Region-specific highway driving scenarios generation for accelerating automated driving systems validation: A large-language-model assisted framework

Ji Zhou✉, Yongqi Zhao, Arno Eichberger

Cite this article: https://doi.org/10.26599/JICV.2026.9210097

ABSTRACT: The safety and reliability of Automated Driving Systems (ADS) must be validated before large-scale deployment, and scenario-based testing is a promising way to improve validation efficiency and reduce cost. However, unidentified cross-regional differences in driving scenarios force manufacturers to repeat extensive validation when deploying ADS-equipped vehicles in new regions. Quantifying these differences, so that already-validated common scenarios need not be retested, remains insufficiently addressed. This work proposes a 14-dimension highway driving behavioral taxonomy and builds an automated framework to compare naturalistic driving datasets from China and Germany, with every contrast computed within matched traffic states. For two context-dependent dimensions whose closed-form definitions are especially fragile, lane-change aggressiveness and interaction danger, a dual-track evaluation combining a deterministic formula with a Large Language Model (LLM) agent is introduced. Using Cliff's delta, significant and practically meaningful cross-regional differences are identified across multiple dimensions and traffic states, and the dual-track method localizes where the two label sets disagree. A regional differential filter then converts the identified divergences into a library of OpenSCENARIO test fragments that are batch-executable on a Hardware-in-the-Loop (HiL) bench, offering a practical way to cut adaptation-validation effort when transferring an ADS between regions.

KEYWORDS: automated driving systems; highway driving scenarios; cross-regional comparison; safety measures; traffic state; large language models

1 Introduction

Automated Driving Systems (ADS) must be validated against driving scenarios that reflect real traffic in each target region, a requirement codified under the safety-of-the-intended-functionality standard (International Organization for Standardization (ISO), 2022). Naturalistic trajectory datasets make empirical cross-regional comparison possible, but existing work rarely provides a unified, stratified metric set that separates real behavioral differences from traffic-state effects. The indicators used in prior studies are also heterogeneous, mixing proximity to conflict, outcome severity and control style. This study addressed these gaps using two public drone-based highway datasets, one is the "Aerial Dataset for China's Congested Highways & Expressways" so called "AD4CHE" (Zhang et al., 2023) and the other is the German highway drone dataset so called "highD" (Krajewski et al., 2018).

Cross-regional comparisons report measurable Time-Headway (THW) and Time-to-Collision (TTC) differences between regions (Liu and Selpi, 2020), matching broader evidence that driving styles vary across cultures (Q. Cheng et al., 2020; Özkan et al., 2006) and motivating naturalistic calibration of Responsibility-Sensitive-Safety rules (Xu et al., 2021). Yet such work typically covers a narrow metric set and does not stratify by traffic state. Reviews of surrogate criticality metrics further show that Time-to-X and headway measures answer a different question from velocity-difference or required-deceleration measures (Mahmud et al., 2017; Westhofen et al., 2022), while surrogate thresholds themselves shift across traffic phases (Ding et al., 2024; Kerner, 2004; Rehborn et al., 2011). Any cross-regional comparison ignoring this structure risks confounding behavior with traffic state.

Lane-change aggressiveness and interaction-timing risk illustrate a second difficulty. Single-threshold rules on one kinematic channel often disagree with human judgement when context matters. Large Language Models (LLMs) are increasingly used for interpreting driving data (Zhao et al., 2026), and frameworks like Chat2Scenario extract structured scenarios from data (Zhao et al., 2024), but stop at the scenario-authoring interface. Their role as analysts reading trajectory excerpts alongside classical baselines for cross-regional comparison is still unspecified.

Institute of Automotive Engineering, Graz University of Technology, Graz 8010, Austria.

✉ Corresponding author. E-mail: ji.zhou@student.tugraz.at

Received: July 26, 2026; Revised: August 16, 2026; Accepted: August 31, 2026

These gaps are addressed by a complete workflow that runs end to end, from the two recorded trajectory datasets to Hardware-in-the-Loop (HiL) bench test execution (Fig. 1). The datasets are field-aligned and stratified into matched Kerner traffic-state cells (Kerner, 1998), so every cross-regional contrast is read within free-flow, synchronized flow, or congestion rather than across a congestion-heavy and a free-flow-dominated recording, and a 14-dimension taxonomy (spanning proximity vs. severity and car-following, lane-change, conflict, and longitudinal control) is evaluated per stratum. The workflow is built around one analytical core and one downstream stage.

The analytical core is an LLM-assisted behavioral assessment. Twelve of the fourteen dimensions reduce to established closed-form surrogate-safety indicators, reported only as a reproducible baseline. The remaining two, lane-change aggressiveness (D7) and interaction-timing risk (D8), resist any single-threshold rule, because "aggressive" and "dangerous" are multi-factor judgements that a scalar cut-off can only render through a within-data quantile. For these two dimensions an LLM is embedded as a second independent analyst, audited against the formula baseline. The Artificial Intelligence agent, LLM model choice, and audit protocol are elaborated in Section 2.5.

Taken together, the contribution of the proposed framework in this work includes three highlights that the existing components do not provide on their own. The first is a comparison protocol in which every cross-regional driving behavioral contrast is computed inside a matched traffic state, so that driving behavior is never compared across traffic regimes. The second is a dual-track audit that makes the fragility of two threshold-based definitions measurable rather than assumed. The third is an automated path from the resulting effect-size matrix to scenario fragments that execute on a production-grade HiL bench. The downstream stage is operational rather than illustrative. Sections 2.7 and 2.8 take the effect-size matrix to a fragment library that imports and runs in batch on a production-grade HiL bench, which validates the test execution feasibility and not a measurement of the tested ADS capability boundary, a limit examined in (Section 4.4).

The remainder proceeds as follows: the Methodology details the two datasets used and then proceeds through stratification, the brief formula track, the LLM-assisted assessment for D7 and D8, the statistical protocol, and the differential filter, fragment library, and HiL bench that complete the pipeline; the Results report the stratified contrasts and the LLM-versus-formula analysis; the Discussion interprets the cross-regional differences, their region-adaptive-validation consequences, and the limitations; and the Conclusion closes the paper.

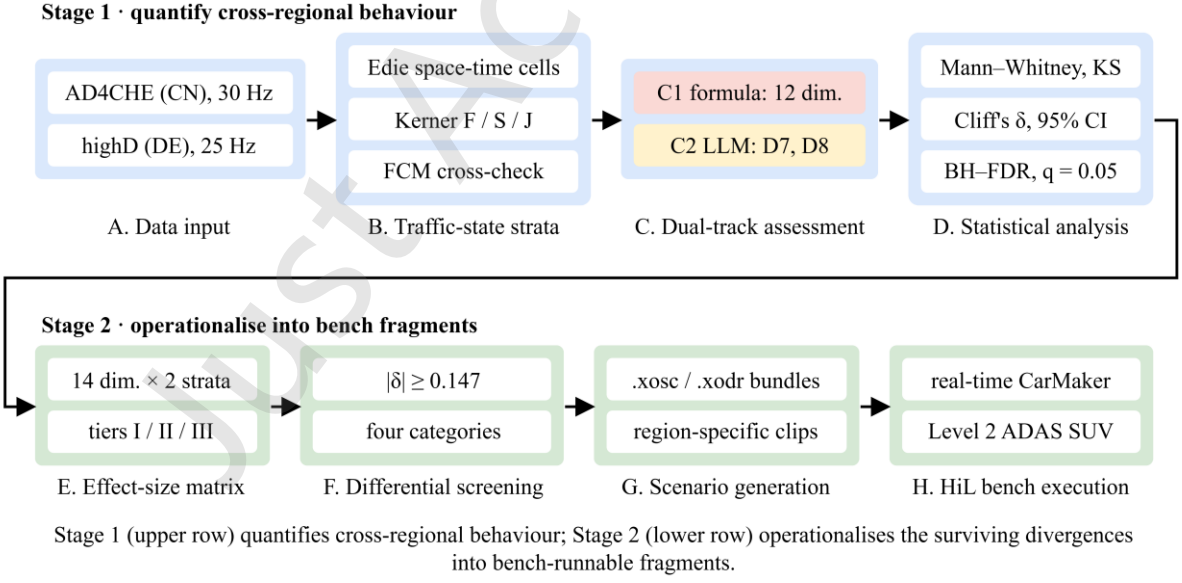


Fig. 1. The complete research workflow as a single input-to-bench chain. Stage 1 quantifies cross-regional driving behavior; Stage 2 turns the surviving divergences of driving scenario into bench-runnable fragments.

2 Methodology

2.1. Datasets, the study input

The input to the entire workflow of Fig. 1 is a pair of drone-based highway trajectory datasets, described first because every later design choice (the dimension set, the stratification grid, the LLM event card) is constrained by what these two datasets share. Several other open-source Chinese trajectory datasets are available, such as SIND for Signalized Intersections (Xu et al., 2022) and the ApolloScape large-scale driving dataset (Huang et al., 2020), but AD4CHE is the only drone-based Chinese dataset whose recording conditions and lane-level field schema line up with highD, so it is the Chinese side of the pair in this

study. Table 1 lists the side-by-side dataset statistics used throughout.

2.1.1 AD4CHE (China)

The AD4CHE dataset (Zhang et al., 2023) contains drone-captured trajectories from congested Chinese expressways; 41 of the 42 released recordings are kept, one being excluded for partial occlusion. Its segments are short but vehicle-dense (Table 1), matching the dataset's intent of capturing stop-and-go and synchronized flow rather than free-flow cruising.

2.1.2 highD (Germany)

The highD dataset (Krajewski et al., 2018) contains naturalistic drone trajectories from six German highway locations; all 60 public recordings are used (Table 1). Its segments are long and predominantly capture free-flowing highway traffic.

