J Intel Connect Veh

Research Article

Probing Large Language Models for Autonomous Driving Behavior

Zhipeng Bao, Wenjie Zhao, Qianwen Li✉

Cite this article: https://doi.org/10.26599/JICV.2026.9210095

ABSTRACT: As large language models (LLMs) are increasingly integrated into autonomous vehicles (AV), understanding their reasoning and behavioral tendencies becomes essential. Trained on vast datasets, LLMs carry behavioral priors and social biases that may shape their driving decisions. Without such insight, developers may struggle to align models for AV needs. To address this, we probe prompt-conditioned high-level action choices of LLMs, each with 1,500 contextual variants. Three widely used LLMs are evaluated with multilingual prompts to select from predefined behavioral options ordered by aggressiveness. An Ordered Logit Model quantifies how contextual factors influence decisions, complemented by thematic analysis to reveal underlying reasoning tendencies. Results show that LLM decisions reflect a mix of model characteristics, linguistic framing, and scenario context. Across conditions, models remain sensitive to rider urgency, traffic complexity, and road-user types. GPT is more conservative, while DeepSeek and LLaMA act more assertively, especially in vehicle interactions. Prompt language also matters. Chinese and French prompts are associated with more assertive behavior than English, with French strongest. Across scenarios, all models shift toward more protective behavior when vulnerable road users appear, reducing aggressiveness and prioritizing safety and smoother flow. These findings characterize model-level behavioral priors relevant to LLM choice and prompt design in AV applications.

KEYWORDS: AI ethics, Decision-making behavior, Driving style, Ordered logit model, Thematic analysis

1 Introduction

Large language models (LLMs) are increasingly adopted as decision-making modules across diverse domains, including politics, finance, and healthcare. However, LLMs generate responses based on prior knowledge encoded during training, which fundamentally shapes their behavioral tendencies and reasoning patterns. Understanding how this knowledge influences model decisions is essential for informed and responsible deployment. Accordingly, researchers across domains are moving beyond output correctness and model utility to actively probe LLM behavior and reasoning across diverse contexts, uncovering implicit assumptions, biases, and domain-specific tendencies. In the political domain, for instance, Rettenberger et al. (2025) have used Germany's Wahl-O-Mat tool to assess how open-source LLMs align with political parties, finding that larger models tend to lean left while smaller ones remain more neutral, reflecting dominant values in the training corpus such as environmentalism, gender equality, and humanitarianism. In finance, Winder et al. (2025) show that LLMs frequently generate high-risk, aggressive investment recommendations, increasing portfolio volatility even when bias-mitigation prompts are applied. In healthcare, Shool et al. (2025) report that LLMs often adopt conservative positions in diagnostic or treatment suggestions, likely due to a risk-averse bias. Moreover, subdomain complexity has been shown to influence the model's decisiveness and safety orientation. These insights are more than descriptive. They provide actionable strategies such as domain-specific fine-tuning, prompt engineering, and risk-aware deployment practices.

Such probing analysis of LLMs in the autonomous driving domain remains largely absent. Most existing studies continue to emphasize output correctness or task-level performance when LLMs are used as high-level decision-making or policy-selection modules, relying on evaluation metrics such as scene interpretation accuracy or driving scores (Cui et al., 2023; Dong et al., 2022; Hu et al., 2025; Kuang et al., 2024; Liao et al., 2024; Peng et al., 2025; Renz et al., 2024). While these metrics quantify task success, they offer limited insight into the model's driving style, decision-making rationale, or sensitivity to contextual nuances. The lack of deeper behavioral evaluation leaves a critical gap in understanding how LLMs apply prior knowledge in dynamic driving scenarios. Without examining why a model chooses one action over another, beyond whether it completes the task, it is not possible to characterize its decision-making patterns, reveal behavioral tendencies such as aggressiveness or conservativeness, or identify inconsistencies across similar situations. This makes it difficult to distinguish between models with similar surface-level performance but fundamentally different driving philosophies. As a result, downstream systems may inherit brittle, opaque, or misaligned behaviors, reducing the robustness, safety, and interpretability of LLM-based autonomous driving systems.

To date, only a few studies, most notably by Takemoto (2024) and Xu et al. (2025), have examined the behavior of LLMs as high-level decision-making or judgment modules in the context of autonomous driving. Takemoto et al. use the Moral Machine framework to present ethical dilemmas involving emergency scenarios, such as deciding between hitting a jaywalking elderly pedestrian or sacrificing the passenger. Their findings show that model decisions are shaped by factors such as species, number of lives, legality, and age, reflecting statistical associations from training data rather than grounded ethical reasoning. Similarly, Xu et al. investigate LLMs' moral judgments using the MIT Moral Machine platform, presenting scenarios where an autonomous vehicle must choose between harmful outcomes. They find that decisions are influenced by attributes such as legality, age, social status, and group size, mirroring human value judgments embedded in training data. While these studies provide valuable insights into LLM behavior in rare, high-

School of Environmental, Civil, Agricultural and Mechanical Engineering, University of Georgia, Athens GA 30606, USA.

✉ Corresponding author. E-mail: Cami.Li@uga.edu

Received: June 16, 2026; Revised: August 6, 2026; Accepted: August 18, 2026

stakes ethical dilemmas, they are fundamentally limited in two key aspects. First, they focus on extreme and low-frequency scenarios that do not represent the majority of real-world driving situations. Second, they primarily analyze moral preferences rather than general driving decision-making processes under everyday traffic conditions.

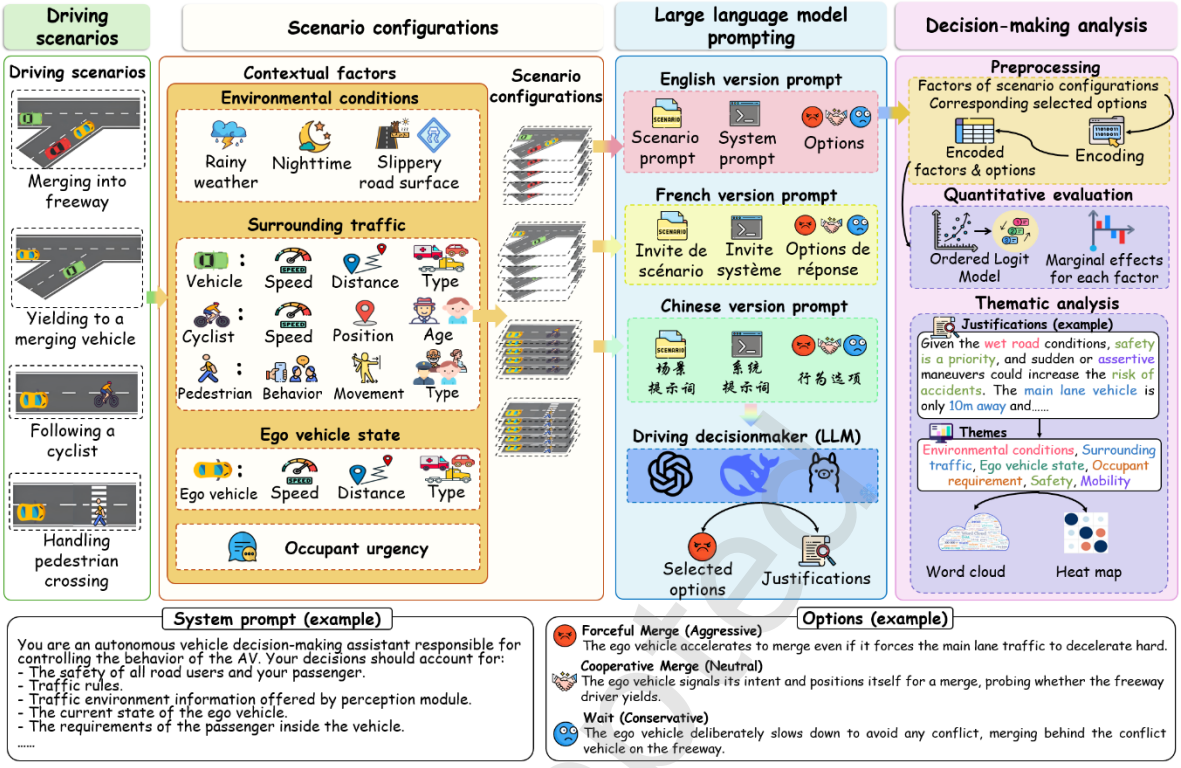


Fig. 1. Pipeline for probing prompt-conditioned LLM action preferences and decision rationales.

