Toward a Privacy-Preserving and Edge-Deployable Intelligent Personal Vehicular Agent

Raphael Alves dos Reis
Department of Computer Science
UFMG
Belo Horizonte, Brazil
cap497@ufmg.br

Prof. Dr. Antonio A. F. Loureiro
Department of Computer Science
UFMG
Belo Horizonte, Brazil
loureiro@dcc.ufmg.br

Prof. Dr. Roberto G. Ribeiro
Department of Computing and Systems
UFOP
Ouro Preto, Brazil
roberto.gomes@ufop.edu.br

***Abstract*—The increasing complexity and digitalization of modern vehicles have transformed the driving environment into a multimodal information ecosystem. Although recent advancements in embedded infotainment systems and natural-language interfaces have enabled limited conversational interaction, current in-vehicle assistants remain fundamentally reactive and constrained: they operate on shallow input modalities, lack contextual awareness, provide inconsistent reasoning, and cannot adapt to user behavior or dynamic driving conditions. This work proposes a comprehensive framework for designing an intelligent *Intelligent Personal Vehicular Agent* (IPVA) capable of multimodal perception, contextual reasoning, safety-aware action, and continuous learning. The proposed IPVA integrates vehicle telemetry, driver voice commands, technical documentation, structured tables, and manual-derived imagery through a hybrid Retrieval-Augmented Generation (RAG) pipeline optimized for automotive knowledge domains, and is designed to interface with existing Advanced Driver-Assistance Systems (ADAS) by providing complementary high-level reasoning and learning capabilities. Additionally, a Perception–Reasoning–Action–Learning (PRAL) architecture is introduced to clearly define the computational and behavioral responsibilities of an intelligent vehicular agent. The system incorporates token-level streaming, federated learning for driving-style classification, multimodal retrieval using BM25, dense embeddings, FAISS indexing, and adaptive policies aligned with automotive safety standards. This paper presents the architectural formulation, design motivations, cognitive workflow, learning paradigms, safety constraints, and implementation considerations required to transition from conventional assistants toward a fully adaptive vehicular agent that augments perception-and-control-oriented ADAS stacks.**

***Index Terms*—Intelligent Personal Vehicular Agent, RAG, Multimodal Retrieval, Safety-Aware AI, Perception–Reasoning–Action–Learning, Telemetry, ASR, Federated Learning, Advanced Driver-Assistance Systems (ADAS), In-Vehicle Human–AI Interaction.**

I. Introduction

Future vehicles will operate not merely as transportation platforms, but as intelligent, connected, perceptive, and adaptive cyber-physical ecosystems. As vehicles incorporate increasingly complex infotainment systems, Advanced Driver-Assistance Systems (ADAS), multimodal sensors, and high-resolution documentation, the ability for drivers to efficiently access, interpret, and act upon technical information becomes critical for safety, usability, and autonomy.

Traditional in-vehicle digital assistants—typically limited to fixed voice commands or limited conversational capabilities—are fundamentally insufficient for addressing these needs. Their limitations are evident across several dimensions: (i) they depend almost exclusively on speech-to-text and simple query matching; (ii) they rarely incorporate live telemetry or environmental context; (iii) they lack access to structured and unstructured technical documentation in real time; (iv) they do not reason about multimodal inputs such as tables, diagrams, or visual signals; and (v) they are unable to improve with continued use.

This paper argues that the next evolutionary step in automotive human–AI interaction is the creation of a *Intelligent Personal Vehicular Agent* (IPVA): an intelligent system capable of autonomously interpreting multimodal inputs, conducting structured reasoning, executing safety-aware actions, and engaging in continuous learning. To support this shift, we introduce a unified Perception–Reasoning–Action–Learning (PRAL) cognitive architecture and a multimodal Retrieval-Augmented Generation (RAG) framework tailored for vehicular knowledge environments.

Contemporary ADAS stacks provide robust perception and control loops for tasks such as lane keeping, adaptive cruise control, and collision avoidance, but they are not designed to perform high-level contextual reasoning or long-term learning about user behavior and documentation. The IPVA is therefore positioned as a complementary layer that reasons over technical manuals and telemetry, explains system behavior, and adapts its interaction policies over time, rather than as a replacement for low-level ADAS control functions.

The proposed IPVA:

- integrates voice queries, telemetry streams, digital manuals, tables, and manual images into a unified multimodal perception layer, and is architected for future integration of camera feeds and richer ADAS signals;
- executes contextual reasoning via paraphrasing, hybrid retrieval, cross-modal reranking, and structured action-graph workflows;
- modulates its actions according to driving behavior, safety policies, and dynamic vehicle state;
- and continually evolves through federated learning, adaptive user modeling, and incremental document indexing.

Rather than viewing the vehicle assistant as a static tool, this work positions the IPVA as a **cognitive agent** operating

under strict real-time and safety constraints. The resulting architecture provides a foundation for the development of next-generation automotive interfaces that are responsive, context-aware, and capable of providing authoritative, safety-aligned guidance across diverse modalities.