Table 1. AD4CHE vs. highD side-by-side dataset statistics

	AD4CHE	highD
Country	China	Germany
Road type	expressway	autobahn
Recording sites	multiple	6 locations
Recordings used	41	60
Total duration (h)	3.20	16.66
Mean recording length (s)	281	999
Frame rate (Hz)	30	25
Total vehicle-km	4,285	44,476
Tracked vehicles	34,807	110,516
Cars / Trucks / Buses (%)	81.4/17.2/1.4	80.7/19.3/0
Dominant traffic state	congested	free-flow
Speed regime (km/h)	0–90	80–130
Lane-marking schema	per-rec. (Zhang et al., 2023)	3–5/dir. (Krajewski et al., 2018)
Inference strata	*S,J*	*S,J*

2.1.3 Scope and comparability

The two datasets match modality (down-looking drone video), trajectory fidelity (sub-meter x/y resolution, lane-level identifiers, per-vehicle class labels), and the lateral-velocity/lateral-acceleration channels that some of the fourteen dimensions (D5 to D9) need. They differ chiefly in dominant traffic state, the confound the stratification protocol of Section 2.3 is designed to remove.

2.2. Comparison framework

The end-to-end pipeline is summarized in Fig. 1. Table 2 organizes the fourteen dimensions (D1–D14) along two axes. Following the criticality-metric reviews of Mahmud et al. (2017) and Westhofen et al. (2022), the metric-semantics axis separates proximity (how far from critical) from severity (how bad the outcome), and the behavioral-scenario axis separates the driving episode into car-following, lane-change, gap acceptance and longitudinal control style; crossing them yields the four categories of Table 2.Each dimension must be computable from the trajectory fields shared by AD4CHE and highD, so indicators needing driver-state, vehicle onboard communication network data, or eye-tracking inputs are excluded by construction. The fourteen dimensions are complementary views of the same driving record rather than statistically independent measurements, and they are not treated as independent here. The per-recording Spearman matrix over the fourteen primary metrics (supplementary material) shows 28 of the 91 distinct pairs at $|\varrho|\geq 0.5$, the strongest reaching $\varrho=+0.90$ between metrics that read the same following geometry through different statistics. The dimensions are therefore reported as a profile of where two regions differ, and the multiplicity this dependence induces is handled by controlling the false discovery rate over the whole family rather than by assuming independent tests.

Their literature anchors are Cat. I car-following proximity (Kesting and Treiber, 2008; Liu and Selpi, 2020; Mahmud et al., 2017; Treiber et al., 2000; Zhao et al., 2024), Cat. II lane-change and lateral interaction (Kesting et al., 2007; Nobukawa et al., 2016; Westhofen et al., 2022; Zhao et al., 2024), Cat. III conflict severity (Westhofen et al., 2022), and Cat. IV longitudinal control and fleet composition, the last including heavy-vehicle lane occupancy as regulated by Germany's truck overtaking ban (Feng et al., 2017; Krajewski et al., 2018; Liu and Selpi, 2020).

Table 2. Core dimensions for cross-regional highway driving comparison

Category	Sub-dimensions	Comparison Goal	Representative Metrics
Following safety margins	D1 TTC; D2 THW; D3 TTC exposure; D4 IDM parameters	Compare routine car-following safety margins	THW/TTC distributions; TTC <1.5s exposure; IDM v_0,T^*,a,b
Lane-change and lateral interaction	D5 cut-in DHW; D6 lane-change frequency; D7 lane-change aggressiveness; D8 PET/ET; D9 gap acceptance	Compare lane-change execution and interaction intensity	Cut-in DHW; lane-changes/km; $A=\lvert v_y\rvert_{\max}\lvert a_y\rvert_{\max}/T_{LC}$ (P75); PET/ET; lead/lag gap
Conflict severity	D10 Δv; D11 RLongA/RA	Compare severity of critical interactions	Δv; required longitudinal deceleration (RLongA/RA)
Longitudinal control and traffic composition	D12 heavy-vehicle leftmost-lane occupancy; D13 acceleration pattern; D14 jerk	Compare control style, fleet mix, and lane-use structure	Heavy-vehicle leftmost-lane occupancy; acceleration distribution; jerk frequency

Taxonomy follows two axes, proximity vs. severity (Mahmud et al., 2017; Westhofen et al., 2022) and behavioural scenario (Kesting et al., 2007; Liu and Selpi, 2020). Full definitions and field mapping for D1–D14 are in Section 2.4.

2.3. Traffic-state stratification

Direct comparison of surrogate safety measures between AD4CHE and highD would confound driving behaviour with traffic state. AD4CHE is congested Chinese expressways (Zhang et al., 2023), highD mostly free-flowing highways (Krajewski et al., 2018), and THW/TTC/headway are governed by traffic state (Ding et al., 2024; Liu and Selpi, 2020), so a single Surrogate Safety Measure (SSM) threshold cannot be applied uniformly across states (Ding et al., 2024). All cross-regional comparisons are therefore conducted within matched traffic-state strata.

Kerner's three-phase scheme (Kerner, 1998, 2004), comprising free flow (F), synchronized flow (S), wide-moving jam (J), is adopted as an operational stratification, not as an endorsement of three-phase over fundamental-diagram models (Coifman, 2015; Treiber et al., 2010). Traffic states follow from Edie's generalized flow/density/speed (Cassidy and Coifman, 1997; Edie, 1963) over 100-m×20-s space–time cells. Cells are then labelled by space-mean speed, with v>80 km/h ⇒ F, 30<v≤80 ⇒ S, v≤30 ⇒ J (Kerner et al., 2004; Schönhof and Helbing, 2007). The 80 km/h boundary aligns with Highway Capacity Manual Levels of Service (Transportation Research Board (TRB), 2016); the 30 km/h boundary matches Kerner's −15 km/h jam-propagation observation (Kerner, 2004). The labels are a speed-based operational proxy. The J stratum is defined by the speed criterion alone, with no test of the spatiotemporal propagation that identifies a wide-moving jam in Kerner's framework, so J is read here as the low-speed stratum rather than as a verified jam identification. The breakpoints are validated by three-cluster fuzzy c-means (FCM) (Bezdek, 1981; Z. Cheng et al., 2020) on the pooled speed–density scatter; if a cluster center deviates from a fixed threshold by more than 10 km/h, the fixed value is retained. A data-driven partition that departs from the textbook value by tens of km/h no longer estimates the phases those values define, and with one dataset per region there is no independent way to establish that the departure is a property of highway traffic rather than of this dataset pair, in which AD4CHE is congestion-oriented and highD free-flow dominated. The published boundaries are therefore retained, and both boundaries were swept from 20 to 50 km/h and from 70 to 100 km/h, that is by up to 20 km/h in either direction. The synchronized stratum clears the 50-cell floor at every setting, and the two strongest contrasts, lane-change rate (D6) and large-jerk share (D14), keep their sign and their significance throughout (δ between +0.86 and +1.00). The jam stratum cannot be swept the same way, since no highD recording is jam-dominant below the +20 km/h setting, so the jam contrast is reported at the adopted cut-offs under the power caveat of Section 3.2. The conclusions therefore do not depend on the particular cut-offs; the full sweep is released in the supplementary material. The cell label is propagated to every vehicle frame via a spatial–temporal join. A stratum holding fewer than 50 cells (about 17 min) on either side is data-insufficient and excluded from cross-regional inference; that floor is set so Mann–Whitney U attains 1−β≥0.80 at medium effect size ($\lvert\delta\rvert$≈0.33, α=0.05). All 14 dimensions are computed per stratum (a 14×2(country)×3(state) matrix) and tested only between matched strata (Section 2.6), which excludes the AD4CHE free-flow stratum.

2.4. Deterministic metric definitions (formula track)

This track is standard. The twelve indicators are published with metrics reused unchanged, and their role here is to be a byte-reproducible baseline against which the LLM track of Section 2.5 can be audited. Twelve of the fourteen dimensions reduce to a textbook indicator (Mahmud et al., 2017; Westhofen et al., 2022), summarized in Table 3; output statistics, data-field anchors, and the lane-change detector

are released in the supplementary material. The remaining two (D7, lane-change aggressiveness; D8, interaction-timing risk) carry a formula baseline for comparability only and receive the dual-track LLM treatment of Section 2.5. Time-derivative metrics are frame-rate-normalized to reconcile AD4CHE (30 Hz) and highD (25 Hz), and a lane-change is registered when `lane Id` switches and holds for ≥1.0 s.

Table 3. Primary metric per dimension. (D7 and D8 additionally receive the dual-track LLM treatment of Section 2.5)

Dim.	Dimension	Primary metric
D1	Car-following Time-to-Collision (TTC)	$TTC=DHW/\Delta v$ (Hayward, 1972)
D2	Time Headway (THW)	$THW=DHW/v_{ego}$ (Vogel, 2003)
D3	TTC exposure	frac. frames TTC<1.5s
D4	Intelligent Driver Model (IDM) calibration	LSQ fit of (v_0,T^*,s_0,a,b) (Kesting and Treiber, 2008)
D5	Cut-in DHW	Headway at Lane-Change (LC) completion into ego lane
D6	LC frequency	changes per veh-km
D7	LC aggressiveness	$A=\lvert v_y\rvert_{max}\lvert a_y\rvert_{max}/T_{LC}$, P75 cut
D8	PET (Post-Encroachment Time)	$PET=t_2-t_1$ (Allen et al., 1978)
D9	Gap acceptance	lead/lag gap at LC initiation
D10	Speed differential	$\Delta v=v_{ego}-v_{pre}$
D11	Required deceleration	$(\Delta v)^2/(2DHW)$, $\Delta v>0$ (Westhofen et al., 2022)
D12	Truck left-lane occ.	frac. frames heavy veh. in leftmost lane
D13	Acceleration	distribution of a_x
D14	Longitudinal jerk	$j=da_x/dt$

2.5. *LLM-assisted driving behavioral assessment*

A methodological core of this work is the use of a large language model (LLM) as a second, independent analyst that reads driving-scenario evidence and labels behavioral difference where a closed-form threshold is demonstrably unsafe. The full assessment agent is shown in Fig. 2 and a single event worked through it in Fig. 3. The argument of this subsection has four parts: why a formula cannot do the job for D7 and D8 (Section 2.5.1), which model is used and why (Section 2.5.2), how the agent is structured so its output stays auditable against the formula (Section 2.5.3), and what the LLM track therefore contributes that the formula cannot (Section 2.5.4).

