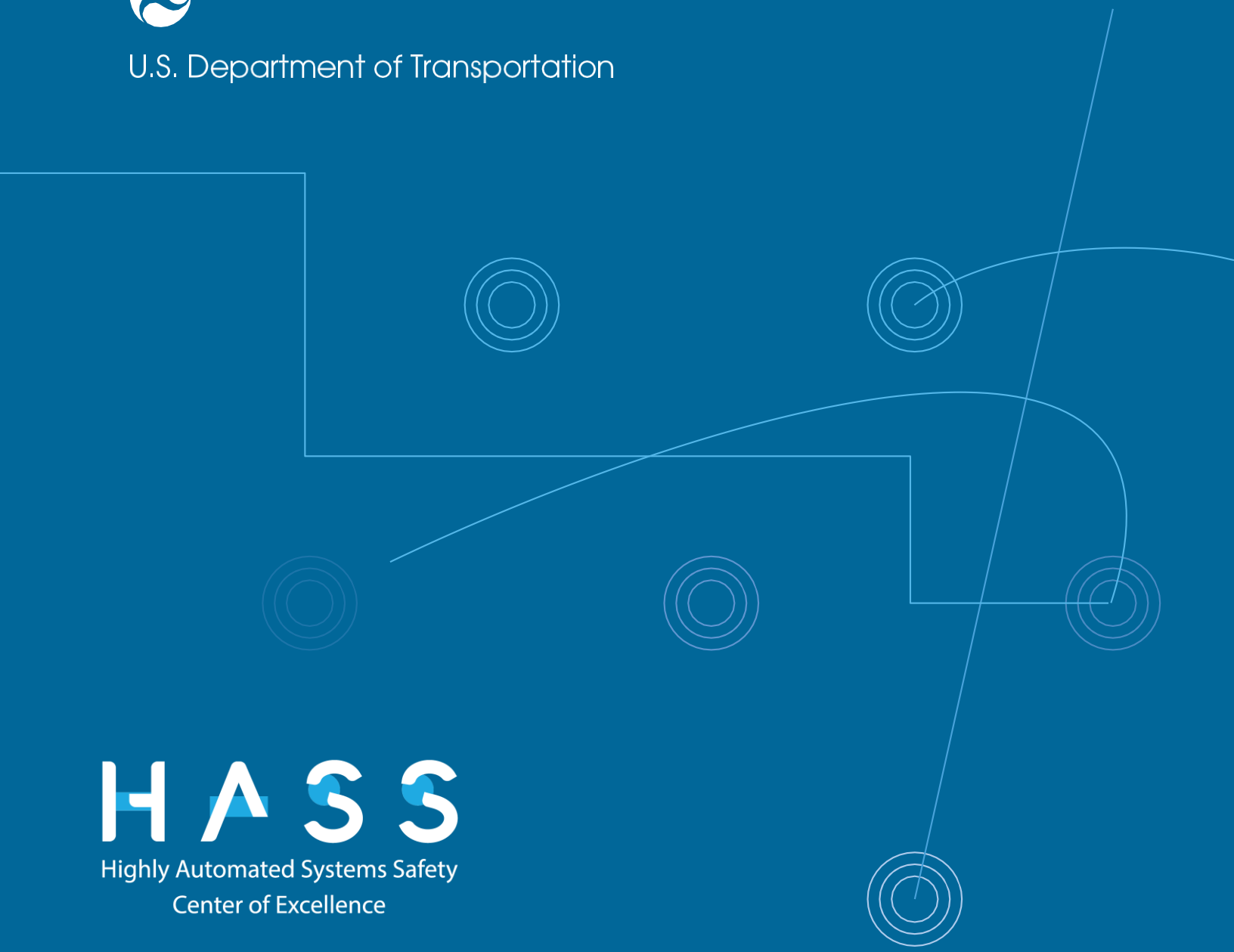
Understanding Safety Challenges of Vehicles Equipped with Automated Driving Systems (ADS) – Analysis of ADS Disengagements

August 2024

Mubassira Khan, Daniel LeMaster, and Wassim G. Najm

U.S. Department of Transportation



HASS

Highly Automated Systems Safety Center of Excellence

Table of Contents

Executive Summary

This paper reports on a study undertaken by the United States Department of Transportation's (U.S. DOT's) Highly Automated Systems Safety Center of Excellence to devise an initial causal factor typology for automated driving system (ADS) disengagements based on California Department of Motor Vehicles (CA DMV) disengagement reports. This study is the first part of an effort that will define the crash problem of ADS-equipped vehicles to assess current limitations in automated multimodal surface transportation systems and identify opportunities for improving their safe deployment into roadway traffic. ADS describes a Level 3, 4, or 5 driving automation system of hardware and software that are collectively capable of performing the entire dynamic driving task on a sustained basis, regardless of whether it is limited to a specific operational design domain. Manufacturers testing ADS in California are required to submit annual reports to share how often their vehicles disengaged from autonomous mode during tests, whether due to technology failure or situations requiring the test driver to take manual control of the vehicle to operate safely. This study analyzed the 2022 and 2023 ADS disengagement reports, where the human driver initiated ADS disengagement at an average of 86 percent of all cases. In some of these cases, drivers disengaged the ADS out of an abundance of caution. Testers provided distinct descriptions of disengagements in only 4 percent of all 2022-2023 reports.

This study created a typology that comprises six ADS-related categories (localization, perception, prediction, planning, control, and system) and one non-ADS related category, including a total of 38 attributes that generally refer to functional insufficiencies of ADS rather than actual root causes given the limitations of disengagement descriptions. Prediction was the most dominant causal factor category, accounting for about 24 percent of all known disengagement reports, followed in a descending order in terms of their relative reported frequency by planning (21 percent), system (17 percent), perception (17 percent), control (11 percent), localization (9 percent), and non-ADS related factors (1 percent). The five most dominant causal factor attributes were incorrect behavior/trajectory estimation of other road users, inaccurate object detection, motion planning issue, mapping discrepancy, and hardware issue. The study also demonstrated the potential of artificial intelligence language models for consistent, efficient ADS disengagement categorization within the domain of automated transportation systems. The results from this test case have intrinsic value and anticipate the potential of language model applications across a wide range of U.S. DOT use cases.

This paper delineates the initial typology of causal factors from interpreting the distinct tester-provided description of facts causing the ADS disengagements as narrated in the CA DMV required report. The CA DMV program did not supply pre-defined causal factor categories and attributes to the participating entities, which resulted in a substantial variation in the provided details by each entity. The typology in this paper could serve as a template for consistent reporting among testing entities in future descriptions of ADS disengagements and crashes. Also, consistent reporting of driving-related information could be helpful by including descriptions of driving scenarios, roadway locations, intersections and traffic control devices, roadway conditions, environmental conditions, and traffic situations. The list of attributes could be expanded to include more details and specifics to the potential root cause of ADS disengagements based on additional information from future reports as well as further consideration of the driving tasks and their challenges.

Introduction

This paper presents the initial analysis and results of a study undertaken by the United States Department of Transportation's (U.S. DOT's) Highly Automated Systems Safety Center of Excellence (HASS COE)[1] to understand the safety challenges of integrating motor vehicles equipped with automated driving systems (ADS) into the surface transportation network. The mission of HASS COE is to ensure federal capacity to review, assess, and validate the safety of automated technologies comprehensively across modes while enabling cross-sector collaboration for a holistic approach to the safe integration of automation in transportation. This study seeks to define the crash problem of ADS-equipped vehicles, herein referred to as automated vehicles (AVs), to assess current limitations in automated multimodal surface transportation systems and identify opportunities for improving the safe deployment of AVs into roadway traffic. The analysis is based on empirical data collected from prototype and production AVs that have been deployed in many parts of the United States. This study will:

1. Devise an initial causal factor typology based on California Department of Motor Vehicles (CA DMV) ADS disengagement reports.[2]

2. Create typologies for common pre-crash scenarios and causal factors of crashes involving AVs using the CA DMV ADS crash data and the National Highway Traffic Safety Administration's (NHTSA's) Standing General Order (SGO) ADS crash reports.[3,4]

3. Quantify the frequency of occurrence and harm measures of AV pre-crash scenarios and concomitant causal factors.

4. Identify AV safety improvement opportunities by prioritizing pre-crash scenarios and associated causal factors.

5. Propose a general framework of problem definition for the safety assessment and validation of automated multimodal transportation technologies.

This paper reports on the analysis and results of the first objective listed above, regarding the initial typology of possible AV crash causal factors based on ADS disengagement reports. The analysis and results of the remaining four objectives will be published in separate technical papers.

The SAE J3016™ *Recommended Practice: Taxonomy and Definitions for Terms Related to Driving Automation*

1 https://www.transportation.gov/hasscoe/about

2 https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/disengagement-reports

3 https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports

4 https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting

Systems for On-Road Motor Vehicles, commonly referenced as the SAE Levels of Driving Automation™, defines six levels of driving automation from Level 0 (no driving automation) to Level 5 (full driving automation) in the context of motor vehicles and their operation on roadways.[5] The term ADS specifically describes a Level 3, 4, or 5 driving automation system of hardware and software that are collectively capable of performing the entire dynamic driving task (DDT) on a sustained basis, regardless of whether it is limited to a specific operational design domain (ODD) (i.e., locations in which the ADS has specifically been designed to function).

This paper provides an example of crash and causal factor typologies of human-driven vehicles (HDVs) based on past advanced driver assistance system (ADAS) research. This is followed by a literature review of recent studies examining potential ADS failures and reported disengagement causes. Further, this paper describes CA DMV ADS disengagement reports and delineates the methodology, tools, and results of this analysis. Finally, this paper concludes with remarks about the findings and recommendations for more detailed data.

5 https://www.sae.org/blog/sae-j3016-update

Crash Typologies of Human-Driven Vehicles

Typologies of common pre-crash scenarios and crash causal factors for HDVs contributed to the identification and definition of functional requirements for effective crash countermeasures, ADAS research and development including performance specifications and test procedures, and estimation of ADAS potential safety benefits. As part of the U.S. DOT Intelligent Transportation Systems program, NHTSA undertook a major research effort to facilitate the development and implementation of cost-effective technologies for improving crash avoidance [1]. The initial step of this NHTSA research effort produced an example typology of crash causal factors based on a case-by-case examination (i.e., clinical analysis) of a sample of 1,183 unsanitized crash reports, as shown in Table 1 [2]. Selected from NHTSA's 1991-1993 General Estimates System (GES) and Crashworthiness Data System national crash databases, these reports represented nine major target crashes: rear-end, backing, lane change, single vehicle roadway departure, opposite direction, straight crossing paths at signalized intersections, straight crossing paths at unsignalized intersections, left turn across path, and reduced visibility. The crash frequency of occurrence was weighted for severity so that these crashes might more closely approximate the national profile. As a result, driving task errors accounted for about 75 percent of all primary causes of target crashes: driver recognition errors (44%), driver decision errors (23%), and erratic actions (8%). Driver physiological impairment led to about 14 percent of target crashes. Low-friction road surface contributed to about 8 percent of target crashes, vehicle defects at approximately 3 percent, and almost negligible reduced visibility.

