Human Crash Baselines for Robotaxis and Robotrucks

Sydney Katz, Harrison Delecki, Jon Qian, Robert Moss
Valgo

1. Overview

Corresponding author: Sydney Katz (sydney@valgo.ai). This document describes the methodology as of June 2026. The tool at `https://humanbaselines.com` reflects the current version.

Autonomous vehicles are now driving millions of miles on public roads, and crash rates have become a common way to describe their safety. One way to contextualize crash rates is to compare them to the crash rates of human drivers. To make these types of comparisons, we have created an interactive tool that provides human crash baselines for robotaxis and autonomous trucks. At the core, computing a crash rate is quite simple: just divide the number of crashes (numerator) by the number of miles driven (denominator). However, producing this baseline is more challenging than it may originally appear for multiple reasons:

- *Data quality and availability*: crash data is often incomplete, inconsistent, and subject to various reporting biases. For example, different states have different reporting requirements, and even within a state, the quality of data can vary across regions and over time.
- *Fragmented sources*: no single dataset contains everything a rate needs and the numerator and denominator often must be stitched together from different agencies. For example, crash counts come from a state police-report system while exposure comes from a separate roadway-inventory or traffic-count program.
- *Sensitivity to choices*: there is rarely one ''right'' way to define the comparison, and it depends on the specific system and operating conditions being evaluated. It is important to compare AV operators with the combination of settings that matches their specific operational design domain (ODD) for a fair assessment of safety.
- *Precision vs. power*: matching a specific operating area tightly and estimating its rate reliably pull in opposite directions: the finer the geographic slice, the fewer crashes fall inside it, and the noisier the estimate becomes.

Rather than resolve these tensions by settling on a single number, our tool makes the underlying choices explicit and reports the range of baselines they produce. The result is not one human crash rate but a family of baselines. A user can match to them to the system and operating conditions they are evaluating. Figure 1 shows a *specification curve* for Austin, TX, which illustrates how the baseline changes across a range of choices.[1]

[1] Specification curves are a common tool in social science for showing how a result varies across different defensible choices. One of their first use cases was to show how varying assumptions affect the hypothesis that hurricanes with female names are deadlier. Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. "Specification Curve Analysis." In: *Nature Human Behaviour* 4.11 (2020), pp. 1208–1214.

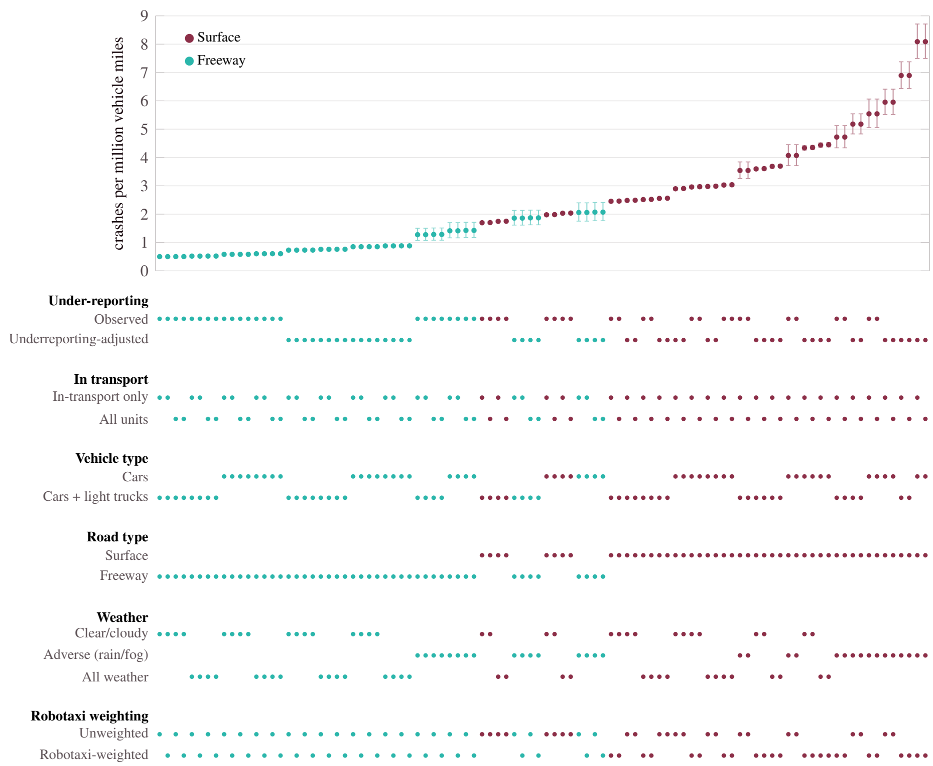

Figure 1. Specification curve for the any-injury human crash rate in Austin (Travis County), 2022. Each of the 96 points is one setting of the six analytical choices we sweep, ranked from lowest to highest. The lower panel marks which choices define each combination. The estimate spans roughly 0.5 to 8 crashes per million vehicle miles across these choices alone. It is important to compare AV operators with the combination that matches their specific ODD for a fair assessment of safety. Color separates surface-street specifications (wine) from freeway ones (teal). Bars are 95% Poisson intervals.

The tool exposes these choices under two modes: geofence (robotaxi) mode, which computes rates over a particular area of a city, and route (autonomous truck) mode, which computes rates along interstate corridors. We currently cover ten metropolitan areas across Texas, California, Arizona, and Nevada, together with interstate corridors across California, Arizona, New Mexico, and Texas, joined into a single connected network on which Interstate 10 runs continuously from Los Angeles to Houston. We propose this tool as a working resource for the industry, regulators, and researchers to use when evaluating the safety of autonomous systems. **Because these choices are judgment calls as much as technical ones, we intend it to grow as a community resource, and we welcome scrutiny, corrections, and contributions to the data sources and methods.** This document describes the methodology we use to compute these baselines, including the data sources, filtering criteria, and assumptions that go into both the numerator and denominator.