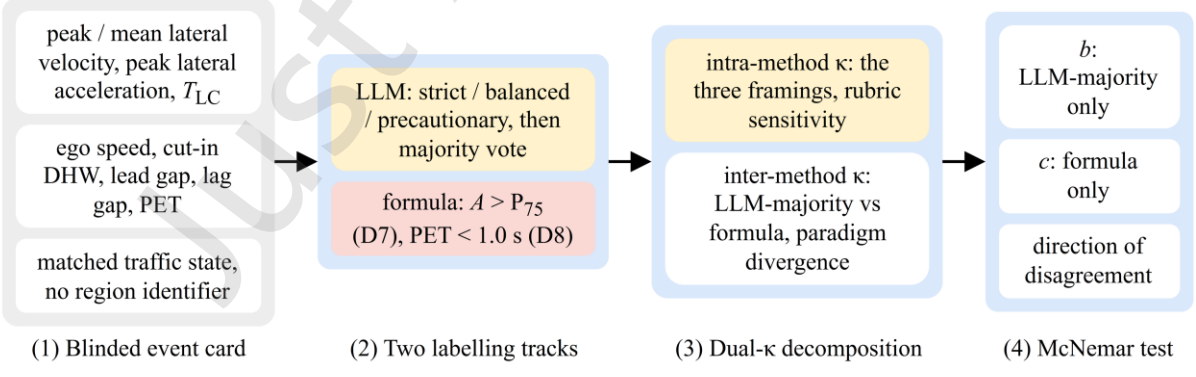


Fig. 2. The LLM assessment agent. One structured event card is the sole observation shared by both tracks. Details in Section 2.5.

2.5.1 Why an LLM is needed where formulas fail

D7 (lane-change aggressiveness) and D8 (interaction-timing risk) resist reduction to a single threshold because the judgement is multi-factorial. Being "aggressive" depends jointly on lateral velocity, accepted gap, density, ego speed and vehicle type, and whether a given Post-Encroachment Time (PET) is "dangerous" depends on closing speed, conflict geometry and traffic state. A supervised alternative is not available either. Both conventional machine learning and deep learning require a training signal, and these are exactly the two dimensions for which no ground-truth label exists, so the formula baseline is retained as the closest available substitute. A scalar cut-off on one channel must

ignore the rest, and this failure is demonstrable rather than theoretical. Published single-feature aggressiveness thresholds (Jahangiri et al., 2018; Satzoda et al., 2015) target SHRP2 emergency events and produce *zero* hits on naturalistic highway data (Westhofen et al., 2022). The composite adopted for the D7 formula baseline, Eq. (1),

$$A = \frac{|v_y|_{\max}\,|a_y|_{\max}}{T_{\mathrm{LC}}}, \tag{1}$$

already collapses three independent channels, peak lateral velocity, peak lateral acceleration, and maneuver duration (time to line crossing, TLC), into one scalar, and obtaining a non-empty positive class from it still requires a within-data quantile cut (overall percentile 75 "P75" used here), which binds the definition of "aggressive" to the joint AD4CHE–highD data distribution and so creates a spurious contrast from a sampling artefact. An LLM scores the full event context against explicit verbal rubric and is, by construction, free of that quantile binding. The other twelve dimensions have stable single-channel definitions and stay on the formula track, so the LLM is applied selectively, with its added cost and opacity incurred only where justified.

2.5.2 Model selection

What the LLM track claims is that a language model can act as a second, auditable analyst inside this workflow for the two dimensions where a scalar threshold demonstrably fails; no result here depends on one model being more capable than another, and none is offered as a model comparison. The assessor is the Databricks-hosted databricks-claude-opus-4-7 endpoint. Two properties of the endpoint matter to the design, and neither is a statement about its relative capability. It sits inside a protected workspace that holds the trajectories, so every call stays in the governed environment the AD4CHE and highD licenses require, and its interface takes no temperature or seed argument, so the rubric perturbation of Section 2.5.3 probes the stability of the concept rather than of the decoder. The model is held fixed across all events and both dimensions, so any measured instability is attributable to the rubric, not a model change.

2.5.3 LLM assessment agent

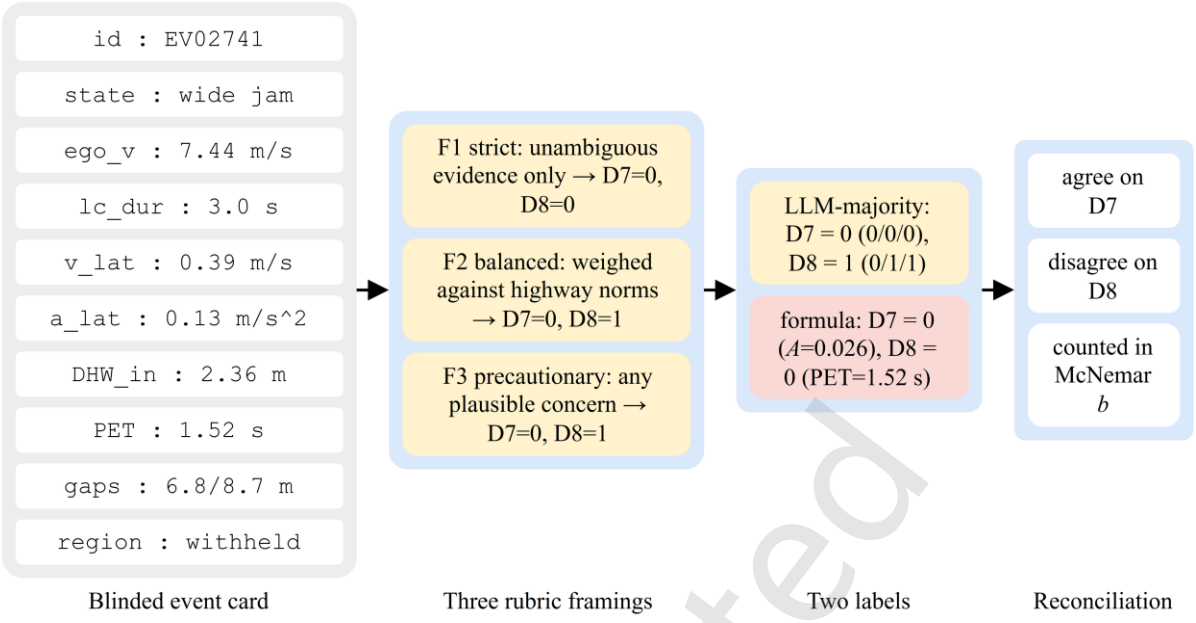
The LLM is embedded in the four-part assessment agent of Fig. 2, so that its output is auditable against the formula and reproducible up to model availability. *(1) Information alignment via an event card.* Each lane-change or cut-in event is compressed into a structured JavaScript Object Notation (JSON) formatted card carrying the lateral-motion, gap, speed and Post-Encroachment Time (PET) fields of the maneuver together with the matched traffic-state tag, tabulated in full in the supplementary material; the card carries no dataset or region identifier, so the label cannot be conditioned on provenance. (2) Three rubric framings plus majority vote. Because the fixed-decoding endpoint cannot be resampled, run-to-run variation is induced deliberately by scoring every card three times, under strict, balanced and precautionary framings. They differ only in the evidential bar they set, not in the factors they read; their cut-off values and verbatim text are released in the supplementary material. The reported label for event e is the majority vote over the three framings, Eq. (2), where $y_e(r) \in \{0,1\}$ is the label under framing $r \in \{\text{strict, bal, prec}\}$, so no single prior can override it. *(3) Dual-κ decomposition.* Intra-method κ is computed for each of the three framing pairs and averaged, and probes rubric sensitivity (a low κ flags an operationally unstable concept, interpreted in Section 4.2); inter-method κ between the formula and the LLM-majority quantifies divergence between the two paradigms. Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) (Byrt et al., 1993) is reported alongside κ against prevalence imbalance. *(4) Directionality via McNemar* (McNemar, 1947). McNemar's exact test on the matched 2×2 discordance table tests whether one track systematically over- or under-flags relative to the other rather than disagreeing at random.

$$\hat{y}_e = \mathbf{1}\left[\sum_r y_e^{(r)} \geq 2\right] \tag{2}$$

2.5.4 What the LLM track contributes over the formula

The two tracks answer the same binary question on the same evidence, so they cross-validate. Agreement raises confidence that a contrast is real, and disagreement localizes where a scalar definition breaks. The LLM removes the quantile binding of Section 2.5.1 and responds to context that no single kinematic channel carries, the closing-in-behind pattern of Fig. 3; both effects are quantified with their evidence in Section 3.3. The cost is a human-authored, prior-sensitive rubric, and the dual-κ design measures that sensitivity, so the LLM is an auditable independent check, not ground truth.

Replication. The dual-track labelling comprises 4,075×3=12,225 calls to the databricks-claude-opus-4-7 endpoint. The released label set is produced from event cards carrying no dataset or region identifier, so no label can be conditioned on provenance; the blinded label file, the raw responses and the identifier map are in the supplementary material. Bit-for-bit reproduction is tied to model availability, but family-level replication (McNemar b/c direction, κ range, sign of the cross-regional δ shift) should survive model drift because the byte-reproducible formula baseline anchors the comparison. Prompt templates, card schema, seed table, and raw responses are archived in the supplementary material.



An AD4CHE wide-moving jam cut-in from the 4,075-event pool: the formula rules it safe on PET, while two of the three framings flag it on the 2.36 m gap at 7.4 m/s, so the majority vote and the formula disagree on D8.

Fig. 3. One event worked through the agent of Fig. 2. A wide-moving-jam cut-in that the formula rules safe on PET (PET = 1.52 s) is flagged by the rubric on a 2.36 m gap at 7.4 m/s.

2.6. Statistical testing protocol

Each cross-regional comparison is conducted within one stratum $s \in \{F,S,J\}$ and one dimension D_i; strata failing N_{min}=50 are directional observations only. The primary test is the two-sided Mann–Whitney U (Mann and Whitney, 1947) at α=0.05, appropriate because most SSMs are skewed and heavy-tailed (Liu and Selpi, 2020; Westhofen et al., 2022); the two-sample Kolmogorov–Smirnov statistic (Massey, 1951) is reported alongside to separate location shifts from tail redistribution. Effect size is **Cliff's** δ (Cliff, 1993), given by Eq. (3),

$$\delta = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left[\operatorname{sgn}(x_i - y_j)\right], \tag{3}$$

the normalized difference between the probabilities that an AD4CHE sample exceeds a highD one and vice versa ($|\delta|$<0.147 negligible, <0.33 small, <0.474 medium, ≥0.474 large), with a 95% bootstrap Confidence Interval (CI) from stratified percentile resampling, at 10,000 iterations for the dimension family below and 2,000 for the D7/D8 label-sensitivity contrast (Efron and Tibshirani, 1993). With the free-flow stratum excluded under the N_{min}=50 floor (Section 2.3), the metric×stratum family is controlled for the False Discovery Rate (FDR) by the Benjamini–Hochberg (BH) procedure at q=0.05. The family comprises 108 tests. The fourteen dimensions contribute between two and six candidate metrics each, and every metric is tested in both admissible strata, so the count exceeds the fourteen headline contrasts reported in the summary matrix (Benjamini and Hochberg, 1995); Bonferroni is reported for transparency but is overly conservative under heterogeneous effects (Feise, 2002).