The causal factor typology in Table 1 formed a foundation for many follow-on activities into crash causation, using the same basic structure with varying frequencies of occurrence. Without a human driving the vehicle, ADS will perform the various driving task actions such as perception, planning, and vehicle control. Consequently, ADS will experience some error types in common with human drivers (i.e., recognition, decision, and control error types) and will likely introduce new ADS-specific error types. Also, like human drivers, ADS will be exposed to driving hazards on low-friction road surfaces (i.e., slippery conditions), under reduced visibility (e.g., fog), in adverse weather (e.g., heavy rain), and with degraded road features (e.g., missing lane markings). Moreover, ADS will suffer system failure and potential defects in vehicle foundational systems (e.g., braking or steering). This study will create a typology that encompasses human driver-like and ADS-specific causal factors.

Table 1. Example of a Crash Causal Typology for Human-Driven Vehicles

Driving Task Errors	Physiological Impairment	Vehicle Defects	Low-Friction Surface	Reduced Visibility
Recognition Errors • Inattention • Looked – Did Not See • Obstructed Vision **Decision Errors** • Tailgating / Unsafe Passing • Misjudged Gap / Velocity • Excessive Speed • Tried to Beat Signal / Other Vehicle **Erratic Actions** • Failure to Control Vehicle • Evasive Maneuver • Violation of Traffic Control Device • Deliberate Unsafe Driving Act	• Under the Influence • Drowsy / Asleep • Illness	• Tires • Engine • Steering • Brakes	• Wet • Snow • Ice	• Atmosphere • Glare

Table 2 presents a typology of 36 distinct pre-crash scenarios that represent the crash population of light vehicles based on data from NHTSA's 2011-2015 GES and Fatality Analysis Reporting System (FARS) national crash databases [3].[6] Pre-crash scenarios describe vehicle movements immediately prior to the crash and the critical event that made the crash imminent (i.e., something occurred that made the crash possible). The 36 pre-crash scenarios accounted for 24,534 (94%) fatal crashes and an estimated 5,020,000 (89%) of all police-reported crashes that involved at least one light vehicle, based respectively on the yearly average of the 2011-2015 FARS and GES crash databases. Maneuver in Table 2 refers to a vehicle passing, parking, turning, changing lanes, merging, or performing a successful corrective action to a previous critical event. Vehicle action includes vehicle maneuver in addition to vehicle decelerating, accelerating, or starting. It is likely that AVs will experience similar pre-crash scenarios as HDVs but with different frequency of occurrence and role (e.g., leading versus following in rear-end crash scenarios).

Table 2. Example of a Pre-Crash Scenario Typology for Human-Driven Vehicles

Control Loss & Road Departure	Vulnerable Road Users	Lane Change & Opposite Direction	Rear-End	Crossing Paths	Other
• Control Loss / Vehicle Action • Control Loss / No Vehicle Action • Road Departure / Maneuver • Road Departure / No Maneuver • Road Departure / Backing	• Pedestrian / Maneuver • Pedestrian / No Maneuver • Pedalcyclist / Maneuver • Pedalcyclist / No Maneuver • Animal / Maneuver • Animal / No Maneuver	• Turning / Same Direction • Lane Change / Same Direction • Drifting / Same Direction • Parking / Same Direction • Opposite Direction / Maneuver • Opposite Direction / No Maneuver	• Striking Maneuver • Lead Vehicle Accelerating • Lead Vehicle Slower • Lead Vehicle Decelerating • Lead Vehicle Stopped	• Right Turn Into Path • Right Turn Across Path • Straight Crossing Paths • Left Turn Across Path / Lateral Direction • Left Turn Into Path • Left Turn Across Path / Opposite Direction	• Vehicle Failure • Backing Into Vehicle • Avoidance / Maneuver • Avoidance / No Maneuver • Non-Collision / No Impact • Object / Maneuver • Object / No Maneuver • Rollovers, Hit & Run, etc.

6 Light vehicles include all passenger cars, vans, minivans, sport utility vehicles, and light pickup trucks with gross vehicle weight ratings less than or equal to 10,000 pounds.

Literature Review of Potential ADS Malfunctions and Disengagement Causes

This section summarizes the results from three relevant studies that addressed potential ADS failures and disengagement causes. The first study generated a theoretical list of potential ADS malfunctions and consequent vehicle-level hazards based on functional safety assessment of a generic Level 4 urban robotaxi. The other two studies derived causal categories and related issues of ADS disengagements based on the analysis of empirical data collected over different years from the CA DMV ADS disengagement reports.

Vehicle-Level Hazard Analysis of a Concept Level 4 ADS Urban Robotaxi

This study applied selected aspects of the functional safety concept phase, Part 3 of ISO 26262,[7] and corresponding parts of the safety of the intended functionality (SOTIF) process, ISO/PAS 21448,[8] to assess the safety of a generic design, Level 4 ADS urban robotaxi that operates in cities and their surrounding areas with a density of human structures [4]. This study adopted a unique approach to the hazard and safety analysis by considering the vehicle in its entirety as the subject of the analysis, rather than considering only a specific system. As a result, this study identified 42 potential vehicle-level hazards using the hazard and operability (HAZOP) and systems theoretic process analysis (STPA) methods. These 42 hazards were organized into a notional hazard structure that reflected three types of the overall driving act (strategic, tactical, and operational efforts), as described in SAE J3016, and two additional categories (physical hazards [e.g., fire] and control transition). The analysis of HAZOP's malfunctions and STPA's unsafe control actions (UCAs) derived the identified vehicle-level hazards.

The HAZOP analysis identified the potential malfunctions in different parts of the relevant vehicle systems, which were correlated to potential vehicle-level hazards. Relevant vehicle systems encompassed on-board environmental sensors, Level 4 ADS functions (fusion, localization, mapping, environmental model and self-perception, and path planning), and other vehicle systems.[9] The HAZOP analysis used seven malfunction guidewords and applied them to high-level functions for the relevant vehicle systems, based on SAE Recommended Practice J2980.[10] The STPA analysis identified UCAs and tied them to the identified hazards, by

7 ISO 26262-1:2018, Road Vehicles Functional Safety, https://www.iso.org/standard/68383.html

8 ISO 21448:2022, Road Vehicles Safety of the Intended Functionality, https://www.iso.org/standard/77490.html

9 Other vehicle systems include in-cabin detection and classification, propulsion, steering, braking, passenger interface, climate control, wiper and washer, exterior lighting and signaling, central locking and entry, seat belt detection and occupant restraint, power window and sunroof, interior lighting, horn, power seat, telematics, and roll stability control.

10 J2980_201804, - Considerations for ISO 26262 ASIL Hazard Classification, https://www.sae.org/standards/content/j2980_201804

applying six guidewords to relevant control actions for the identified vehicle systems. Note that this study treated each individual system as a "black box" and focused on the control commands issued between systems that included power window and sunroof system, central locking and entry system, braking system, emergency management system, passenger, Level 4 ADS (path planning subsystem), powertrain system, and roll stability control subsystem.

Table 3 lists the attributes of malfunctions and UCAs under five categories, which might contribute to potential vehicle-level hazards in Level 4 ADS urban robotaxis. Malfunctions of "dynamic/static objects" include the on-board environmental sensor not detecting, intermittently detecting, or stuck reporting same dynamic or static objects surrounding the AV. An example of "vehicle conspicuity" malfunction is the failure of ADS to determine the need to enhance vehicle conspicuity with signaling.

Table 3. Attributes of Malfunctions and Unsafe Control Actions in Level 4 ADS Urban Robotaxis

Fusion, Localization, & Mapping	On-board Environmental Sensor	Environmental Model & Self-Perception	Path Planning	Control
• Road-Level Vehicle Location • Lane-Level Vehicle Location • Vehicle Position • Sensor Data Fusion	• Dynamic / Static Objects • Weather Conditions	• Path of Surrounding Dynamic Objects • Environmental Context & Scene • In-Path Objects • Road Signage • Free Space	**Vehicle Maneuver** • Parking • Acceleration, Deceleration, Cruising, or Maintaining Speed • Following Lead Vehicle • Navigating Roundabout • Turning Left / Right / U • Lane Centering / Keeping, Lane Changing, Merging, or Overtaking • Avoidance Maneuver **Vehicle Conspicuity**	**Foundational Vehicle Controls** • Braking • Steering • Propulsion **Vehicle Maneuver** • Stopping Distance • Parking • Acceleration, Deceleration, …, or Lead Vehicle Following • Lane Centering / Keeping, Lane Changing, Merging, or Overtaking • Left / Right / U Turn • Avoidance Maneuver

Exploring the Who, What, and Why of ADS Disengagements

Table 4 shows the results of a study that categorized causes of reported ADS disengagements as perception discrepancy, planning discrepancy, control discrepancy, environmental conditions and other road users, and hardware and software discrepancy [5]. This study identified and quantified the initiator of disengaging the system (who), the cause of the disengagement (what and why), the maturity of the system, and the AV location of the transition. This study examined disengagements, crashes, and vehicle miles traveled (VMTs) of AVs by manufacturers testing their ADS capabilities in complex real-world environments, as part of CA DMV's Autonomous Vehicle Tester (AVT) Program. Vehicles restricted from testing on public roadways included trailers, motorcycles, vehicles with operating authority, vehicles with a gross weight of greater than 10,001 pounds, and hazardous vehicles. From September 2014 to November 2018, AVs traveled 3,669,472 miles, experienced 124 crashes, and disengaged 159,840 times, resulting in an average of 4.35 disengagements per 100 VMTs.

This study removed from the dataset the causes of disengagements that: were recorded as planned testing and validation of new features, occurred in a parking facility, were reported as indeterminable by the manufacturer, had an indeterminable initiator, and were caused by the vehicle being outside the ODD. In addition, about 147,000 disengagement reports by two testers were removed from the dataset due to lack of variation among human-initiated disengagements, as all were noted as "operator takeover" while the ADS-initiated disengagements were occurring due to a control, perception, planning, or hardware/software discrepancy. Consequently, this study analyzed a subset totaling 5,731 disengagements, showing that the human driver initiated disengagements in 75 percent of the records. There were two instances when the AV was disengaged by a remote operator for a planning anomaly and software/hardware discrepancy. Planning discrepancy accounted for 35 percent of all 5,731 disengagements, followed in a descending order by relative frequency: software and hardware discrepancy (26%), perception issue (21%), environmental or other road user (12%), and control discrepancy (7%).

Table 4. Breakdown of ADS Disengagement Causes by Five Categories

Perception	Planning	Control	Environment & Other Road Users	Software & Hardware
• Traffic Light Detection • Invalid Object • Delayed Perception • Perception Issue	• Unwanted Vehicle Maneuver • Vehicle Localization & Planning • Improper • Motion Planning • Planner Not Ready • Complete Lane Change	• Improper Acceleration or Deceleration • Hard Braking • Cruise Control • Steering Issue • Improper Gap • Irregularity in Controls	• Weather Conditions • Poor Lane Markers • Emergency Vehicle • Blocked Lane • Construction • Road Debris or Rough Surface • Other Road Users	• Communications • Stock Vehicle • Basic Vehicle Requirements • Hardware Discrepancy • Software Discrepancy • System Discrepancy • System Tuning • System Health & Readiness

Characterization and Mitigation of ADS Insufficiencies

The goal of this study was to formulate a generic architectural design pattern, compatible with existing methods and ADS, to improve the mitigation of system functional insufficiencies (FIs) that undermine passenger safety and lead to hazardous situations on the road [6]. FIs arise from insufficiencies of specifications and performance limitations in sensors, actuators, and algorithm implementations, including neural networks and probabilistic calculations. This study analyzed the 2021 CA DMV ADS disengagement reports, showing that disengagements were five times more often caused by FIs rather than by system faults (reports clearly claiming that a software, hardware, or other systematic fault had occurred). In addition, this study made a comprehensive list of FIs and their characteristics by analyzing over 10 hours of publicly available road test videos. As a result, this analysis identified insufficiency types in four major categories: world model, motion plan, traffic rule, and ODD.

