GRASP

A Scalable RAG-Enhanced Multi-LLM Architecture for Intention-Aware Autonomous Vehicle Decision-Making

Dr. Mohit Garg
Assistant Professor, School of Computing
National College of Ireland, Dublin

91%
Collision Reduction

0.945
Best Weighted Score

98.2%
Regulatory Compliance

9
LLMs Evaluated

Sudeep Aryan Gaddameedi, Mohit Garg, Eric Gyamfi, "A Scalable RAG-Enhanced Multi-LLM Architecture for Intention-Aware Autonomous Vehicle Decision-Making", in ATT workshop at AAMAS 2026, Cyprus, May 25-26, 2026.

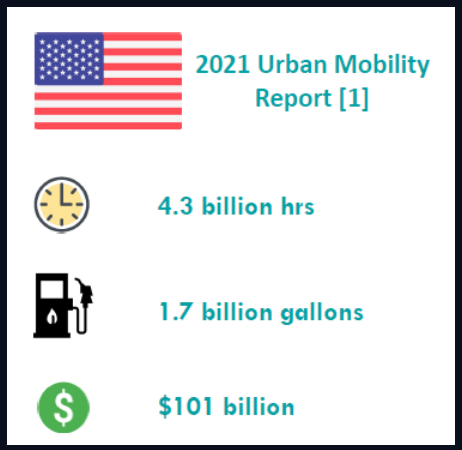
Motivation

❖ **1.35 million deaths worldwide every year***

- Most road accidents occur due to human perception errors and large reaction times.
- Human-driven vehicles produce stop-and-go waves due to their large reaction times and perception errors, resulting in reduced traffic safety and efficiency.

❖ **Autonomous Vehicles (AVs)**

- Can exploit information from multiple surrounding vehicles via V2V communication links
- Are widely expected to improve both traffic safety and efficiency.



*D. Schrank, B. Eisele, T. Lomax, and J. Bak, "2021 urban mobility report," Texas A&M Transportation Institute with cooperation from INRIX, Tech. Rep., 2021.
*World Health Organization, "Global status report on road safety," 2018.

The Problem: Why Do Self-Driving Cars Need Better AI?

- LLMs use static knowledge — traffic laws and rules change, but models don't update
- Vendor lock-in — current systems are tied to one AI provider (e.g., only GPT)
- Cars can't share intentions — they share position data but not WHY they're turning
- No standard architecture — every team builds from scratch, wasting effort

Our Solution: GRASP

A modular framework that gives self-driving cars updatable knowledge (RAG), lets them share intentions in plain language (A2A), and works with any AI model (MCP).

Related Work — Three Research Threads

GRASP integrates what each stream lacked individually

LLMs in AVs

GPT-Driver · LanguageMPC · DiLu
SACA · SafeDrive · AccidentGPT

▸ Research Gap

Gap: Static knowledge; tied to one provider; no semantic communication

RAG for AVs

LangChain · LlamaIndex
R^2AG · RAG-KG-IL · MiA-RAG

▸ Research Gap

Gap: General-purpose — not designed for traffic laws, HD maps, or AV scenarios

Multi-Agent Comms

MAPPO · ISDT · Q-FAIR-MIX
SAE J2735 · ETSI ITS · C-V2X

▸ Research Gap

Gap: V2X shares telemetry only; MARL needs 2–5M training steps; opaque policies

GRASP Fills All Gaps

RAG + A2A + MCP
First unified framework

▸ Research Gap

Dynamic knowledge · Semantic intent sharing · Vendor-agnostic deployment

No prior work integrates RAG knowledge grounding + semantic A2A + MCP standardization in one AV architecture