A result is declared robust only when (a) raw p<0.05, (b) BH-adjusted p*<0.05, and (c) the 95% bootstrap CI of δ excludes (−0.147,+0.147); (a)- or (a)+(b)-only results are flagged "marginal". Two units must be distinguished. The observation unit is the level at which the raw quantity exists, which is the vehicle-frame for D1–D4, D10–D11 and D13–D14, the lane-change event for D5–D9, and the vehicle for D12. The testing unit is the recording. Each observation set is aggregated to one value per recording per stratum before the Mann–Whitney test, so repeated frames or events within a recording are not treated as independent observations. Where a stratum holds too few events for that aggregation to be meaningful, a pooled event-level test is used instead and is flagged as such. The observation unit, the testing unit, the sample size per traffic state, and the use of the pooled fallback are tabulated per dimension in the supplementary material, which also carries the per-event follower deceleration used as the external criterion in Section 4.2.

For D7 and D8, three agreement statistics are reported, Eqs. (4)–(6), on the matched 2×2 table of formula vs. LLM-majority labels, with discordant off-diagonal counts b and c:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

$$\mathrm{PABAK} = 2p_o - 1 \tag{5}$$

$$\chi^2_{\mathrm{McN}} = \frac{(b-c)^2}{b+c}, \tag{6}$$

where p_o is observed and p_e chance agreement. Cohen's κ (Cohen, 1960) is computed both **intra-method** (across the three rubric framings, probing rubric sensitivity) and **inter-method** (formula vs. LLM-majority); the same coefficient is the standard measure for LLM-versus-human annotation agreement (Nasution and Onan, 2024). PABAK (Byrt et al., 1993) guards against prevalence imbalance, and McNemar's exact test (McNemar, 1947), the established test for comparing two classifiers on a matched sample (Dietterich, 1998; Rainio et al., 2024), resolves the *direction* of disagreement. The cross-regional Mann–Whitney is re-run on both label paradigms as a sensitivity analysis.

2.7. *Differential filter and fragment library*

The stratified Cliff's δ matrix of Section 2.6 is the starting point for a fragment library. A differential filter screens the matrix by the Cliff (1993) negligibility band ($|\delta_{i,s}| \geq 0.147$), keeping only divergences that clear the robustness rule of Section 2.6 and partitioning them into the four behavioural categories of Table 2. Each surviving divergence is then routed back into the real AD4CHE or highD trajectories at its native granularity (event, lane-change window or Edie cell), most divergent first, and exported as OpenSCENARIO replays (ASAM e.V., 2020) over a shared OpenDRIVE network (ASAM e.V., 2021), with ego, neighbors and ambient vehicles driven by FollowTrajectoryAction from recorded kinematics. Each fragment ships with the catalogue, track slice and manifest a bench needs to replay it.

Each fragment is emittable with the ego as either the VICTIM (follower/yielding vehicle, via the recorded `followingId`) or the ACTOR (manoeuvring vehicle), covering both the "can the ADS avoid this" and "will the ADS reproduce this" regression questions. The pipeline is implemented and produces the bundle, whose concreteness on a production-grade bench is established in Section 2.8. Closing a region-tuned controller and reactive ambient traffic onto this replay is future work (Section 4.4).

2.8. *Bench-replay evaluation of Stage 1 fragments*

The bench step establishes that the Stage 1 fragment library is concrete. Fragments spanning the China- and Germany-side seeds and the four behavioural categories are replayed on a vehicle Electrical & Electronic System integration HiL test bench (Fig. 4) whose ego is a CarMaker (IPG Automotive GmbH, 2024) vehicle-dynamics model on a real-time host, instrumented for a physically real perception path (Fig. 4) so a production Electronic Control Unit (ECU) can later be closed onto the ego.**Fig. 5** These runs exercise the import-and-execute path of this instrument, not that ECU loop. Each fragment is driven open-loop through the `FollowTrajectoryAction` replay of Section 2.7, with the ego under a reference longitudinal controller.4.4

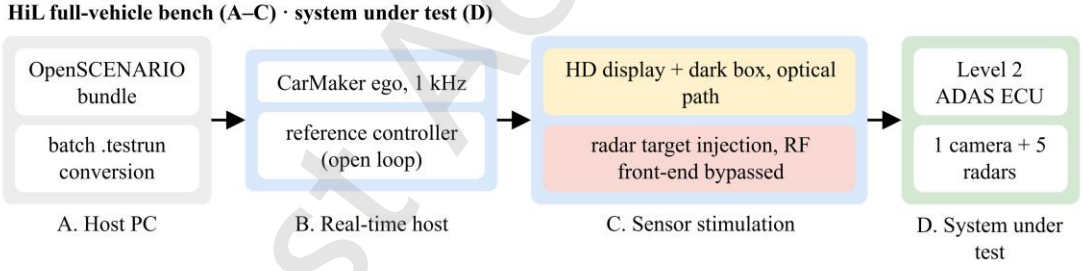


Fig. 4. Hardware-in-the-Loop (HiL) bench for replaying Stage 1 fragments. The conversion path, the stimulation paths and the system under test are described in Section 2.8.

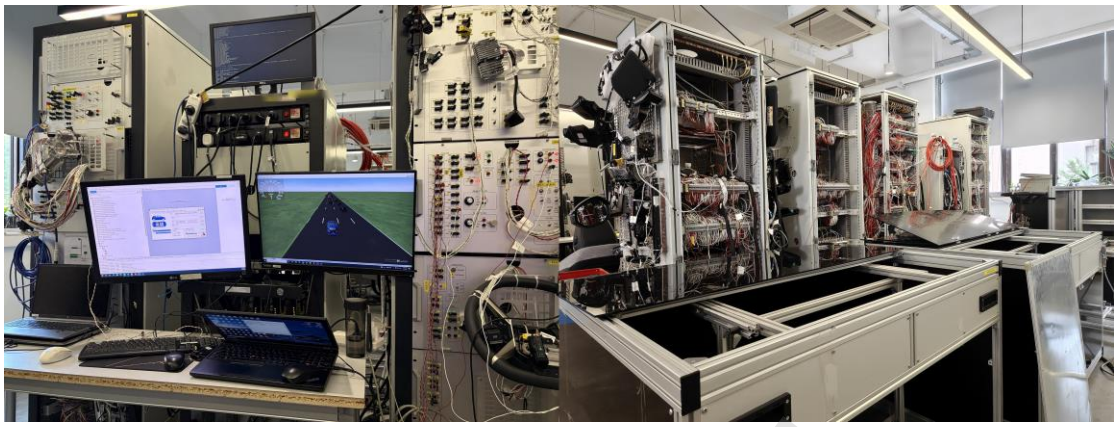


Fig. 5. Physical realization of the HiL bench of Fig. 4: the operator station (left) and the full HiL bench (right).

The exported OpenSCENARIO fragments are converted to the bench's proprietary `.testrun` format in batch through the CarMaker command-line interface, so the whole library is staged for real-time execution in one pass without per-scenario manual import; before a run the same bundle can also be previewed in esmini (Knabe et al., 2024) without occupying the bench. Each run logs a continuous multi-channel measured trace (ego kinematics, longitudinal deceleration, brake-actuation pressure, and front-object range) at 1 kHz, read directly from the `.dat` channels. The import-and-execute readout of this protocol on a six-fragment batch is reported in Section 3.5.

3 Results

3.1. Stratum-level sample adequacy

The three-cluster FCM validation of the Kerner breakpoints (Section 2.3) returns data-driven boundaries of 66.1 and 107.4 km/h on the pooled (k,v) scatter (m=2, `seed`=42), both deviating from the fixed 30/80 thresholds by more than the 10 km/h tolerance. The AD4CHE-jam / highD-free skew under-populates the low-speed cluster and pulls both upper centres above 100 km/h. The tolerance rule therefore rejects this output and the fixed 30/80 km/h Kerner thresholds (Edie, 1963; Kerner, 1998, 2004; Schönhof and Helbing, 2007; TRB, 2016) are retained.**Fig. 6**

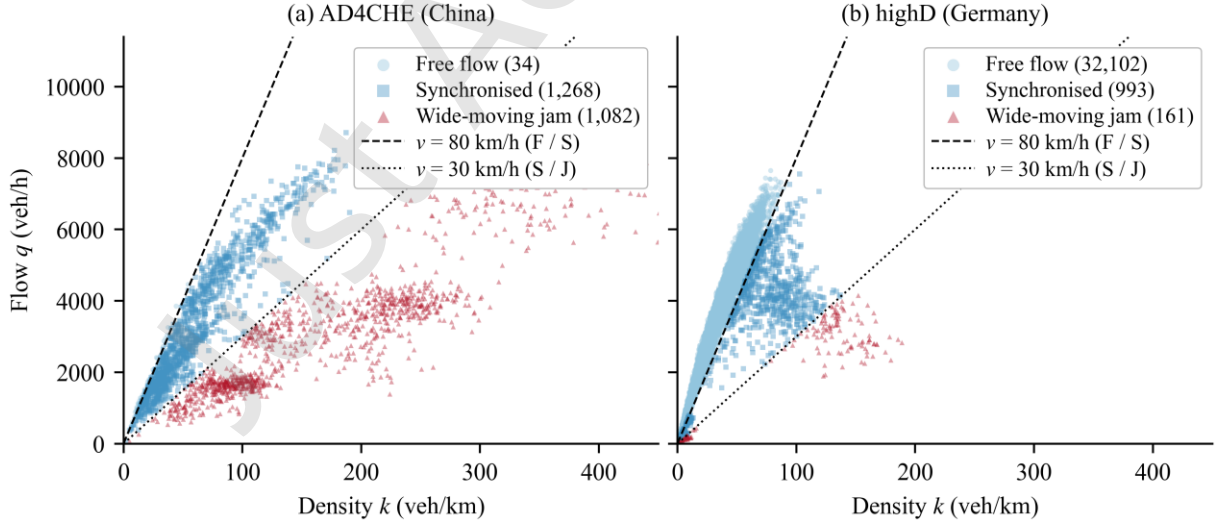


Fig. 6. Flow–density scatter of 20 s Edie cells coloured by three-phase label. Dashed and dotted rays mark the fixed speed cut-offs $v_{F/S}$=80 km/h and $v_{S/J}$=30 km/h.