This study focused on the SOTIF's output insufficiencies (OIs) to comprehensively characterize FIs in ADS, such as missed object or false object detection and incorrect predictions of trajectories. OIs can be easily attributed to the few major internal ADS functions, such as perception or path planning. This study performed statistical analysis of over 2,500 disengagement reports to learn about reasons for disengagements and classified the causes and frequencies of their occurrence. While the CA DMV reports provide a large amount

of data, the text reports are often ambiguous because different companies provide different levels of details in the description of facts causing disengagement. Therefore, video recordings of AVs driving on public roads with real diverse traffic were studied as well to obtain a more intuitive understanding of the disengagement causes and consequences. The distribution of disengagement causes was 69 percent insufficiencies, 14 percent faults, 9 percent unclear cause, and 8 percent out of scope. The "out of scope" category means the disengagements were correctly and automatically triggered by the ADS without leading to any hazardous situation.

Table 5 presents 16 distinct OI types, divided into four categories by color coding (world model, traffic rule, motion plan, and ODD) and grouped by the ADS module responsible for the OI (e.g., localization). Distribution patterns of OIs per OI category were similar in both the 2021 disengagement reports and the real-world video study of road tests. OIs related to the world model accounted for 50 percent of the CA DMV reports, followed in a descending order by relative frequency: motion plan (43%), traffic rule (6%), and ODD (1%).

Table 5. Breakdown of ADS Insufficiencies by Unique Categories and ADS Modules

Localization	Map	Perception	Prediction	Motion Planning	ODD Checker
• Wrong Ego-Vehicle Localization*	• Wrong Map*	• Missed Object* • Ghost Object* • Wrong Object Position, Orientation, or Dimension* • Wrong Object Classification* • Wrong Drivable Space Identification* • Wrong Traffic Sign, Light, Lane Marking, or Operator Recognition**	• Wrong Object Trajectory*	• Violation of Traffic Regulation (e.g., Right of Way)** • Counter-Intuitive Motion Plan† • Indeterminate Motion Plan† • Unsafe Planned Trajectory†	• Wrong Weather Classification†† • Wrong Road Classification†† • Wrong Traffic Classification††

*World Model **Traffic Rule †Motion Plan ††ODD

Description of ADS Disengagement Reports

AV manufacturers that are testing ADS in the CA DMV AVT Program are required to submit annual reports to share how often their vehicles disengaged from autonomous mode during tests (whether due to technology failure or situations requiring the test driver/operator to take manual control of the vehicle to operate safely). In compliance with this requirement, each testing entity had its own definition for what counts as reportable disengagement when the safety drivers took over. For instance, Waymo runs simulations of what would have happened had there not been an intervention, and disregards disengagements where nothing bad would have happened.[11] Generally, ADS disengagements could occur for the following reasons: naturally occurring situations requiring urgent attention by the safety driver; driver caution, judgement, or preference; courtesy to other road users; or ADS limitations or software errors.[12]

CA DMV lists and defines key terms to help the public learn more about AV technology and gain a better understanding of AV testing and deployment.[13] The disengagement reports contain the following information:

- Manufacturer, Permit Number, Date, and Vehicle Identification Number
- Vehicle Is Capable of Operating without a Driver: Yes or No
- Driver Present: Yes or No
- Disengagement Initiated by: AV System, Test Driver, Remote Operator, or Passenger
- Disengagement Location: Freeway, Interstate, Expressway, Highway, Street, or Parking Facility.[14] Interstate and Expressway locations were added to the 2023 data list.
- Description of Facts Causing Disengagement (i.e., narrative)

The CA DMV AVT program does not provide pre-defined categories for the manufacturers to enter on the submitted disengagement reports. Instead, each disengagement report has a field for manufacturers to describe the facts causing a disengagement by using a phrase, sentence, or sentences (last bullet in above list). There is substantial variation in the details provided by each manufacturer due to the lack of predefined

11 https://www.forbes.com/sites/bradtempleton/2021/02/09/california-robocar-disengagement-reports-reveal-about-tesla-autox-apple-others/)

12 https://medium.com/cruise/the-disengagement-myth-1b5cbdf8e239

13 https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-definitions

14 Highway refers to a major roadway between towns or cities, intersecting with side streets and private driveways that provide motorists with continual entry points. An interstate is a highway serving two or more states. Like highways, a freeway provides a faster, more-direct route between destinations but does not intersect with other streets or is lined by private business and homes. An expressway is a divided highway with partial control of access.

standards. The apparent variation among manufacturers and years indicates the need to group disengagement causes into broader categories [6]. This study analyzed the 2022 and 2023 ADS disengagement datasets. Table 6 provides the statistics of these two datasets by different descriptors. The human driver initiated the disengagement of ADS in about 89 and 84 percent of all cases in 2022 and 2023, respectively. In many of these cases, safety drivers disengaged the ADS out of an abundance of caution since they were instructed to disengage if they had any doubts to assure safety. Disengagement occurred on a street location in about 60 and 24 percent of the cases in 2022 and 2023, respectively—remarkably different between the two years. Distinct descriptions of disengagements provided by the testers comprised a small ratio of 3 percent of the total 2022 reports and 5 percent of the total 2023 reports.

Table 6. Breakdown of 2022-2023 CA DMV ADS Disengagement Reports

Descriptor	Year	
	2022	2023
Total Reports	8,216	6,562
Disengagement Initiated by		
AV / ADS	937	1,087
Driver	7,279	5,495
Disengagement Location		
Freeway	2,455	2,150
Interstate	-	1,364
Expressway	-	13
Highway	200	1,348
Street	3,908	1,580
Parking Facility	-	2
Distinct Descriptions	238	304
ADS Manufacturers	23	18

In 2022, Apple Inc. submitted about 73 percent of the total 8,216 disengagement reports. Figure 1 provides the percent distribution of the ADS disengagement reports and the 238 distinct descriptions among the 23 testing entities submitted in 2022. Similarly, in 2023, Apple Inc. submitted the most ADS disengagement reports totaling 3,194 or about 49 percent of all 6,562 reports. Figure 2 provides the percent distribution of the ADS disengagement reports and the 304 distinct descriptions submitted in 2023 by 18 ADS testers. A new entrant in 2023, Bosch, provided the most with 137 or about 45 percent out of 304 distinct descriptions of ADS disengagements.

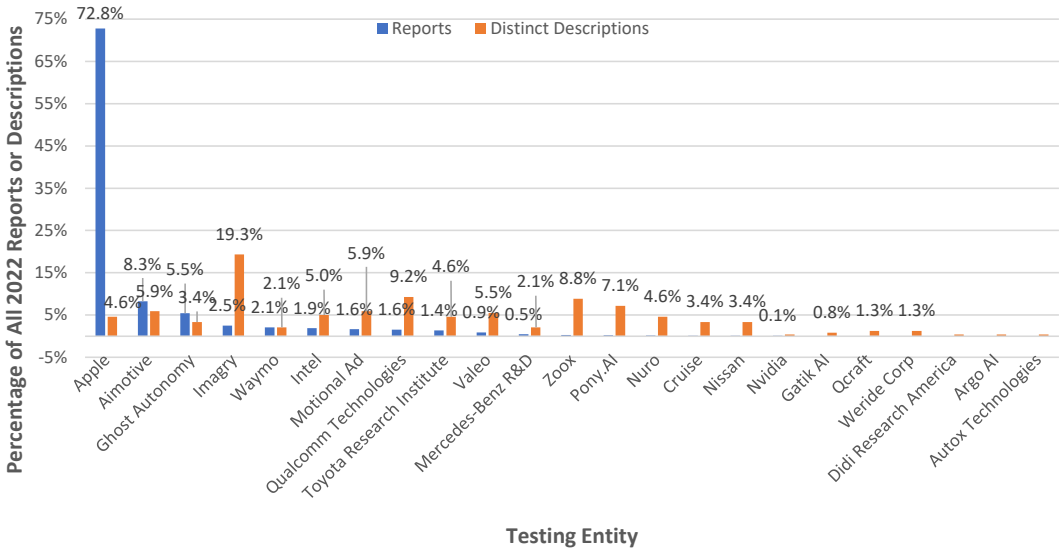


Figure 1. Distribution of 2022 ADS Disengagement Reports and Distinct Descriptions by Testing Entity

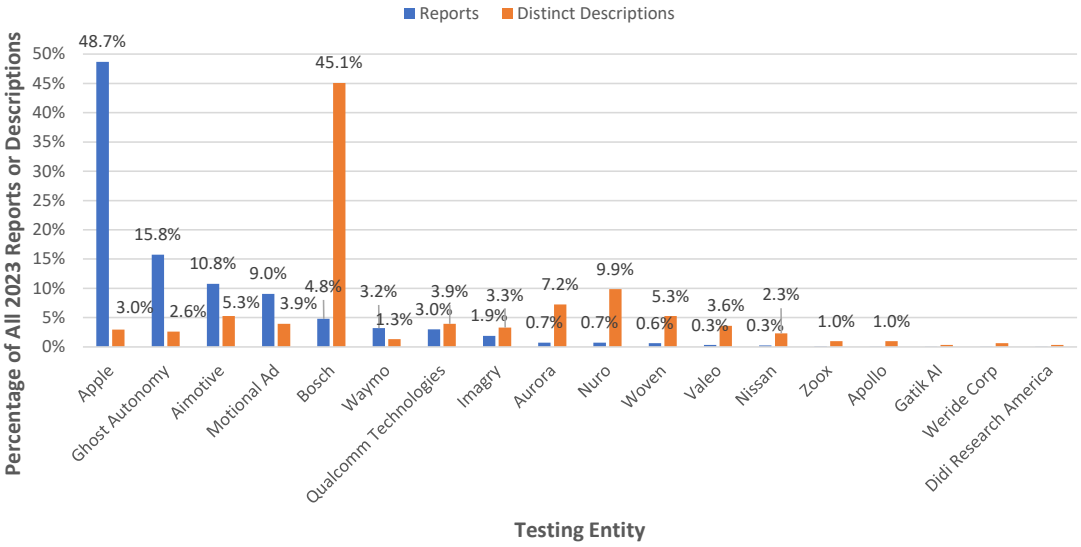


Figure 2. Distribution of 2023 ADS Disengagement Reports and Distinct Descriptions by Testing Entity

Disengagement Data Analysis

This study adopted an initial typology of causal factor categories for ADS disengagements that correspond to the categories of the HDV's crash causal factors. As noted in Table 1, crash causal factors consist of driver-related factors, including driving task errors and driver physiological impairment, and non-driver-related factors comprising vehicle defects, low-friction surface, and reduced visibility. Similarly, this initial typology consists of ADS-related and non-ADS-related factors. ADS-related factors are based on ADS functionality that performs the following actions [7]:

1. *Localization:* Determine location.
2. *Perception:* Perceive relevant static and dynamic objects in proximity to the AV.
3. *Prediction:* Predict the future behavior of relevant objects.
4. *Planning:* Create a collision-free and lawful driving plan.
5. *Control:* Correctly execute and actuate the driving plan.
6. *Communication:* Communicate and interact with other (vulnerable) road users.
7. *System:* Determine if specified nominal performance is not achieved.