GRASP Architecture — Three Pillars

RAG — Knowledge Layer

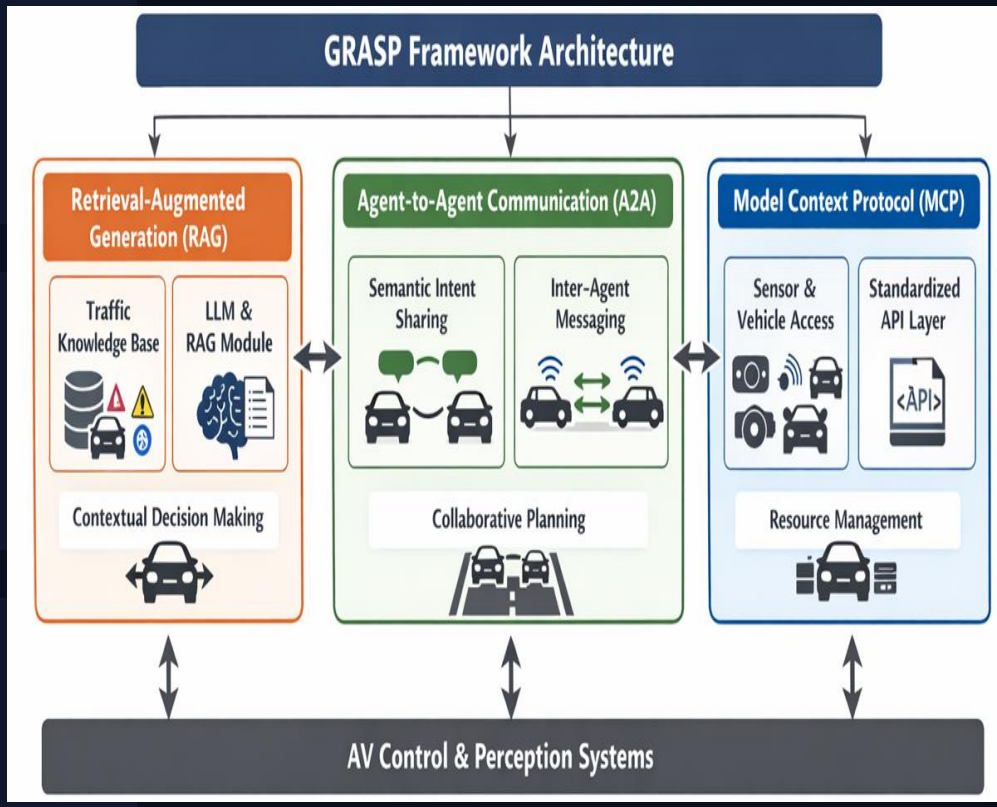
Retrieves traffic laws, HD maps & past scenarios from a vector database in real-time (~45ms)

A2A — Communication Layer

Vehicles share planned actions, reasoning & confidence in natural language

MCP — Integration Layer

Standardized interface to swap AI models (GPT-5, Claude 4.5, Gemini 3) without code changes



$a^*_t = \text{argmax}\ P(a \mid K_t, \Gamma_t, \Phi_t; \theta) \quad | \quad \Sigma\, \Delta\tau_i \leq 200\text{ms}$

System Architecture (GRASP)

GRASP Decision-Making Pipeline — Full Architecture

Inputs → RAG-enhanced CoT reasoning → 3-layer safety validation → V2V action → DSPy optimisation | all within 200 ms

Optimized Weights ↲ feed back into Context Builder — every fleet trip improves the next decision

Sensors/Perception · Knowledge/MCP · Context & CoT · Safety · Action & V2V · Optimization · Optimized Weights

Pipeline Walk-Through — Layer by Layer

How each stage contributes to safe, adaptive, and cooperative driving

① INPUT LAYER

Sensors → Perception (15 ms): fuses LiDAR, camera, radar. Knowledge Bases → RAG Engine (45 ms): retrieves matching traffic laws & past scenarios from a vector store. Vehicle Resources → MCP Interface (10 ms): standardised tool access — any LLM can plug in without code changes.

② CONTEXT & CHAIN-OF-THOUGHT REASONING

Context Builder merges all three inputs into one rich prompt. Chain-of-Thought forces the LLM to reason step-by-step (e.g. 'speed limit is 30 km/h; adjacent vehicle is braking → I should decelerate now') — making every decision auditable and explainable.