Table 4. Cell counts per traffic-state phase after stratification (100 m × 20 s cells; fixed 30/80 km/h thresholds).

Phase	AD4CHE cells	highD cells	Total
F (free-flow)	34	32,102	32,136
S (synchronised flow)	1,268	993	2,261
J (wide-moving jam)	1,082	161	1,243
Total	2,384	33,256	35,640

The per-stratum cell counts (Table 4) confirm the complementary skew (AD4CHE concentrates in *S* and *J*, highD overwhelmingly in *F* with sparse *S*/*J*) and set the inference eligibility. AD4CHE contributes only 34 free-flow cells (a consequence of the congestion-oriented recording design (Zhang et al., 2023)), below the N_{min}=50 floor (Section 2.3). The *F* stratum is therefore dropped from country-level inference, and hypothesis testing is limited to the synchronised and wide-moving jam strata, where both sides have enough cells.

3.2. Stratified contrasts, D1–D14

For every dimension the observations are aggregated to the recording level within each stratum, so that each highway recording contributes one point per indicator per stratum (Section 2.6), and the two datasets are then compared on those recording-level values. With the free-flow stratum excluded (Section 3.1), inference runs over the two admissible strata. Per-dimension effect sizes are consolidated into the summary matrix of Section 3.4, and the full table is released as a companion file (`stratified_stats.csv`). Positive δ indicates AD4CHE exceeds highD on the stated metric.

Statistical-power caveat. The wide-moving-jam stratum is highD-sparse, so the Mann–Whitney null is coarse and the bootstrap envelopes on Cliff's δ are wide. Effects that fail q<0.05 in jams are therefore read as "inconclusive under this design" rather than as null. For the event-sparse lane-change dimensions (D5, D7, D8, D9) the recording-level jam cell cannot be tested and falls back to the pooled event-level Mann–Whitney of Section 2.6, marked with a dagger (†) in the summary matrix of Section 3.4. A leave-one-recording-out check (loo_jam_results.csv) leaves the main J-cells stable. Dropping either highD recording behind the event-level D7/D8 contrast (Table 5) preserves the sign and significance of the LLM-label effect under the reported labelling, so no main conclusion rests on a single recording.

The prose below states only the main direction of each category and the few anchor magnitudes that support the main findings; the full effect-size matrix is in Section 3.4 and the companion .csv file.

Car-following proximity (D1–D4). Once traffic state is matched, AD4CHE follows more tightly at the risk tails (shorter minimum TTC in synchronized flow and higher TTC-hazard exposure in jams), while its preferred steady-state spacing (THW, the Intelligent Driver Model (IDM) desired-gap proxy) is statistically indistinguishable from highD. Routine spacing therefore tracks the traffic stratum, not the country; the regional difference is in the tails, not the set-point.

Table 5. Stratified D7/D8 cross-regional contrast under formula vs. LLM-majority labels (4,075-event pool). n_{CN}=1,358/2,385 and n_{DE}=302/30 in *S*/*J*. Positive δ = AD4CHE exceeds; 95% CI from 2,000-resample bootstrap.

Dim.	Stratum	Label	Positive CN	Positive DE	δ [95% CI]	p
D7	*S*	formula	0.057	0.318	−0.260 [−0.314,−0.207]	$<10^{-4}$
D7	*S*	LLM-majority	0.022	0.050	−0.028 [−0.054,−0.003]	7.6×10^{-3}
D7	*J*	formula	0.005	0.333	−0.329 [−0.526,−0.162]	$<10^{-4}$
D7	*J*	LLM-majority	0.008	0.167	−0.159 [−0.294,−0.028]	$<10^{-4}$
D8	*S*	formula	0.127	0.126	+0.002 [−0.041,+0.044]	0.94
D8	*S*	LLM-majority	0.151	0.235	−0.084 [−0.137,−0.032]	4×10^{-4}
D8	*J*	formula	0.003	0.000	+0.003 [+0.001,+0.005]	0.78
D8	*J*	LLM-majority	0.018	0.133	−0.116 [−0.250,−0.014]	$<10^{-4}$

Lane-change and interaction behavior (D5–D9). Lane-change rate (D6) is the strongest and most consistent contrast in the study. AD4CHE changes lanes roughly an order of magnitude more often than highD per vehicle-kilometer in both admissible strata (δ≈+1.0), and the gap survives per-recording veh-km normalization, so it is not a road-length artefact. Lane count is recorded for highD but not for AD4CHE, so the layout question is answered within highD. Its 4-, 6- and 7-lane recordings all sit an order of magnitude below the AD4CHE median on this metric (0.00 to 0.17 against 1.38 lane changes per vehicle-kilometer), against a cross-regional δ of +0.98, so three lanes of difference inside one country do not approach the cross-regional gap; the full check is in the supplementary material. Yet the individual merges are more conservative, not less. In jams AD4CHE carries longer PET (D8) and larger selected gaps (D9), so the regional pattern is frequently merging with larger accepted gaps. The binary-threshold "German lane-changes are more

aggressive" contrast does not reproduce under the continuous per-event metric (D7). It is a highD-side P75-saturation artefact, the failure the LLM audit of Section 3.3 resolves.

Longitudinal control style (D10–D14). Under congestion AD4CHE platoons run more uniformly (lower speed differential and peak acceleration) yet demand a markedly higher required-deceleration reserve in synchronized flow, consistent with a tighter, more reactive following style. Truck left-lane occupancy (D12) is near-zero for AD4CHE under Chinese lane-use rule. The most notable single result is the large-jerk share (D14), where Cliff's δ saturates at +1.0 in both strata. AD4CHE runs at an order-of-magnitude higher jerk density than highD under matched flow. This contrast is not a frame-rate artefact. Resampling AD4CHE from 30 to highD's 25 Hz before the finite-difference jerk estimate leaves the large-jerk share essentially unchanged.

Taken together, once traffic state is controlled, AD4CHE drivers follow more tightly at the risk tails, change lanes far more frequently but individually more conservatively, apply harder reactive braking in synchronized flow, and run at systematically higher jerk density, while their preferred steady-state spacing is essentially the German one. Two dimensions ($\delta \approx +1.0$ throughout, lane-change rate and jerk density) account for the main cross-regional gap; the rest are either flow-state-modulated or null, a structure developed tier-by-tier in Section 4.1.

3.3. *LLM vs. formula, D7 and D8 in depth*

Event pool. The dual-track instantiated in Section 2.5 runs on the full formula-valid pool of 4,075 lane-change events, China-heavy by construction (per-stratum counts in Table 5), under the formula and LLM-majority labels defined there (Gilardi et al., 2023; Zheng et al., 2023).

Reliability. The agreement statistics (Fig. 7(a), Fig. 8) split the two dimensions sharply. On D7 intra-LLM agreement is only slight (Landis and Koch, 1977), the rubrics rarely agreeing on what counts as "aggressive," whereas on D8 it rises to fair. Inter-method agreement is high on both dimensions (percent agreement 95.1%/96.3%), yet the McNemar test is significant in *opposite* directions (both exact $p<10^{-17}$). On D7 the formula flags markedly more events aggressive than the LLM, while on D8 the LLM-majority label flags substantially more events dangerous than the formula. Relabeling the full pool from cards stripped of every dataset and region identifier reproduces 98.8% of the D7 labels and 98.2% of the D8 labels and preserves the sign of all eight effects of Table 5, so the comparison is not carried by provenance cues (supplementary material). The figures reported here are those of the original labelling; the blinded run is released as a consistency check, and under it the D8 jam-stratum contrast falls below significance while the D7 jam contrast remains significant on the 30-event highD jam cell.

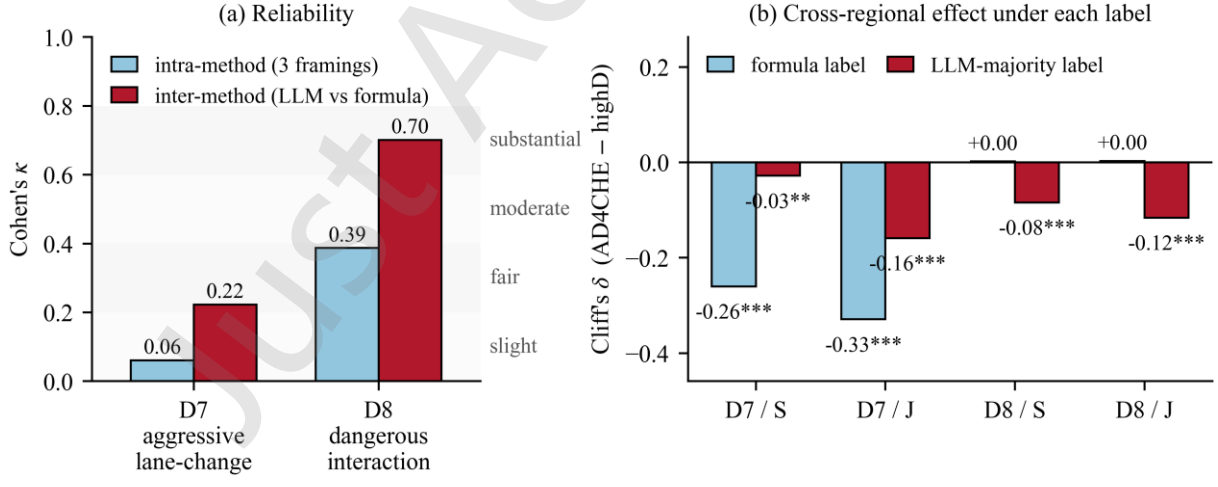


Fig. 7. D7/D8 dual-track summary (4,075-event pool). *(a)* intra-LLM and inter-method κ against Landis–Koch bands; *(b)* cross-regional Cliff's δ per stratum under formula vs. LLM-majority labels (*$p<0.05$, **$p<0.01$, ***$p<0.001$).