2. Methodology

The rate we aim to estimate is the number of crashes per million vehicle miles traveled (VMT), calculated as

$$\text{Crash Rate} = \frac{\text{Number of Crashes}}{\text{Million Vehicle Miles Traveled}} \tag{1}$$

where the numerator is the count of crashes that meet a set of criteria and the denominator is the estimated miles driven by the population of vehicles that could have been involved in those crashes. A key principle of our methodology is that both the numerator and denominator are measured over the same population of vehicles, roads, and other environmental conditions. Therefore, the choices we make in defining the numerator directly inform the choices we make in defining the denominator, and vice versa.

2.1 Numerator

We compute the numerator using publicly available crash data. Most states maintain an extract of police-reported crashes at the crash, vehicle, and person level. We apply a series of filters to these datasets to identify the crashes that are relevant for comparison, following the benchmark methodology of Scanlon et al.[2] Each filter narrows the set of crash-involved vehicles toward the population that is comparable to the autonomous systems we are evaluating. For example, we may filter to crashes to a particular geofenced area, surface streets, in-transport passenger vehicles (those moving in traffic, not parked), and dry weather to make a robotaxi comparison (figure 2). Each of these filters has a mirror on the denominator side, where we estimate the exposure of the comparable population of vehicles under the same conditions. Most of these filters are available as fields in the crash data, but they typically still require additional processing as outlined in the rest of this section.

[2] John M. Scanlon, Kristofer D. Kusano, Yin-Hsiu Chen, Timothy L. McMurry, and Trent Victor. "Benchmarks for Retrospective Automated Driving System Crash Rate Analysis Using Police-Reported Crash Data." In: *Traffic Injury Prevention* 25.sup1 (2024), S51–S65. DOI: 10.1080/15389588.2024.2380522.

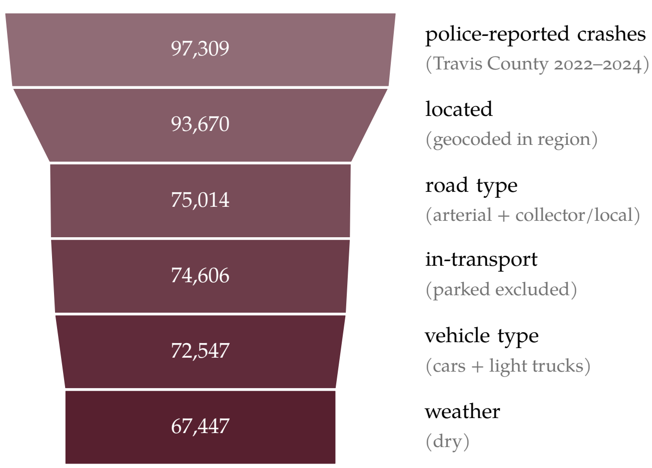

Figure 2. Example of a numerator filter cascade for a specific set of operating conditions, shown for Austin (Travis County), 2022–2024. Each step narrows the set of crash-involved vehicles toward the comparable population.

Outcome and underreporting correction. The crash outcome is defined as how severe an event must be to qualify. Police reports grade injuries on the KABCO scale.[3] Our outcomes represent successively stricter subsets of all police-reported crashes: any reported injury (K, A, B, or C), then serious or fatal (K or A), then fatal (K), following Kusano et al.[4] We also use airbag deployment as a physical-severity proxy that does not rely on the injury coding. A stricter outcome trades statistical power

[3] KABCO is the standard police injury scale: **K** fatal, **A** suspected serious, **B** suspected minor, **C** possible injury, and **O** no injury. American National Standards Institute. *Manual on Classification of Motor Vehicle Traffic Crashes*. Tech. rep. ANSI D16-2017. American National Standards Institute, 2017.

[4] Kristofer D. Kusano, John M. Scanlon, Yin-Hsiu Chen, Timothy L. McMurry, and Trent Victor. "Comparison of Waymo Rider-Only Crash Rates by Crash Type to Human Benchmarks at 56.7 Million Miles." In: *Traffic Injury Prevention* (2025). DOI: 10.1080/15389588.2025.2499887.

for a cleaner signal, since severe crashes are rarer but more consistently reported. The second choice corrects for under-reporting. Police records miss a large share of minor crashes, so a police-reported count understates how often crashes actually occur. Following Blincoe et al.,[5] the count can be scaled by severity-specific factors estimated from national data to estimate the total rather than the reported number of crashes. Property-damage-only crashes are scaled by 2.48 and non-fatal injury crashes by 1.47, while fatal crashes, which are essentially always reported, are left unchanged. Whether to apply this correction is left to the user, since it trades a directly observed count for a modeled one.

[5] Lawrence Blincoe et al. *The Economic and Societal Impact of Motor Vehicle Crashes, 2019 (Revised)*. Tech. rep. DOT HS 813 403. National Highway Traffic Safety Administration, 2023.

Location. We count only crashes inside the region of interest, so each crash must be placed in space. Most records arrive already geocoded into latitude and longitude by the reporting agency. Where coordinates are missing, we estimate them from the street and cross-street in the report by matching against the OpenStreetMap road network, with the per-state details deferred to section 4.[6] Crashes that cannot be placed are excluded. The shape of the region of interest and the spatial unit it is divided into depend on the mode, and are described in section 3.

[6] OpenStreetMap contributors. *OpenStreetMap*. Data available under the Open Database License (ODbL). 2026. URL: `https://www.openstreetmap.org/copyright`.