③ SAFETY LAYER — 3-tier guard

Guardrails & Assertions check: (a) output schema, (b) RAG-verified legal compliance, (c) physical feasibility. VIOLATION → Retry / Self-Correct sends the error hint back to CoT for re-reasoning (loop). PASS → decision proceeds downstream.

④ ACTION & V2V BROADCAST

Action Decision selects the optimal manoeuvre. Intention Sharing Protocol broadcasts it in natural language — not raw telemetry — to nearby vehicles via V2V. Result: vehicles cooperate proactively, not reactively. This single change delivers 91 % collision reduction.

⑤ CONTINUOUS OPTIMIZATION LOOP

V2V logs feed a Metric Evaluator (safety, compliance, latency). DSPy Teleprompter (MIPRO / Bootstrap) automatically tunes prom pt weights based on fleet outcomes. Optimized Weights flow back into the Context Builder — every trip across the fleet makes the next trip safer.

How It Works — 7-Step Real-Time Pipeline

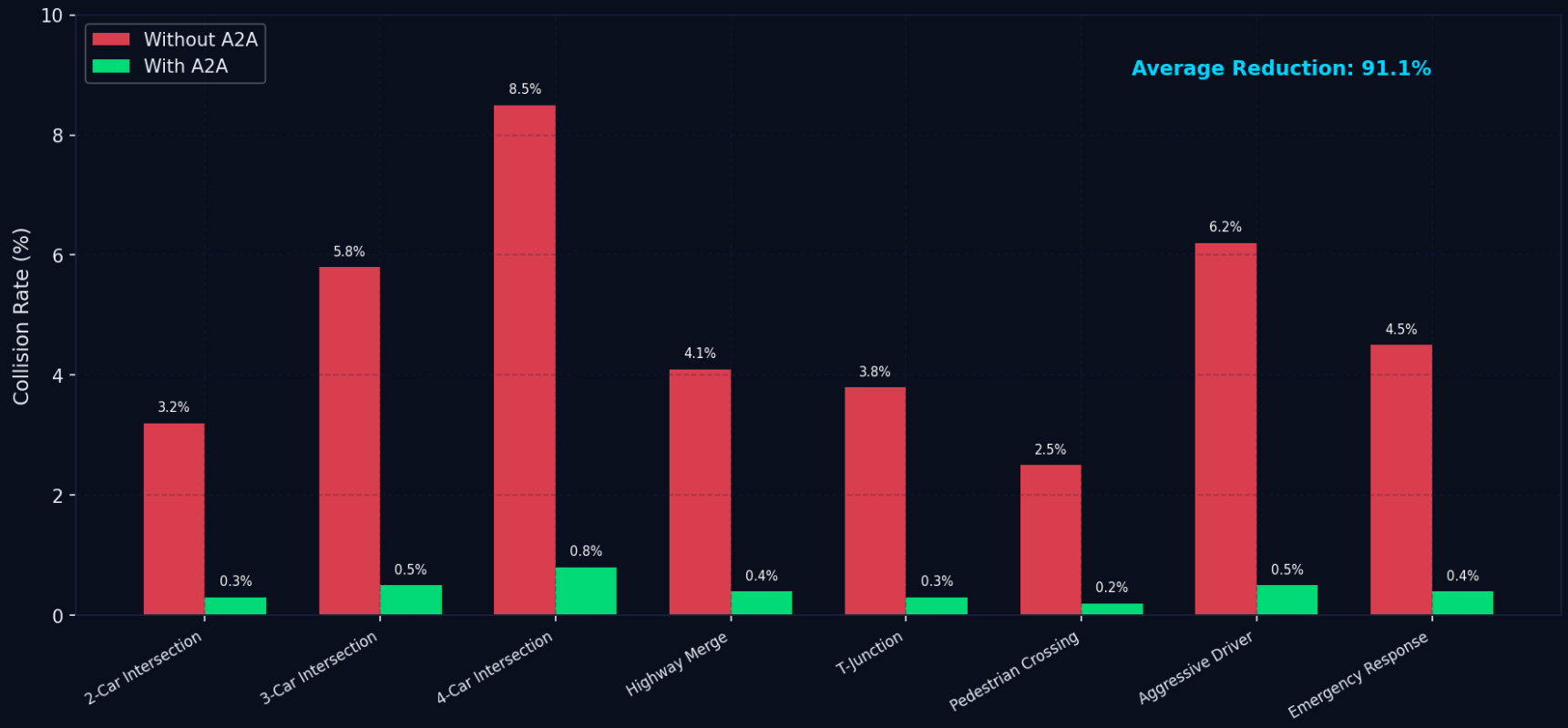
Every driving decision goes through these 7 steps in under 200ms