Table 7 presents the causal factor typology for ADS disengagement based on the examination of the individual 2022-2023 distinct descriptions from the CA DMV AVT program. This typology comprises six ADS-related categories and one non-ADS category, which includes a total of 38 attributes that generally refer to FIs and OIs rather than actual root causes given the limitations of disengagement descriptions. The six ADS-related categories correspond to basic ADS functions listed above, except for communication due to the lack of related information in the distinct descriptions. The seventh non-ADS category includes factors that pose a safety challenge to the perception, prediction, planning, and control of ADS functionality, such as another road user violating the traffic rules or encroaching onto the lane of the AV without signaling or safe maneuvering space. In addition to the five attributes listed in Table 7, the non-ADS Factors category may include vehicle defects, adverse driving conditions, and other factors that would be mentioned in future ADS disengagement and crash reports.

Table 7. Typology of ADS Disengagement Causal Factors Based on 2022-2023 Distinct Descriptions

Localization	Perception	Prediction	Planning	Control	System	Non-ADS Factors
• Mapping (Discrepancy) • Lane Detection Issue • Navigation /	• Sensor Issue • Inaccurate Road / Road Lines Perception	• Incorrect Behavior / Trajectory Estimation of Other Road	• Motion Planning Issue • Failure to Yield / Give Priority to Another	• Lane Keeping Issue • Maneuver Issue • Maintaining Desired Path	• Hardware Issue • Software Issue • System Performance Issue	• Another Vehicle Making an Illegal Maneuver • Other Road

Localization	Perception	Prediction	Planning	Control	System	Non-ADS Factors
Localization Issue • Positioning Error (Discrepancy)	• Inaccurate Intersection Perception • Inaccurate Road / Road Lines Perception at Intersection • Inaccurate Traffic Light Detection at Intersection • Inaccurate Traffic Sign Detection at Intersection • Inaccurate Object Detection • Inaccurate Car Detection • Lane Detection Issue • Inaccurate Traffic Light or Stop Line Perception	Users	Road User • Traffic Situation • Inaccurate Path Planning • Incorrect Self Trajectory Plan • Lane Change Planning Issue	• Braking Issue • Speed Control • ODD Issue • Acceleration Issue • Steering Issue	• Sensor Input Delay	User Behaving Poorly • Obstacle in Path • Construction Zone • Unexpected Road Conditions

By comparison to the typology in Table 4 based on 2014-2018 reports, the non-ADS factors in this typology don't include weather conditions and emergency vehicles because the 2022-2023 descriptions did not clearly indicate such factors. Similarly, the ADS System (software and hardware) factor does not include communications, stock vehicle, basic vehicle requirements, and system tuning due to the lack of clear reference in the descriptions. The list of attributes in Table 7 will likely expand to include such missing factors as follow-on research looks into future ADS disengagement reports and AV crashes. It should be noted that Table 7 captures all other ADS-related attributes in Table 4 and Table 5. In comparison to the results of the theoretical analysis of the Level 4 ADS urban robotaxi in Table 3, Table 7 does not include specific location error details such as road- or lane-level vehicle location errors under the localization category, specific vehicle maneuvers such as turning or avoidance maneuver under the planning and control categories, or vehicle conspicuity.

This study aims to correlate information gleaned from the ADS disengagement reports to AV crash data. In addition to the coded AV location in ADS disengagement reports, the distinct descriptions provide some insight into the driving scenarios, atmospheric conditions (e.g., weather), specific roadway characteristics (e.g., intersection and ramp), and other driving circumstances. Table 8 delineates the driving scenarios that this study identified from a few narratives available in the 2022 CA DMV ADS disengagement reports. By comparison to Table 2 listing 36 scenarios, Table 8 shows a total of 12 distinct driving scenarios (light gray cells) along with specific information about variations in each scenario. Analysis of 2023 and future

disengagement reports will likely identify more driving scenarios that correspond to the 36 pre-crash scenarios in Table 2.

Table 8. Driving Scenarios of ADS Disengagements Based on 2022 Reports

Control Loss & Road Departure	Vulnerable Road Users	Lane Change & Opposite Direction	Rear-End	Crossing Paths	Other
Road Departure / No Maneuver • Unintended AV lane departure	*Pedestrian / No Maneuver* • Pedestrian walking alongside road	*Lane Change / Same Direction* • HDV Cut in • AV changing lanes – HDV fast approaching • Bus merging in front of AV • AV changing to lane with stopped / slow traffic without slowing down	*Lead Vehicle Slower* • HDV following AV too closely *Lead Vehicle Stopped* • HDV parked too close to lane / double parked	*Right Turn Into Path* • AV going straight with right of way *Right Turn Across Path* • Truck turning across multiple lanes *Straight Crossing Paths* • HDV running red light / stop sign • AV not stopping at intersection *Left Turn Into Path* • AV going straight with right of way *Left Turn Across Path / Opposite Direction* • AV going straight with right of way	*Backing Into Vehicle* • HDV backing into AV *Object / No Maneuver* • Blocked lane in construction zone

Approach to Quantifying the Occurrence of Causal Factor Categories and Attributes

This analysis quantified the frequency of occurrence of each causal factor category and each related attribute in distinct descriptions and in all reported descriptions using 2022 and 2023 CA DMV ADS disengagement reports. Figure 3 illustrates the classical machine learning approach to analyze and quantify the disengagement data, including:

- Identification of distinct descriptions from the ADS disengagement reports and creating the logic to track the frequency of these distinct descriptions.
- Manual examination of each distinct description from 2022 and 2023 reports to identify the categories and attributes (i.e., possible root causes, FIs, and OIs).
- Exploration of automated processing techniques to analyze the data in a more efficient and consistent manner, and their application to current and future disengagement reports, to predict causal factor of disengagement categories, using:
    - Classical machine learning (ML) models:
        - Naïve Bayes: a supervised ML algorithm used for classification tasks such as text classification.
        - Support Vector Machine (SVM): a ML approach used for classification and regression tasks. This study used SVM for classification.

- Generative artificial intelligence (AI) tools:
  - Large language models (LLMs) and small language models (SLMs)
  - Retrieval Augmented Generation (RAG) and fine tuning (FT)
  - With and without Chain-of-Thought (CoT) prompting
- 2022 distinct disengagement descriptions to train the ML and Generative AI models and 2023 distinct descriptions to test these models for identifying and quantifying the reported frequency of the seven causal factor categories in Table 7.

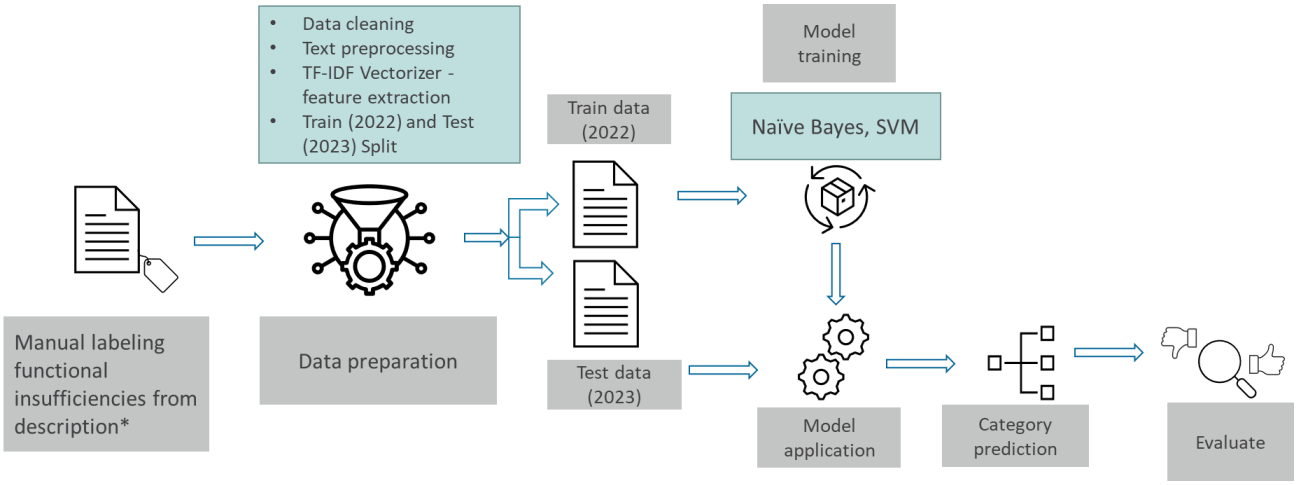


Figure 3. Analysis Approach and Workflow

Causal Factor Categories

This section provides the analysis results of 2022-2023 ADS disengagements for the seven causal factor categories listed in Table 7, including the counts/relative counts of each category using distinct ADS disengagement descriptions and all disengagement reports, accuracy assessment of applied ML and AI tools, and results comparison among four selected testing entities.

Disengagement Categorization Based on Manual Examination of Individual Cases

Figure 4 shows the distribution of all distinct descriptions of ADS disengagements by causal factor categories and compares their relative counts between 2022 and 2023. This figure also includes the relative counts of 15 descriptions with unknown causes and two descriptions of a driver mistake (unintended ADS disengagement by the driver). Planning was the most cited issue in 2022 at 29 percent of all distinct descriptions, which fell to the second most cited in 2023 at 18 percent. On the other hand, control was the fourth reported issue at 11 percent in 2022 but climbed to the first on the list in 2023 at 28 percent. The new 2023 entrant, Bosch, cited control issues in most of their distinct descriptions that accounted for almost half of all reported distinct descriptions.

Figure 5 presents the distribution of ADS disengagement reports by causal factor categories and compares their relative counts between 2022 and 2023. In 2022, the three most reported issues in descending order were planning (30%), perception (21%), and prediction (19%). By comparison, the three most reported issues in 2023 were prediction (30%), system (24%), and control (17%).