Unlike prior work that centers on ethical dilemmas, this study shifts the focus to everyday driving scenarios, which represent the majority of real-world traffic exposure. We investigate LLMs' prompt-conditioned high-level decision-making tendencies across four representative scenarios, each enriched with varied contextual factors. The predefined high-level action options are ranked by level of aggressiveness, reflecting real-world variations in driving style (Bolovinou et al., 2014) rather than directly executed vehicle control. Three widely used LLMs are prompted in three languages to select a high-level action option and provide an explanation. We apply the Ordered Logit Model (OLM) to quantify decision tendencies and the effects of contextual factors, and conduct a thematic analysis to reveal reasoning patterns in model justifications.

In summary, the contributions of this paper are as follows. First, we analyze prompt-conditioned high-level decision-making tendencies of LLMs in routine driving scenarios, addressing the gap in existing work that primarily focuses on rare and extreme cases. Second, we identify the key factors that influence LLMs' high-level action choices, providing insight into how models weigh contextual information during decision-making. Third, we conduct an evaluation across multiple LLMs and prompt languages, revealing how model-specific priors and linguistic variations jointly shape prompt-conditioned decision tendencies.

The remainder of this manuscript is organized as follows. Section 2 introduces the methodological framework, including the design of traffic scenarios, contextual factors, driving actions, and the OLM used for quantitative analysis. Section 3 presents the model estimation results along with a thematic analysis of LLM explanations across the four representative traffic scenarios. Section 4 concludes the study.

2 Method

This study investigates prompt-conditioned high-level action selection by LLMs across representative traffic scenarios (**Fig. 1**). To evaluate model responses under diverse conditions, we systematically vary contextual factors—including environmental conditions, surrounding traffic, ego vehicle state, and occupant urgency—to generate multiple case variations for each scenario. Selected LLMs are prompted in multiple languages and instructed to choose an option representing a driving style (aggressive, neutral, or conservative) and provide a justification, following a fixed output format. The resulting choices represent policy-level behavioral preferences elicited from textual prompts; they do not constitute executed trajectories or embodied, closed-loop driving behavior. The experiment adopts a single-step, one-shot, open-loop, and single-agent text-based design. An OLM is employed to quantify the influence of each factor on style selection, and marginal effects are computed to assess directional relationships. To complement this quantitative analysis, we conduct an explanation-based evaluation by extracting thematic patterns from LLM justifications, visualized through word clouds and a heat map to reveal the emphasis placed on different factors across models.

2.1. Data generation: The design of scenarios, factors, and behavioral options

Scenarios: We design four test scenarios to probe the driving decision-making of LLMs in traffic situations that conventionally involve interaction with other road users, including merging into freeway, yielding to a merging vehicle, following a cyclist, and handling pedestrian crossings. These scenarios are selected for their ubiquity in

real-world driving, the presence of negotiation or social inference, and the need for subtle behavioral trade-offs. Each case provides a static snapshot of a traffic situation with interaction-relevant contextual information, requiring decisions that balance assertiveness, caution, and courtesy. Unlike rare or extreme edge cases, these scenarios offer repeated, nuanced opportunities to evaluate how LLMs interpret context, weigh priorities, and express driving style. Collectively, they provide a controlled testbed for analyzing whether LLMs exhibit context-sensitive high-level action preferences across diverse textual traffic situations.

Factors: Prompt-conditioned high-level action choices may be shaped not only by the scenario description itself but also by the surrounding contextual information included in the prompt. To capture the full range of driving styles expressed by LLMs, we introduce systematic variations across key contextual factors within each scenario. This allows us to create diverse, realistic case configurations that test the sensitivity of LLM responses to changing conditions. We categorize contextual factors into four groups: (1) environmental conditions (e.g., rainy weather, slippery road surface), (2) surrounding traffic (e.g., nearby agents' type and speed), (3) ego vehicle state (e.g., current speed, lane position), and (4) occupant urgency. The first three categories reflect widely recognized influences on human and AV driving behavior (Lee and See, 2004). The fourth is specific to LLM-powered AVs that may adjust their behavior based on the urgency of passenger requests. In the main experiment, "emergency vehicle" refers only to the vehicle category, with no assumption that its lights or sirens are active, as a preliminary test found that explicitly activating them produced no material change in the response patterns. Details of the contextual factors are summarized in Table 1.

Beyond scenario-specific factors, we vary the driving decisionmaker, represented by three LLMs (GPT, Deepseek, Llama). These models are widely used large-scale LLMs with multilingual capabilities, but they differ substantially in architecture, parameterization, openness, training procedures, and alignment strategies (Zhao et al., 2023), allowing us to assess whether model-specific traits lead to systematic differences in prompt-conditioned action choices under identical conditions. We also vary the prompt language (including English, Chinese, and French), recognizing that LLM responses may differ across languages even with equivalent meaning (Wang et al., 2025). Given the global nature of AV deployment, it is important to evaluate whether LLM behavior remains consistent across languages or if linguistic variation introduces unintended behavioral shifts. To ensure that the observed cross-language differences are not attributable to translation artifacts, we further conduct a back-translation validation and tone/framing consistency analysis, as reported in Appendix A1.

Options: For each driving scenario, we define three ordered behavioral options—aggressive, neutral, and conservative. Rather than being defined solely by physical driving metrics, these categories are grounded in decision-making strategies that reflect how the ego vehicle balances its own objectives with interactions with other road users. Aggressive behavior prioritizes the ego vehicle's own efficiency and right-of-way, typically proceeding without accommodating others or engaging in negotiation. Neutral behavior represents a cooperative strategy, where the vehicle probes, communicates, or adapts to surrounding agents before committing to an action. Conservative behavior prioritizes safety and the interests of other road users, often yielding early or avoiding potential conflicts. Descriptions of the specific driving actions are provided in Section 3. This categorization is informed by prior research on driving style classification, which commonly distinguishes drivers based on their risk tolerance, response to surrounding agents, and willingness to assert right-of-way (Bolovinou et al., 2014). These options enable systematic comparison of LLM behavior across scenarios and highlight how models balance safety, efficiency, and social interaction.

Table 1. Driving scenario contextual factors.

Factors	Value	Description
Rainy weather	0: sunny, 1: rainy	/
Nighttime	0: day, 1: night	/
Slippery road surface	0: dry, 1: slippery	/
Space ahead of the conflict point	20m , 50m , 100m, 300m	The available gap in front of the conflict point, affecting the merging feasibility.
Arrival time difference	1: <-2s, 2: -2s~2s, 3: >2	Time offset between the ego and the interactive vehicle (mainstream or merging) arriving at the conflict point, estimated based on vehicles' distance to the conflict point and their speed.
Space ahead of the bicycle	20m, 50m, 100m, 300m	Space ahead of the bicycle, indicating available room for the ego vehicle to overtake.
Mainstream vehicle type	1: emergency, 2: truck, 3: normal	Type of vehicle on the main road—emergency vehicle, truck, or normal car.
Mainstream vehicle speed	35km/h, 60km/h, 100km/h	The speed of the mainstream vehicle.
Mainstream vehicle distance to the conflict point	10m, 50m, 100m, 150m, 200m	Distance from the mainstream vehicle to the end of the ramp.
Merging vehicle type	1: emergency, 2: truck, 3: normal	Type of the vehicle merging into freeway—emergency vehicle, truck, or normal car.
Merging vehicle speed	35km/h, 60km/h, 100km/h	The speed of the merging vehicle.
Merging vehicle distance to the conflict point	10m, 50m, 100m, 150m, 200m	Distance from the merging vehicle to the end of the ramp.
Bicycle in the middle of the lane	0: right, 1: middle	Lateral position of the bicycle in the lane—middle or right side.