👁️	**Perception**	15ms	Fuse LiDAR, camera, radar data
📚	**RAG Retrieval**	45ms	Retrieve traffic laws & similar past scenarios
🔌	**MCP Abstraction**	10ms	Standardize sensor data for any LLM
🧠	**DSPy Decision**	80ms	LLM reasons with Chain-of-Thought + safety guardrails
📡	**A2A Sharing**	25ms	Broadcast intention to nearby vehicles
🛡️	**Safety Check**	15ms	3-layer validation: schema + laws + physics
🔄	**Optimization**	async	Continuous learning from fleet data

Key Result: Intention Sharing Reduces Collisions by 91%

Core hypothesis validated — semantic communication is fundamentally better than raw telemetry

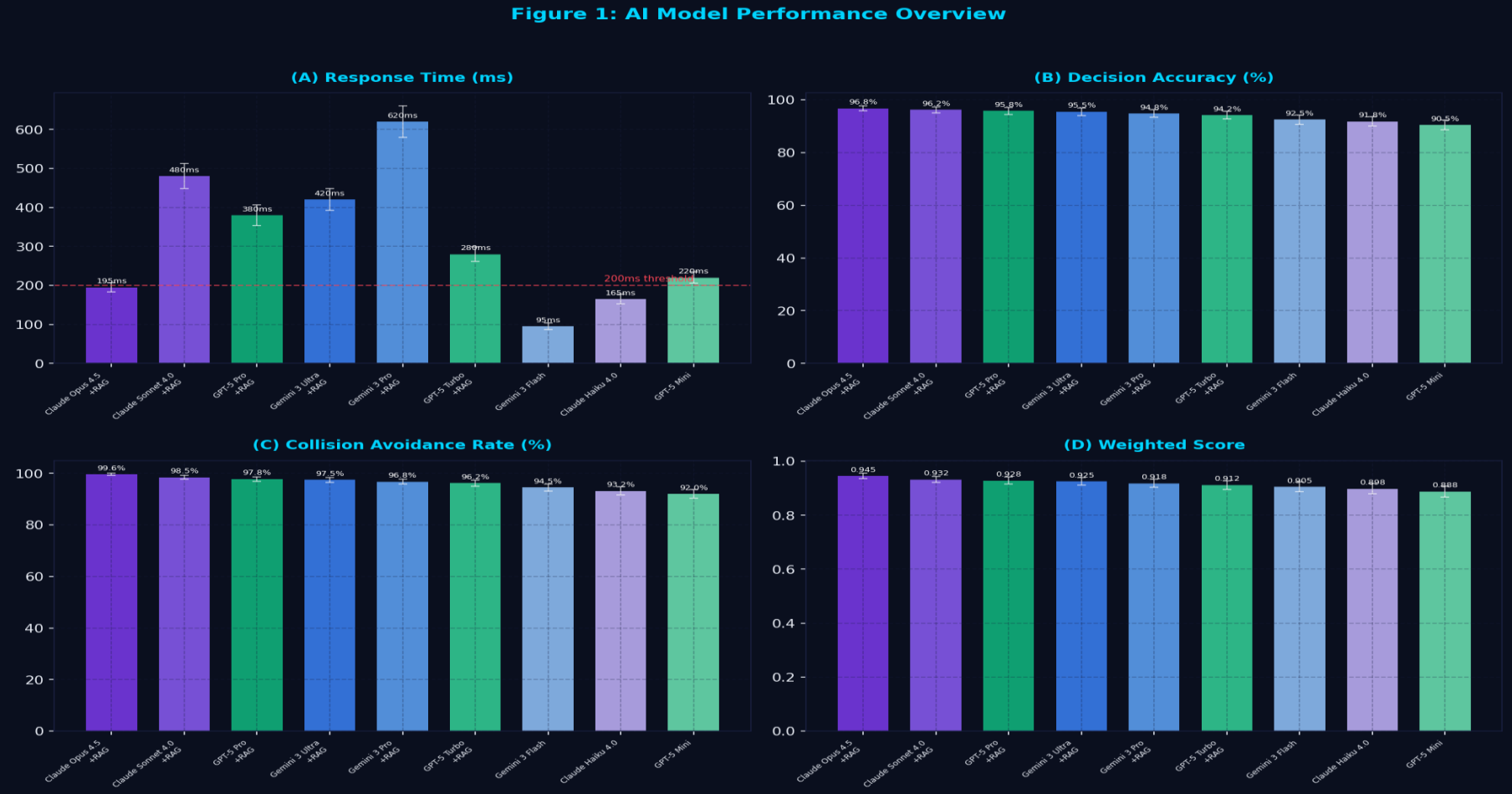
Figure 3: Collision Rate — With vs Without Intention Sharing (A2A)



