applied. This step allowed fuzzy matching of extracted entities with the predefined list of cities in the graph, ensuring robust route computation even in cases of imperfect recognition.

Approximately five to ten distinct audio samples were tested, demonstrating robust performance in the majority of cases. Once extracted, the cities were mapped onto a graph representing around 40–50 Galician cities, constructed to reflect realistic geographic distributions. The routing was computed using a Greedy Best-First Search (GreedyBFS) algorithm, prioritizing computational efficiency and responsiveness. Visual representations of the computed routes were generated to enhance interpretability and user understanding.

C. Object Detection with YOLOv8

The third module focused on developing visual perception capabilities through object detection models. YOLOv8 was employed to train models capable of detecting pedestrians and traffic signs, using datasets obtained from Roboflow. Each dataset contained between 4,000 and 40,000 annotated images, enabling comprehensive model training.

Training was performed locally on a MacBook Pro M1 equipped with 16 GB of RAM, utilizing CPU resources. The models were evaluated during and after training through systematic observation of detection outputs on unseen data, confirming their ability to generalize and accurately recognize relevant objects under various conditions. Although traffic light detection was initially considered, the focus shifted to pedestrian and traffic sign detection to ensure higher reliability and model performance.

IV. Results Analysis

The experimental evaluation of the system covered the three developed modules: conversational assistance, audio-based route optimization, and object detection. Each component was analyzed independently to assess its behavior, strengths, and observed limitations.

A. Conversational Assistant with LLMs

The chatbot assistant based on the Mistral model successfully processed user inputs and generated coherent, contextually appropriate recommendations regarding the vehicle's condition and road environment. The sequential prompt aggregation strategy proved effective in compensating for the model's limited context window, enabling the LLM to generate a holistic feedback message after receiving all responses at once.

The feedback provided by the assistant was understandable and relevant in most cases. The interaction design, based on predefined questions followed by a single comprehensive response, enhanced clarity and avoided confusion that could arise from partial or fragmented conversations.

B. Audio-Based Route Extraction and Optimization

The system demonstrated the ability to transcribe spoken audio with high accuracy using Whisper, capturing the majority of city names mentioned by the user. The Named Entity Recognition (NER) module effectively identified geographic entities from the transcribed text, enabling successful extraction of origin and destination points.

The GreedyBFS routing algorithm produced satisfactory paths across the graph representing Galician cities. Visual inspection of the computed routes confirmed that the selected paths were plausible and consistent with the underlying graph topology. Although occasional transcription or recognition errors were observed in challenging audio samples, the overall performance validated the feasibility of combining speech recognition, NER, and graph-based routing in real-time applications.

C. Object Detection with YOLOv8

The pedestrian and traffic sign detection models trained with YOLOv8 achieved a good level of performance, as observed through qualitative testing on unseen images. The models were able to accurately detect and localize pedestrians and various types of traffic signs under diverse conditions, including different lighting and backgrounds.

Training on large, annotated datasets from Roboflow contributed to the robustness of the detection models despite the limitations of CPU-only training hardware. Visual inspections confirmed that the models generalized well, with minimal false positives and missed detections in the test samples. The shift from traffic light detection to traffic sign detection allowed the project to maintain high detection reliability while focusing on elements critical for driving assistance.

V. Conclusion

This project demonstrates the feasibility of integrating Natural Language Processing, Speech Recognition, Route Optimization, and Computer Vision techniques into a modular autonomous driving assistance system. By decomposing the problem into three complementary components—conversational analysis, audio-based routing, and visual perception—each technological challenge was addressed independently while contributing to a coherent overall architecture.

The conversational assistant based on a locally deployed LLM (Mistral) effectively generated relevant recommendations through structured dialogue, showcasing the potential of lightweight language models for vehicle-driver interaction without relying on cloud resources. The audio processing and route planning module successfully combined Whisper transcription, Named Entity Recognition, and graph-based search algorithms to extract and visualize optimized routes from natural speech, although minor challenges in audio robustness were observed. The pedestrian and traffic sign detection models, developed using YOLOv8 and trained on large-scale datasets, provided reliable visual perception capabilities, ensuring the system could identify critical elements of the driving environment.

Despite hardware limitations, the exclusive use of local processing throughout the system reinforced its applicability in scenarios requiring privacy, autonomy, and low-latency responses. The modularity of the design facilitates future extensions, such as improving the chatbot's adaptive reasoning, enhancing route calculation with more sophisticated algorithms, or expanding object detection capabilities to cover additional classes like traffic lights and dynamic obstacles.

Overall, the project highlights the potential and practicality of combining multimodal AI technologies to support intelligent driving systems, offering a strong foundation for continued research and development in autonomous vehicle assistance.

References

References

[1] OpenAI. (n.d.). *GitHub - openai/whisper: Robust Speech Recognition via Large-Scale Weak Supervision*. GitHub. https://github.com/openai/whisper

[2] AventIQ-AI. (n.d.). *roberta-named-entity-recognition*. Hugging Face. https://huggingface.co/AventIQ-AI/roberta-named-entity-recognition

[3] Habib Abdurrasyid. (n.d.). *Red-Green-Blue Detection of a Traffic Light Object Detection Dataset*. Roboflow. https://universe.roboflow.com/habib-abdurrasyid/red-green-blue-detection-of-a-traffic-light

[4] School. (2023, December 21). *MachineProject Object Detection Dataset*. Roboflow. https://universe.roboflow.com/school-yjaef/machineproject

[5] Leo Ueno. (2025, April 4). *People Detection Object Detection Dataset and Pre-Trained Model*. Roboflow. https://universe.roboflow.com/leo-ueno/people-detection-o4rdr

[6] Fsevval. (2023, December 23). *Traffic sign Object Detection Dataset and Pre-Trained Model*. Roboflow. https://universe.roboflow.com/fsevval/traffic-sign-guy19