

Online Adaptive Probabilistic Safety Certificate with Language Guidance

Zhuoyuan Wang*

Carnegie Mellon University

ZHUOYUAW@ANDREW.CMU.EDU

Xiyu Deng*

Carnegie Mellon University

XIYUD@ANDREW.CMU.EDU

Hikaru Hoshino*

University of Hyogo

HOSHINO@ENG.U-HYOGO.AC.JP

Yorie Nakahira

Carnegie Mellon University

YORIE@CMU.EDU

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Achieving long-term safety in uncertain/extreme environments while accounting for human preferences remains a fundamental challenge for autonomous systems. Existing methods often trade off long-term guarantees for fast real-time control and cannot adapt to variability in human preferences or risk tolerance. To address these limitations, we propose a language-guided adaptive probabilistic safety certificate (PSC) framework that guarantees long-term safety for stochastic systems under environmental uncertainty while accommodating diverse human preferences. The proposed framework integrates natural-language inputs from users and Bayesian estimators of the environment into adaptive safety certificates that explicitly account for user preferences, system dynamics, and quantified uncertainties. Our key technical innovation leverages probabilistic invariance—a generalization of forward invariance to a probability space—to obtain myopic safety conditions with long-term safety guarantees. We validate the framework through numerical simulations of autonomous lane-keeping with human-in-the-loop guidance under uncertain and extreme road conditions, demonstrating enhanced safety–performance trade-offs, adaptability to changing environments, and personalization to different user preferences. Code is available at https://github.com/hoshino06/adaptive_lane_keeping.

Keywords: Safety; Adaptation; Stochastic systems; LLMs; Language-guided control.

1. Introduction

Ensuring long-term safety in autonomous systems operating under uncertain and evolving environments is a fundamental challenge. In many real-world scenarios such as autonomous driving, both the environment and the human user’s preferences can vary over time. For instance, a driver may prefer specific driving styles and comfort levels while the road friction and traffic conditions remain unknown. Despite such variability, safety must always remain the top priority. This motivates the need for adaptive safe control frameworks that can efficiently adjust to diverse human preferences and uncertain dynamics, while maintaining provable long-term safety guarantees. However, existing safe control methods face several limitations. First, most existing approaches lack systematic mechanisms to account for language guidance and users’ preferences and risk tolerance, leading to suboptimal performance and alignment. Second, approaches based on adaptive control barrier functions often trade off long-term safety guarantees for real-time efficiency (Taylor and Ames, 2020;

* Equal contribution.

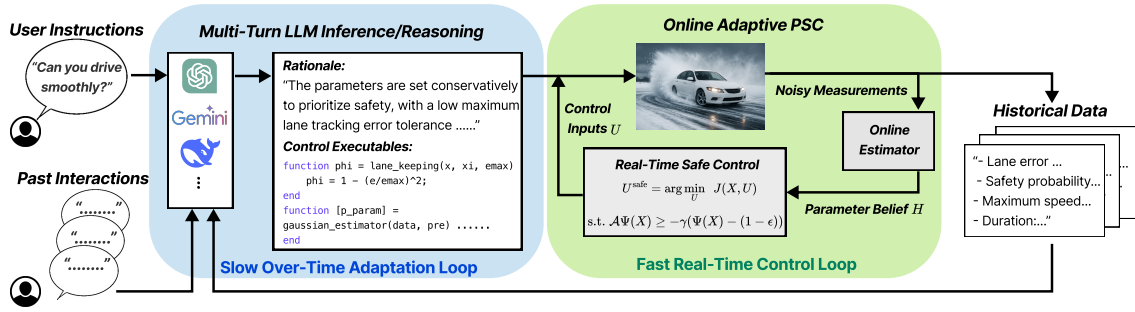


Figure 1: The proposed language-guided adaptive probabilistic safety certificate (PSC) framework.

Xiao et al., 2021), while probabilistic reachability-based techniques typically require expensive on-line computation to remain valid under environment uncertainty (Abate et al., 2008; Chapman et al., 2019; Kariotoglou et al., 2013).