Vehicle type. We count only the class of vehicle the automated system belongs to, corresponding to passenger vehicles (cars and light trucks) for a robotaxi and heavy trucks for an autonomous truck. We sort each unit into a class from the report's body-style code, with pickups and trucks split into light and heavy by their gross vehicle weight rating. The most challenging portion of vehicle type filtering is dealing with units the report leaves ambiguous: a body-style code that is missing or coded unknown, or a pickup with no weight recorded to tell a light truck from a heavy one. Dropping these would bias the count, because missing information is not spread evenly across vehicle classes. Instead, we keep every ambiguous unit and resolve it using whatever evidence the record carries. When a vehicle identification number is present, decoding it against a federal database pins the unit to a single class.[7] When it is not, we fall back to a distribution and assign the unit fractional weights across the classes it might belong to. If the vehicle's make is present, we use the class distribution observed for that make. Otherwise, we use the overall distribution across all classified vehicles. This treatment of not-fully-specified units follows Scanlon et al.[8] The result is a count that reflects the uncertain units in proportion to what they most likely were, rather than discarding them.

[7] National Highway Traffic Safety Administration. *Vehicle Product Information Catalog (vPIC) VIN Decoder*. URL: `https://vpic.nhtsa.dot.gov/decoder/`.

[8] John M. Scanlon, Timothy L. McMurry, Yin-Hsiu Chen, Kristofer D. Kusano, and Trent Victor. "From Stoplights to On-Ramps: A Comprehensive Set of Crash Rate Benchmarks for Freeway and Surface Street ADS Evaluation." In: *SAE International Journal of Transportation Safety* 14.2 (2026). SAE 09-14-02-0003.

Road type. We count only crashes on the kind of road the automated system drives. Roads are graded by their federal functional class, which sorts them from interstates down through arterials, collectors, and local streets. The challenge for this filter is that the road class recorded on the report often cannot be taken at face value. It is left blank for a large share of city-street crashes, and crashes on a freeway's frontage road are coded under the parent freeway rather than the surface street the vehicle was actually on. Taken as given, we would file many surface crashes as freeway and drop many others that do not have a road type assigned. We instead recover the true class from where the crash happened, snapping each crash to the nearest segment of the roadway-inventory network and reading the class off that segment, with the per-state details deferred to section 4. Crashes that cannot be placed on a segment fall back to the surface-street pool.

Weather. We can also narrow to crashes that happened under a given weather condition: dry, rain, or fog. The condition is read from the weather field the officer recorded on the report.

2.2 Denominator

Once we have the crash count, we need to estimate the miles driven by the population of vehicles that could have been involved in those crashes. The denominator is the VMT of the comparable population, which we refer to as its exposure, measured over the same roads and time period as the numerator. We construct it as a spatial shape that distributes VMT across cells or segments from a roadway network, rescaled to match a published magnitude of total VMT for the region. Many states maintain a roadway inventory that includes VMT estimates for each segment, and we use these where they are complete. Many states also publish reports on the total VMT by vehicle class and functional class, which we use to carve the magnitude into the relevant class and road type. Where a state does not have a complete inventory, we fall back to the Highway Performance Monitoring System (HPMS), which provides a national-standard roadway network with VMT estimates.[9]

For filters such as road type and vehicle class, we apply the same criteria to the denominator as we do to the numerator. For example, if we are counting crashes on surface streets involving passenger vehicles, we restrict the denominator to the VMT of passenger vehicles on surface streets to match. To restrict by vehicle class, we use the Federal Highway Administration's VM-4 travel fractions, which estimate the share of VMT by vehicle class for each functional class.[10] To restrict by road type, we stratify the VMT by functional class and collapse it into surface and freeway categories to match the numerator's road-type filter. Weather is handled differently, because the roadway data does not record the conditions travel occurred in. We instead scale the exposure by the fraction of travel that took place under the selected condition, estimated from nearby weather-station observations.[11]

The denominator is subject to more uncertainty than the numerator, especially for local roads where traffic counts are sparse and the share of travel must be estimated rather than measured. We annualize all exposure to the same span of years as the crash count, so that the two sides of the rate cover the same period, and defer the per-state choices of roadway inventory, magnitude source, and local-road handling to section 4.

[9] HPMS is the Federal Highway Administration's national repository of public-road data. Each state annually reports standardized attributes, including traffic volumes and functional class, for its road network, which the agency assembles into a public, nationally consistent geospatial dataset. We use the most recent full-extent national release for 2023. Federal Highway Administration. *Highway Performance Monitoring System (HPMS) 2023*. U.S. Department of Transportation, Bureau of Transportation Statistics, National Transportation Atlas Database. 2024.

[10] VM-4 is a table in the agency's annual Highway Statistics series giving the percentage of vehicle-miles traveled by vehicle type for each functional class, reported separately for rural and urban roads. Federal Highway Administration. *Highway Statistics 2023, Table VM-4: Distribution of Annual Vehicle Distance Traveled, Percentage by Vehicle Type*. Tech. rep. U.S. Department of Transportation, Federal Highway Administration, 2024.

[11] National Centers for Environmental Information. *Integrated Surface Database (ISD)*. National Oceanic and Atmospheric Administration. URL: `https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database`.

2.3 Rate

The rate is computed as the matched crash count over the matched exposure, calculated for each outcome level and for each spatial unit (cell or segment). Because the defensible choices vary, the tool produces a family of baselines. Sweeping across these choices produces the specification curve introduced in figure 1, which illustrates how the baseline changes across a range of defensible assumptions.

3. Mode Specific Methodology

The methodology presented so far has been mode-agnostic. Every baseline is a matched crash count divided by matched exposure, reported across a family of defensible assumptions. The two modes differ only in how that process is instantiated. Geofence (robotaxi) mode estimates a rate over an area of a city, while route (autonomous truck) mode estimates a rate along an interstate corridor. Table 1 shows the specific choices that define each mode. Vehicle class is a user-selectable filter, and the table lists the default robotaxi and autonomous truck comparison. The spatial unit, exposure, rate normalization, rate unit, and confidence interval are specific to the mode itself. Road type is deliberately absent: it does not split cleanly by mode, since a geofence rate can include highways and an autonomous-truck route includes the off-interstate roads used to reach depots. The remainder of this section expands on these mode-specific choices.