II. Background and Motivation

A. Limitations of Current In-Vehicle Assistants

Most commercial in-vehicle assistants are built around narrow-domain automatic speech recognition and rule-based command execution. These systems lack access to the full breadth of vehicle documentation, cannot interpret or cross-reference technical content, and provide shallow linguistic understanding. They cannot incorporate visual cues (e.g., dashboard icons, wiring diagrams), structured information (e.g., torque tables, fluid charts), or dynamic telemetry signals. Their lack of contextual reasoning severely limits their utility in real-world scenarios in which the driver requires accurate, concise, and contextually appropriate technical guidance.

B. Need for Multimodal, Context-Aware Vehicular Agents

Vehicle manuals now contain thousands of pages of information, spanning textual descriptions, diagrams, exploded views, procedural steps, electrical schematics, and maintenance charts. In practice, drivers rarely consult these resources during operation due to physical and cognitive constraints. Meanwhile, telemetry streams already provide real-time signals that can dramatically enhance the agent's understanding of the driving context, and future integration with camera feeds or richer ADAS-level semantic signals can extend this capability even further. A IPVA must therefore unify these modalities and synthesize them into actionable knowledge.

C. Retrieval-Augmented Generation for Automotive Knowledge

RAG architectures combine symbolic retrieval mechanisms with generative language models, providing improved factual grounding and interpretability. However, conventional RAG systems focus strictly on textual documents. Automotive knowledge systems require **multimodal RAG**, incorporating:

- lexical BM25 search for high-recall text matching;
- dense semantic vector search for contextual retrieval;
- FAISS-based approximate nearest-neighbor indexing for scalability;
- visual embeddings for diagram and symbol retrieval;
- structured table extraction for mechanical specifications and procedures.

This combination enables the IPVA to retrieve authoritative passages from manuals, identify relevant images, and cross-reference structured data before generating a response.

D. PRAL: A Cognitive Framework for Vehicular Agents

The PRAL architecture provides a principled structure for designing intelligent agents that operate in dynamic, safety-critical settings. While conversational assistants operate primarily under a query–response paradigm, intelligent agents must continuously perform:

1) **Perception** — capturing and interpreting multimodal inputs;
2) **Reasoning** — synthesizing retrieved knowledge and context into coherent understanding;
3) **Action** — generating safe, adaptive, and situationally aware outputs or interventions;
4) **Learning** — refining internal models and index structures based on user interaction and environmental feedback.

This framework is particularly well-suited for vehicular environments, where safety, real-time constraints, multimodality, and continual adaptation are essential.

E. Motivating Use Cases

The proposed IPVA directly addresses several high-impact use cases:

- explaining the meaning and implications of vehicle status indicators or diagnostic codes by grounding OEM- or ADAS-provided signals in the technical manual;
- retrieving torque specifications, battery locations, or fuse diagrams from manuals based on ambiguous natural-language queries;
- modulating verbosity and response structure according to driving style and current vehicle motion;
- detecting risky driver behavior (e.g., aggressive maneuvers) and adapting communication strategies proactively.

These motivations guide the design of the IPVA's architectural components presented in subsequent sections.

III. System Architecture

The Intelligent Personal Vehicular Agent (IPVA) is designed as a multimodal, safety-aware, and continuously adaptive system capable of integrating heterogeneous data sources into a unified cognitive pipeline. The architecture is organized into five major layers: (1) multimodal perception, (2) hybrid retrieval, (3) reasoning and contextual synthesis, (4) action execution and safety modulation, and (5) continuous learning. Each layer interfaces with the others through well-defined representations that enable scalability, interpretability, and modular substitution of components.

Figure 1 conceptualizes the information flow among system modules. The agent operates entirely on-device, ensuring low-latency inference, privacy preservation, and full functionality in connectivity-constrained environments.

A. Layered Overview

Table I summarizes the five-layer structure.

Each layer is described in detail in the following sections.

B. Multimodal Perception Layer

The Perception layer captures and interprets all external and internal signals relevant to the vehicle, environment, and user. The IPVA integrates several distinct modalities:

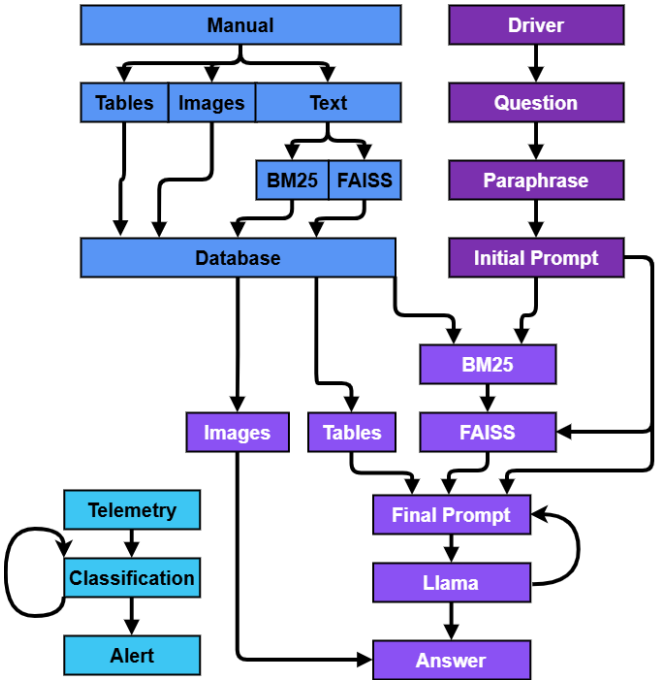

Fig. 1: System architecture.