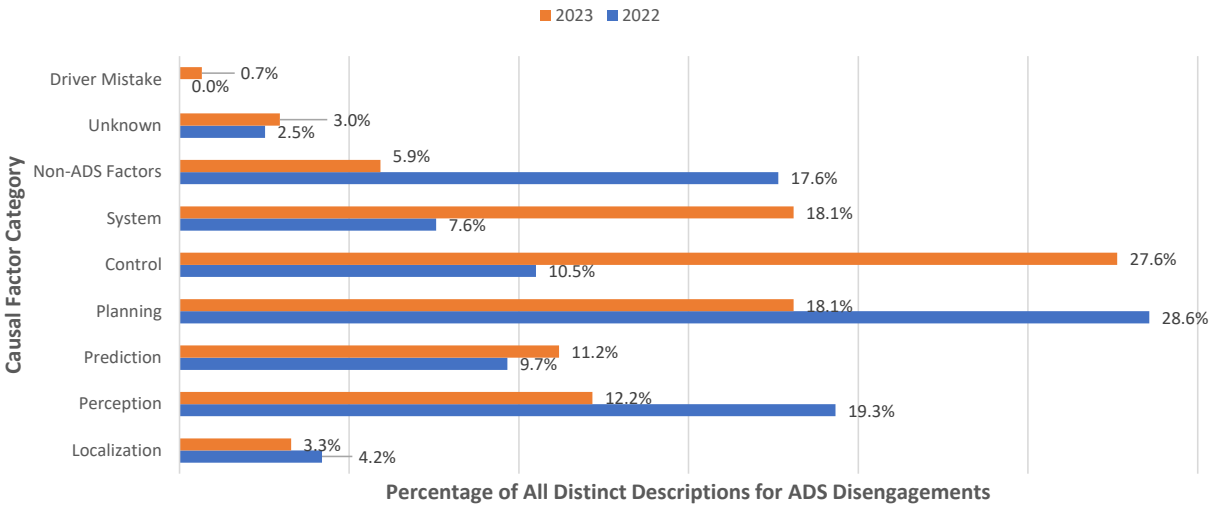


Figure 4. Distribution of Distinct Descriptions by Causal Factor Category for 2022-2023 ADS Disengagement Reports

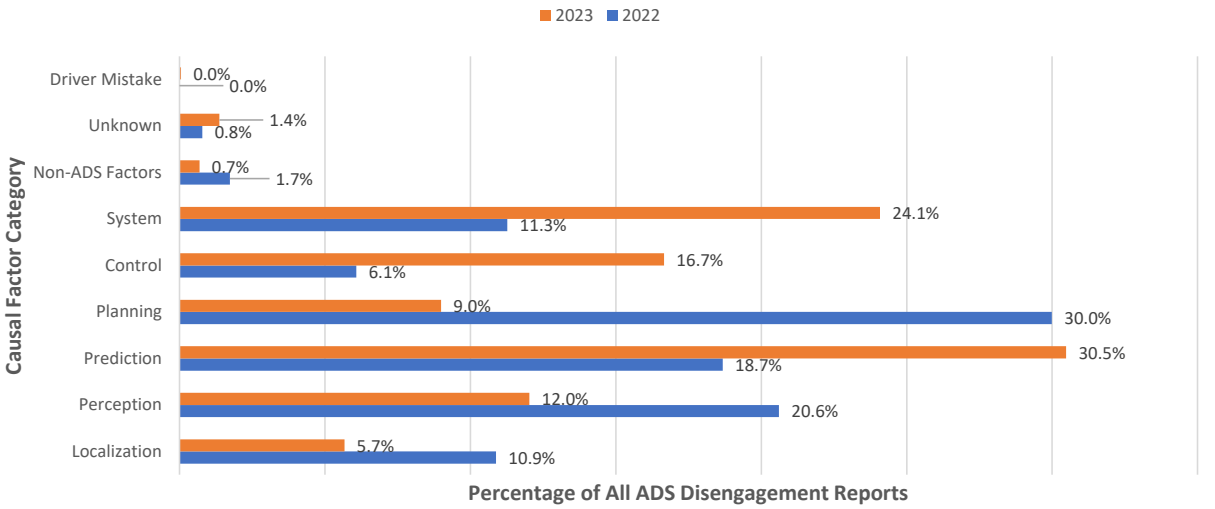


Figure 5. Distribution of 2022-2023 ADS Disengagement Reports Distinct by Causal Factor Category

The relative frequency of occurrence of each causal factor category noticeably changed between all 2022 and 2023 ADS disengagement reports as seen in Figure 5 due to various reasons, such as contribution of new entrants in 2023, withdrawal of some 2022 entities, and changes by remaining entities in exposure, system update, and description of facts causing ADS disengagement. To better understand differences in relative

frequency and to rank prominent causal factor categories based on 2022 and 2023 disengagement reports, this study selected five testing entities for further analysis. Figure 6 illustrates the results of this analysis, which facilitates the comparison of prominent causal factor categories within and across entities between 2022 and 2023. The selected entities contributed 85 and 72 percent to all 2022 and 2023 ADS disengagement reports, respectively. The following observations are noted from Figure 6 for each causal factor category that was reported the most by entities:

- *Localization:* Thirty percent of aiMotive reports in 2022, but none in 2023. Nine percent of Apple reports in 2023, down from 11 percent in 2022.
- *Perception:* Forty-five percent of Waymo reports in 2022, down to 25 percent in 2023. Eighty-five percent of aiMotive reports in 2023, up from 41 percent in 2022.
- *Prediction:* Fifty-two percent of Zoox reports in 2022 and 50 percent in 2023. Sixty percent of Apple reports in 2023, up from 25 percent in 2022.
- *Planning:* Thirty-two percent of Apple reports in 2022, down to 7.7 percent in 2023. 7.9 percent of aiMotive reports in 2023, down from 18 percent in 2022.
- *Control:* Fifty-four percent of Waymo reports in 2022 and 51 percent in 2023.
- *System:* Ninety percent of Motional Ad reports in 2022 and 95 percent in 2023.
- *Non-ADS Factors:* Twenty-nine percent of Zoox reports in 2022 and up to 50 percent in 2023.

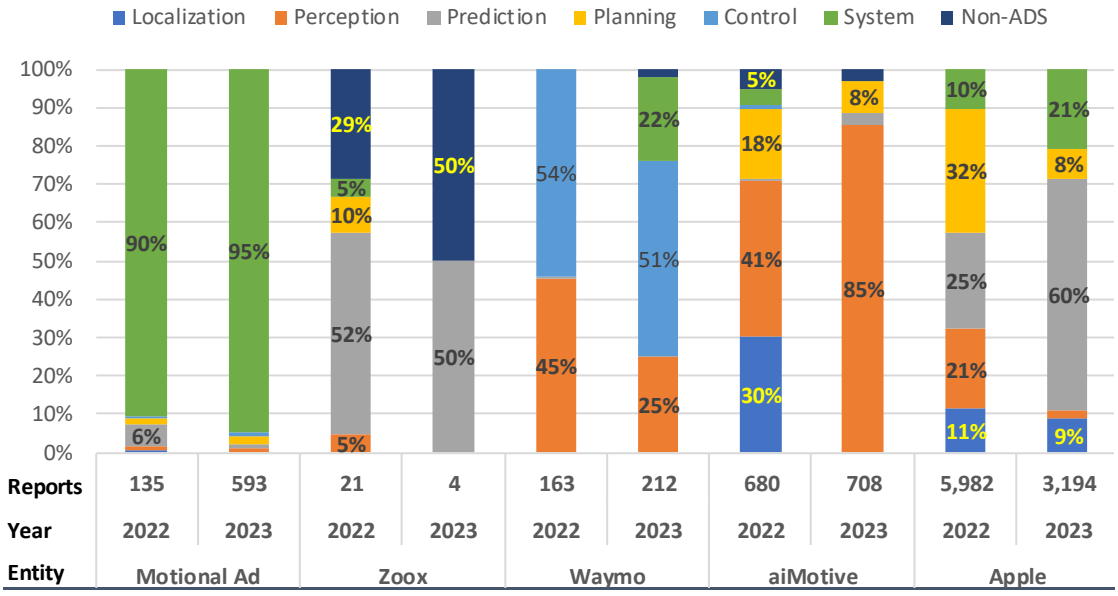


Figure 6. Distribution of Causal Factor Categories by Selected Entities Based on 2022-2023 ADS Disengagement Reports

Disengagement Categorization Using ML and Generative AI Language Models

In this section, the central hypothesis is that AI-based language models can surpass prior-generation ML

models in accuracy and efficiency for ADS disengagement categorization tasks. Furthermore, the research suggests that language models may achieve performance approaching human reviewers (where there will always be some disagreement) while offering significant cost and time savings. This capability could establish language models as a foundation for a wide range of applications involving text and narrative processing within the U.S. DOT. This goal is potentially within reach but, for now, this exploration emphasizes the importance of careful selection among model options when tackling various aspects of this challenge.

Experiment Design: Examining Model and Workflow Variations

This study's approach is to treat the disengagement data itself as the stage for an "experiment on the experiment." The primary objective is to understand how variations in model choices and workflows influence ADS disengagement characterization outcomes. Figure 7 shows the span of model choices investigated so far.

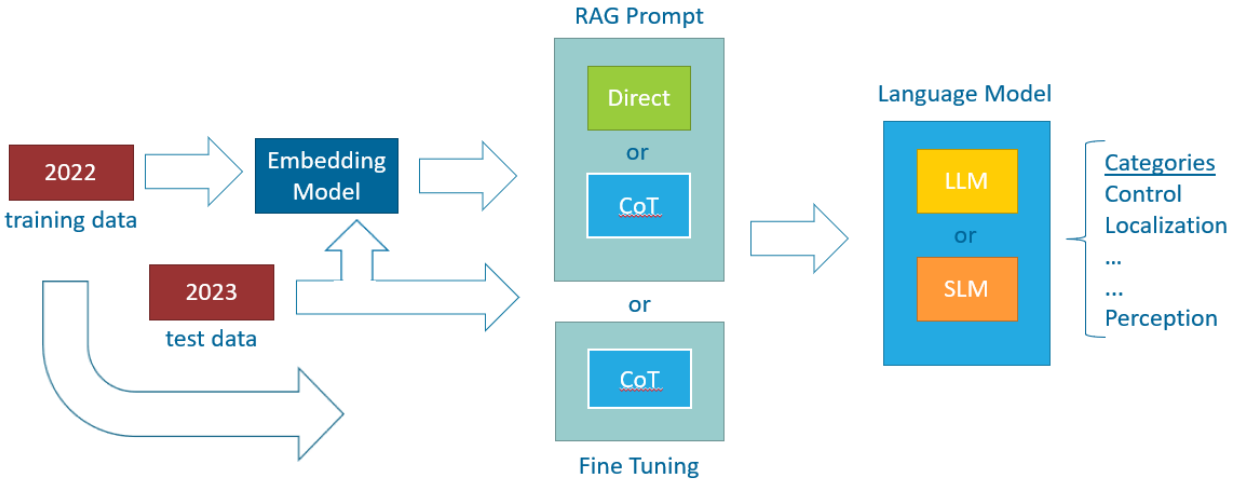


Figure 7. Language Model Workflow with All Variations Shown

Several potential combinations of these elements have been explored so far. Model size is the first parameter that we varied. This investigation includes two distinct classes of language models: an LLM (Gemini Pro) and an SLM (phi-2). While many other models exist at intermediate scales, phi-2 was chosen for its very small size and adequate benchmarking performance, along with ease of fine tuning.

For both models, this study explored various techniques for leveraging 2022 ADS disengagement examples to categorize ADS disengagements from 2023. The first technique, RAG, utilizes a different AI tool known as an embedding model as a preprocessing step to identify examples from 2022 that exhibit the most semantic similarity to the description being categorized in 2023. This approach goes beyond simple word pattern matching, focusing on the underlying meaning conveyed by the text. The identified examples are then used to construct a prompt that guides the language model under test in generating a prediction. We differentiate between two prompting methods used with RAG: CoT and direct categorization. CoT prompts provide examples with descriptions, root causes for disengagement, and the corresponding categories. Direct categorization prompts bypass the root cause step, directly linking descriptions to categories.

For the SLM, this study also attempted to improve performance through FT. In contrast to RAG, which provides the model with a pattern to follow without altering its underlying structure, FT involves making small adjustments to the SLM's neural network itself. The research investigates a technique called Quantized Low-Rank Adaptation (QLoRA) for fine tuning the selected SLM. While FT capabilities exist for Gemini Pro, utilizing the tuned model through the API necessitates additional authentication beyond a standard key, presenting a current obstacle within the development environment. This issue will be resolved in future work.