	Geofence (robotaxi)	Route (autonomous truck)
Rate unit	Per million vehicle miles	Per million vehicle traversals
Spatial unit	S2 grid cells	Corridor segments
Exposure	Road VMT per cell	Truck VMT per segment
Vehicle class	Passenger vehicles	Combination trucks
Confidence interval	Exact Poisson	Empirical Bayes

Table 1. The two modes apply the same crash-count-over-exposure recipe with different vehicles, spatial units, and exposure bases. Vehicle class is the default robotaxi and autonomous-truck selection; the remaining rows follow from the mode.

3.1 Geofence (Robotaxi) Mode

For geofence mode, we estimate a crash rate over a user-defined area of a city, measured as crashes per million vehicle miles traveled. By default the comparison is restricted to surface streets, where robotaxi operation has so far been concentrated, and excludes freeways and their frontage roads. County-wide rates use the crash record's county directly. We count every qualifying crash in the county and divide by the county's vehicle miles. To compare against a system that operates in only part of a county, we need to know where within the county each crash and each mile occurred. Following Chen et al.,[12] we discretize the area into level-13 S2 cells[13] and assign each crash to the cell containing its coordinates, so a rate can be computed over any sub-region the user draws.

[12] Yin-Hsiu Chen, John M. Scanlon, Kristofer D. Kusano, Timothy L. McMurry, and Trent Victor. "Dynamic Benchmarks: Spatial and Temporal Alignment for ADS Performance Evaluation." In: *Transportation Research Record: Journal of the Transportation Research Board* (2025). DOI: 10.1177/03611981251398744.

[13] S2 indexes the globe by recursively subdividing it into cells, with each level splitting the previous into four. Level-13 cells average roughly a square kilometer. S2 Geometry Developers. *S2 Geometry Library*. 2017. URL: https://s2geometry.io.

Within each cell, we count the crashes and divide by the VMT on the road segments that fall inside it. Aggregated over the selected area, this baseline weights each location by how much human traffic it carries, which assumes a robotaxi drives the same spatial mix as the average vehicle. Robotaxis do not: they concentrate in certain neighborhoods and along certain routes, so their mix of driving miles looks different from the area average. The user can optionally apply a spatial weighting multiplier, again following Chen et al., that reweights the rate toward the cells a robotaxi is more likely to drive in.[14] We compute confidence intervals using the exact Poisson interval, which is appropriate for count data and does not rely on large-sample approximations. It stays valid even when the selected area contains no observed crashes, returning a one-sided upper bound rather than collapsing to zero.

[14] We currently use the publicly available Waymo data to determine the robotaxi driving patterns, but the method is general and can be applied to any dataset of robotaxi driving. `https://waymo.com/safety/impact/`

3.2 Route (Autonomous Truck) Mode

For route mode, we estimate a crash rate along an interstate route, measured as crashes per million traversals, where one traversal is a single end-to-end drive of the route. The basic comparison runs between two points on the interstate mainline. Because freight trips begin and end at depots that sit off the interstate, we also let the user drop two depot locations and add the legs that connect each depot to the interstate, so the quoted risk can cover the entire trip.

We break the route into short pieces, each of which contributes to the expected number of crashes for one traversal. Each piece carries a local human crash rate calculated as the crashes recorded on it divided by the combination-truck miles driven on it. We weight each piece's rate by the miles a truck actually drives through it and add these contributions across the whole path, giving the expected number of crashes on a single traversal.

We define the break the route into pieces differently for the interstate mainline and the off-interstate access legs. On the interstate mainline, the pieces are one-mile segments. Interstates are indexed by milepost and carry per-segment combination-truck traffic counts, so each one-mile segment has its own crash count and its own truck miles. The route between the two endpoints is the shortest path through these segments, and each segment on it contributes its per-mile crash rate multiplied by its length.

The off-interstate access legs have no comparable milepost index, so we fall back to the same level-13 S2 cells used in geofence mode. Each cell carries the combination-truck miles driven on its non-interstate roads, estimated from the state roadway inventory. We compute a driving route from each depot to an in-scope on-ramp with the Open Source Routing Machine over the OpenStreetMap road network.[15] The resulting route passes through a sequence of cells, each contributing its per-mile crash rate times the miles the route spends inside it. An access leg is therefore geofence mode stretched along a line: the same cells and the same per-cell exposure, measured over the miles a single trip drives through them rather than over an area.

[15] Dennis Luxen and Christian Vetter. "Real-time routing with OpenStreetMap data." In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2011. DOI: `10.1145/2093973.2094062`.

The full depot-to-depot trip consists of three of these paths joined end to end (an access leg from the first depot to its on-ramp, the interstate segments in the middle, and a second access leg from the exit ramp to the second depot). Because every piece is already expressed as expected crashes for one traversal, the legs combine by simple addition. The access paths are precomputed for every cell, and the engine selects the on-ramp best suited to the direction of each trip.

Because a traversal sums the crash counts of many pieces with different exposures, the trip total is a weighted sum of Poisson counts, which is not itself Poisson, so we cannot use the exact Poisson formulation to compute confidence intervals. We instead take a Bayesian approach and treat each piece's underlying crash rate as uncertain, and assign it a Gamma prior fit from an aggregation of the rate of like pieces grouped by road type. A piece with little exposure of its own therefore borrows strength from its group rather than reading as implausibly safe or implausibly dangerous on the strength of one or two crashes. Combining each piece's observed crashes with this prior yields a Gamma posterior for its rate. Summed across the traversal, these posteriors give the expected crashes for one trip, but their weighted sum is no longer Gamma, so we approximate it with a single Gamma matched to its mean and variance and use the resulting distribution to compute confidence bounds.