TABLE I: Architectural Layer Overview

Layer	Primary Responsibilities
Perception	ASR, telemetry, manual imagery, table parsing, manual segmentation
Retrieval	BM25, dense embeddings, FAISS, multimodal reranking
Reasoning	Query reformulation, reasoning templates, context synthesis
Action	Token-level streaming, safety-aware responses, adaptive communication
Learning	Federated learning, incremental indexing, adaptive prompting

1) Driver Voice Input: A local Automatic Speech Recognition (ASR) module transcribes spoken queries into text. Unlike cloud-based ASR systems, the local version ensures resilience in offline contexts, maintains user privacy, and remains robust to vehicle cabin acoustic noise. Transcriptions are normalized into a canonical linguistic form before entering the retrieval pipeline.

2) Technical Manual Processing: Vehicle manuals often exceed hundreds of pages and include:

- textual descriptions of components,
- tables detailing specifications and tolerances,
- diagrams and exploded views,
- symbolic dashboard indicators,
- maintenance procedures and troubleshooting steps.

Manual preprocessing consists of:

1) **PDF segmentation** into semantically coherent chunks;
2) **BM25 indexing** of text segments;
3) **dense embedding generation** using a sentence-transformer encoder;
4) **FAISS index construction** for scalable similarity search;
5) **table extraction** using structural heuristics and OCR when necessary;
6) **image extraction** at page level for multimodal retrieval.

This multimodal representation ensures that queries referencing diagrams, tables, or textual passages are treated uniformly during retrieval.

3) Telemetry and Driving Behavior: The vehicle's CAN bus produces continuous streams of telemetry signals, such as steering angle, throttle/brake position, speed, and yaw rate. These signals feed a federated learning (FL) classifier that infers one of three driving styles: *calm*, *normal*, or *aggressive*. The agent uses this classification to adapt verbosity, timing, and content of responses.

4) Manual Image Perception and Future Camera Integration: Page-level images and diagrams extracted from the technical manuals are encoded using a vision encoder to produce high-dimensional embeddings compatible with the FAISS index. This enables queries that benefit from visual grounding in manual figures (e.g., component layouts, diagrams) rather than text alone. The same infrastructure is intended to support future integration of real-time camera streams or ADAS-provided visual semantics, but the current prototype operates solely on manual-derived imagery.

5) User Interaction Context: Interaction history—previous questions, corrections, preferred wording, audio profile—is tracked to construct a personalized behavioral model. This model allows the IPVA to tailor explanations and anticipate user needs more effectively.

C. Hybrid Retrieval Layer

The retrieval layer constitutes the core of the IPVA's grounding mechanism. Unlike pure LLM-based assistants, the IPVA relies on explicit, verifiable information retrieved from authoritative sources. Retrieval is performed in three stages:

1) Lexical Retrieval via BM25: BM25 provides high-recall candidate selection based on keyword similarity. This stage is essential to ensure that rare automotive terminology, part names, or diagnostic codes are not overlooked due to semantic embedding limitations.

2) Semantic Retrieval via Dense Embeddings: A transformer-based encoder maps each manual segment and table row into a high-dimensional vector space. Queries are similarly embedded, and cosine similarity retrieval provides semantically relevant passages even when user phrasing differs significantly from manual terminology.

3) FAISS-Based Approximate Nearest Neighbor Search: FAISS accelerates semantic retrieval by enabling scalable nearest-neighbor search over thousands of manual segments. Both text embeddings and visual embeddings (diagrams, dashboard symbols) reside in FAISS, unifying retrieval across modalities.

4) Cross-Modal Reranking: BM25 and FAISS candidates are intersected and reranked using an LLM-based embedding similarity metric. This ensures that the final context window:

- fits within the LLM's token budget,
- contains multimodal evidence (text + images + tables),

- maximizes semantic coherence with the user query.

5) Context Window Construction: Selected passages, images, and structured table entries are concatenated into a final context prompt. A maximum token budget is enforced to satisfy real-time constraints, with priority given to:

1) passages directly referencing the query,
2) troubleshooting procedures,
3) safety-related sections,
4) table values such as torques, capacities, and tolerances.

D. Reasoning and Contextual Synthesis Layer

Once retrieval provides the grounding material, the Reasoning layer

IV. PRAL Cognitive Architecture

The transition from a traditional vehicle assistant to a fully autonomous vehicular agent requires a structured cognitive model. To that end, the IPVA is formulated under the Perception–Reasoning–Action–Learning (PRAL) framework, which decomposes intelligent behavior into well-defined capabilities. Unlike a conventional conversational assistant, which simply maps user queries to predefined responses, a PRAL-based vehicular agent interprets multimodal signals, engages in contextual reasoning, executes safety-aware actions, and continuously adapts to the driver and the environment.

This section formalizes how each PRAL component is implemented within the IPVA and how the cognitive workflow collectively enables real-time, context-aware vehicular intelligence.

A. Perception

Perception encompasses the mechanisms by which the agent acquires information from the environment, the user, and the vehicle itself. In automotive settings, perceptual capabilities must integrate disparate, asynchronous, and heterogeneous signals. The IPVA's perception subsystem is multimodal, incorporating voice, manual-derived imagery, structured data, telemetry, and historical user behavior.