Child cyclist	0: adult, 1: children	Indicates whether the cyclist is an adult or a child.
No following vehicle	0: following vehicle present, 1: following vehicle absent	Whether a vehicle is present behind the ego vehicle.
Type of the following vehicle	1: emergency, 2: truck, 3: normal	Type of the vehicle behind the ego vehicle—emergency, truck, or normal.
Pedestrian type	1: children, 2: college student, 3: elderly, 4: police, 5: pregnant	Identity of the pedestrian, such as a child, student, elderly, police, or pregnant individual.
Pedestrian behavior	1: looking around, 2: talking to others, 3: watching the phone	Behavior of the pedestrian—looking around, talking to others, or watching a phone.
Pedestrian movement	0: stopping near the crosswalk, 1: walking to the crosswalk	Pedestrian movement—either pausing or walking toward the crosswalk.
Stopped ego vehicle	0: no, 1: yes	Whether the ego vehicle has come to a stop near the merging ramp.
Ego vehicle distance to the conflict point	10m, 50m, 100m	Distance between the ego vehicle and the end of the ramp.
Ego vehicle distance to the pedestrian	10m, 50m, 100m	Distance between the ego vehicle and the pedestrian.
Ego vehicle distance to the bicycle	10m, 50m, 100m	Distance between the ego vehicle and the bicycle ahead.
Ego vehicle speed in freeway operations	30km/h, 50km/h, 80km/h	Ego vehicle's speed in merging and yielding scenarios.
Ego vehicle speed when interacting with pedestrians and bicycles	25km/h, 50km/h	Ego vehicle's speed in cyclist and pedestrian scenarios.
Type of ego vehicle	1: emergency, 2: truck, 3: normal	Type of ego vehicle—emergency, truck, or normal.
Occupant urgency	0: none, 1: hurry, 2: urgent	Urgency level of the occupant's travel purpose, ranging from none to hurry (e.g., rushing to work) to urgent (e.g., medical emergency).
Prompt language	1: English, 2: Chinese, 3: French	The language used to prompt the LLM—English, Chinese, or French.
Driving decisionmaker	1: GPT, 2: Deepseek, 3: Llama	LLM used for driving decision-making—GPT, Deepseek, or Llama.

2.2. Ordered logit model

Given the ordinal and discrete nature of the behavioral options, we employ an OLM to analyze LLM driving decisions (Marcoux et al., 2018). Both ordered logit and ordered probit models are suitable for ordinal outcomes; in this study, we adopt the logit specification due to its interpretability and widespread use in transportation and discrete-choice research. Specifically, we treat q $(q = 1,2, \dots, Q)$ as the index for observations and $k(k = 1,2,3)$ as the index for the ordered behavioral options. k takes the value of aggressive style option $(k = 1)$, neutral style option $(k = 2)$, and conservative style option $(k = 3)$. In the traditional ordered outcome model, the discrete behavioral option (y_q) is assumed to be associated with an underlying continuous latent variable (y_q^*). This latent variable is typically specified as the following linear function:

$$y_q^* = \alpha x_q + \varepsilon_q, y_q = k \quad if\ \psi_{k-1} < y_q^* < \psi_k, \tag{1}$$

where y_q^* is mapped to the behavioral option y_q by the ψ thresholds $(\psi_0 = -\infty\ and\ \psi_2 = +\infty)$ in the usual ordered-outcome fashion. x_q is a column vector of factors that influence the propensity associated with behavioral option selection. α is a corresponding column vector of coefficients and $\varepsilon_q$is an idiosyncratic random error term assumed to be identically and independently standard logistic distributed across observation q. The validity of this independence assumption in the present setting is justified by the data generation procedure described in Section 3.1 and further supported by robustness checks using cluster-robust standard errors, as reported in Appendix A3. The parallel-regressions assumption is rejected by the Brant test in all three scenarios; we retain the ordered specification because the options are ordered by construction, and Appendix A6 shows that the findings are robust to removing the ordering restriction.

Given these relationships across the different parameters, the resulting probability expressions for observation q and alternative k for the OLM take the following form:

$$P_q(k) = \Lambda(\psi_k - \alpha x_q) - \Lambda(\psi_{k-1} - \alpha x_q), \tag{2}$$

where $\Lambda(\cdot)$ represents the standard logistic cumulative distribution function.

The log-likelihood function $\ell(\alpha, \psi)$ for the OLM is given by:

$$\ell(\alpha, \psi) = \sum_{q=1}^{Q} \sum_{k=1}^{3} \mathbb{I}(y_q = k) \cdot \log[\Lambda(\psi_k - \alpha x_q) - \Lambda(\psi_{k-1} - \alpha x_q)]. \tag{3}$$

To quantify the marginal effect of each factor on the probability of selecting a particular behavioral option, we adopt a finite difference approach. For each observation q, we perturb one element of the factor vector x_q by a small constant ε in both directions, while holding all other elements constant. The marginal effect of the d-th factor (i.e., the d-th element of x_q) on the probability of choosing option k is approximated by:

$$\frac{\partial P(y_q = k)}{\partial x_{qd}} \approx \frac{P(y_q = k | x_q^{d+}) - P(y_q = k | x_q^{d-})}{2\varepsilon}, \tag{4}$$

where $x_q^{(d+)}$ and $x_q^{(d-)}$ denote versions of the vector x_q with the d-th element increased or decreased by ε, respectively. We set $\varepsilon = 1 \times 10^{-4}$ on the standardized scale; because the central difference has $O(\varepsilon^2)$ truncation error, the estimates are insensitive to this choice. All predictors are standardized prior to estimation, so effects are reported per one standard deviation of the predictor, and the derivative is evaluated

at every observation and averaged over the sample, yielding average marginal effects rather than effects at the sample mean. **Eq. (4)** applies directly to the ordered factors entering the model linearly, namely the spatial gaps, speeds, occupant urgency, and the arrival-time-difference code. All multi-level categorical attributes are one-hot encoded before estimation with one level held out as the reference category, so no derivative is taken with respect to an unordered variable; for the resulting binary indicators, **Eq. (4)** gives the local slope per one standard deviation rather than a 0-to-1 discrete change, which is approximately twice as large. Signs and the relative ordering of effects are identical under either quantity. For each scenario, the initial specification included all factors in Table 1 applicable to that scenario, since Table 1 enumerates the union of the four scenario design spaces rather than a single common set. Variables with $p > 0.05$ were considered for removal, and the reduced specification was retained when it produced a lower or comparable AIC.

3 Experiment

In this section, we evaluate differences in the prompt-conditioned driving-style preferences of three LLMs across four textual driving scenarios. Using the OLM, we quantitatively assess which factors significantly influence the behavioral options selected by the LLMs. We further examine the marginal effect of each factor to evaluate both the direction and magnitude of its impact on the likelihood of selecting options representing more aggressive or more conservative high-level interaction strategies. Pearson correlation matrices for the final OLM variables indicate no multicollinearity concerns (Hair et al., 2019), with all coefficients below 0.5 (Appendix A4). All models are additionally re-estimated without the ordering restriction in Appendix A6, with the direction and significance of every effect reported in this section unchanged. In addition, we conduct a thematic analysis, employing the word cloud and the heat map to examine the key themes emphasized in the LLMs' justifications and how these themes relate to the selected driving options.

3.1. Details

We select three widely used LLMs: GPT-4o (Hurst et al., 2024), DeepSeek-V3-0324 (Liu et al., 2024), and Llama 3.1-405B (Dubey et al., 2024). All are large-scale models with strong language understanding capabilities and multilingual support. GPT-4o is closed-source, while DeepSeek-V3 and Llama 3.1-405B are open-source. Detailed models' information is provided in Appendix A5. All LLMs are prompted using the same system instruction and a default temperature setting of 1.0 to preserve natural response variability. Each model selects from predefined behavioral options and provides a structured output including the selected option and justification.

For each scenario, we enumerate the complete combinatorial space defined by the contextual factors active in that scenario and draw 1,500 case variations uniformly at random without replacement from this space, yielding a balanced random sample rather than a full-factorial or fractional-factorial design. The 1,500 cases are then divided into nine subsets assigned to unique combinations of three LLMs and three languages (English, Chinese, and French), and we verify that each contextual factor is distributed comparably across the nine subsets. These languages span distinct language families and cultural contexts that may influence how LLMs interpret driving scenarios. Each observation is generated through an independent API call with no shared memory or conversational context between calls. Because each case corresponds to a distinct combination of factor values, the input prompts differ substantively across observations, supporting the independence assumption required by the OLM. A robustness check using cluster-robust standard errors is reported in Appendix A3.

Thematic focus is extracted from LLM-generated justifications using zero-shot Natural Language Inference (Yin et al., 2019). Each justification is evaluated against six predefined theme hypotheses using facebook/bart-large-mnli (Lewis et al., 2020) in multi-label mode, with a theme marked as present when the entailment score exceeds 0.5. Each behavioral option is down-sampled to the minimum class count per scenario–model pair before aggregation, and theme counts are column-normalized to percentages summing to 100%.