4. State Specific Methodology

The methodology described so far applies to every state, but each state's crash and roadway data have with their own gaps, and filling them required a variety of judgement calls specific to each state. This section records those decisions in terms of what each source provided, what was missing, and what we used in its place.

4.1 Texas

In Texas we cover four metropolitan areas: Austin, Dallas, Houston, and San Antonio, each geofenced to its county (Travis, Dallas, Harris, and Bexar), together with the interstate corridors that connect them.

Crash data. Crashes come from the TxDOT Crash Records Information System (CRIS) public extract, a release of officer crash report data with personal information removed, recorded at the crash, vehicle, and person level.[16] The city geofences use crash years 2022 through 2024, while the interstate corridors are pinned to 2022. Two parts of the report data required additional processing beyond reading a field. CRIS records a vehicle's body style but not its weight, so a pickup or truck cannot be sorted into a light or heavy class from the report alone. We recover the class by decoding the vehicle identification number where it is present, and fall back to the make-level and overall distributions otherwise. The reported road class is also blank for many city-street crashes and files frontage-road crashes under the

[16] Texas Department of Transportation. *Crash Records Information System (CRIS) Public Query*. URL: `https://cris.dot.state.tx.us/public/Query`.

parent freeway, so we recover the true class by snapping each crash to the TxDOT roadway-inventory network rather than trusting the coded value. Coordinates are present on most records, and the minority without them are geocoded from the street and cross-street. CRIS also records airbag deployment at the person level, so the airbag outcomes are available directly.

Exposure. Texas exposure comes from the TxDOT Roadway Inventory, the state DOT's own record of traffic volumes across the road network.[17] We use it directly. The per-segment vehicle-miles are allocated to cells by the inventory's own geometry. The inventory already spans the full network, local roads included, so there is no gap to fill and we do not fall back to HPMS. The one modeling assumption for geofence mode is the VM-4 split from section 2.2. The inventory's traffic is all-vehicle, so we scale each road class by its passenger-vehicle share to match the numerator. The interstate corridors instead draw their per-segment combination-truck volumes from HPMS, consistent with the Arizona and New Mexico mainlines (section 4.3); HPMS's combination counts are complete on the Texas mainline and track the inventory's own combination volumes to within a fraction of a percent.

[17] Texas Department of Transportation. *TxDOT Roadway Inventory*. URL: `https://www.txdot.gov/inside-txdot/division/transportation-planning/roadway-inventory.html`.

4.2 California

In California we cover four metropolitan areas: Los Angeles, San Francisco Bay Area, Sacramento, and San Diego, each geofenced to its county or counties (Los Angeles; San Francisco, San Mateo, and Santa Clara; Sacramento and Yolo; and San Diego), together with the interstate corridors that connect them and continue east into Arizona.

Crash data. California crashes come from the Statewide Integrated Traffic Records System (SWITRS), the California Highway Patrol's statewide record of reported crashes.[18] The city geofences use crash years 2022 through 2024. The defining challenge in California is location. Crashes reported by the Highway Patrol carry coordinates, but the many crashes reported by local agencies (municipal police and the county sheriff) often do not, so a large share of records (far larger than in Texas) must be geocoded before they can be placed in a cell. We geocode these from the primary road, cross street, and city named on the report against the OpenStreetMap network. Vehicle classification also works differently. SWITRS records no vehicle identification number. Therefore, instead of decoding one, we read the class directly from the vehicle-type code on the report, which already separates cars from light trucks, and resolve the ambiguous units by make as described earlier. The reported road class is recovered as in Texas, by inferring the roadway function from where the crash occurred rather than trusting the coded value. Airbag deployment is recorded at the party and passenger level, so the airbag outcomes are available.

[18] California Highway Patrol. *Statewide Integrated Traffic Records System (SWITRS)*. California Open Data Portal. URL: `https://data.ca.gov/dataset/ccrs`.

Exposure. California has no statewide per-segment record of local traffic, so there is no single inventory to read directly. We build the denominator with the general construction from section 2.2 as a spatial shape rescaled to a published magnitude. The shape comes from HPMS, which covers the California network down to local roads. The magnitude comes from California Public Road Data (CPRD), Caltrans's annual roadway statistics, whose totals are the only complete account of local vehicle-miles in the state.[19] We distribute each county's CPRD total across cells in proportion to the HPMS volumes, then apply the VM-4 passenger-vehicle split described in section 2.2.

[19] California Department of Transportation. *California Public Road Data*. URL: `https://dot.ca.gov/programs/research-innovation-system-information/highway-performance-monitoring-system`.

CPRD reports its vehicle-miles by who maintains the road (jurisdiction) with a state-highways total covering the interstate, freeway, and state-route system, and the remainder maintained by cities and counties. We form the surface magnitude by taking the county's all-jurisdiction total and subtracting the state-highways portion, which leaves the city- and county-maintained roads. We use the state-highways portion as the freeway magnitude.

This leaves one tension between the two sides of the rate. The numerator sorts crashes into surface and freeway by functional class, while CPRD sorts vehicle-miles by jurisdiction. The two mostly agree except for surface state routes. A US highway or state-numbered arterial is a surface road by function yet state-maintained by jurisdiction, so its miles count toward the freeway denominator while its crashes count toward the surface numerator. Rather than force a single answer, we expose the magnitude source as a switch. The default uses CPRD, anchored to Caltrans's published totals and inclusive of the local-road travel that HPMS undercounts, accepting the small surface-state-route mismatch. The alternative uses HPMS classified by functional class, which aligns the denominator's road types exactly with the numerator's and erases the mismatch. However, this technique inherits California's incomplete HPMS arterial counts, where some segments record no traffic. The default trades a clean road-type match for a trustworthy total, and the alternative makes the opposite trade.