1) Voice Perception: Driver commands are captured through the ASR module, which operates locally to ensure low latency and privacy. The model is tuned for vehicular cabin acoustics and produces transcriptions robust to engine noise, wind interference, and microphone placement. The ASR output undergoes:

- linguistic normalization,
- filler-word removal,
- temporal segmentation for streaming,
- canonical reformulation for retrieval.

These steps homogenize spoken utterances into structures compatible with the retrieval and reasoning layers.

2) Document Perception: The technical manual, a primary knowledge source for vehicle operation and maintenance, is decomposed into perceptual units:

- raw text segments,
- tables with structured values,
- diagrams and icons extracted as page-level images,
- hierarchical metadata such as section titles and component IDs.

A dedicated preprocessing pipeline maps each unit to vector embeddings, lexical indexes, and visual signatures. This multimodal representation allows the agent to retrieve evidence across text, tables, and imagery simultaneously.

3) Telemetry Perception: The vehicle's real-time operational state is continuously monitored through telemetry streams originating from the vehicle network. Signals such as acceleration, steering angle, yaw rate, and braking intensity feed:

- a federated learning classifier for driving style,
- risk assessment heuristics (e.g., harsh braking detection),
- safety gating mechanisms that constrain agent behavior.

This perceptual channel is essential for safety-aware interaction, enabling the IPVA to modulate response structure and complexity.

4) Visual Perception: Visual information currently comes from page-level images and diagrams extracted from the manuals. A vision encoder maps these images to vectors aligned with the text embeddings, enabling cross-modal matching between user queries and visual content such as component layouts or schematic diagrams. The architecture is designed so that, in future deployments, the same mechanism can ingest camera streams or ADAS-derived visual semantics, but the present prototype does not perform direct perception over live camera feeds or dashboard symbols.

5) User Behavioral Perception: Patterns in past interactions—frequent queries, linguistic preferences, driving routines—form a long-term perceptual memory. This information guides personalization, adaptive prompting, and targeted model refinement.

B. Reasoning

Reasoning transforms perceptual signals into coherent internal representations and actionable knowledge. In the IPVA, reasoning spans query reformulation, multimodal retrieval, semantic alignment, temporal context tracking, and structured cognitive planning.

1) Query Reformulation and Intent Refinement: Natural language queries posed by drivers are often ambiguous, incomplete, or compressed due to cognitive load during driving. The IPVA performs automatic paraphrasing to:

- map colloquial utterances to technical terminology,
- expand shorthand descriptions into full semantic propositions,
- disambiguate component references (e.g., "the left side light"),
- normalize synonyms and multi-language terms.

This step significantly increases retrieval precision for automotive documentation.

2) Multimodal Retrieval and Evidence Aggregation: The Reasoning layer integrates results from the hybrid retrieval module. BM25 identifies lexically similar segments, while FAISS provides semantically aligned passages and diagrams.

The agent intersects and reranks the candidate set using LLM-derived similarity measures, enforcing relevance across:

- textual explanations,
- structured specifications,
- diagrammatic representations,
- extracted tables.

3) Reasoning Templates and Procedural Graphs: Safety-critical reasoning requires stability and structure. Therefore, the IPVA follows predefined templates for:

- definitions and clarifications,
- safety warnings,
- troubleshooting flows,
- procedural guides,
- diagnostic chains.

Action graphs encode conditional reasoning patterns, ensuring consistency across scenarios such as component identification or resolving fault codes.

4) Contextual Synthesis with Grounded LLMs: The LLM synthesizes retrieved evidence into a coherent answer. To preserve trust and safety:

- hallucination risks are reduced by explicit context binding,
- only retrieved evidence is used to construct the response,
- the agent maintains explicit references to diagrams or table entries.

The result is an authoritative explanation aligned with technical documentation and driving context.

C. Action

Action defines how the agent responds and intervenes, integrating safety, timing, communication modality, and user adaptation.

1) Token-Level Streaming Action: The agent delivers responses incrementally at token-level granularity. Streaming:

- reduces latency,
- enables real-time feedback,
- preserves situational awareness,
- aligns with human conversational flow.

2) Telemetry-Adaptive Behavior: Based on the inferred driving style:

- during aggressive driving, responses become shorter and audibly emphasized;
- during calm driving, the agent provides more detailed guidance;
- under critical vehicle maneuvers, responses may be deferred or simplified.

3) Safety-Gated Action Policies: Following safety principles analogous to UNECE R155/R156, the IPVA:

- suppresses distracting content during hazardous conditions,
- prioritizes warnings over informational responses,
- avoids step-by-step procedures when the vehicle is in motion,
- ensures no action contradicts real-time telemetry.

4) Command Execution: When connected to the vehicle network, the agent can assist with:

- infotainment control,
- navigation adjustments,
- user reminders,
- maintenance notifications.

D. Learning

Learning enables long-term adaptation, personalization, and continuous performance improvement.

1) Federated Driving-Style Classification: A federated learning loop updates the driver-behavior classifier without transmitting raw telemetry. This preserves privacy and enables personalization based on local data distributions.

2) Adaptive Prompt Evolution: The IPVA modifies its prompting strategies to reflect:

- drivers' preferred levels of detail,
- typical question patterns,
- linguistic tendencies,
- frequent failure modes.