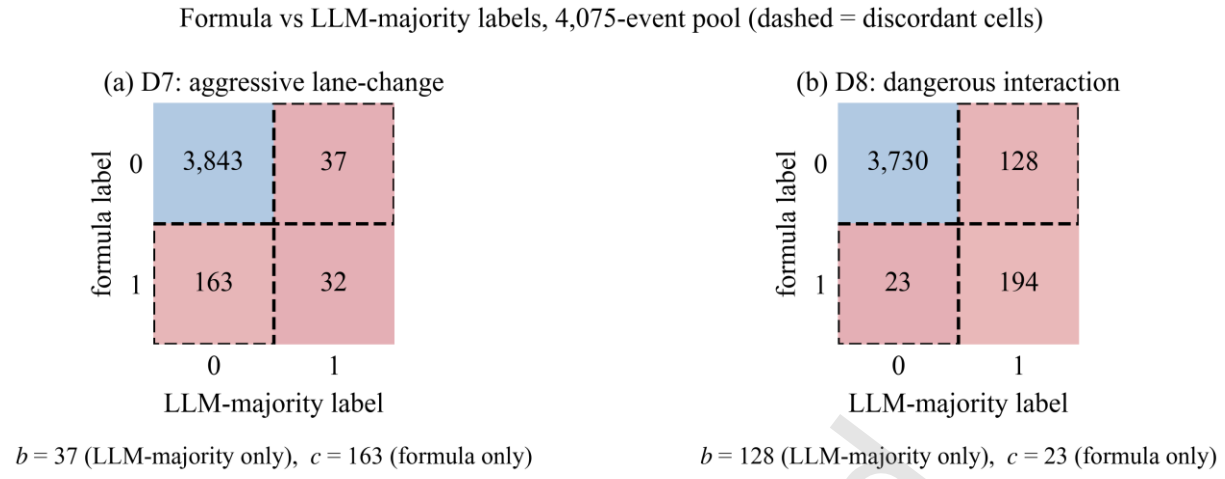


Fig. 8. Matched formula-vs-LLM-majority 2×2 tables (off-diagonal in red = discordance). The McNemar asymmetry (*b* LLM-only, *c* formula-only) flips direction between D7 (formula over-flags) and D8 (formula under-flags).

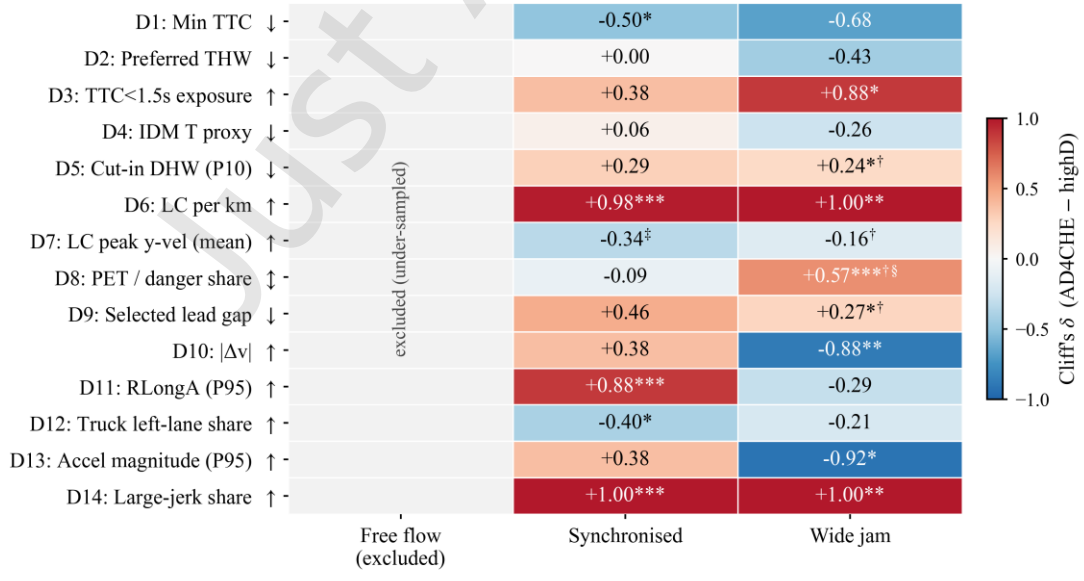
The disagreement carries a clear country signature. The formula-only-aggressive events on D7 are predominantly highD, whereas the LLM-only-dangerous events on D8 are predominantly AD4CHE and follow a closing-in-behind pattern (small lag gap at high lag speed).

Sensitivity of the cross-regional effect under the LLM labels. Table 5 and Fig. 7(b) repeat the per-stratum AD4CHE-vs-highD contrast under both label paradigms. On D7 both labels agree that highD exceeds AD4CHE, but the LLM shrinks the formula effect; on D8 the two diverge on the inference itself, the formula reading null in both strata while a significant highD-exceeds-AD4CHE effect is present under the LLM-majority label. The operational reading of this split is developed in Section 4.2.

In the main matrix (Fig. 9), D7 and D8 are reported under the formula for cross-dimension comparability, with both cells annotated by the dual-track finding of Table 5; the operational reading of these two cells is developed in Section 4.2.

3.4. *Cross-dimensional summary matrix*

Fig. 9 consolidates the fourteen dimensions into a single effect-size matrix (rows D1–D14, columns the Kerner strata *F*/*S*/*J*, each cell the primary-metric Cliff's δ with the star and dagger codes defined in the caption). The three-tier split it reveals (stratum-invariant, flow-state-modulated, no-shift) and its consequences for region-adaptive validation are developed in Discussion Section 4.1–Section 4.3.



↑ larger value = higher risk; ↓ larger value = lower risk; ↕ the two strata use quantities of opposite polarity.
† pooled event-level test. ‡ formula effect not reproduced under the LLM rubric. § pooled PET in seconds, so its sign convention is inverted.

Fig. 9. Cross-regional effect-size matrix: Cliff's δ (AD4CHE − highD) per dimension within the synchronized (S) and wide-moving-jam (J) strata; positive means AD4CHE exceeds highD. Free flow is excluded (Section 3.1).

3.5. Bench-replay feasibility of Stage 1 fragments

Table 6 reports six replays staged through the HiL protocol of Section 2.8, spanning both regional seeds and the conflict and lane-change categories with entry speeds from 15.7 to 68.9 km/h. All six pass schema validation against the OpenSCENARIO importer, then import, convert to `.testrun`, and execute to completion on the real-time host. Three of the six (runs 001/004/008) replay one source fragment under an identical bench configuration; their $a_{\min}$ spread of −2.26 to −2.49 m/s² (*coefficient of variation* ≈5%) bounds run-to-run repeatability.4.4

Table 6. Per-run execution readout for the six replayed Stage 1 fragments. Values are read directly from the logged channels.

Run	Seed	Cat./str.	v_0 (km/h)	dur. (s)	$a_{\min}$ (m/s²)	$p_{brk}^{\max}$ (bar)	$R_{\min}$ (m)
001	CN	conflict/*J*	27.7	20.5	−2.26	19.8	62.7
002	CN	conflict/*J*	17.8	20.5	−2.30	19.9	49.0
004	CN	conflict/*J*	27.3	20.3	−2.37	17.2	62.5
005	CN	lane-change/*S*	15.7	20.5	−3.59	8.8	97.0
006	DE	lane-change/*S*	68.9	17.5	−0.79	0.0	3.1
008	CN	conflict/*J*	27.3	20.5	−2.49	20.4	62.7

Note: runs 001, 004 and 008 replay one source fragment under an identical bench configuration, which bounds run-to-run repeatability. Symbols: v_0 entry speed, a_{min} peak longitudinal deceleration, p^{brk} peak brake-actuation pressure, R_{min} minimum front-object range.

The region-paired form of these replays is illustrated in Fig. 10. For each of the three typical maneuvers, the AD4CHE-extracted and traffic-state-matched highD-extracted OpenSCENARIO fragments are executed on the CarMaker bench, so a single row contrasts the two regions on the same maneuver under a matched traffic state.

Turning this feasibility readout into a region-resolved behavioral verdict requires the closed-loop production-ECU path and a controlled key-performance-indicator campaign, deferred to future work (Section 4.4).

4 Discussion

4.1. Interpretation of cross-regional differences

The matrix of Fig. 9 splits the fourteen indicators into three behaviorally distinct tiers, each carrying a different validation consequence.

Tier I. Invariant-of-stratum national shift. Lane-change rate (D6) and large-jerk share (D14) show a significant AD4CHE-vs-highD effect in both synchronized and wide-moving-jam strata ($\delta \approx +1.0$, $q < 10^{-3}$; D6 robust to the veh-km normalization, D14 robust to the 25-Hz resampling check of Section 3.2). AD4CHE thus runs at a higher lateral and longitudinal rate than highD independent of local traffic state. Cultural, regulatory, and infrastructural factors (Q. Cheng et al., 2020; Özkan et al., 2006; Rehborn et al., 2011) may contribute, but the design isolates a stable effect, not a stable cause.

Tier II. Flow-state-modulated effects. Minimum TTC (D1) and required-deceleration tail (D11) shift only in synchronized flow, while TTC-hazard exposure (D3), speed differential (D10), and peak-acceleration tail (D13) shift only in jams. The split is coherent. Synchronized flow is where car-following discipline dominates, so a tighter AD4CHE style appears as shorter tail TTC and higher braking reserve, whereas jams compress inter-vehicle kinematics into a narrow slow band, leaving only the occasional hard acceleration or closing maneuver as a differential, which is precisely why traffic state is not pooled. The event-level D5/D8/D9 jam results agree. AD4CHE merges are more frequent yet individually target *larger* gaps and *longer* PET, offsetting higher lateral activity with more conservative per-event margins.

Tier III. No national shift once stratified. Preferred THW (D2), the IDM *T* proxy (D4), and D12 in jams show no FDR-significant shift. Matched-state steady-state spacing is the same on both sides, so a longitudinal controller's *set-point* needs no regional retuning; adaptation lives in the *tails* (Tier II) and the *event rate* (Tier I). D7 sits between tiers. The continuous per-event peak is not FDR-significant, but the LLM-audited binary label (Section 4.2) recovers a moderate highD-exceeds-AD4CHE jam effect while erasing the synchronized-flow effect, so D7 reads as Tier-II in jams and Tier-III in synchronized flow.