Results and Analysis Using ML and Generative AI Language Models

Figure 8 and Figure 9 show the performance in the direct characterization and CoT cases, respectively. Blue bars represent percentage accuracy for all reports while orange bars are percentage accuracy for only distinct descriptions (i.e., each distinct description is weighted equally, even if it only appears once). Labels are used to describe which model was used (LLM versus SLM), the means for supplying examples from 2022 (RAG or FT), and numbers to indicate either the number of examples provided for RAG or the number of training iterations (epochs) when fine tuning. Figure 8 compares the direct categorization results between the ML and Generative AI models and shows significant performance improvements of the Generative AI model predictions compared to ML model predictions.

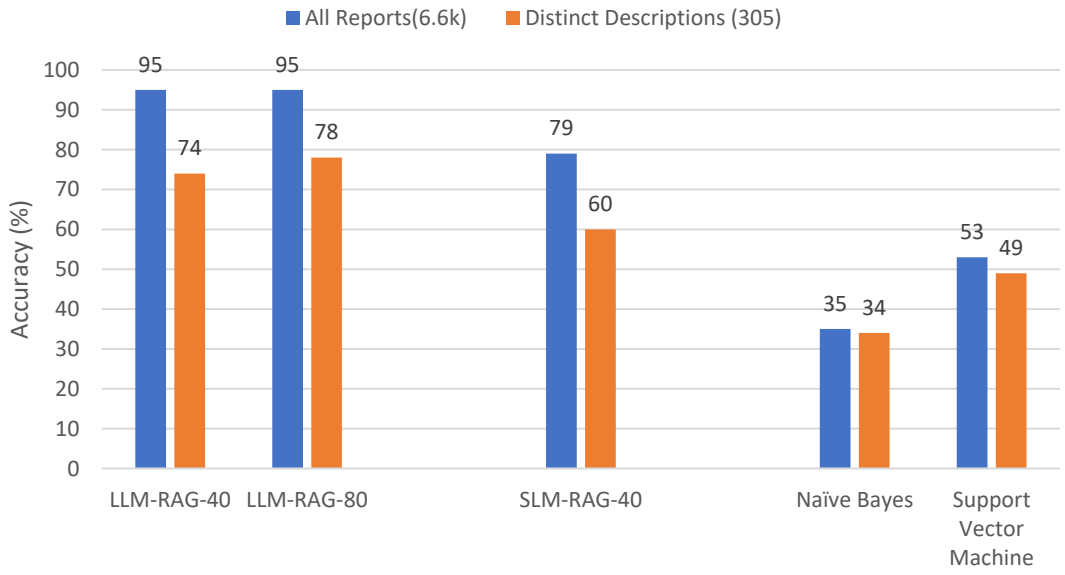


Figure 8. Large and Small Language Model Results for Direct Categorization

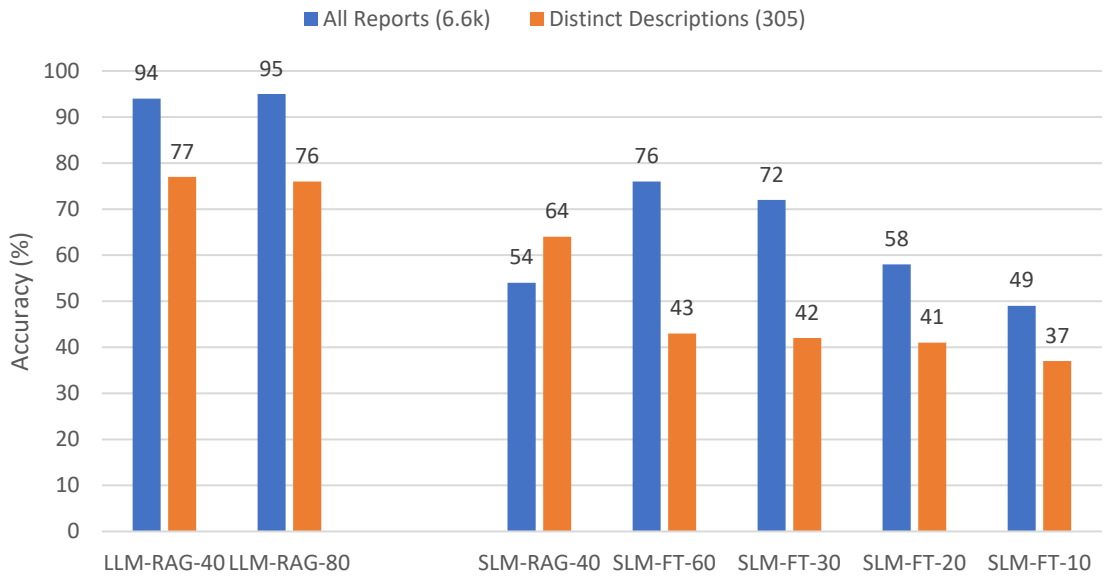


Figure 9. Large and Small Language Model Results Using CoT

Note that, in the RAG case, the SLM cannot process more than 40 examples on average due to its smaller context window. The analysis reveals that for the SLM, CoT prompts hinder performance compared to direct categorization. This does not mean that CoT is a bad approach; rather, it suggests the need to explore alternative CoT formulations. As expected, the LLM outperforms the SLM. Interestingly, doubling the number of examples from 40 to 80 for the LLM yielded minimal improvement. This observation suggests that the embedding model effectively selects relevant examples from 2022 that closely match the semantic content of each 2023 report. Contrary to initial expectations, fine tuning the SLM did not surpass the performance of direct categorization with RAG. As in the CoT case, this result suggests the need for further exploration, especially considering the many adjustable parameters available when fine tuning. Appendix A provides more detailed information about the accuracy and confusion matrices of using ML and AI models in this study.

Analysis Results of Causal Factor Attributes

This section provides the analysis results of 2022-2023 ADS disengagements for the 38 causal factor attributes listed in Table 7, including the counts/relative counts of each attribute using all disengagement reports. Table 9 and Table 10 present statistics on the frequency of occurrence of each attribute in each causal factor category based on 2022 and 2023 ADS disengagement reports, respectively, excluding reports with unknown information. These statistics are expressed in terms of the total count of ADS disengagement reports, count relative to the frequency of occurrence of their category, and count relative to the overall number of known reports. Figure 10 displays the most occurring attributes in the 2022 and 2023 disengagement reports, with three percent or greater count relative to the overall number of known reports. As seen in Figure 10, there were six and 10 such attributes reported in 2022 and 2023, respectively. The following five causal factor attributes were dominantly reported in both years:

1. *Incorrect Behavior/Trajectory Estimation of Other Road Users* of the *Prediction* category at about 19 percent and 31 percent of all known reports in 2022 and 2023, respectively.

2. *Inaccurate Object Detection* of the *Perception* category at about 17 percent and 5 percent of all known reports in 2022 and 2023, respectively.

3. *Motion Planning Issue* of the *Planning* category at about 13 percent and 4 percent of all known reports in 2022 and 2023, respectively.

4. *Mapping Discrepancy* of the *Localization* category at about 8 percent and 4 percent of all known reports in 2022 and 2023, respectively.

5. *Hardware Issue* of the *System* category at about 8 percent and 10 percent of all known reports in 2022 and 2023, respectively.

The *Failure to Yield/Give Priority to Another Road User* of the *Planning* category accounted for about 12 percent of all known ADS disengagement reports in 2022 and only 1 percent in 2023. On the other hand, *Software Issue* of the *System* category accounted for about 11 percent of all known reports in 2023 and only two percent in 2022. The difference in the most-commonly reported causal factor attributes between 2022 and 2023 could be due to the influence of new entrants in 2023 and the change in distinct descriptions by the same tester from 2022 to 2023. The *Non-ADS Factors* category only contributed to about 2 percent and 1 percent of all known ADS disengagement reports in 2022 and 2023, respectively.

Figure 11 displays the statistics of the dominant causal factor attributes from Figure 10 in terms of their counts relative to the total number of known reports in each category.

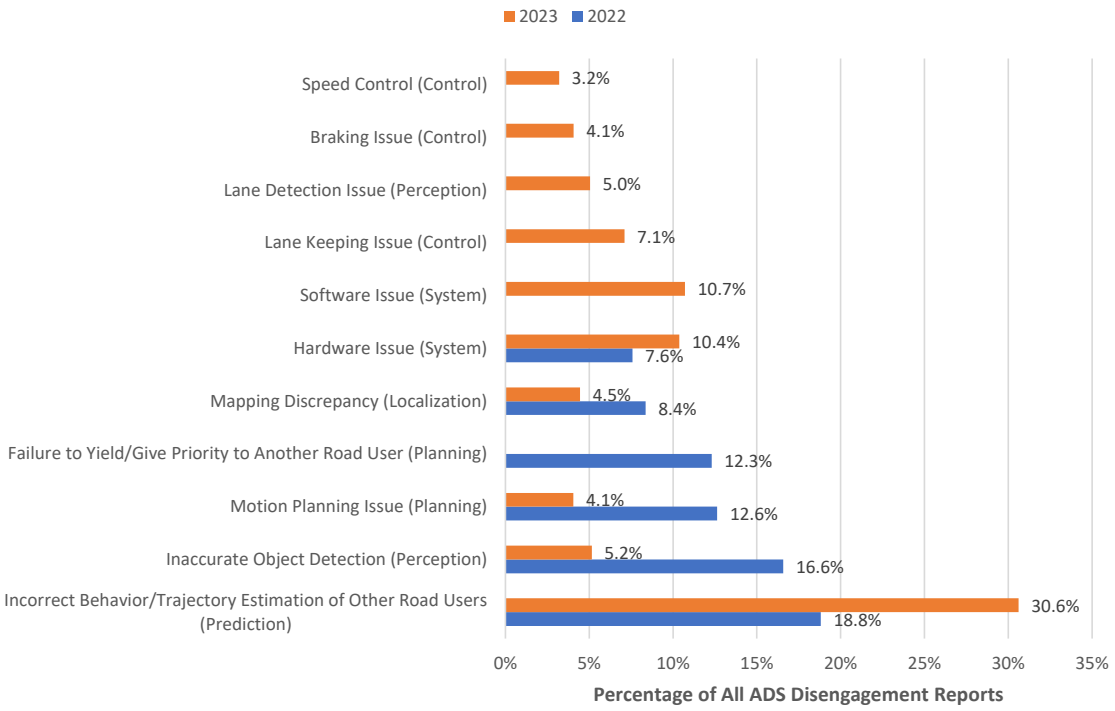


Figure 10. Dominant Causal Factors in All 2022-2023 ADS Disengagement Reports

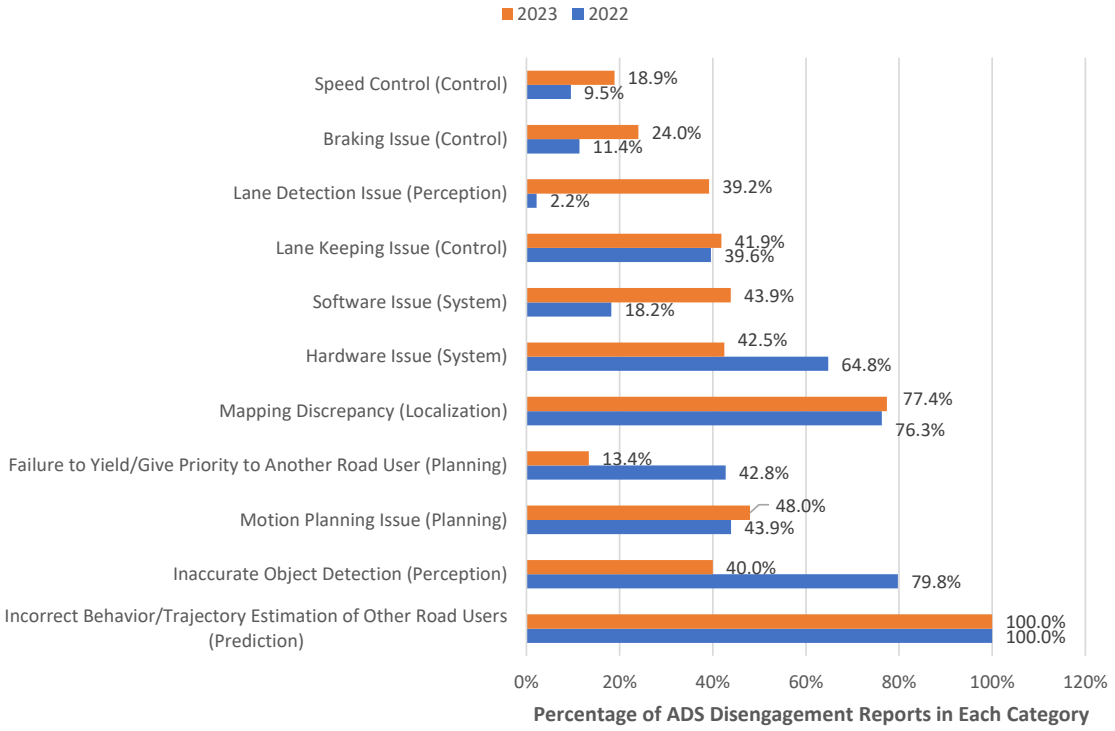


Figure 11. Dominant Causal Factors in Each Category of 2022-2023 ADS Disengagement Reports