3.2. Merging into freeway

In this scenario, the LLM is tasked with assisting an AV that is attempting to merge into the freeway. We provide three behavioral options for LLMs: Merge Forcefully (Aggressive)—the ego vehicle accelerates to merge even if it forces the main lane traffic to decelerate hard; Merge Cooperatively (Neutral)—the ego vehicle signals its intent and positions itself for a merge, probing whether the freeway driver yields; and Wait (Conservative)—the ego vehicle deliberately slows down to avoid any conflict, merging behind the conflict vehicle on the freeway. The selected options are then evaluated using the OLM, with the results summarized in **Table 2** and the marginal effects of each factor plotted in **Fig. 2**. The OLM converges at a log-likelihood of –941.210, with an AIC of 1910.000 and a BIC of 1985.000.

Table 2 Ordered logit model estimation results for the "merging into freeway" scenario: aggressive (y=0, n_0=403), neutral (y=1, n_1=476), and conservative (n_2=621).

Independent variable	Coefficient	Std. Err	z	P>|z|	95% CI (Lower)	95% CI (Upper)
Space ahead of the conflict point	-0.428	0.078	-5.490	<0.001	-0.581	-0.275
Occupant urgency	-1.982	0.087	-22.821	<0.001	-2.153	-1.812
Arrival time difference	-0.162	0.076	-2.124	0.034	-0.312	-0.013
Rainy weather	2.793	0.144	19.420	<0.001	2.511	3.075
Nighttime	1.018	0.084	12.184	<0.001	0.855	1.183
Slippery road surface	0.808	0.072	11.252	<0.001	0.668	0.950
Emergency mainstream vehicle	0.405	0.063	6.439	<0.001	0.282	0.529
Emergency ego vehicle	-1.364	0.074	-18.323	<0.001	-1.510	-1.219
No following vehicle	-0.145	0.061	-2.384	0.017	-0.266	-0.026
GPT as the decisionmaker	2.483	0.136	18.230	<0.001	2.217	2.751
Chinese prompts	-0.580	0.077	-7.582	<0.001	-0.730	-0.430

French prompts	-0.726	0.087	-8.337	<0.001	-0.897	-0.556
Threshold (0/1)	-2.195	0.103	-21.285	<0.001	-2.397	-1.993
Threshold (1/2)	1.080	0.044	24.670	<0.001	0.995	1.167
Number of observations	1500					
Log-likelihood function at convergence, ℓ	-941.210					
AIC	1910.000					
BIC	1985.000					

Note. Binary indicators are one-hot encoded with one level omitted as the reference category: sunny weather, daytime, and dry road surface for the environmental indicators; non-emergency vehicle types for the emergency-vehicle indicators; DeepSeek and LLaMA pooled for the model indicator; and English for the prompt-language indicators.

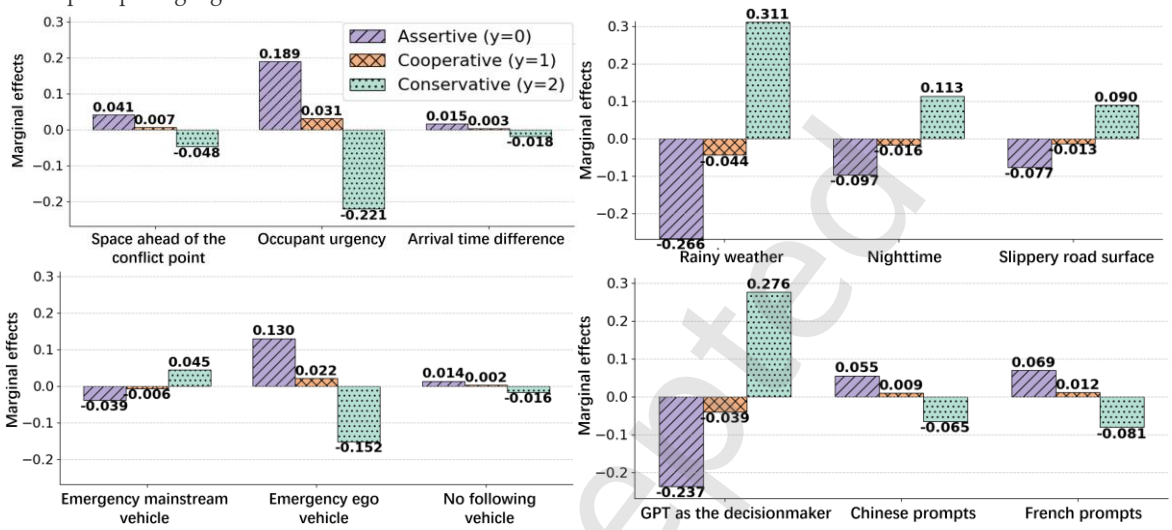


Fig. 2. Marginal effects of each factor in the freeway merging scenario. Reported values are average marginal effects per one standard deviation of the predictor; for binary indicators they are local slopes of the fitted probability surface rather than discrete 0-to-1 changes.

Unless otherwise noted, all marginal effects reported in this and the following sections are average marginal effects per one standard deviation of the predictor; for binary indicators these are local slopes rather than 0-to-1 discrete changes (see Section 2.2). Based on the results in **Table 2** and **Fig. 2**, several variables of the environmental condition factor significantly influence the behavioral option choices made by LLMs. Specifically, when the weather is rainy, the time of day is night, or the road surface is slippery, LLMs are more likely to select the conservative option. The corresponding increases in the probability of choosing the conservative option are +0.311, +0.113, and +0.090, respectively. Among the Surrounding traffic factors, several variables show significant effects on LLMs' choices. Specifically, a one-standard-deviation increase in the space ahead of the conflict point decreases the likelihood of selecting the conservative option by −0.048, while increasing the probability of choosing the aggressive option by +0.041. This indicates that greater spatial clearance is associated with a higher probability of the assertive option. A one-standard-deviation increase in the arrival time difference—where the ego vehicle arrives earlier at the conflict point—similarly reduces the probability of the conservative option (−0.018), indicating that the arrival-time advantage is among the factors associated with option selection. When the mainstream vehicle is identified as an emergency vehicle, the likelihood of aggressive responses drops (−0.039), while the conservative option becomes more likely (+0.045). Additionally, the absence of a following vehicle slightly raises the probability of the assertive option (+0.014). When the ego vehicle itself is an emergency vehicle, the probability of selecting the assertive option increases markedly, by +0.130. Similarly, occupant urgency—which ranges from no urgency through being late for work to needing to reach a hospital—shifts behavior toward assertiveness, with a one-standard-deviation increase raising the probability of aggressive behavior by +0.189. Regarding the model type, GPT selects the conservative option substantially more often than the other two models. The marginal effect results show that being GPT increases the probability of selecting a conservative option by +0.276. For the Prompt language, both Chinese and French prompts lead to more aggressive driving choices relative to English. Specifically, using Chinese increases the likelihood of aggressive behavior by +0.055, while French increases it even further by +0.069. These results indicate that prompt language is associated with systematic differences in option selection.

LLM behavioral patterns are further supported by the thematic analysis in **Fig. 3**, which offers qualitative insight into the models' decision rationales. The word clouds show that all three models frequently reference environmental conditions (e.g., rainy weather, road condition), surrounding traffic (e.g., main lane, truck, gap), and occupant urgency (e.g., hospital, urgency). Notably, Deepseek and LLaMA mention assertive-related terms (e.g., assertive, forcing) more often than GPT, consistent with their higher rate of assertive option selection. This aligns with the OLM results and indicates that, while all three models respond to context, their option distributions differ under the same scenario.

The heatmap further reveals that when aggressive behavior is selected, all three models place relatively less weight on environmental conditions and focus more on surrounding traffic and safety. GPT emphasizes

surrounding traffic (17.3%), occupant urgency (17.0%), and safety (17.3%), while Deepseek and LLaMA prioritize surrounding traffic, ego vehicle state, and either mobility or safety (each around 18%). For neutral