The interstate corridors again use per-segment combination-truck volumes paired with the Class-8 numerator. Within California we take these from Caltrans truck-count stations, which report five-or-more-axle volumes as the combination-truck proxy, rather than from HPMS, whose combination counts on these interstates are unreliable with blank stretches and implausible values. We assign each one-mile segment the count from the nearest station on the same route. This is an interpolation. Stations sit every couple of miles along the busy corridors and thin to ten or fifteen miles apart on remote rural interstate, where combination volume is nearly constant anyway. These corridors continue east into Arizona, where the exposure is drawn from a different source described in section 4.3.

4.3 Arizona

In Arizona we cover the Phoenix metropolitan area, geofenced to Maricopa County, together with the Arizona portions of the interstate corridors that run east from California.

Crash data. Arizona crashes come from the Arizona Department of Transportation's ALISS crash records, exported as linked crash, unit, and person files.[20] We take Maricopa County for the years 2022 through 2024. The records carry coordinates, so there is little to geocode. The export records no vehicle identification number, so we read the class from the body-style code on the report. Ambiguous units are resolved by make as described earlier. The reported road class needs the most work. Arizona's records carry no freeway flag, so we infer surface from freeway by the route named on the report, treating interstates as freeway and numbered city streets as surface. We resolve the state and US routes that run as freeway in some stretches and surface in others by whether the crash falls within a few hundred meters of a known freeway segment. Airbag deployment is recorded, so the airbag outcomes are available.

[20] Arizona Department of Transportation. *Arizona Crash Information (ALISS)*. URL: `https://azdot.gov/planning/traffic-safety/arizona-motor-vehicle-crash-facts`.

Exposure. Arizona publishes vehicle-miles only at the state level, with no county total to anchor to, so for Maricopa we build the magnitude ourselves. We start from the ADOT traffic-count dataset,[21] which gives an annual average daily traffic count for each road segment, multiply each count by the segment's length, and attach a functional class to each segment from HPMS. Totalling by functional class gives the surface and freeway magnitudes, which we distribute across cells in proportion to the HPMS volumes and split to passenger vehicles with the VM-4 fractions.

[21] Arizona Department of Transportation. *Average Annual Daily Traffic (AADT)*. URL: `https://services6.arcgis.com/clPWQMwZfdWn4MQZ/arcgis/rest/services/ADOT_AverageAnnualDailyTraffic_2023/FeatureServer`.

Because the magnitude is bucketed by functional class from the start, the surface and freeway denominators line up with the way the numerator classifies roads, so Arizona does not have the jurisdiction-versus-function mismatch that California carries. However, the denominator does carry one gap. Arizona's HPMS network has no local roads, and the counts reach only about a third of the state's local travel, so local-road exposure is substantially undercounted. The effect is bounded since arterials rather than local streets carry most of the travel in the operating area. With no county total to check against, we can only compare the aggregate to the statewide figure, which our version reaches to about ninety-three percent, and the shortfall is almost entirely the missing local travel.

The interstate corridors use the same per-segment combination-truck construction as Texas and California. We use the combination-truck volumes from HPMS, which is complete and reliable on the Arizona mainline.

4.4 New Mexico

New Mexico enters only as a connector. Its stretch of Interstate 10, from the Arizona line near San Simon east through Lordsburg, Deming, and Las Cruces to the Texas line at El Paso, is the physical link that joins the western (California and Arizona) corridors to the Texas ones, so a single route can run from Los Angeles to Houston on one continuous interstate. We carry no New Mexico geofence city and no other New Mexico interstate, since I-25 and I-40 there would be corridors with no in-scope connection.

Crash data. New Mexico crashes come from a New Mexico Department of Transportation data request, a statewide release of crash and vehicle records with personal information removed. Unlike the other states this release is for 2024 rather than 2022, the only year provided; the rate is computed per segment over matched single-year exposure, so each leg is internally consistent in time, and the New Mexico leg's weather exposure is drawn from the same year. The release has no person table, so we read airbag deployment from the vehicle record (a small share is blank, which we treat as unknown) and injury severity from the crash-level injury counts. Vehicle body style is recorded as decoded text, which we map to the class taxonomy directly. Coordinates are present on every record, but the route name is filled on under half, so we recover the road from where the crash happened by snapping to the segment network rather than from the reported route.

Exposure. As on the Arizona and Texas mainlines, the New Mexico interstate corridor draws its per-segment combination-truck volumes from HPMS, whose combination counts are complete on the New Mexico mainline.

4.5 Nevada

In Nevada we cover a single metropolitan area, Las Vegas, geofenced to Clark County.

Crash data. Nevada crashes come from the Nevada Department of Transportation's statewide crash export, which we scope to Clark County for the years 2022 through 2024.[22] Because the crashes already carry coordinates, there is little to geocode. The export has no vehicle identification number and no numeric body-style code, so we read the class from a descriptive vehicle-type string and resolve the unspecified units by make as described earlier. The road class is more challenging. Nevada's records do not carry freeway flag or route name, so we cannot infer the class from the route. Instead we snap each crash to the nearest segment of the road network and take its class. One filter that we apply was unavailable in the crash data export. Specifically, the export has no reliable record of whether a vehicle was parked, so we cannot exclude parked vehicles and instead keep every unit. Airbag deployment is recorded for each person, so the airbag outcomes are available.

[22] Nevada Department of Transportation. *Nevada Crash Data*. URL: `https://www.dot.nv.gov/`.

Exposure. The Nevada Department of Transportation publishes annual vehicle-miles for Clark County, broken out by functional class and including local roads.[23] We use it directly as the magnitude, distribute it across cells by the HPMS shape, and apply the VM-4 passenger-vehicle split. Because the published total is already split by functional class, the surface and freeway denominators line up with the numerator with no further processing. However, local roads require additional processing. HPMS carries their geometry but no traffic counts. To place the published local-road travel in space we spread it across cells in proportion to each cell's length of local road, assuming roughly uniform local volume.