Table 9. Statistics of Causal Factor Categories and Attributes of ADS Disengagements Based on 2022 CA DMV Reports

2022 Disengagement Reports																	
8,151																	
Localization			**Perception**			**Prediction**			**Planning**			**Control**			**System**			**Non-ADS Factors**		
894	100%	11%	1,694	100%	21%	1,534	100%	19%	2,345	100%	**29%**	588	100%	7%	954	100%	12%	142	100%	2%
Mapping (Discrepancy)			**Inaccurate Object Detection**			**Incorrect Behavior/Trajectory Estimation of Other Road Users**			**Motion Planning Issue**			**Lane Keeping Issue**			**Hardware Issue**			**Other Road User Behaving Poorly**		
682	**76%**	8%	1,351	**80%**	17%	1,534	**100%**	**19%**	1,030	**44%**	13%	233	**40%**	3%	618	**65%**	8%	72	**51%**	1%
Lane Detection Issue			**Inaccurate Traffic Light Detection at Intersection**						**Failure to Yield/Give Priority to Another Road User**			**Maneuver Issue**			**Software Issue**			**Unexpected Road Conditions**		
123	14%	2%	197	12%	2%				1,003	43%	12%	138	23%	2%	174	18%	2%	22	15%	0.3%
Navigation/ Localization Issue			**Inaccurate Road/Road Lines Perception**						**Traffic Situation**			**Maintaining Desired Path**			**System Performance Issue**			**Obstacle in Path**		
87	10%	1%	39	2%	0.5%				128	5%	2%	74	13%	1%	140	15%	2%	21	15%	0.3%
Positioning Error (Discrepancy)			**Lane Detection Issue**						**Inaccurate Path Planning**			**Braking Issue**			**Sensor Input Delay**			**Construction Zone**		
2	0.2%	0.02%	37	2%	0.5%				111	5%	1%	67	11%	1%	22	2%	0.3%	17	12%	0.2%
			Sensor Issue						**Incorrect Self Trajectory Plan**			**Speed Control**						**Another Vehicle Making an Illegal Maneuver**		
			35	2%	0.4%				59	3%	1%	56	10%	1%				10	7%	0.1%
			Inaccurate Car Detection						**Lane Change Planning Issue**			**ODD Issue**								
			22	1%	0.3%				14	1%	0.2%	9	2%	0.1%						
			Inaccurate Intersection Perception									**Acceleration Issue**								
			5	0.3%	0.1%							6	1%	0.1%						
			Inaccurate Traffic Sign Detection at Intersection									**Steering Issue**								
			5	0.3%	0.1%							5	1%	0.1%						
			Inaccurate Road/Road Lines Perception at Intersection																	
			3	0.2%	0.04%															

Total count	Percent of Category	Percent of All Reports

Table 10. Statistics of Causal Factor Categories and Attributes of ADS Disengagements Based on 2023 CA DMV Reports

2023 Disengagement Reports																	
6,466																	
Localization			**Perception**			**Prediction**			**Planning**			**Control**			**System**			**Non-ADS Factors**		
372	100%	6%	832	100%	13%	1,980	100%	**31%**	546	100%	8%	1,096	100%	17%	1,580	100%	24%	60	100%	1%
Mapping (Discrepancy)			**Inaccurate Object Detection**			**Incorrect Behavior/Trajectory Estimation of Other Road Users**			**Motion Planning Issue**			**Lane Keeping Issue**			**Software Issue**			**Unexpected Road Conditions**		
288	**77%**	4%	333	**40%**	5%	1,980	**100%**	**31%**	262	**48%**	4%	459	**42%**	7%	693	**44%**	11%	27	**45%**	0.4%
Navigation/Localization Issue			**Lane Detection Issue**						**Incorrect Self Trajectory Plan**			**Braking Issue**			**Hardware Issue**			**Other Road User Behaving Poorly**		
42	11%	1%	326	39%	5%				149	27%	2%	263	24%	4%	671	42%	10%	21	35%	0.3%
Positioning Error (Discrepancy)			**Inaccurate Traffic Light or Stop Line Perception**						**Failure to Yield/Give Priority to Another Road User**			**Speed Control**			**System Performance Issue**			**Obstacle in Path**		
42	11%	1%	124	14.9%	2%				73	13%	1%	207	19%	3%	171	11%	3%	6	10%	0.1%
Lane Detection Issue			**Sensor Issue**						**Inaccurate Path Planning**			**Maneuver Issue**			**Sensor Input Delay**			**Another Vehicle Making an Illegal Maneuver**		
-	0%	0%	26	3%	0.4%				48	9%	1%	160	15%	2%	45	3%	1%	5	8%	0.1%
			Inaccurate Road/Road Lines Perception						**Lane Change Planning Issue**			**Steering Issue**						**Construction Zone**		
			20	2%	0.3%				14	3%	0.2%	5	0%	0.1%				1	2%	0.02%
			Inaccurate Car Detection						**Traffic Situation**			**Acceleration Issue**								
			2	0%	0.03%				-	0%	0%	2	0%	0.03%						
			Inaccurate Intersection Perception									**Maintaining Desired Path**								
			1	0.1%	0.02%							-	0%	0%						
			Inaccurate Traffic Light Detection at Intersection									**ODD Issue**								
			-	0%	0%							-	0%	0%						
			Inaccurate Traffic Sign Detection at Intersection																	
			-	0.0%	0.0%															
			Inaccurate Road/Road Lines Perception at Intersection																	
			-	0.0%	0.00%															

Total count	Percent of Category	Percent of All Reports

Concluding Remarks

This analysis developed an initial typology of causal factors that led to ADS disengagements based on 2022-2023 CA CMV AVT program data. The goal of this study is to eventually understand the crash problem of AVs by creating typologies for their crash causal factors and common pre-crash scenarios, similar to prior research on HDV crashes for effective ADAS development. The next step of this study involves the analysis of AV crashes, which will examine the CA DMV and NHTSA SGO crash data and correlate it to CA DMV ADS disengagement reports. Common elements to correlate the two datasets may include, but are not limited to, causal factors, driving scenarios and dynamic interactions between AVs and other road users, roadway locations where the event occurred, relation to intersections and types of traffic control device, roadway conditions, environmental conditions, traffic situations, AV type, and AV manufacturer/tester.

This analysis derived the initial typology of causal factors from interpreting the distinct tester-provided description of facts causing the ADS disengagements as narrated in the CA DMV required report. The CA DMV AVT program did not supply pre-defined causal factor categories and attributes to the participating entities, which resulted in a substantial variation in the provided details by each entity. During the conduct of this analysis, the researchers in this study observed that:

- The safety driver provided the disengagement descriptions using prompts supplied by the testing entities, which widely varied in detail among the ADS developers.
- Interactions with other road users and context during the disengagement events needed to be inferred from the description by the safety drivers.
- Changes in the testing entity-supplied prompts were observed between the 2022 and 2023 ADS disengagement reports.

Due to the noticeable difference in the description of ADS disengagements among the testing entities and between 2022 and 2023 datasets, this analysis selected six causal factor categories based on basic ADS functions and one category for factors not related to ADS functionality. Based on the combined 2022-2023 ADS disengagement reports, excluding reports with driver mistakenly disengaging ADS and with unknown information, Figure 12 shows the contribution of each category in terms of their relative frequency of all known disengagement reports.

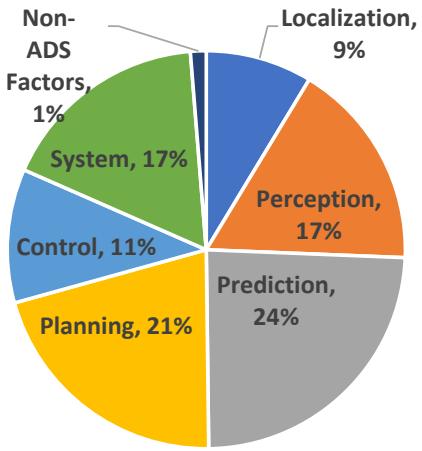


Figure 7. Relative Frequency Distribution of ADS Disengagement Categories

The causal factor typology comprised 38 attributes that contributed to ADS disengagements based on the narratives reviewed in the 2022-2023 reports. The detailed list in Table 7 could serve as a template for consistent reporting among testing entities in future descriptions of ADS disengagements and AV crashes. Also, consistent reporting of driving-related information could be helpful by including descriptions of driving scenarios, roadway locations, intersections and traffic control devices, roadway conditions, environmental conditions, and traffic situations. The list of attributes could be expanded to include more details and specifics to the potential root cause of ADS disengagements based on additional information from future reports as well as further consideration of the driving tasks and their challenges. For instance, better insight into ADS disengagements could benefit from delineating:[15]

- Detection errors between static and dynamic obstacles.
- Erroneous prediction of dynamic obstacle trajectory between cars and pedestrians.
- Perception challenges of sensor uncertainty, occlusion and reflection, drastic illumination changes and glare, atmospheric visibility, and precipitation.
- Long-term, short-term, and immediate planning errors.
- Vehicle control issues specific to lateral control, longitudinal control, and object and event detection and response.
- Miscellaneous driving task actions such as signaling and vehicle conspicuity.

Based on the combined 2022-2023 ADS disengagement reports, excluding reports with the driver mistakenly disengaging ADS and with unknown information, the most dominant causal factor attribute was incorrect behavior/trajectory estimation of other road users of the prediction category, which accounted for about 24 percent of all known disengagement reports. Inaccurate object detection of the perception category was ranked second at 12 percent of all known disengagement reports, followed in a descending order by:

15 https://www.coursera.org/learn/intro-self-driving-cars

planning's motion planning issue (9%), system's hardware issue (9%), planning's failure to yield/give priority to another road user (8%), localization's mapping discrepancy (7%), system's software issue (6%), and control's lane keeping issue (5%). Each of the remaining 30 causal factor attributes accounted for under three percent of all known ADS disengagement reports.