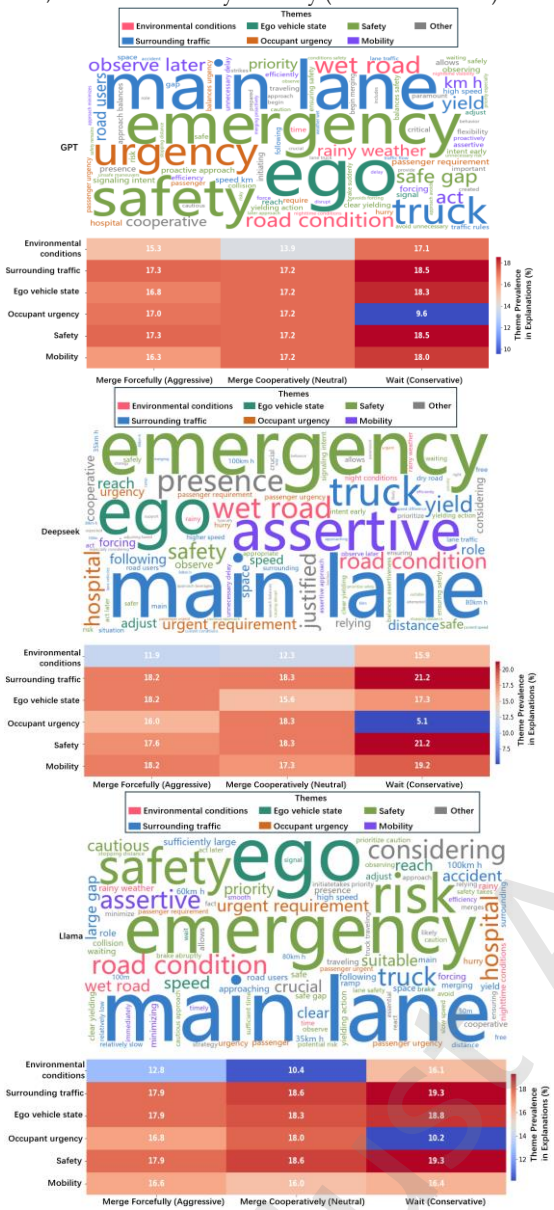


Fig. 3. Thematic analysis of LLM-generated explanations in the "merging into freeway" scenario when prompted in English.

choices, environmental cues remain low, with GPT distributing focus evenly, Deepseek reducing attention to ego state, and LLaMA assigning less weight to mobility. In conservative decisions, attention shifts away from urgency, most notably for Deepseek, which emphasizes safety and traffic (21.2% each). Overall, these findings show that the selected options are associated with both contextual inputs and model identity.

3.3. Yielding to a merging vehicle

In this scenario, the LLM-powered AV serves as the mainstream vehicle, not the merging one. We define three behavioral options for the LLMs: Do Not Yield (Aggressive)—the ego vehicle maintains its speed and lane, clearly signaling no intention to accommodate the merging vehicle; Signal No Yield but Yield if Pressed (Neutral)—the ego vehicle initially holds speed and lane, indicating it will not yield, but backs off if the merging vehicle asserts itself; and Yield Proactively (Conservative)—the ego vehicle slows down or changes lanes in advance to make room for the merging vehicle. These options are evaluated using an OLM, with results summarized in **Table 3** and marginal effects shown in **Fig. 4**. The model converged with a log-likelihood of −733.150, an AIC of 1494.000, and a BIC of 1569.000.

Based on the findings in **Table 3** and **Fig. 4**, the results for the "yielding to a merging vehicle" scenario largely align with those from the "merging into freeway" case, particularly regarding the influence of environmental conditions, occupant urgency, and surrounding traffic. However, two key differences emerge. First, the variable arrival time difference is no longer significant; instead, merging vehicle speed becomes significant. A one-standard-deviation increase in merging vehicle speed slightly increases the likelihood of the conservative option (+0.020), indicating that higher merging-vehicle speed is associated with a higher probability of proactive yielding. Second, there is a shift in the influence of prompt language. While Chinese and French were significant in the previous scenario, here English and Chinese prompts are associated with more conservative responses (+0.085 and +0.075, respectively), implying that French continues to lead to more assertive decisions by comparison.

Table 3. Ordered logit model estimation results for the "yielding to a merging vehicle" scenario: aggressive (y=0, n_0=214), neutral (y=1, n_1=343), and conservative (n_2=943).

Independent variable	Coefficient	Std. Err	z	P>|z|	95% CI (Lower)	95% CI (Upper)
Merging vehicle speed	0.196	0.071	2.777	0.005	0.058	0.334
Space ahead of the conflict point	-0.522	0.071	-7.405	<0.001	-0.661	-0.384
Occupant urgency	-2.029	0.103	-19.734	<0.001	-2.231	-1.828
Rainy weather	3.343	0.187	17.934	<0.001	2.994	3.692
Nighttime	1.137	0.106	10.681	<0.001	0.928	1.346
Slippery road surface	0.615	0.079	7.803	<0.001	0.461	0.770
Emergency merging vehicle	0.370	0.073	5.078	<0.001	0.227	0.513
Emergency ego vehicle	-1.436	0.084	-17.071	<0.001	-1.602	-1.272
No following vehicle	0.613	0.073	8.398	<0.001	0.471	0.756
GPT as the decisionmaker	0.865	0.102	8.483	<0.001	0.664	1.067
English prompts	0.853	0.102	8.354	<0.001	0.653	1.054
Chinese prompts	-1.756	0.096	-18.306	<0.001	-1.945	-1.568
Threshold (0/1)	-3.995	0.163	-24.443	<0.001	-4.316	-3.675

Threshold (1/2)	1.030	0.051	20.126	<0.001	0.930	1.131
Number of observations	1500					
Log-likelihood function at convergence, ℓ	-733.150					
AIC	1494.000					
BIC	1569.000					

Note. Binary indicators are one-hot encoded with one level omitted as the reference category: sunny weather, daytime, and dry road surface for the environmental indicators; non-emergency vehicle types for the emergency-vehicle indicators; DeepSeek and LLaMA pooled for the model indicator; and French for the prompt-language indicators.

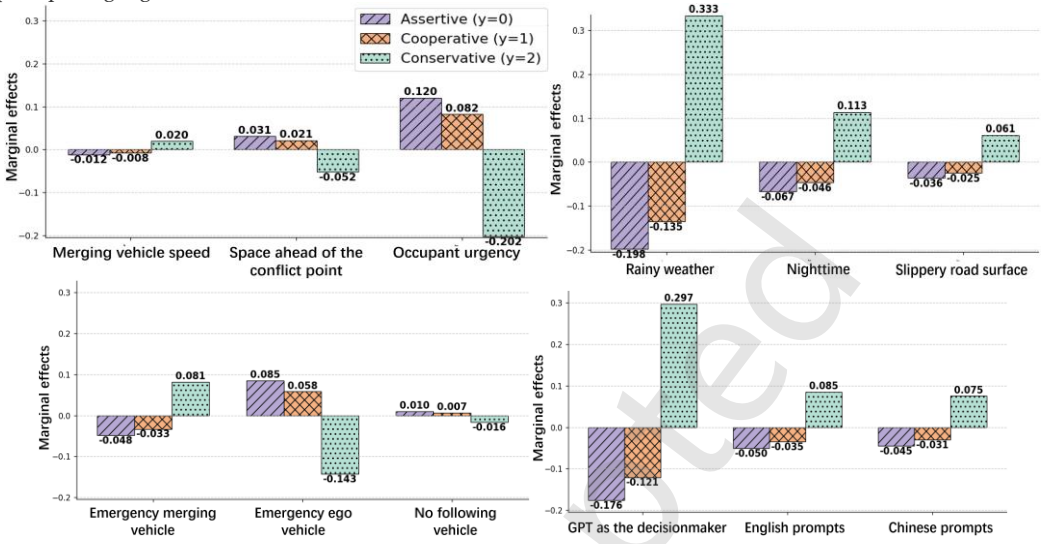
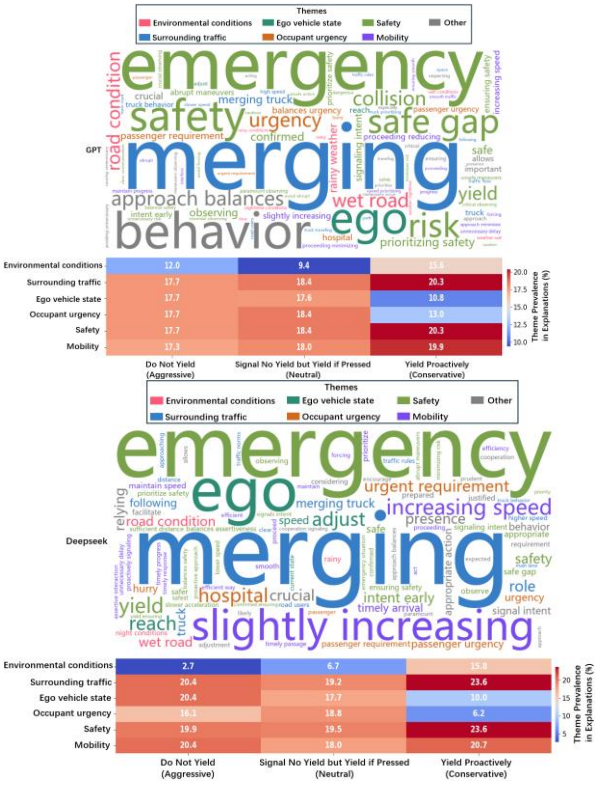


Fig. 4. The marginal effects of each factor across the three behavior categories: aggressive (y=0), neutral (y=1), and conservative (y=2). Reported values are average marginal effects per one standard deviation of the predictor; for binary indicators they are local slopes of the fitted probability surface rather than discrete 0-to-1 changes.