[23] Nevada Department of Transportation, Roadway Systems Division. *Annual Vehicle Miles of Travel (AVMT) Publication, 2023 HPMS Data Year*. URL: `https://www.dot.nv.gov/home/showpublisheddocument/22863/638605184064300000`.

5. RAVE Checklist

This section walks through the RAVE checklist for the methodology presented in this paper.[24]

[24] The RAVE (Retrospective Automated Vehicle Evaluation) checklist is a set of fifteen recommendations for conducting and evaluating retrospective studies that compare automated driving systems against human crash rates. John M. Scanlon, Eric R. Teoh, David G. Kidd, Kristofer D. Kusano, Jonas Bärgman, Geoffrey Chi-Johnston, et al. "RAVE checklist: Recommendations for overcoming challenges in retrospective safety studies of automated driving systems." In: *Traffic Injury Prevention* 26.5 (2025), pp. 608–621. DOI: `10.1080/15389588.2024.2435620`.

1. *Ensure compatibility of the ADS and benchmark data for both crashes and exposure.* The guiding principle of our methodology is that the numerator and denominator are measured over the same population of vehicles, roads, and conditions, so every crash filter has a matching restriction on the exposure. We apply the four alignment methods the checklist describes.

2. *Prioritize methodological accuracy but default to conservative analysis.* Rather than settle on one set of choices, we report the full family of baselines, so the specification curve doubles as a built-in sensitivity analysis. Our defaults favor directly measured values over modeled values, e.g., with the underreporting correction off unless the user enables it.

3. *Favor outcomes that are directly measurable.* Our outcome levels read directly off the police-reported KABCO injury scale: any reported injury, suspected serious or fatal, and fatal. We also offer airbag deployment as a physical-severity proxy that does not depend on the injury coding at all.

4. *Consider outcome data transformations to help overcome limitations.* We offer two outcome data transformations that are both documented and user-controlled. The underreporting correction scales reported counts by severity-specific factors to estimate the crashes that never reached the data, and the probabilistic weighting of ambiguous vehicle units recovers their likely class rather than dropping them. Each is a toggle so that the user can see the observed and modeled versions side by side.

5. *Quantify uncertainty of the estimates with statistical methods.* Every baseline carries a 95% confidence interval, computed by a method matched to how the rate is aggregated (section 3). For geofence mode, we use an exact weighted-Poisson interval. For route mode, we use an empirical Bayes Gamma-Poisson interval, which captures the extra variability between road segments.

6. *Cite all data sources used in the study.* We draw on public sources only and cite each dataset where it is used. The tool lists all sources in-app alongside the baselines they support.

7. *Provide descriptive statistics of the ADS and benchmark data composition.* The tool is itself a descriptive-statistics explorer. Every filter (area, road type, vehicle class, weather, outcome, and year) is selectable. The crash count and exposure recompute for each slice, so the user can see how the population breaks down along any dimension.

8. *Describe the ADS deployment being evaluated.* Not applicable. We supply only the human benchmark, so describing the deployment is the evaluator's responsibility. However, the tool's flexibility allows the user to tailor the baseline to match the deployment as closely as possible.

9. *Clearly document analysis decisions and steps in the methods section.* This white paper documents every filter and modeling choice.

10. *Document limitations in the study's data, methods, scope, and interpretation.* We document limitations where they arise. The denominator discussion flags where exposure is estimated rather than measured, most acutely on local roads, and the state-specific section (section 4) records per-area caveats such as a missing parked-vehicle filter or sparse coverage of a given field.

11. *Build upon past research and justify the scope for what was selected to be studied.* Our methodology follows the established crash-benchmark literature, cited throughout, and extends it to new areas and to the autonomous truck setting. The scope (which metropolitan areas and corridors we cover) is set by where automated systems operate and where suitable public data exists.

12. *Develop research questions that are clear, concise, and specific.* Every baseline answers a single, fully-specified question. The tool requires each ODD decision to be set, so results are precise to the specified conditions.

13. *Ensure conclusions accurately reflect the study design.* We draw no safety conclusion ourselves. We publish the human baseline with its confidence interval and caveats. Reporting the full family of baselines rather than a single number keeps any conclusion drawn against it tied to the specific choices that produced it.

14. *Exercise caution in relating findings between different severity levels.* We never extrapolate across severities. Each outcome level, from all police-reported crashes down to fatal, is computed from its own crashes and reported separately, so each severity stands on its own rather than being inferred from another.

15. *Present rates in incidents per exposure units.* Every baseline is reported as crashes per million vehicle miles traveled, an incidents-per-exposure rate by construction.

References

1. Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. "Specification Curve Analysis." In: *Nature Human Behaviour* 4.11 (2020), pp. 1208–1214.

2. John M. Scanlon, Kristofer D. Kusano, Yin-Hsiu Chen, Timothy L. McMurry, and Trent Victor. "Benchmarks for Retrospective Automated Driving System Crash Rate Analysis Using Police-Reported Crash Data." In: *Traffic Injury Prevention* 25.sup1 (2024), S51–S65. DOI: `10.1080/15389588.2024.2380522`.

3. American National Standards Institute. *Manual on Classification of Motor Vehicle Traffic Crashes*. Tech. rep. ANSI D16-2017. American National Standards Institute, 2017.

4. Kristofer D. Kusano, John M. Scanlon, Yin-Hsiu Chen, Timothy L. McMurry, and Trent Victor. "Comparison of Waymo Rider-Only Crash Rates by Crash Type to Human Benchmarks at 56.7 Million Miles." In: *Traffic Injury Prevention* (2025). DOI: `10.1080/15389588.2025.2499887`.