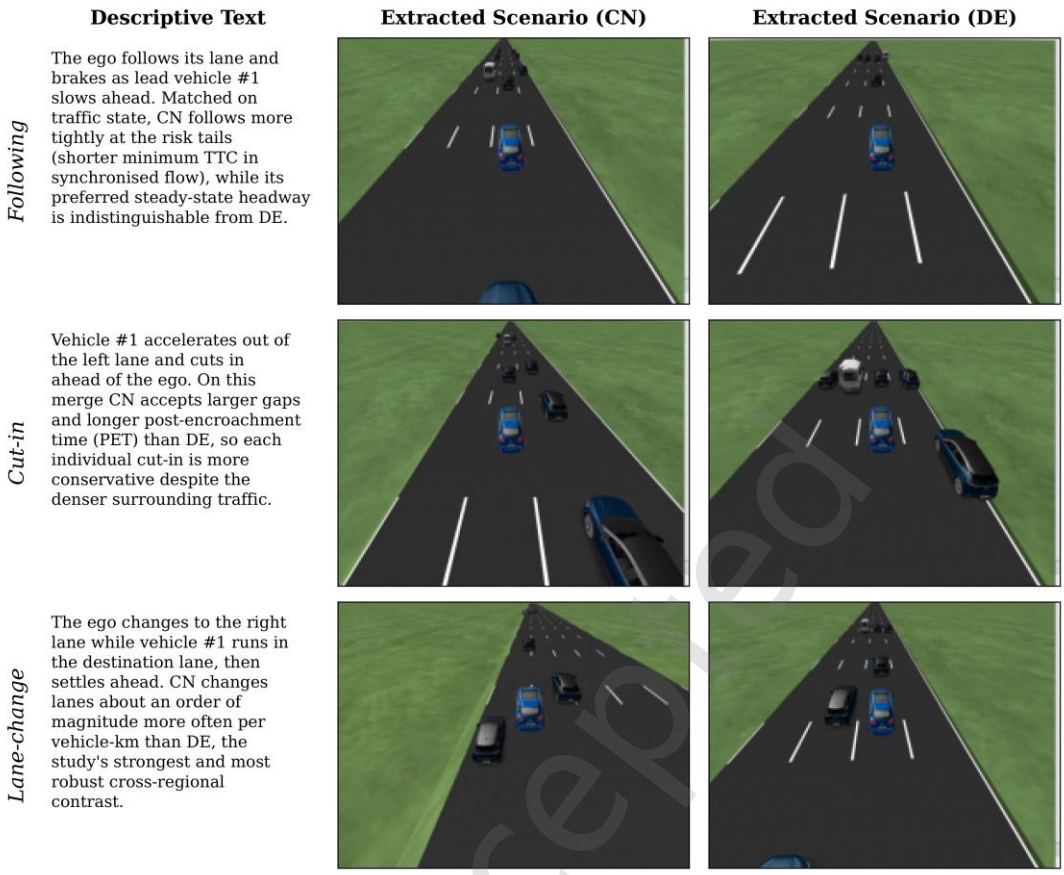


Fig. 10. Region-paired scenario fragments for three typical highway maneuvers. Each row places the AD4CHE fragment (center) beside its traffic-state-matched highD counterpart (right).

4.2. LLM assessment reliability and limitations

The D7/D8 dual-track results of Section 3.3 (Figs. 7 and 8 and Table 5) carry three messages worth separating from the numbers themselves.

The two methods agree where the phenomenon is unambiguous. On D8 the residual disagreement is one-directional (Fig. 8). The LLM additionally flags closing-in-behind events not flagged by the *PET*<1.0 *s* rule, because a scalar cut-off treats small PET as the only failure mode while a multi-factor rubric also responds to the kinematic context above that threshold. The cross-regional contrast therefore changes sign between the two labels (Table 5). Neither label is a ground truth, but the two are not the only available criteria. Both are scored from the state at the lane-change frame, so the braking the affected follower performs afterwards is external to both. It is measured from the trajectories and is visible to neither track. Over the 4,075 events, the peak deceleration reached by that follower within 3 s separates the two label sets differently. On D8 the LLM-majority separates the criterion more sharply than the formula (Cliff's δ=+0.26 against +0.20), and the contrast is sharper still on the events where the two disagree. Those flagged only by the LLM-majority are followed by markedly harder braking than the events both tracks call safe, while those flagged only by the formula are not. On D7 the result runs the other way, with the formula separating marginally better, consistent with the rubric being the more conservative of the two on that dimension. The per-track and disagreement-cell figures, the evasive-braking rates and the 2 s, 3 s and 5 s windows are released in the supplementary material. Hard braking shows that the maneuver had a real effect on the vehicle behind, and with no reference labelling neither label can be scored for correctness. It does show that the extra events the rubric flags on D8 are not arbitrary. What the two tracks establish between themselves remains systematic agreement and disagreement, not the accuracy of either one.

*The two methods partially disagree where the scalar formulation obscures multi-factor evidence. On D7 the McNemar discordance flips to formula-heavy and the divergence is stratum-specific (*Table 5*). The excess formula-positives in synchronized flow clear the P75 threshold on a single factor rather than on a consistent aggressiveness signal, so the contrast nearly vanishes under the rubric, while in jams it persists at a smaller magnitude. The dual-track thus separates the artefact stratum from the signal one.*

*Intra-LLM agreement is far lower on D7 than on D8 (*Fig. 7*(a)). A low intra-κ admits two readings, an ill-posed concept or unstable rubric wording, which the follower-response criterion cannot separate because it speaks to the labels and not to the*

rubric. The former is the more plausible, since D8 stays well-posed under the same perturbation while D7 does not, but it is a diagnostic hypothesis, not a settled result. The D7 cross-regional contrast is therefore treated as exploratory, pending the validation deferred below.

Two caveats are specific to the D7/D8 extrapolation. The human-authored rubric means the three-framing perturbation probes parameter-level rather than rule-structure variation, and the agreement statistics measure consistency between two rubrics rather than against a human gold label, so a two-annotator coded subsample is left to future work.

4.3. Implications for region-adaptive automated driving

The tier decomposition of Section 4.1 maps onto a concrete procedure for region-adaptive validation, consistent with ISO 21448 (ISO, 2022). Tier III dimensions justify a *shared* envelope, since preferred THW, IDM desired-gap proxy, and peak LC lateral velocity populate from either corpus without loss, as driver-model rather than country parameters (Treiber et al., 2000; Xu et al., 2021); Tier II dimensions need *stratum-conditional* envelopes (a China-typical synchronized card tightens D1 and increases D11, a jam card increases D3 and reduces D10/D13); and Tier I dimensions call for *invariant*-per-stratum shifts, e.g. a higher nominal LC-per-km rate for any China-tuned pipeline regardless of traffic state.2.7

4.4. Limitations and future work

The extracted fragment library is deliberately not a full representative pool of highway driving scenarios. It is the complement of one, collecting the scenarios where two regions differ enough that a system validated in one cannot be assumed equivalent to the other, to be run alongside an existing regression suite rather than in place of it. Four caveats qualify the results. *(i) Dataset asymmetry.* AD4CHE is congestion-oriented, so the F stratum is excluded at the adopted cut-offs and the small highD jam stratum is read through the power caveat of Section 3.2. *(ii) Two-dataset generalization.* With one dataset per country, site-specific factors (geometry, location, time of day) can only partially represent the two regions' driving scenario differences, further study with more datasets from each region will be needed in future work. *(iii) Metric robustness.* D6 depends on the harmonized lane-change detector (Section 2.4) and D14 on the finite-difference jerk definition, though the D14 contrast survives the 25-Hz resampling check of Section 3.2. *(iv) Aerial perspective and LLM scope. Drone recordings miss driver state, other road users, and weather, so conclusions apply to daytime dry interior-lane driving; the D7/D8 dual-track is one labelling paradigm, and rerunning under ordinal or continuous risk scores is the immediate next step. Whether the labels are specific to one model family is likewise open, since the data licenses confine the trajectories to an environment serving a single family (Section 2.5.2), so repeating the protocol across model families needs either a dataset whose license permits external inference or an environment hosting several, and it would test the labels rather than the design of the audit.*

The central direction for future work is to turn the bench's feasibility readout (Section 2.8) into a measured capability boundary for the system under test. This requires closing the production-ECU loop of Fig. 4 onto the CarMaker ego with the real ADAS sensor set and control hardware physically in the loop (not the reference controller used here) and replaying a balanced two-sided fragment set under a controlled, region-resolved KPI campaign that quantifies where a given ADS passes or fails on the two regions' scenarios. Two further extensions support that goal: reactive ambient traffic via IDM+MOBIL agents (Kesting et al., 2007; Treiber et al., 2000) would replace recorded-kinematics playback with closed-loop interaction, and a second European dataset would broaden the seeded fragment set beyond the single highD jam stratum. Scenario synthesis is a third direction. The library replays recorded fragments, so diffusing each one into parameter-space variants, generalizing and generating novel scenarios that stay inside the region envelope remains to be done.

5 Conclusion

A 14-dimension, traffic-state-stratified taxonomy compared Chinese (AD4CHE) and German (highD) highway driving on matched Kerner phases, sorting the dimensions into stratum-invariant, flow-modulated and no-shift tiers. For the two dimensions a scalar threshold cannot capture, an LLM assessment agent audited the formula baseline and localized where the two labels disagree. The surviving divergences became an OpenSCENARIO fragment library shown to be concrete and batch-runnable, importing and executing to completion on a production-grade HiL bench. The framework therefore offers a practical route from naturalistic trajectory data to an executable, region-specific scenario library, whose use is to concentrate validation efforts on the scenarios that differ when an ADS is transferred between regions.

Author contributions

Ji Zhou: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Yongqi Zhao:** Writing – review & editing, Formal Analysis, Methodology. **Arno Eichberger:** Writing – review & editing, Supervision, Methodology, Funding acquisition.