Fig. 5 shows that all three models consider a broad range of themes when explaining their decisions. Deepseek and LLaMA reference more mobility-related phrases (e.g., increasing speed, maintaining speed), consistent with a greater relative weight on traffic flow. This aligns with the OLM findings that both models select assertive options more often than GPT. The heatmaps reveal that environmental conditions receive relatively little attention across all behavioral choices. However, Deepseek (2.7%–15.8%) and LLaMA (4.4%–12.9%) assign even less weight to this theme than GPT (9.4%–15.6%), corresponding to a greater relative weight on dynamic traffic elements and performance-related factors. For aggressive behavior, GPT and LLaMA show relatively balanced attention across non-environmental themes, while Deepseek places less emphasis on occupant urgency. In neutral responses, LLaMA references surrounding traffic, ego vehicle state, and safety most frequently. GPT and Deepseek distribute their references more evenly across themes. When selecting conservative actions, all three models reduce attention to the ego vehicle state and occupant urgency. This indicates that conservative selections are accompanied by fewer references to internal or goal-related themes and more references to surrounding traffic and safety.

3.4. Following a cyclist

In this scenario, the LLM-powered AV follows a cyclist in a shared lane. We define three behavioral options: Overtake with Signal (Aggressive)—the ego vehicle lightly honks and accelerates to pass the cyclist; Negotiate the Overtake (Neutral)—the vehicle closes the gap slightly without honking, positioning itself to signal intent and gauge the cyclist's response; and Follow without Overtaking (Conservative)—the vehicle slows down and maintains a safe distance, not attempt to pass. These options are evaluated using an OLM, with results summarized in **Table 4** and marginal effects shown in **Fig. 6**. The model converged with a log-likelihood of −398.280, an AIC of 822.600, and a BIC of 891.600.



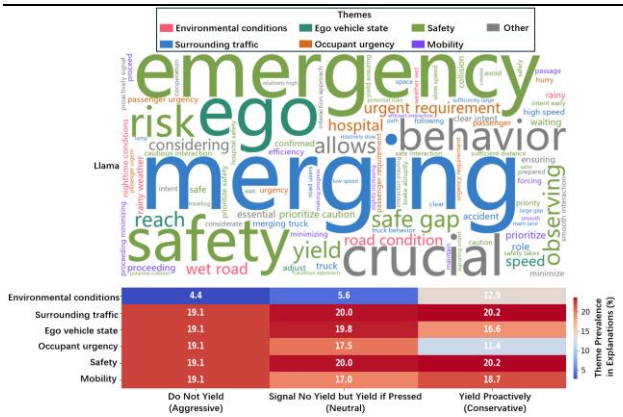


Fig. 5. Thematic analysis of LLM-generated explanations in the "yielding to a merging vehicle" scenario when prompted in English.

Table 4. Ordered logit model estimation results for the "following a cyclist" scenario: aggressive (y=0, n_0=163), neutral (y=1, n_1=210), and conservative (n_2=1127).

Independent variable	Coefficient	Std. Err	z	P>|z|	95% CI (Lower)	95% CI (Upper)
Space ahead of the bicycle	-0.604	0.105	-5.730	<0.001	-0.811	-0.398
Occupant urgency	-1.901	0.148	-12.883	<0.001	-2.191	-1.612
Rainy weather	4.222	0.308	13.673	<0.001	3.617	4.828
Nighttime	1.630	0.172	9.466	<0.001	1.293	1.968
Slippery road surface	0.681	0.149	4.583	<0.001	0.388	0.972
Bicycle in the middle of the lane	0.293	0.104	2.827	0.005	0.090	0.496
Child cyclist	0.707	0.109	6.465	<0.001	0.493	0.922
Emergency ego vehicle	-2.203	0.146	-15.249	<0.001	-2.486	-1.920
Emergency following vehicle	-0.446	0.106	-4.212	<0.001	-0.723	-0.243
GPT as the decisionmaker	3.650	0.265	13.756	<0.001	3.134	4.171
English prompts	0.623	0.126	4.930	<0.001	0.375	0.871
Threshold (0/1)	-4.911	0.268	-18.305	<0.001	-5.438	-4.384
Threshold (1/2)	-0.220	0.120	-1.844	0.065	-0.456	0.014
Number of observations	1500					
Log-likelihood function at convergence, ℓ	-398.280					
AIC	822.600					
BIC	891.600					

Note. Binary indicators are one-hot encoded with one level omitted as the reference category: sunny weather, daytime, and dry road surface for the environmental indicators; non-emergency vehicle types for the emergency-vehicle indicators; DeepSeek and LLaMA pooled for the model indicator; and Chinese and French pooled for the prompt-language indicator.

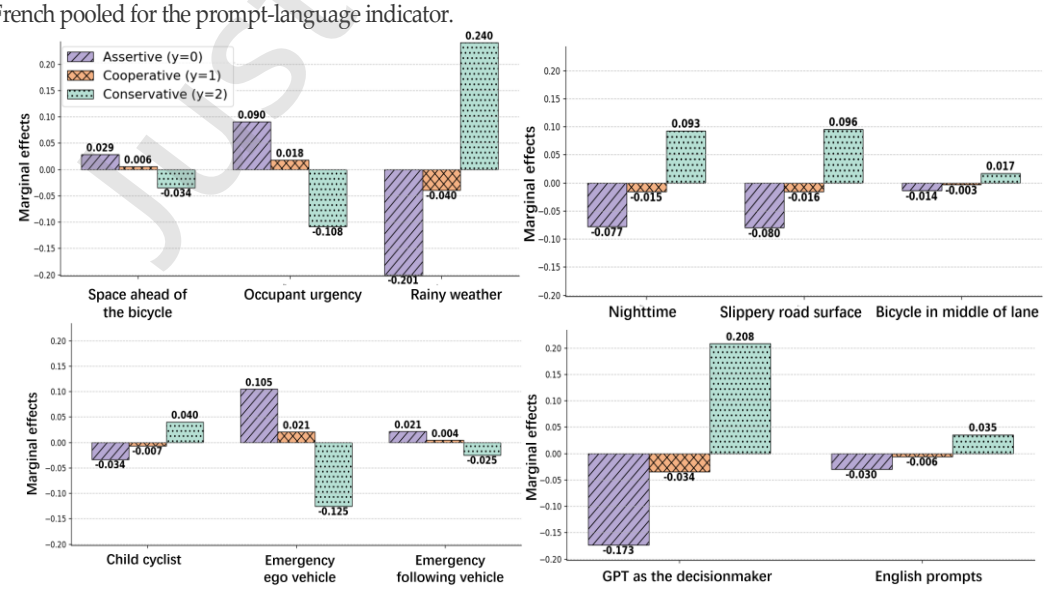


Fig. 6. The marginal effects of each factor across the three behavior categories: aggressive (y=0), neutral (y=1), and conservative (y=2). Reported values are average marginal effects per one standard deviation of the predictor; for binary indicators they are local slopes of the fitted probability surface rather than discrete 0-to-1 changes.