3) Incremental Knowledge Expansion: Newly extracted tables, updated manual versions, or user-provided images are integrated incrementally into the retrieval index, extending the IPVA's knowledge base without prohibitively expensive reprocessing.

4) Multimodal Model Adaptation: The ASR model, image encoder, and language model adapt through:

- fine-tuning for cabin acoustics,
- visual grounding based on user-provided photos,
- reinforcement from successful interactions.

E. Summary of PRAL Integration

Table II maps components of the IPVA to the PRAL layers.

TABLE II: Mapping of IPVA Components to PRAL Layers

Perception	ASR, telemetry, manual images, manuals, tables, interaction history
Reasoning	Query paraphrasing, hybrid retrieval, reranking, reasoning templates
Action	Streaming output, safety gating, telemetry-adaptive communication
Learning	Federated driving-style updates, incremental indexing, adaptive prompting

Together, these capabilities constitute a fully operational cognitive agent capable of providing reliable, grounded, personalized, and safety-aware assistance during vehicle operation.

V. Implementation Details

The Intelligent Personal Vehicular Agent (IPVA) prototype is implemented as a modular software stack designed for local execution on consumer-grade hardware. This section details the components, engineering strategies, and integration decisions that enable real-time performance, multimodal retrieval, safety-aware interaction, and adaptive learning.

A. System Infrastructure

The system is built around a local server composed of:

- a Python backend coordinating multimodal retrieval and LLM inference;
- a browser-based frontend rendering incremental text streaming;
- local LLM execution via an API-compatible inference engine;
- persistent on-disk indices for BM25, FAISS, and asset databases.

All computation, including ASR, embedding generation, and LLM inference, executes locally without dependence on remote cloud services. This design aligns with vehicular safety and privacy constraints by eliminating third-party exposure of telemetry, voice data, or manual-derived content.

B. Manual Preprocessing Pipeline

A dedicated preprocessing script handles the transformation of technical manuals into structured, multimodal knowledge resources. The pipeline includes:

1) Text Extraction and Segmentation: PDF pages are OCR-processed when needed and segmented into semantic units (e.g., paragraphs, procedure steps, part descriptions). Each segment receives:

- a unique identifier,
- source metadata (manual section, page number),
- classification tags when applicable (maintenance, safety, electrical).

2) BM25 Index Construction: All segments are indexed with BM25 to support high-recall lexical retrieval, particularly important for vehicle-specific terminology such as component codes, torque specifications, and fuse labels.

3) Dense Embedding Generation: A sentence-transformer encoder produces dense embeddings for:

- text segments,
- table entries (converted to sentence-like form),
- page-level images after captioning or embedding extraction.

The embeddings feed into FAISS for efficient vector search.

4) FAISS Index Construction: The FAISS index accelerates similarity search across tens of thousands of embeddings. The index structure is optimized for:

- sub-millisecond query performance,
- mixed embedding types (textual + visual),
- support for incremental updates without full rebuild.

5) Table Parsing: Tables containing torque values, fluid capacities, tolerances, and wiring diagrams are parsed into structured representations. Each entry is tokenized and encoded as both text and key–value pairs to improve retrieval robustness.

C. Voice Processing and Streaming

ASR is implemented using a local whisper-compatible transcription engine. Incoming voice data is chunked and transcribed incrementally, allowing the interface to submit queries before the user completes full sentences. The output passes through:

- punctuation restoration,
- noise filtering,
- semantic normalization.

Streaming LLM output is delivered back to the user interface through an HTTP chunked-response channel. Each token is appended to an on-screen message bubble, reducing cognitive load during driving by enabling early consumption of partial answers.

D. Telemetry Integration

Telemetry arrives through a dedicated interface that processes CAN-bus signals in real time. The system aggregates the following measurements:

- vehicle speed,
- steering wheel angle,
- throttle and brake positions,
- lateral and longitudinal acceleration.

A federated learning classifier identifies the current driving style (*calm*, *normal*, *aggressive*). This model is trained locally using incremental updates derived from driving sessions, with no raw telemetry leaving the system.

Telemetry also informs:

- response verbosity,
- safety gating,
- prioritization of critical alerts.

E. Multimodal Retrieval Engine

The multimodal RAG engine combines BM25, FAISS, and LLM-based reranking to construct a context window for grounded generation.

1) Candidate Generation: Both BM25 and FAISS independently generate top-k candidate segments. Their intersection set captures lexical accuracy and semantic relevance simultaneously.

2) Cross-Modal Reranking: An LLM-based scoring model ranks candidates according to:

- textual alignment,
- visual similarity (when diagrams apply),
- matching structured attributes (e.g., tables).

3) Context Window Assembly: The reranked list is truncated according to a token budget. Images are embedded as references or metadata links, tables as formatted textual descriptions, and text segments as paragraphs.

F. Safety Enforcement Engine

The IPVA implements a safety enforcement policy inspired by UNECE R155/R156 principles, which are widely adopted in production vehicles for cybersecurity and software update governance. By aligning the agent's behavior with these regulations, the IPVA is designed to be coupled with off-the-shelf advanced driver-assistance systems (ADAS) as a higher-level, explainable interaction layer rather than as an independent safety controller. The policy defines:

- **when** explanations should be simplified,
- **when** responses should be deferred,
- **when** warnings should override user queries,
- **how** to ensure grounding and prevent hallucinations.