Without A2A: vehicles guess each other's plans → collisions | With A2A: vehicles share plans in natural language → proactive coordination

First Benchmark: GPT-5 vs Claude 4.5 vs Gemini 3

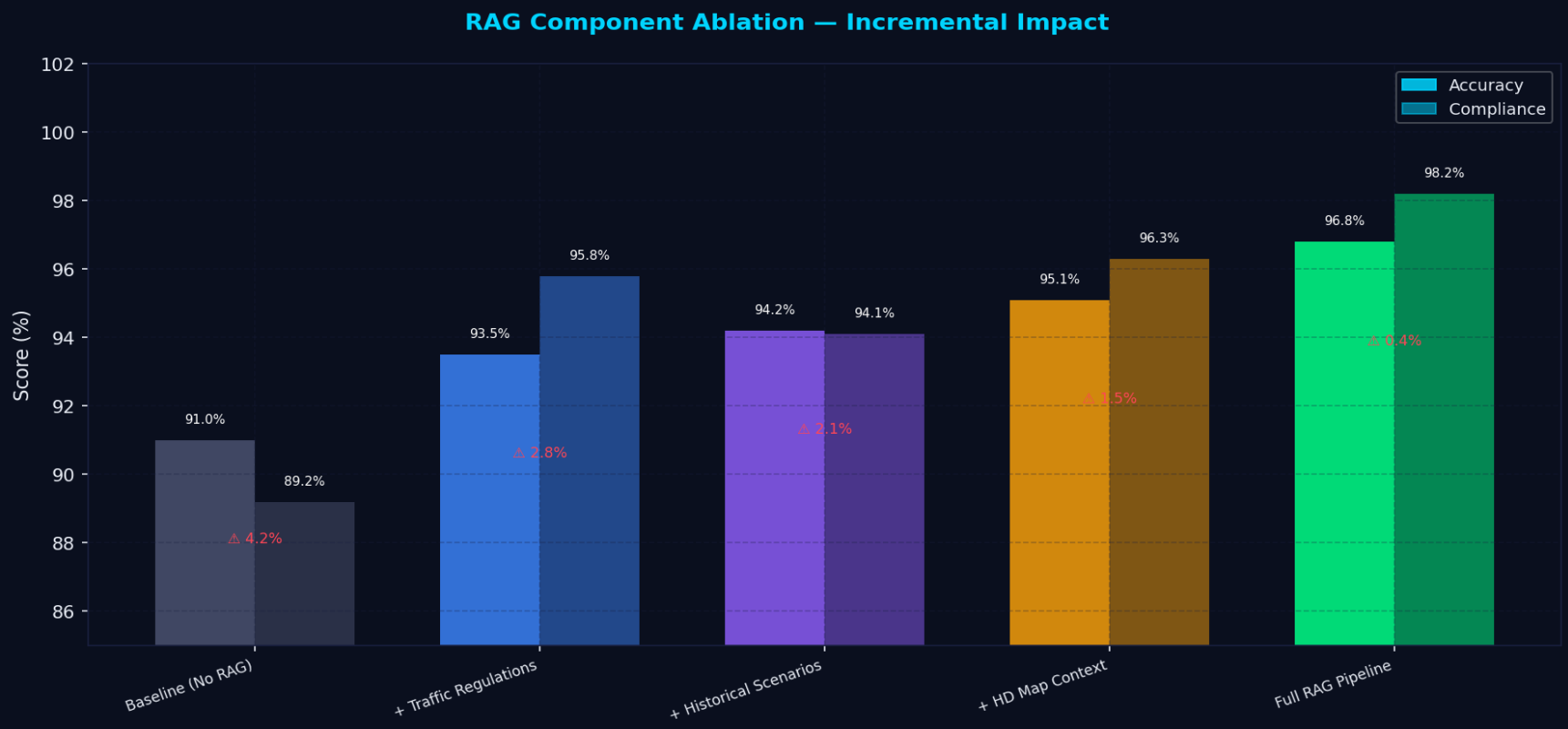
9 models evaluated across 59 scenarios and 2,160 trials