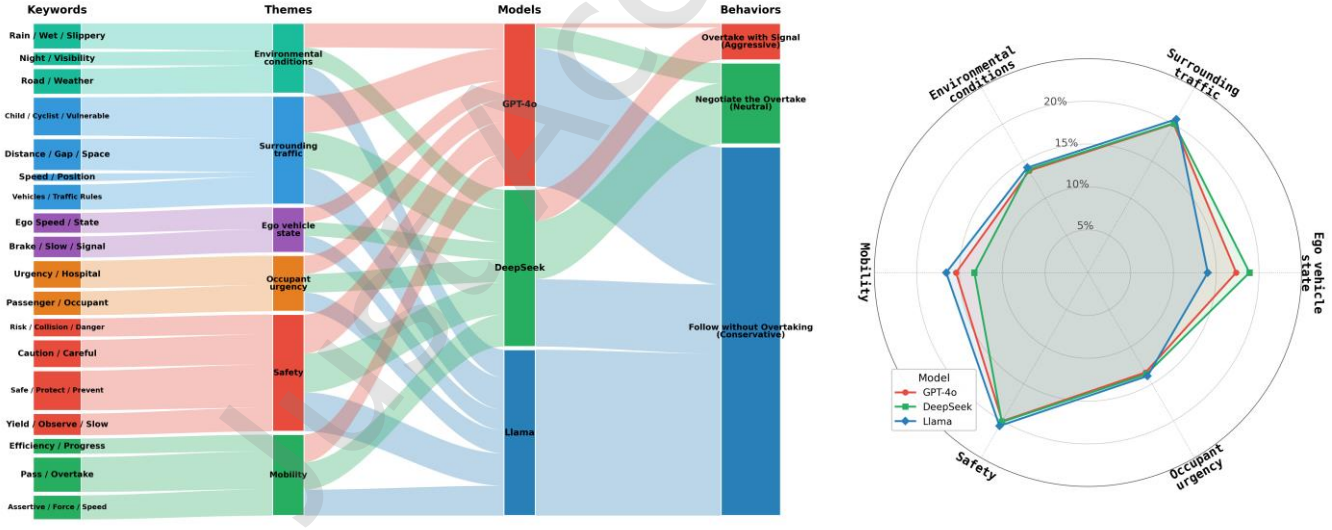
The OLM results in **Table 4** and the marginal effects in **Fig. 6** show how spatial context, road-user attributes, and external pressure are associated with the option selected in the cyclist scenario. A one-standard-deviation increase in the space ahead of the cyclist increases the probability of the assertive overtaking option (+0.029), indicating that greater clearance ahead is associated with a higher probability of overtaking. However, when the cyclist rides in the center of the lane, the probability of the conservative option increases (+0.017). The presence of a child cyclist increases the probability of the conservative option further (+0.040), the largest effect among the road-user attributes in this scenario. In contrast, when an emergency vehicle is following the ego vehicle, the probability of the assertive option increases (+0.105), and the justifications in this condition make frequent reference to clearing the lane. Model identity and prompt language are again associated with the selected option. GPT selects the conservative option markedly more often (+0.208), consistent with the preceding scenarios. English prompts are similarly associated with a higher probability of the conservative option (+0.035).

Unlike the previous two scenarios, the bicycle-following scenario shows a clear shift toward more conservative behavior, most notably with Llama selecting only a single behavioral option. We use a Sankey diagram instead of a heat map to better illustrate how keywords are aggregated into themes and then mapped to model-specific behavioral choices. **Fig. 7a** reveals a clear thematic concentration across all three models, with surrounding traffic, safety, and mobility serving as the dominant drivers of decision-making. Keywords related to cyclist interaction (e.g., cyclist/vulnerable, distance/gap/space), risk awareness (e.g., collision, danger, safe/protect), and overtaking feasibility (e.g., pass/overtake, efficiency) are consistently aggregated into these themes, indicating a high degree of overlap in the themes the three models reference for this scenario. Specifically, the justifications of all three models are dominated by references to the vulnerable road user, collision avoidance, and traffic progress, with comparatively few references to ego vehicle state and occupant requirement.

Despite this thematic overlap, the option distributions differ substantially. DeepSeek shows the widest spread across options, with mobility-related themes appearing in its assertive selections more frequently than for the other two models. GPT-4o selects the conservative option predominantly, with a limited number of transitions to other options. Llama selects the conservative option exclusively, with all themes mapping to that single outcome. These results indicate that although the three models reference similar themes in cyclist interactions, the distribution of selected options differs substantially across them.

3.5. Handling pedestrian crossing

In this scenario, the LLM-powered AV approaches a crosswalk with a nearby pedestrian. Three behavioral options are provided: Proceed Without Yielding (Aggressive)—the ego vehicle maintains speed, assuming the pedestrian will not cross; Approach and Probe Cautiously (Neutral)—the vehicle slows down while continuing forward to assess the pedestrian's intent; and Yield Proactively (Conservative)—the vehicle slows significantly or stops to invite the pedestrian to cross. Across all three LLMs, the Approach and Probe Cautiously (Neutral) option was



(a) (b)

Fig. 7. Thematic analysis of LLM-generated explanations in the "following a cyclist" scenario and the "handling pedestrian crossing" scenario when prompted in English.

selected almost exclusively. Deepseek selected it 1,428 times, compared to 17 Aggressive and 55 Conservative responses, while GPT and LLaMA chose Neutral exclusively. To rule out the possibility that this concentration is caused by predefined option framing, we conduct an additional open-ended analysis in Appendix A2, and the results remain broadly consistent with the main experiment.

Due to the lack of significant variation in behavioral choices, a quantitative analysis using the ordered logit model is not applicable. Instead, the analysis focuses on qualitative interpretation through thematic analysis.

Because behavioral choices in the pedestrian-crossing scenario are identical across models, we use a radar chart to compare their relative

thematic emphasis. **Fig. 7b** shows that all three models consistently prioritize surrounding traffic and safety, while placing less emphasis on mobility and occupant urgency. This pattern indicates that, when faced with ambiguous pedestrian intent, LLMs adopt a more cautious and socially attentive stance centered on pedestrian-related interaction and risk avoidance. At the same time, differences emerge mainly in the relative weighting of ego vehicle state and mobility. DeepSeek places greater emphasis on ego vehicle state and less on mobility, whereas Llama shows the opposite tendency; GPT-4o remains between the two and exhibits the most balanced profile. These results suggest that, as alignment-trained models, all three LLMs exhibit a consistent emphasis on pedestrian safety in this scenario, reflecting a shared prioritization of vulnerable road users that aligns with human moral and ethical expectations. However, the observed differences in ego vehicle state and mobility indicate that the models still vary in how they internally balance self-motion monitoring and traffic progression, likely possibly reflecting differences in training data and alignment strategies. Notably, in this scenario, variations in thematic emphasis do not translate into differences in behavioral choice, suggesting that the influence of pedestrian-related safety considerations dominates and suppresses alternative action tendencies. This contrasts with the earlier scenarios, where differences in thematic focus were more directly reflected in behavioral variation.

4 Discussion

The observed tendency for French and Chinese prompts to elicit more assertive behavior may be interpreted from several perspectives, though definitive causal explanations remain difficult given the black-box nature of LLMs. First, safety-alignment coverage is uneven across languages: alignment and safety tuning are conducted primarily in English (Ouyang et al., 2022), and non-English languages are underrepresented in training data (Bender et al., 2021), so guardrails and risk boundaries may be less densely specified for non-English inputs. Second, tokenizers compress languages at differing rates, and because the contextual factors varied across our cases are distributed throughout the prompt, differences in token granularity could shift how attention is allocated among those factors; token counts are measurable, but attention weights are not accessible for the models evaluated here. Third, French-language training data may reflect cultural norms that differ from English-language sources; cross-cultural research has documented that French drivers exhibit relatively more assertive driving patterns compared to the European average (Fondation VINCI Autoroutes, 2024), and if such patterns are present in LLM training corpora, they could contribute to shaping responses activated by French-language prompts. These three mechanisms cannot be separated within the present design, since changing the prompt language alters alignment coverage, tokenization, and corpus composition simultaneously; controlled experiments on open-weight models would be required to disentangle them. In the absence of publicly available information on LLM training data composition, these interpretations are informed by our empirical observations and existing literature rather than verified causal mechanisms. A fourth mechanism, differences in the semantic or emotional intensity with which assertive maneuvers are described across languages, can be assessed directly and is not supported. Because our prompts are produced by forward translation, any systematic intensification in Chinese or French would produce exactly the shift we observe. As reported in Appendix A1, back-translation yields cosine similarity above 0.96 for the system prompts and BERTScore F1 no lower than 0.942 for the behavioral options, and the tone assessment in Appendix A1.3 scored command strength, politeness, and negotiation framing at 0 in all eight comparison pairs, with no pair receiving the maximum score on any dimension. The observed language effects therefore reflect how the models respond to the prompt language itself rather than differences in what the prompts assert.

Another notable finding is that GPT-4o consistently exhibits a more conservative decision tendency compared to DeepSeek-V3 and Llama 3.1. One possible explanation is the difference in alignment approaches across models. GPT-4o was developed with substantial emphasis on post-training alignment, red teaming, and risk evaluation (Hurst et al., 2024), which may encourage more cautious responses in risk-sensitive driving situations. DeepSeek-V3 emphasizes large-scale pretraining and reasoning-oriented post-training (Liu et al., 2024), while Llama 3.1 highlights the role of downstream system-level safeguards, suggesting a more developer-configurable safety posture (Dubey et al., 2024). These differences may contribute to distinct behavioral priors, though because training data, alignment procedures, and deployment safeguards are intertwined and not fully observable, this interpretation should be treated as suggestive rather than causal. This comparison concerns the models' text-based decision outputs only and does not speak to how any of these models would perform within a complete driving system.