Replication and data sharing

The source datasets are third-party and free for academic use on application to their owners: highD (Krajewski et al., 2018) and AD4CHE (Zhang et al., 2023). The stratified

effect-size matrix (stratified_stats.csv) and the robustness tables (loo_jam_results.csv) are released as companion files. The region-specific OpenSCENARIO fragment library, its generating code and a replication package explanatory file covering the data format, computer environment and simulation files are deposited on ETS Data; the DOI will be assigned upon acceptance and inserted here. A demonstration video is included in the same package.

Acknowledgements

The Hardware-in-the-Loop test bench used in this research work was from Stellantis group company. The publication fee (Article Publishing Charge (APC) will be supported by the TU Graz Open Access Publishing Fund if needed.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Allen, B.L., Shin, B.T., Cooper, P.J., 1978. Analysis of traffic conflicts and collisions. Transp Res Rec, 667, 67–74.

ASAM e.V., 2020. ASAM OpenSCENARIO: Dynamic content of driving simulation scenarios. Standard, version 1.0.0. https://www.asam.net/standards/detail/openscenario-xml/

ASAM e.V., 2021. ASAM OpenDRIVE: Logical road network description. Standard, version 1.6.0. https://www.asam.net/standards/detail/opendrive/

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B, 57, 289–300.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. New York, NY, USA: Plenum Press.

Byrt, T., Bishop, J., Carlin, J.B., 1993. Bias, prevalence and kappa. J Clin Epidemiol, 46, 423–429.

Cassidy, M.J., Coifman, B., 1997. Relation among average speed, flow, and density and analogous relation between density and occupancy. Transp Res Rec, 1591, 1–6.

Cheng, Q., Jiang, X., Wang, W., Dietrich, A., Bengler, K., Qin, Y., 2020. Analyses on the heterogeneity of car-following behaviour: Evidence from a cross-cultural driving simulator study. IET Intell Transp Syst, 14, 834–841.

Cheng, Z., Wang, W., Lu, J., Xing, X., 2020. Classifying the traffic state of urban expressways: A machine-learning approach. Transp Res Part A, 137, 411–428.

Cliff, N., 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin, 114, 494–509.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37–46.

Coifman, B., 2015. Empirical flow–density and speed–spacing relationships: Evidence of vehicle length dependency. Transp Res Part B, 78, 54–65.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10, 1895–1923.

Ding, S., Abdel-Aty, M., Wang, Z., Wang, D., 2024. Insights into vehicle conflicts based on traffic flow dynamics. Scientific Reports, 14, 1536.

Edie, L.C., 1963. Discussion of traffic stream measurements and definitions. In: Proc. 2nd Int. Symp. Theory of Traffic Flow. Almond, J., Ed. Paris, France: OECD, 139–154.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall/CRC.

Feise, R.J., 2002. Do multiple outcome measures require p-value adjustment? BMC Medical Research Methodology, 2, 8.

Feng, F., Bao, S., Sayer, J.R., Flannagan, C., Manser, M., Wunderlich, R., 2017. Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. Accid Anal Prev, 104, 125–136.

Gilardi, F., Alizadeh, M., Kubli, M., 2023. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci USA, 120, e2305016120.

Hayward, J.C., 1972. Near-miss determination through use of a scale of danger. Highway Research Record, 384, 24–34.

Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R., 2020. The ApolloScape open dataset for autonomous driving and its application. IEEE Trans Pattern Anal Mach Intell, 42, 2702–2719.

International Organization for Standardization (ISO), 2022. Road vehicles—safety of the intended functionality. ISO 21448:2022. https://www.iso.org/standard/77490.html

IPG Automotive GmbH, 2024. CarMaker: Virtual test driving for vehicle dynamics and ADAS/AD. Software,

version 13. https://www.ipg-automotive.com/en/products-solutions/software/carmaker/

Jahangiri, A., Berardi, V.J., Machiani, S.G., 2018. Application of real field connected vehicle data for aggressive driving identification on horizontal curves. IEEE Trans Intell Transp Syst, 19, 2316–2324.

Kerner, B.S., 1998. Experimental features of self-organization in traffic flow. Phys Rev Lett, 81, 3797–3800.

Kerner, B.S., 2004. The Physics of Traffic: Empirical Freeway Pattern Features, Engineering Applications, and Theory. Berlin, Germany: Springer.

Kerner, B.S., Rehborn, H., Aleksic, M., Haug, A., 2004. Recognition and tracking of spatial–temporal congested traffic patterns on freeways. Transp Res Part C, 12, 369–400.

Kesting, A., Treiber, M., 2008. Calibrating car-following models by using trajectory data: Methodological study. Transp Res Rec, 2088, 148–156.

Kesting, A., Treiber, M., Helbing, D., 2007. General lane-changing model MOBIL for car-following models. Transp Res Rec, 1999, 86–94.

Knabe, E., et al., 2024. esmini: Environment simulator minimalistic — an OpenSCENARIO player. https://github.com/esmini/esmini

Krajewski, R., Bock, J., Kloeker, L., Eckstein, L., 2018. The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2118–2125.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics, 33, 159–174.

Liu, T., Selpi, 2020. Comparison of car-following behavior in terms of safety indicators between China and Sweden. IEEE Trans Intell Transp Syst, 21, 3696–3705.

Mahmud, S.M.S., Ferreira, L., Hoque, M.S., Tavassoli, A., 2017. Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. IATSS Research, 41, 153–163.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18, 50–60.

Massey, F.J., Jr., 1951. The Kolmogorov–Smirnov test for goodness of fit. J Am Stat Assoc, 46, 68–78.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153–157.

Nasution, A.H., Onan, A., 2024. ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks. IEEE Access, 12, 71876–71900.

Nobukawa, K., Bao, S., LeBlanc, D.J., Zhao, D., Peng, H., Pan, C.S., 2016. Gap acceptance during lane changes by large-truck drivers—An image-based analysis. IEEE Trans Intell Transp Syst, 17, 772–781.

Özkan, T., Lajunen, T., Chliaoutakis, J.E., Parker, D., Summala, H., 2006. Cross-cultural differences in driving behaviours: A comparison of six countries. Transp Res Part F, 9, 227–242.

Rainio, O., Teuho, J., Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14, 6086.

Rehborn, H., Klenov, S.L., Palmer, J., 2011. An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany. Physica A, 390, 4466–4485.

Satzoda, R.K., Gunaratne, P., Trivedi, M.M., 2015. Drive quality analysis of lane change maneuvers for naturalistic driving studies. In: 2015 IEEE Intelligent Vehicles Symposium (IV), 654–659.

Schönhof, M., Helbing, D., 2007. Empirical features of congested traffic states and their implications for traffic modeling. Transp Sci, 41, 135–166.

Transportation Research Board (TRB), 2016. Highway Capacity Manual, Sixth Edition: A Guide for Multimodal Mobility Analysis. Washington, DC, USA: National Academies of Sciences, Engineering, and Medicine.

Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Phys Rev E, 62, 1805–1824.

Treiber, M., Kesting, A., Helbing, D., 2010. Three-phase traffic theory and two-phase models with a fundamental diagram in the light of empirical stylized facts. Transp Res Part B, 44, 983–1000.

Vogel, K., 2003. A comparison of headway and time to collision as safety indicators. Accid Anal Prev, 35, 427–433.

Westhofen, L., Neurohr, C., Koopmann, T., Butz, M., Schütt, B., Utesch, F., et al., 2022. Criticality metrics for automated driving: A review and suitability analysis of the state of the art. Arch Comput Methods Eng, 30, 1–35.

Xu, X., Wang, X., Wu, X., Hassanin, O., Chai, C., 2021. Calibration and evaluation of the Responsibility-Sensitive Safety model of autonomous car-following maneuvers using naturalistic driving study data. Transp Res Part C, 123, 102988.

Xu, Y., Shao, W., Li, J., Yang, K., Wang, W., Huang, H., et al., 2022. SIND: A drone dataset at signalized intersection in China. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 2471–2478.

Zhang, Y., Wang, C., Yu, R., Wang, L., Quan, W., Gao, Y., et al., 2023. The AD4CHE dataset and its application in typical congestion scenarios of traffic jam pilot systems. IEEE Trans Intell Veh, 8, 3312–3323.

Zhao, Y., Xiao, W., Mihalj, T., Hu, J., Eichberger, A., 2024. Chat2Scenario: Scenario extraction from dataset through utilization of large language model. In: 2024 IEEE Intelligent Vehicles Symposium (IV), 559–566.

Zhao, Y., Zhou, J., Bi, D., Mihalj, T., Hu, J., Eichberger, A., 2026. A survey on the application of large language models in scenario-based testing of automated driving systems. IEEE Trans Intell Transp Syst, 27, 5001–5023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al., 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In: Advances in Neural Information Processing Systems 36 (NeurIPS), 46595–46623.

Author biography

Ji Zhou received the bachelor's degree and master's degree (M.Sc.) in automotive engineering from Jilin University, Changchun, China, in 2007 and in 2009. He is currently pursuing the Ph.D. degree with the Institute of Automotive Engineering, Graz University of Technology, Graz, Austria. He has worked with Ricardo, former PSA Group, and is with Stellantis Group for more than 15 years. He has rich expertise in powertrain control software development, validation and calibration, vehicle electrical and electronic (EE) architecture integration validation, and EE product industrialization. His current research interests include advanced methodologies of automated driving systems integration and validation.

Yongqi Zhao received the bachelor's degree from China University of Petroleum (East China), Qingdao, China, in 2019, and the master's degree from the Technical University of Braunschweig, Braunschweig, Germany, in 2022. He is currently pursuing the Ph.D. degree with the Institute of Automotive Engineering, Graz University of Technology, Graz, Austria, with a research focus on virtual testing of automated driving systems. While pursuing his master's degree, he gained practical experience through internships with Momenta, Stuttgart, Germany; and Volkswagen Group, Wolfsburg, Germany.

Arno Eichberger received the degree in mechanical engineering and the Ph.D. degree (Hons.) in technical sciences from Graz University of Technology, Graz, Austria, in 1995 and 1998, respectively. From 1998 to 2007, he was with Magna Steyr Fahrzeugtechnik AG & Company, Graz, where he dealt with different aspects of active and passive safety. Since 2007, he has been with the Institute of Automotive Engineering, Graz University of Technology, dealing with driver assistance systems, vehicle dynamics, and suspensions. Since 2012, he has been an Associate Professor holding a "venia docendi" of automotive engineering.