5. Lawrence Blincoe et al. *The Economic and Societal Impact of Motor Vehicle Crashes, 2019 (Revised)*. Tech. rep. DOT HS 813 403. National Highway Traffic Safety Administration, 2023.

6. OpenStreetMap contributors. *OpenStreetMap*. Data available under the Open Database License (ODbL). 2026. URL: `https://www.openstreetmap.org/copyright`.

7. National Highway Traffic Safety Administration. *Vehicle Product Information Catalog (vPIC) VIN Decoder*. URL: `https://vpic.nhtsa.dot.gov/decoder/`.

8. John M. Scanlon, Timothy L. McMurry, Yin-Hsiu Chen, Kristofer D. Kusano, and Trent Victor. "From Stoplights to On-Ramps: A Comprehensive Set of Crash Rate Benchmarks for Freeway and Surface Street ADS Evaluation." In: *SAE International Journal of Transportation Safety* 14.2 (2026). SAE 09-14-02-0003.

9. Federal Highway Administration. *Highway Performance Monitoring System (HPMS) 2023*. U.S. Department of Transportation, Bureau of Transportation Statistics, National Transportation Atlas Database. 2024.

10. Federal Highway Administration. *Highway Statistics 2023, Table VM-4: Distribution of Annual Vehicle Distance Traveled, Percentage by Vehicle Type*. Tech. rep. U.S. Department of Transportation, Federal Highway Administration, 2024.

11. National Centers for Environmental Information. *Integrated Surface Database (ISD)*. National Oceanic and Atmospheric Administration. URL: `https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database`.

12. Yin-Hsiu Chen, John M. Scanlon, Kristofer D. Kusano, Timothy L. McMurry, and Trent Victor. "Dynamic Benchmarks: Spatial and Temporal Alignment for ADS Performance Evaluation." In: *Transportation Research Record: Journal of the Transportation Research Board* (2025). DOI: `10.1177/03611981251398744`.

13. S2 Geometry Developers. *S2 Geometry Library*. 2017. URL: `https://s2geometry.io`.

14. Dennis Luxen and Christian Vetter. "Real-time routing with OpenStreetMap data." In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2011. DOI: 10.1145/2093973.2094062.

15. Texas Department of Transportation. *Crash Records Information System (CRIS) Public Query*. URL: https://cris.dot.state.tx.us/public/Query.

16. Texas Department of Transportation. *TxDOT Roadway Inventory*. URL: https://www.txdot.gov/inside-txdot/division/transportation-planning/roadway-inventory.html.

17. California Highway Patrol. *Statewide Integrated Traffic Records System (SWITRS)*. California Open Data Portal. URL: https://data.ca.gov/dataset/ccrs.

18. California Department of Transportation. *California Public Road Data*. URL: https://dot.ca.gov/programs/research-innovation-system-information/highway-performance-monitoring-system.

19. Arizona Department of Transportation. *Arizona Crash Information (ALISS)*. URL: https://azdot.gov/planning/traffic-safety/arizona-motor-vehicle-crash-facts.

20. Arizona Department of Transportation. *Average Annual Daily Traffic (AADT)*. URL: https://services6.arcgis.com/clPWQMwZfdWn4MQZ/arcgis/rest/services/ADOT_AverageAnnualDailyTraffic_2023/FeatureServer.

21. Nevada Department of Transportation. *Nevada Crash Data*. URL: https://www.dot.nv.gov/.

22. Nevada Department of Transportation, Roadway Systems Division. *Annual Vehicle Miles of Travel (AVMT) Publication, 2023 HPMS Data Year*. URL: https://www.dot.nv.gov/home/showpublisheddocument/22863/638605184064300000.

23. John M. Scanlon, Eric R. Teoh, David G. Kidd, Kristofer D. Kusano, Jonas Bärgman, Geoffrey Chi-Johnston, Luigi Di Lillo, Francesca M. Favarò, Carol A. Flannagan, Henrik Liers, Bonnie Lin, Magdalena Lindman, Shane McLaughlin, Miguel A. Perez, and Trent Victor. "RAVE checklist: Recommendations for overcoming challenges in retrospective safety studies of automated driving systems." In: *Traffic Injury Prevention* 26.5 (2025), pp. 608–621. DOI: 10.1080/15389588.2024.2435620.

Further reading

24. Noah J. Goodall. "Comparability of Driving Automation Crash Databases." In: *Journal of Safety Research* (2025).

25. Jiayu Joyce Chen and Steven E. Shladover. "Initial Indications of Safety of Driverless Automated Driving Systems." In: *arXiv 2403.14648* (2024).

26. Luigi Di Lillo, T. Gode, X. Zhou, John M. Scanlon, R. Chen, and Trent Victor. *Do Autonomous Vehicles Outperform Latest-Generation Human-Driven Vehicles? A Comparison to Waymo's Auto Liability Insurance Claims at 25 Million Miles*. Tech. rep. Waymo LLC, 2024.

27. Mary L. Cummings. *Assessing Readiness of Self-Driving Vehicles*. 2024.

28. Carol A. Flannagan, Andrew Leslie, Raymond Kiefer, Scott Bogard, Geoff Chi-Johnston, Laura Freeman, Rayman Huang, David Walsh, and Antony Joseph. *Establishing a Crash Rate Benchmark Using Large-Scale Naturalistic Human Ridehail Data*. Tech. rep. University of Michigan Transportation Research Institute (UMTRI), 2023.

29. Noah J. Goodall. "Potential Crash Rate Benchmarks for Automated Vehicles." In: *Transportation Research Record: Journal of the Transportation Research Board* (2021). DOI: 10.1177/03611981211009878.

30. Nidhi Kalra and Susan M. Paddock. *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* Tech. rep. RR-1478-RC. RAND Corporation, 2016.