A further implication concerns system-level deployment. Because the models differ systematically in disposition while retaining a stochastic component, safety cannot rest on prompt design or model selection alone. A defensible architecture would situate the language model within a hard safety envelope enforced by a deterministic supervisory layer, such as a safety shield or a control barrier function, which admits an intent only if the resulting state remains within a certified safe set and otherwise projects it onto the nearest admissible action. The model would then condition preference among behaviors already verified safe rather than determine safety itself, so that the guarantee derives from the supervisory layer and holds irrespective of which model is queried or how it is prompted. Within such an envelope, the behavioral differences documented here become a design resource rather than only a source of inconsistency: models with distinct dispositions could underpin selectable or context-dependent driving modes, favoring caution under adverse weather or near vulnerable road users and efficiency in benign conditions or where occupant urgency is high. Such switching would modulate preference inside the envelope without altering the envelope itself. Realizing either arrangement would require closed-loop validation beyond the scope of this study.

5 Limitations

Two limitations should be noted. First, because all cases within a language share an identical system prompt template, prompt formulation is held constant across models and scenarios. This protects the comparative claims made here, but leaves open how the absolute distribution over options would shift under alternative phrasings.

Second, this study probes textual intent selection rather than closed-loop driving. A discrete intent such as merging forcefully specifies a behavioral disposition, not a trajectory, so transferring this approach to a simulator such as CARLA or nuPlan would require an intermediate layer

translating each option into parameterized trajectory constraints. Two difficulties follow. A selected intent may be kinematically infeasible under the prevailing state, and where a planner clips an aggressive intent to preserve a minimum TTC or headway, the executed behavior reflects the safety layer rather than the model, which complicates attributing closed-loop performance to the language model. The behavioral priors documented here are therefore best understood as inputs to a decision layer whose closed-loop consequences depend on planner design and safety supervision.

6 Conclusions

In this study, we prompt three LLMs with multilingual textual instructions to select high-level action options across four everyday driving scenarios. We then use the Ordered Logit Model and thematic analysis to investigate their prompt-conditioned driving-style preferences and decision rationales. GPT consistently favors conservative options, reflecting a more safety-oriented preference, while DeepSeek and LLaMA demonstrate more assertive tendencies. English prompts are associated with more cautious responses, whereas Chinese and French prompts are associated with greater assertiveness; these are observed regularities whose origins cannot be established from the models' undisclosed training and alignment procedures. When interacting with vulnerable road users such as cyclists and pedestrians, all models place greater emphasis on safety, resulting in fewer aggressive selections. These patterns suggest that model-specific characteristics, linguistic framing, and traffic context jointly shape prompt-conditioned high-level action preferences.

These findings offer guidance for the design and preliminary evaluation of LLM-based high-level decision modules in AV systems. Model selection should consider the desired policy-level driving style and risk tolerance, as different models exhibit distinct behavioral preferences. Prompt design can also be used to reinforce desired tendencies, while the observed sensitivity to multilingual prompts highlights the need for careful prompt standardization and localization in international deployment. This study characterizes textual decision tendencies under controlled prompts and does not validate LLMs for direct vehicle control or any safety-critical actuation.

Author contributions

Zhipeng Bao: Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Wenjie Zhao:** Writing – review & editing, Methodology. **Qianwen(Cami) Li:** Writing – review & editing, Methodology, Supervision, Formal Analysis.

Replication and data sharing

All data, code, and supplementary appendices (A1–A5) used in this study are openly available at https://github.com/hzxbzp/Probing-LLMs-Autonomous-Driving to ensure transparency, reproducibility, and ease of future extension.

Declaration of competing interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Declaration of generative AI and AI-assisted technologies in the writing process

The author utilized OpenAI's ChatGPT to assist with grammar and language improvements while preparing this manuscript. Its use was limited to enhancing the readability and fluency of the text, with no involvement in the scientific content, analysis, or interpretation. After utilizing this tool/service, the author(s) thoroughly reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency,

Bolovinou, A., Amditis, A., Bellotti, F., & Tarkiainen, M. (2014). Driving style recognition for co-operative driving: A survey. The Sixth International Conference on Adaptive and Self-Adaptive Systems and Applications,

Cui, C., Yang, Z., Zhou, Y., Ma, Y., Lu, J., & Wang, Z. (2023). Large language models for autonomous driving: Real-world experiments. CoRR.

Dong, J., Chen, S., Miralinaghi, M., Chen, T., & Labi, S. (2022). Development and testing of an image transformer for explainable autonomous driving systems. Journal of Intelligent and Connected Vehicles, 5(3), 235–249.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., & Fan, A. (2024). The llama 3 herd of models. arXiv e-prints, arXiv: 2407.21783.

Hu, Z., Xu, M., & Cheng, Q. (2025). Multimodal Large-Language model empowering Next-Generation autonomous driving systems. Journal of Intelligent and Connected Vehicles, 8(2), 9210059–9210051–9210059–9210053.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., & Radford, A. (2024). Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Kuang, S., Liu, Y., Wang, X., Wu, X., & Wei, Y. (2024). Harnessing multimodal large language models for traffic knowledge graph generation and decision-making. *Communications in Transportation Research*, *4*, 100146.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50–80.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th annual meeting of the association for computational linguistics.

Liao, H., Shen, H., Li, Z., Wang, C., Li, G., Bie, Y., & Xu, C.

(2024). GPT-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models. *Communications in Transportation Research*, *4*, 100116.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., & Ruan, C. (2024). Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.

Marcoux, R., Yasmin, S., Eluru, N., & Rahman, M. (2018). Evaluating temporal variability of exogenous variable impacts over 25 years: An application of scaled generalized ordered logit model for driver injury severity. Analytic methods in accident research, 20, 15–29.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730–27744.

Peng, M., Guo, X., Chen, X., Chen, K., Zhu, M., Chen, L., & Wang, F.-Y. (2025). Lc-llm: Explainable lane-change intention and trajectory predictions with large language models. Communications in Transportation Research, 5, 100170.

Renz, K., Chen, L., Marcu, A.-M., Hünermann, J., Hanotte, B., Karnsund, A., Shotton, J., Arani, E., & Sinavski, O. (2024). Carllava: Vision language models for camera-only closed-loop driving. arXiv preprint arXiv:2406.10165.

Rettenberger, L., Reischl, M., & Schutera, M. (2025). Assessing political bias in large language models. Journal of Computational Social Science, 8(2), 1–17.

Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., & Tara, M. (2025). A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Medical Informatics and Decision Making, 25(1), 117.

Takemoto, K. (2024). The moral machine experiment on large language models. Royal Society open science, 11(2), 231393.

Wang, Q., Pan, S., Linzen, T., & Black, E. (2025). Multilingual Prompting for Improving LLM Generation Diversity. arXiv preprint arXiv:2505.15229.

Winder, P., Hildebrand, C., & Hartmann, J. (2025). Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors. PloS one, 20(6), e0325459.

Xu, Z., Sengar, N., Chen, T., Chung, H., & Oviedo-Trespalacios, O. (2025). Where is morality on wheels? Decoding large language model (LLM)-driven decision in the ethical dilemmas of autonomous vehicles. Travel Behaviour and Society, 40, 101039.

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP),

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., & Dong, Z. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2).

Author biography

Zhipeng Bao received the B.S. degree in Automotive Engineering from Wuhan University of Technology, and the M.S. degree in Vehicle Engineering from Jilin University, China. He is currently pursuing a Ph.D. candidate in Mechanical Engineering at the University of Georgia, USA. His research interests lie in embodied AI, the application of vision language models in autonomous driving. He has worked on vehicle trajectory prediction, drone trajectory planning, 3D computer vision and SLAM systems.

Wenjie Zhao received her master's degree in Systems Science and her bachelor's degree from Beijing Jiaotong University, Beijing, China. She is currently pursuing a Ph.D. in Civil Engineering at the University of Georgia, US. Her research focuses on intelligent transportation systems, with particular emphasis on transportation modeling and autonomous vehicle applications.

Qianwen(Cami) Li is an Assistant Professor in the School of Environmental, Civil, Agricultural and Mechanical Engineering (ECAM) at the University of Georgia. She received his Ph.D. in Civil Engineering (Transportation) from the University of South Florida. Her research focuses on enhancing transportation safety and efficiency using emerging technologies such as vehicle automation, communication, electrification, and smart infrastructure.