🏆 **Best: Claude Opus 4.5 (0.945)** ⚡ **Fastest: Gemini 3 Flash (95ms)** 💰 **Best Value: GPT-5 Mini ($0.18/1K)**

RAG Ablation: Each Component Adds Value

Collision rate drops from 4.2% → 0.4% with full pipeline (only +45ms latency)



Config	Accuracy	Collision	Latency
No RAG	91.0%	4.2%	150ms
+ Traffic Laws	93.5%	2.8%	168ms
+ History	94.2%	2.1%	175ms
+ HD Maps	95.1%	1.5%	182ms
Full Pipeline	96.8%	0.4%	195ms

Deep Dive: Complexity, Latency & Cost Analysis

RAG-enhanced models stay robust; adaptive selection cuts cost 37%

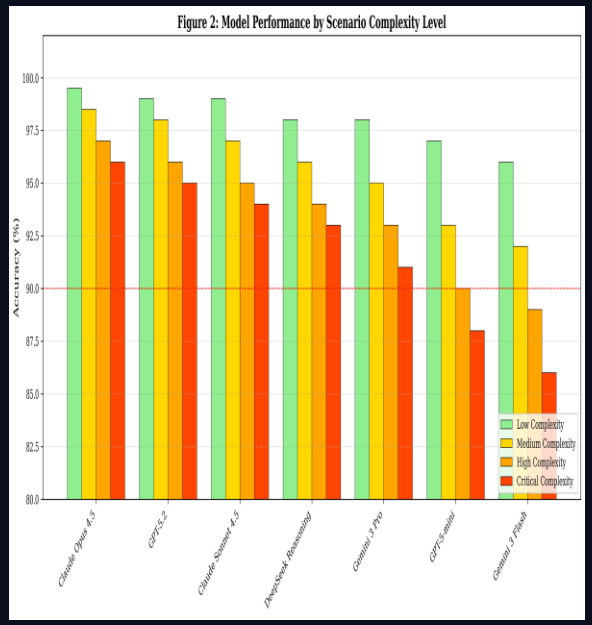


Fig 2: Accuracy vs. Scenario Complexity

RAG models maintain >95% in Critical scenarios; non-RAG degrade steeply

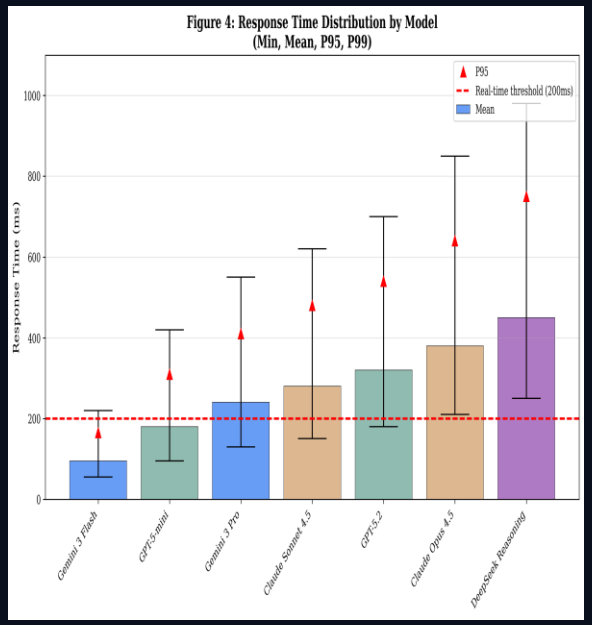


Fig 4: Response Time Distribution

Tail latency (P95, P99) critical for safety; Gemini Flash stays under 200ms

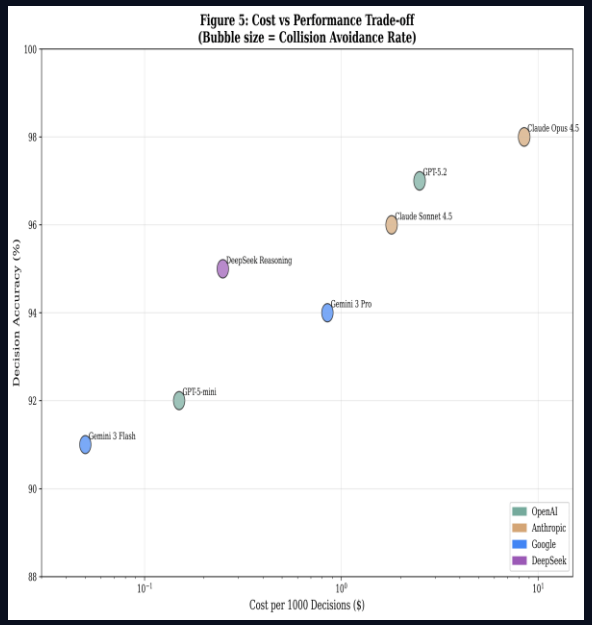


Fig 5: Cost vs. Performance Tradeoff

GPT-5 Mini: best value ($0.18/1K); Claude Opus: best safety ($3.45/1K)

Adaptive model selection: fast models for simple decisions, premium for critical intersections → 37% cost reduction with 96.2% accuracy

Key Contributions & Future Work

1. GRASP Framework — first to integrate RAG + A2A + MCP for autonomous vehicles
2. First benchmark of GPT-5, Claude 4.5, Gemini 3 in safety-critical AV scenarios
3. MCP enables 92% code reuse, reduces integration from weeks → hours
4. Hybrid safety: DSPy guardrails + RAG compliance + deterministic fallback
5. 91% collision reduction validated across 59 scenarios, 2,160 trials

Future Research:

- Multi-modal RAG (vision + language)
- Federated learning for privacy-preserving knowledge sharing
- Edge-optimized models for sub-50ms latency
- Formal verification for ISO 26262 ASIL-D certification

Thank You!

Questions & Discussion

References

- [1] D. Schrank, B. Eisele, and T. Lomax, “Urban mobility report,” The Texas A&M Transportation Institute with cooperation from INRIX, Tech. Rep., 2021.
- [2] World Health Organization, “Global status report on road safety,” 2018.