To address these challenges, we propose a novel language-guided adaptive probabilistic safety certificate (PSC) framework, which ensures long-term safety of stochastic systems in uncertain and adaptive settings while accounting for evolving human specifications and preferences. The overall framework is illustrated in Fig. 1. Our method enforces forward invariance in the probability space to guarantee long-term safety, achieved by integrating a system parameter estimator that updates the uncertainty representation with a safety certificate applying invariance conditions based on the current estimate. We derive a sufficient safety condition that guarantees long-term safety remains above a user-specified threshold, and show that this condition yields one-step control constraints that can be seamlessly incorporated into optimization-based controllers including model predictive control (MPC) with arbitrary planning horizons. Unlike conventional control barrier function (CBF) approaches that operate in the state space and rely on hand-crafted barrier functions, the proposed adaptive PSC requires only a characterization of the desired safe set, enabling multi-turn large language models (LLMs) (Yi et al., 2024; Zhang et al., 2025) to effectively translate natural language instructions into formal safety specifications without compromising control performance. This design facilitates efficient online adaptation to evolving user guidance while ensuring long-term probabilistic safety under model uncertainty. The main contributions of the paper are as follows.

- We propose a novel adaptive probabilistic safety certificate that enforces forward invariance in the probability space, to systematically ensure long-term safety of uncertain systems with user-desired risk levels (Theorem 2).
- We develop a holistic multi-turn LLM-based adaptive safe control framework that can translate natural language guidance into formal safety specifications, and adapt to changing environments and user preferences while maintaining long-term safety (Fig. 1, 6 and Table 1, 2).
- We demonstrate the effectiveness of the proposed method in autonomous lane-keeping scenarios under uncertain and extreme road conditions. The results show significantly improved safety vs. efficiency trade-offs, effective adaptation to evolving human preferences, and sustained long-term safety with reduced online computational cost (Fig. 2, 5 and Table 1, 2).

2. Related Work

Adaptive Safe Control. A wide range of safe control techniques has been developed for systems subject to noise and uncertainties when the safety specifications are given and fixed. For example,

robust control barrier functions (CBFs) have been proposed to handle worst-case bounded disturbances in deterministic systems (Cosner et al., 2021; Jankovic, 2018; Prajna et al., 2007). To further address model uncertainties, adaptive CBFs (Taylor and Ames, 2020; Xiao et al., 2021) and their robust (Lopez et al., 2020; Castaneda et al., 2021; Long et al., 2022) and high-order (Xiao and Belta, 2021; Cohen and Belta, 2022; Nguyen and Sreenath, 2016) variants have been introduced. Many of these methods consider bounded disturbances and worst-case framework, and may not be immediately applicable for stochastic systems.

For stochastic systems, methods such as probabilistic CBFs (Luo et al., 2020; Lyu et al., 2021; Fan et al., 2020) and conditional value-at-risk (CVaR)-based approaches (Ahmadi et al., 2021) have been developed. On the other hand, stochastic reachability or probabilistic reachability techniques are developed to estimate the likelihood that the system trajectory remains within safe regions or reaches desired targets in the presence of unbounded noise and uncertainties (Fisac et al., 2018; Abate et al., 2008; Chapman et al., 2019; Kariotoglou et al., 2013; Abate et al., 2006; Vasileva et al., 2020). These techniques are often subject to a stringent tradeoff between assuring long-term safety vs. reducing real-time computation and may not be easily adapted to changing environments and human preferences (Wang et al., 2026). In this paper, we allow the safe sets to depend on language guidance from human users, and the safety certificates to adapt to changing environments.

Model predictive control (MPC) methods have also been widely studied for decision-making under uncertainty, including chance-constrained, stochastic, and probabilistic MPC formulations (Farina et al., 2016; Heirung et al., 2018; Mesbah, 2016). The proposed safety certificate can be seamlessly incorporated into MPC as additional constraints. Unlike conventional MPC approaches, where the planning horizon must exceed the safety constraint horizon to maintain long-term safety, the proposed safety certificate can enable longer-term safety with shorter MPC horizons. With augmented risk information, the proposed method can decouple the computational complexity of MPC from the long-term safety horizon (see Fig. 2-4).