Telemetry and environmental context are continuously fused with the query intent to determine allowable interaction patterns.

Figure 2 illustrates the prototype interface developed for the offline Intelligent Personal Vehicular Agent. The left section allows the user to choose among the five top-selling vehicles in the Brazilian domestic market, enabling quick switching between different preprocessed manuals. Immediately below, real-time machine-learning classifiers display the inferred driving style (calm, normal, or aggressive), reflecting behavioral patterns extracted from local telemetry. Further down, the interface presents real-time CAN-bus sensor graphs, although only one signal is shown in this printed version. The right side of the interface contains the chatbot panel, where the user submits questions and receives grounded responses enriched with retrieved excerpts and images from the vehicle manual. In the example shown, the system also triggers a safety alert due to aggressive driving, and the generated answer is intentionally concise because the driving-style classifier labeled the current behavior as aggressive, activating the safety-aware response policy.

VI. Experimental Methodology

A comprehensive evaluation methodology is required to validate multimodal retrieval, safety-aware reasoning, user experience, and learning efficiency. This section outlines the planned experiments.

A. Dataset Construction

To evaluate the IPVA, datasets are constructed from:

- entire automotive manuals including text, tables, and diagrams;
- synthetic voice queries reflecting realistic driving conditions;
- telemetry samples collected across diverse driving profiles;
- annotated manual diagrams of mechanical components and system layouts.

Queries are grouped into categories:

- component identification,
- procedure lookup,
- safety warnings,
- troubleshooting,
- specification retrieval (e.g., torque, capacity).

B. Evaluation of Retrieval Performance

Retrieval accuracy will be measured using:

- Recall@k for BM25,
- Recall@k for FAISS,
- reranking precision,
- multimodal retrieval accuracy for images and tables.

Manually annotated relevance judgments will serve as ground truth.

C. Evaluation of Streaming Responsiveness

Token-level streaming is evaluated by:

- average time-to-first-token (TTFT),
- average chunk latency,
- end-to-end response time,
- perceived latency in user studies.

Benchmarking occurs under different CPU/GPU constraints.

D. Evaluation of Safety Policies

Safety-aware action gating is tested in driving simulation scenarios with:

- calm cruising,
- aggressive maneuvering,
- emergency braking,
- lane-change episodes,
- distracted driving simulations.

Metrics include:

- inappropriate response suppression rate,
- timely warning issuance rate,
- distraction-minimizing behavior.

E. Evaluation of Driving-Style Classifier

The federated classifier is evaluated using:

- classification accuracy,
- personalization improvement over time,
- privacy-preservation metrics,
- convergence rate under non-IID data.

F. User Experience Evaluation

A user study will assess:

- clarity and helpfulness of responses,
- perceived cognitive load,
- satisfaction with timing and streaming,
- trust in safety-aware modifications.

Participants interact with the IPVA during simulated or controlled driving sessions.

VII. Planned Benchmarking Suite

To provide systematic comparison across IPVA variants and baselines, a custom benchmark suite will be developed:

A. Multimodal Retrieval Benchmark

Combines textual, visual, and structured queries reflecting real manuals:

- lookup of specifications via table queries,
- diagram identification through image-to-text matching,
- troubleshooting chains requiring reasoning across multiple manual sections.

B. Safety-Oriented Interaction Benchmark

Measures the stability of the Action layer under:

- varying telemetry,
- simulated hazards,
- rapid-query bursts.