Finally, this study demonstrated the potential of AI language models for the consistent and efficient categorization of ADS disengagement within the domain of automated transportation systems. The results from this test case have intrinsic value and also anticipate the potential of language model applications across a wide range of U.S. DOT use cases. This study highlighted the importance of model selection and workflow design in achieving optimal performance. While FT did not yield significant benefits in this initial exploration, the effectiveness of RAG techniques and the overall encouraging results from both language models warrant further investigation on all fronts. By continuing to explore variations in models, workflows, and human labeling practices, this study team expects to leverage language models to unlock efficient and accurate text classification at low cost, ultimately contributing to the safety and advancement of automated transportation systems.

Appendix: Accuracy and Confusion Matrices of ML and AI Models

Confusion matrices provide a visualization tool to interpret the performance of classification models. In this study, confusion matrices were used to evaluate the performance of both the classical ML models (i.e., Naïve Bayes and SVM) and language models (i.e., LLMs and SLMs). The diagonal entries of the confusion matrix represent correctly classified instances, while off-diagonal entries represent misclassifications. Figure 8 shows the confusion matrices for the Naïve Bayes and SVM models. Analysis of the confusion matrix for both ML models revealed that they performed poorly in predicting the correct category of ADS disengagement. Both models predicted either perception or planning for most of the disengagement categories and a lot of errors piled up where control and prediction categories were predicted as planning. Very similar texts are used to describe many ADS disengagement reports and classical models failed to extract inherent semantic meaning from the extracted features from the text descriptions.

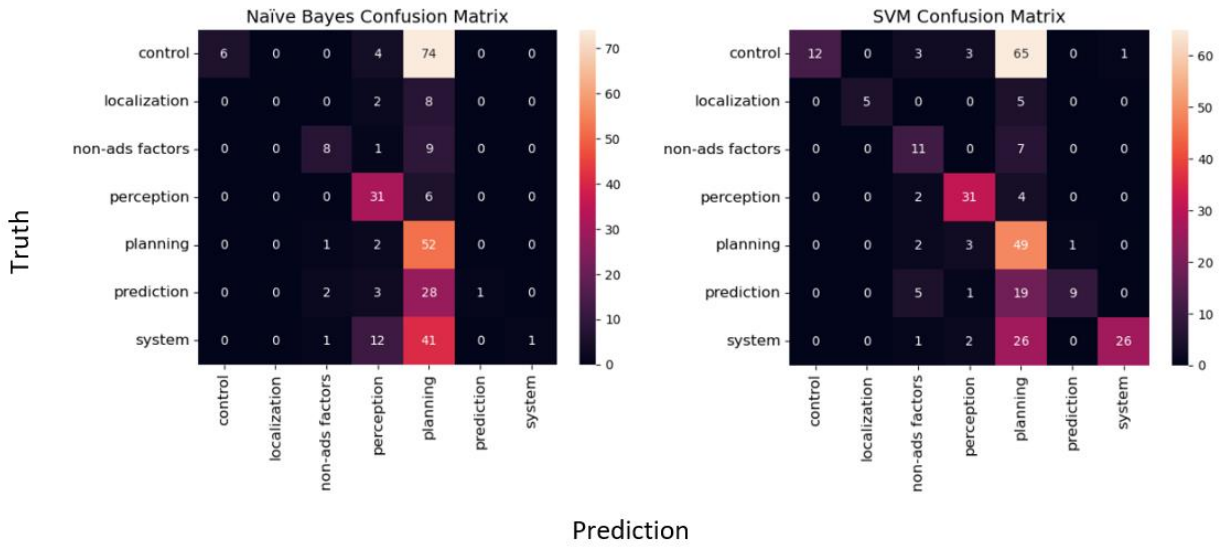


Figure 8. Confusion Matrices for the Naïve Bayes and SVM Classifiers

Figure 9 shows the confusion matrices for the CoT-40 cases. The failure modes of the SLM are indicated with a solid oval in Figure 9. Analysis of the SLM's confusion matrix revealed a five percent failure rate, where the model failed to select any predefined category and instead generated an irrelevant response. Further investigation is needed to determine the cause of this behavior. Potential solutions include modifying the prompt or incorporating follow-up questions to address these cases.

The confusion matrix also offered insights into the categorization scheme itself. Both models exhibited errors

when predicting "other road users" or "control" when the ground truth was "planning" or "prediction." These errors are indicated with long dash circles in Figure 9. This could indicate inconsistency in human evaluation or excessive overlap within the category definitions. In essence, the model's performance not only provides predictions, but also highlights areas for improvement within the categorization framework.

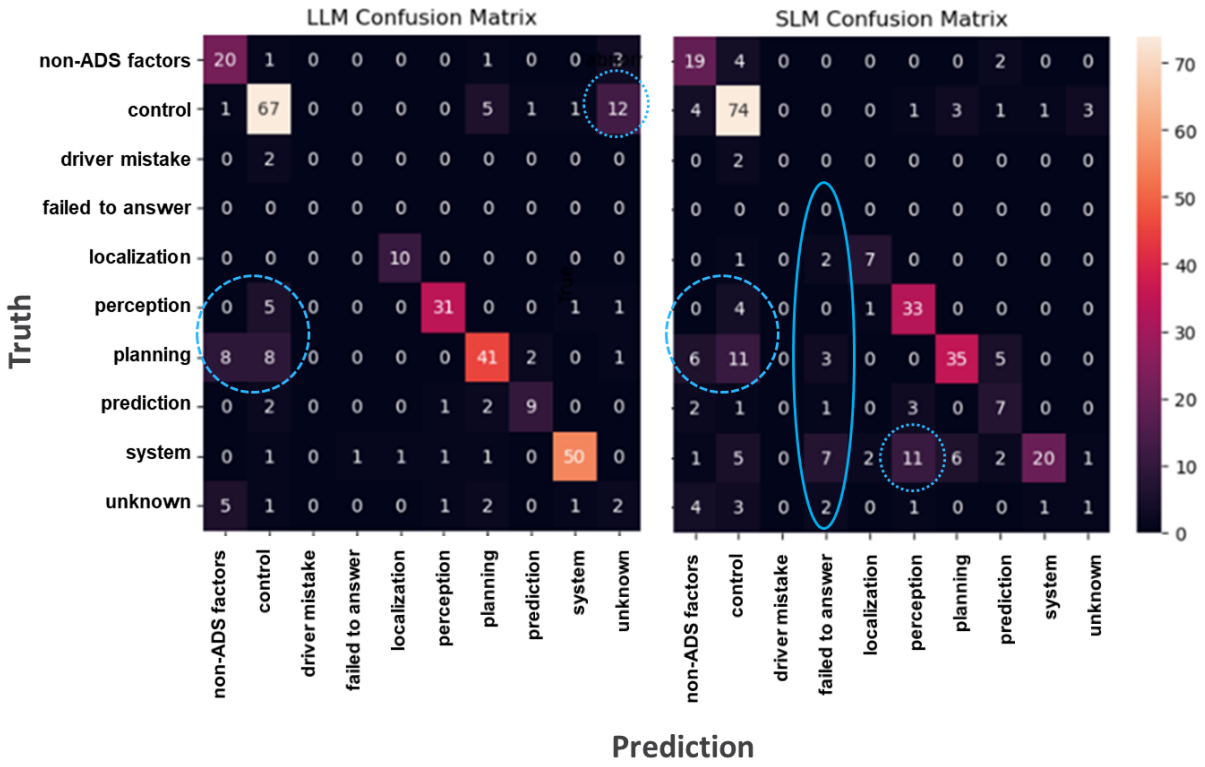


Figure 9. Confusion Matrices for the CoT-40 cases

Finally, the confusion matrix analysis identified specific error hotspots, such as the LLM frequently predicting "unknown" for "control" cases and the SLM predicting "perception" for "system" cases (indicated by small dash circles in the figure). By examining these specific instances, one can delve deeper to understand the reasoning behind these discrepancies and explore methods to mitigate these issues in a generalizable manner.

List of Acronyms

ADAS	Advanced Driver Assistance Systems
ADS	Automated Driving Systems
AI	Artificial Intelligence
AV	Automated Vehicle
AVT	Autonomous Vehicle Tester
CA DMV	California Department of Motor Vehicles
CoT	Chain of Thought
DDT	Dynamic Driving Task
FARS	Fatality Analysis Reporting System
FI	Functional Insufficiency
FT	Fine Tuning
GES	General Estimates System
HASS COE	Highly Automated Systems Safety Center of Excellence
HAZOP	Hazard and Operability
HDV	Human-Driven Vehicle
LLM	Large Language Model
ML	Machine Learning

NHTSA	National Highway Traffic Safety Administration
ODD	Operational Design Domain
OI	Output Insufficiency
QLoRA	Quantized Low-Rank Adaptation
RAG	Retrieval Augmented Generation
SGO	Standing General Order
SLM	Small Language Model
SOTIF	Safety Of The Intended Functionality
STPA	Systems Theoretic Process Analysis
SVM	Support Vector Machine
U.S. DOT	United States Department of Transportation
UCA	Unsafe Control Action
VMT	Vehicle Mile Traveled

References

[1] W.A. Leasure and A.L. Burgett, *NHTSA's IVHS Collision Avoidance Research Program: Strategic Plan and Status Update*. Report No. 94S3001, https://rosap.ntl.bts.gov/view/dot/2759, January 1994.

[2] W. Najm, M. Mironer, J. Koziol, J.-S. Wang, and R.R. Knipling, *Synthesis Report: Examination of Target Vehicular Crashes and Potential ITS Countermeasures*. DOT HS 808 263, National Highway Traffic Safety Administration, Washington, DC, June 1995.

[3] E. Swanson, F. Foderaro, M. Yanagisawa, W.G. Najm, and P. Azeredo, *Statistics of Light-Vehicle Pre-Crash Scenarios Based on 2011-2015 National Crash Data*. DOT HS 812 745, National Highway Traffic Safety Administration, Washington, DC, August 2019.

[4] C. Becker, J. Brewer, and S. Huelsman, *Vehicle-Level Hazard Analysis of a Concept Level 4 Automated Driving System: Urban Taxi.* DOT-VNTSC-NHTSA-10-07, Volpe National Transportation Systems Center, Cambridge, MA, September 2021.

[5] A.M. Boggs, R. Arvin, and A.J. Khattak, *Exploring the Who, What, When, Where, and Why of Automated Vehicle Disengagements*. Accident Analysis & Prevention, Volume 136, March 2020.

[6] Y. Fu, J. Seemann, C. Hanselaar, T. Beurskens, A. Terechko, E. Silvas, and M. Heemels, *Characterization and Mitigation of Insufficiencies in Automated Driving Systems*. Paper No. 27ESV-000110, 27th International Technical Conference on the Enhanced Safety of Vehicles, Japan, April 3-6, 2023.

[7] *Safety First for Automated Driving*, 2019. https://group.mercedes-benz.com/documents/innovation/other/safety-first-for-automated-driving.pdf

Disclaimer

The views and opinions expressed in this document are the authors and do not necessarily reflect those of the U.S. Department of Transportation (U.S. DOT). The contents do not necessarily reflect the official policy of the U.S. DOT.

The U.S. Government does not endorse products, manufacturers, or outside entities. Trademarks, names, or logos appear here only because they are considered essential to the objective of the presentation. They are included for informational purposes only and are not intended to reflect a preference, approval, or endorsement of any one product or entity.