Language-Guided Control. Large language models (LLMs) have demonstrated the effectiveness of multi-turn prompting for generating contextually consistent and coherent responses through iterative refinement (Yi et al., 2024; Zhang et al., 2025). To incorporate human intent and preferences into control, LLMs have been combined with model predictive control (MPC) frameworks to interpret and integrate natural-language guidance into control objectives (Maher, 2025; Miyaoka et al., 2024; Sanyal and Roy, 2025). LLMs have also been used to generate context-aware behaviors (Deng et al., 2024; Khan et al., 2025; Wu et al., 2025) or to couple language reasoning with reachability analysis for safe motion planning (Bayat et al., 2025; Hafez et al., 2025). These approaches highlight the potential of LLMs as adaptive reasoning modules that translate high-level language inputs into control-relevant information.

Recent studies have explored LLM-guided safe control, where language models assist in constructing control barrier functions (CBFs) and defining corresponding safety conditions for deterministic systems (Khan et al., 2025; Sanyal and Roy, 2025; Brunke et al., 2025). However, since multiple CBF formulations can satisfy the same safety requirement yet induce markedly different control behaviors, these approaches often rely on manual design of CBFs by human engineers or unreliable automatic generation of barrier functions by LLMs. In contrast, the proposed adaptive probabilistic safety certificate (PSC) framework leverages LLMs solely to generate high-level safe set specifications, which are subsequently translated into formal safety conditions using stochastic system models and Bayesian estimators. This yields a principled and reliable process for constructing safety certificates that ensure long-term safety and alignment in stochastic, non-stationary environments.

3. Problem Formulation

Preliminary. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ represent that f is a mapping from space \mathcal{X} to space \mathcal{Y} . Let $\mathbb{1}\{\mathcal{E}\}$ be an indicator function, which takes 1 when condition \mathcal{E} holds and 0 otherwise. Given events \mathcal{E} and \mathcal{E}_c , let $\mathbb{P}(\mathcal{E})$ be the probability of \mathcal{E} and $\mathbb{P}(\mathcal{E}|\mathcal{E}_c)$ be the conditional probability of \mathcal{E} given the occurrence of \mathcal{E}_c . Given random variables X and Y , let $\mathbb{E}[X]$ be the expectation of X and $\mathbb{E}[X|Y = y]$ be the conditional expectation of X given $Y = y$. We use upper-case letters (e.g., Y) to denote random variables and lower-case letters (e.g., y) to denote their specific realizations.

We consider adaptive safe control problems with language guidance, where human user provides task-related commands and preferences, and an adaptive controller is designed to follow human instructions, while maintaining *long-term* safety of the system despite system uncertainties. Specifically, we consider a discrete-time stochastic control system with unknown parameters and human-specified safety requirements. The system evolves according to

$$X_{k+1} \sim F(dX_{k+1} | X_k, U_k; \xi), \quad (1)$$

where $X_k \in \mathcal{X} \subset \mathbb{R}^n$ denotes the system state, $U_k \in \mathcal{U} \subset \mathbb{R}^m$ is the control input, and $\xi \in \mathbb{R}^l$ is an *unknown system parameter*. The transition kernel $F : \mathcal{B}(\mathcal{X}) \times \mathcal{X} \times \mathcal{U} \times \mathbb{R}^l \rightarrow [0, 1]$ characterizes the stochastic dynamics, with $\mathcal{B}(\mathcal{X})$ denoting the Borel σ -algebra of \mathcal{X} (Hernández-Lerma and Lasserre, 2012). A human-specified safety requirement is described by a function $\phi : \mathbb{R}^n \times \mathbb{R}^l \rightarrow \mathbb{R}$, and the corresponding safe set is defined as

$$\mathcal{C}(\xi) := \{x \in \mathbb{R}^n : \phi(x, \xi) \geq 0\}, \quad (2)$$

where ϕ captures safety constraints. While prior works manually design ϕ to improve safe control performance (Ames et al., 2019), our approach is agnostic to different yet equivalent representations of ϕ , enabling seamless integration with large language models (LLMs) for automating its construction from human specifications of safety requirements.¹ We then define an indicator function for long-term safety over a finite horizon $\tau_k = \{k, k+1, \dots, k+T\}$ as

$$\mathbb{1}_{\text{LS}}(X_{k:k+T}, \xi) := \begin{cases} 1, & \text{if } X_{k'} \in \mathcal{C}(\xi), \forall k' \in \tau_k, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where $X_{k:k+T} := \{X_k, X_{k+1}, \dots, X_{k+T}\}$ denotes the trajectory segment.