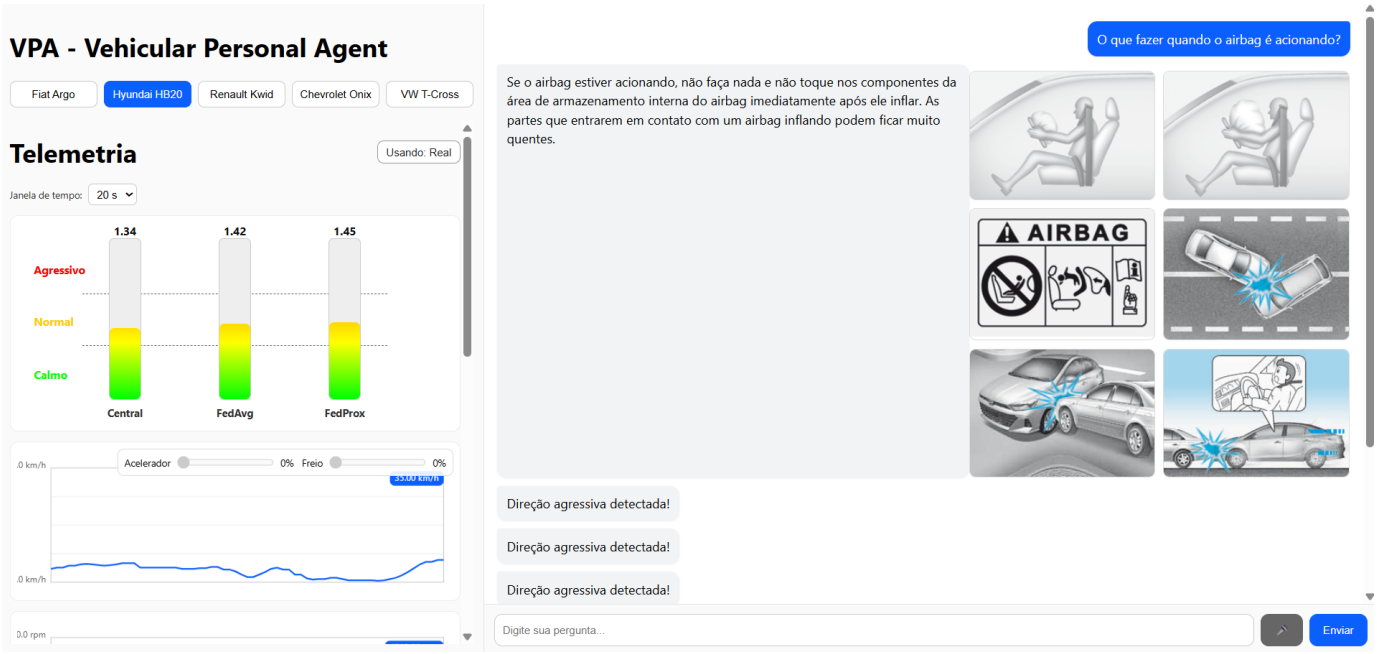

Fig. 2: Prototype interface of the offline Intelligent Personal Vehicular Agent.

C. Learning Efficiency Benchmark

Evaluates adaptation over time:

- improved retrieval precision,
- enhanced driving-style classification,
- personalized response structuring,
- reduced need for manual lookup of documentation in repeated scenarios.

VIII. Discussion

The Intelligent Personal Vehicular Agent (IPVA) presented in this work demonstrates how multimodal retrieval, contextual reasoning, safety-aware response generation, and continuous adaptation can be unified under a Perception–Reasoning–Action–Learning (PRAL) framework. The design choices reflect constraints inherent to automotive environments: limited attention availability from the user, safety priorities, real-time interaction requirements, on-device computational considerations, and heterogeneous data modalities (voice, telemetry, imagery, structured tables, and technical documentation).

A central advantage of the proposed architecture is the explicit grounding of LLM outputs in retrieved evidence from technical manuals. This substantially reduces hallucination likelihood and increases trustworthiness—an essential requirement for safety-critical domains. The hybrid retrieval mechanism further ensures high coverage and precision by combining lexical, semantic, and visual search strategies.

The PRAL model provides a systematic lens through which to design and analyze agent behavior. In the domain of vehicular assistants, this model highlights the differences between a system that merely answers queries and a cognitive agent capable of interpreting context, adapting behavior, executing safety-aware actions, and learning continuously. This distinction becomes essential as vehicles incorporate more automation and drivers increasingly expect intelligent, domain-aware assistance.

IX. Limitations

Despite promising results, the proposed IPVA architecture presents several limitations that must be acknowledged.

A. Model Size and Local Execution Constraints

Executing ASR, embedding models, visual encoders, and LLMs locally requires significant computational resources. While feasible on high-performance edge devices, mid-range hardware may face latency bottlenecks, especially under multimodal workloads.

B. Limited Visual Understanding

The current system uses page-level images and diagrams from manuals as its only visual source. However, diagnosing complex physical conditions requires either high-resolution real-world perception or rich OEM/ADAS-provided semantic signals. Extending the current pipeline to operate over live camera streams or structured vehicle-state data remains future work, and visual perception is therefore intentionally limited in the present prototype.

C. Dependency on Manual Quality

Technical manuals vary widely in formatting, notation, diagram clarity, and organizational consistency. Manuals with poor OCR structure or inconsistent labeling may introduce retrieval noise, reducing the quality of grounded responses.

D. Safety Validation Complexity

While the IPVA incorporates safety-aware response modulation, formal verification of behavior under all driving conditions poses challenges. Ensuring that no response induces distraction or contradicts real-time telemetry requires further validation under controlled tests.

E. Learning Risks

Although learning modules (e.g., federated learning, prompt adaptation, ASR fine-tuning) improve performance over time, they may also introduce drift or unintended behaviors. Without robust monitoring, personalization could degrade general robustness.

X. Future Work

The following research directions aim to enhance the capabilities and reliability of the Intelligent Personal Vehicular Agent.

A. Advanced Multimodal Perception

Future iterations should incorporate:

- tighter integration with OEM ADAS perception stacks and semantic vehicle-state signals (e.g., standardized status flags instead of purely visual dashboard symbols),
- fine-tuned vision models for automotive components when camera inputs are available,
- point-cloud processing for 3D perception where relevant,
- sensor fusion with radar, lidar, or other on-board sensors when exposed through standardized interfaces.

B. Contextual Memory and Long-Term Profiling

The current system stores only short-term interaction context and simplified user profiles. Future iterations should expand this capability toward a persistent vehicle–driver memory layer that captures salient events ("highlight moments") across the vehicle's operational history. Such a memory subsystem would support:

- episodic storage of relevant vehicle states, anomalies, or user actions;
- long-term preference modeling for individualized assistance;
- multi-driver differentiation and identity-aware behavior adaptation;
- hierarchical memory for multi-step procedures and long-running tasks;
- standardized interfaces for external tools (e.g., garage shops, OEM service tools) to access or contribute structured diagnostic insights.

This memory layer positions the IPVA not only as an in-cabin assistant but also as a basis for future automotive AI frameworks capable of coordinating information across multiple systems, users, and tools.

C. Interactive Troubleshooting Flows

Although reasoning templates encode procedures, future versions may support interactive, multi-step troubleshooting where the agent:

- asks clarifying questions,
- validates progress visually,
- guides the driver through diagnostic flows dynamically.

D. Safety Certification and Compliance

To deploy the IPVA in production vehicles, formal compliance frameworks must be developed. These frameworks should integrate:

- functional safety analysis,
- cybersecurity risk assessment,
- interpretability guarantees for LLM outputs,
- simulation-based validation.

E. Extended Learning Strategies

Future work may incorporate:

- reinforcement learning from real-world interactions and long-term driver trajectories;
- self-supervised multimodal alignment for manuals, diagrams, and—for future versions—real-time perception streams;
- cross-driver knowledge sharing using privacy-preserving aggregation to enable cooperative improvement across fleets;
- incremental domain adaptation for new vehicle models and updated service documentation;
- multi-agent communication protocols enabling IPVA instances to exchange safety alerts, road-condition updates, or contextual knowledge in vehicular social network scenarios;
- a unified agent-framework interface allowing external automotive tools (e.g., garage-shop diagnostics, fleet-management systems, OEM service platforms) to integrate with the IPVA cognitive and memory layers.

These directions transform the IPVA from a standalone assistant into a reusable software framework for future automotive multi-agent ecosystems.

XI. Conclusion

This work presents a comprehensive architecture for an Intelligent Personal Vehicular Agent (IPVA) anchored in the Perception–Reasoning–Action–Learning (PRAL) framework. By integrating multimodal perception from voice, telemetry, manual imagery, manuals, and structured tables into a unified Retrieval-Augmented Generation pipeline, and by being designed to interface with existing ADAS stacks, the proposed system bridges the gap between conventional digital assistants and adaptive vehicular agents.

The architecture supports token-level streaming for real-time interaction, safety-aware behavior modulation informed by telemetry, multimodal retrieval through BM25, dense embeddings, and FAISS indexing, and continuous improvement through federated learning and adaptive prompting.

The IPVA thus represents a significant step toward vehicular AI systems capable of understanding complex contextual cues, grounding responses in authoritative documentation, adapting to driver behavior, and operating within safety-critical constraints, while complementing ADAS perception-and-control loops with explainable reasoning and learning. Continued work on perception capabilities, learning mechanisms, and safety verification will further advance the feasibility and robustness of vehicular cognitive agents in real-world environments.

Acknowledgement

We express our sincere gratitude to FUNDEP for their generous support, which was instrumental in the successful completion of this research. Their commitment to advancing scientific and technological innovation in Brazil provided essential resources for this Agent prototype. We deeply appreciate FUNDEP's dedication to fostering impactful projects that address real-world challenges, such as improving accessibility for vehicular manuals in Brazil's automotive sector.

References

[1] T. Medeiros, M. Medeiros, M. Azevedo, et al., "Analysis of Language-Model-Powered Chatbots for Query Resolution in PDF-Based Automotive Manuals," *Vehicles*, vol. 5, no. 4, pp. 1384–1399, 2023.

[2] F. Liu, Z. Kang, X. Han, "Optimizing RAG Techniques for Automotive Industry PDF Chatbots: A Case Study with Locally Deployed Ollama Models," unpublished.

[3] S. B. Islam, et al., "OPEN-RAG: Enhanced Retrieval-Augmented Reasoning with Open-Source Large Language Models," *arXiv preprint arXiv:2410.12374*, 2025.

[4] Y. Gao, et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint arXiv:2312.10997*, 2023.

[5] ChatDOC, "Document Processing Framework," 2024. [Online]. Available: https://chatdoc.com

[6] FAISS, "A Library for Efficient Similarity Search and Clustering of Dense Vectors," 2023. [Online]. Available: https://github.com/facebookresearch/faiss

[7] LangChain, "Framework for Adaptive Applications," 2024. [Online]. Available: https://www.langchain.dev

[8] Hugging Face, "Open Portuguese LLM Leaderboard," 2025. [Online]. Available: https://huggingface.co/spaces/open-llm-leaderboard

[9] Cognigy, "AI Agents for Automotive," 2023. [Online]. Available: https://www.cognigy.com/solutions/automotive

[10] Salesforce, "Agentic AI for Connected Cars," 2025. [Online]. Available: https://www.salesforce.com/news/stories/agentic-ai-for-connected-cars/

[11] J. Lin, A. K. Singh, Y. Zhang, "Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition," unpublished.

[12] T. Strohmann, D. Siemon, and S. Robra-Bissantz, "Designing Virtual In-vehicle Assistants: Design Guidelines for Creating a Convincing User Experience," *AIS Trans. on HCI*, vol. 11, no. 2, pp. 54–78, 2019.

[13] E. Lugano, "Virtual Assistants and Self-Driving Cars," in *Proc. IEEE Intelligent Vehicles Symposium (IV)*, pp. 1532–1537, 2017.

[14] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs," unpublished.

[15] J. Kirkpatrick, et al., "Overcoming catastrophic forgetting in neural networks," *Proc. of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[16] Rapid Innovation, "AI Agents in Automotive 2024 Ultimate Guide," 2024. [Online]. Available: https://www.rapid2023.io/post/ai-agents-for-the-automotive-industry