To capture epistemic uncertainty in the unknown parameter ξ , we define a sequence of random variables $\{\Xi_i\}_{i \in \mathbb{Z}_+}$ with corresponding probability density functions $\{H_i\}_{i \in \mathbb{Z}_+}$, each reflecting the current estimate of the uncertain parameter at time step i . These distributions characterize uncertainty arising from limited or noisy measurements and can be estimated using many estimators, including Bayesian neural networks (Hafner et al., 2018; Gal and Ghahramani, 2016) and Gaussian processes (Shahriari et al., 2015; Pan et al., 2017). The estimator and its configuration can be determined based on human specifications, encoding user preferences and desired adaptation behavior. Then, for a given nominal controller $\pi : \mathcal{X} \times \mathbb{R}^l \rightarrow \mathcal{U}$ and parameter belief H_i , the long-term safety probability from state x is defined as

$$\begin{aligned} \Psi_{H_i}^\pi(x) &= \mathbb{E}_{\Xi_i \sim H_i} [\mathbb{1}_{\text{LS}}(X_{k:k+T}, \Xi_i) | X_k = x] \\ &= \int_{\mathcal{B}(\mathcal{X})^T \times \mathbb{R}^l} \mathbb{1}_{\text{LS}}(X_{k:k+T}, \Xi_i) \mathbb{P}(dX_{k+1:k+T} | X_k = x, \pi, \Xi_i) H_i(\Xi_i) d\Xi_i, \end{aligned} \quad (4)$$

1. See Appendix H for ablation studies demonstrating consistent performance of the proposed method across diverse forms of (LLM-generated) barrier functions.

where the control inputs U_k is computed based on the policy π and the trajectory distribution is given by

$$\mathbb{P}(dX_{k+1:k+T} \mid X_k = x, \Xi_i) = \prod_{k'=k}^{k+T-1} F(dX_{k'+1} \mid X_{k'}, \pi(X_{k'}, \hat{\xi}_i); \Xi_i). \quad (5)$$

Here, $\hat{\xi}_i = \int_{\mathbb{R}^l} \Xi_i H_i(\Xi_i) d\Xi_i$ represents the current estimate of the unknown parameter, which is used in the nominal controller $\pi(X_k, \hat{\xi}_i)$. Since the dynamics (1) is time-invariant, $\Psi_{H_i}^\pi(x)$ does not depend on k .

The goal is to design an adaptive safe control policy π_{safe} that updates both the control and estimation strategy to ensure that, at every time step, the expected long-term safety probability maintains higher than a risk tolerance threshold $1 - \epsilon$, *i.e.*,

$$\mathbb{E}_{\pi_{\text{safe}}} [\Psi_{H_k}^\pi(X_k)] \geq 1 - \epsilon, \quad \forall k \in \mathbb{Z}_+. \quad (6)$$

The expectation is taken over the distribution of X_k conditioned on the initial state X_0 with the closed-loop policy $U_t = \pi_{\text{safe}}(X_t, \hat{\xi}_t)$ for $t \in \{k, \dots, k+T\}$. In addition to maintaining long-term safety (6), the controller minimizes a performance objective $J(X_k, U_k)$, which may vary based on user preferences. For instance, one may consider $J(X_k, U_k) = \mathbb{E}[\sum_{t=k}^{k+T} c(X_t, U_t)]$, where $c(\cdot, \cdot)$ is a cost function on states and controls. This formulation captures the interaction between human-specified safety requirements, estimator preferences, and performance objectives under a unified adaptive control framework that guarantees long-term safety in stochastic systems with unknown parameters. See Wang et al. (2026) for a detailed discussion of the advantages of the long-term safety formulation compared to existing stochastic safe control methods, such as Luo et al. (2020).

4. Proposed Method

In this section, we present the proposed safe control framework with language guidance. The overall architecture is illustrated in Fig. 1. The framework consists of two interconnected loops: an inner safe control loop, which operates in *real time* based on the adaptive probabilistic safety certificate (PSC), and an outer multi-turn LLM loop, which evolves at a *slower timescale* to incorporate human instructions and historical control data. The inner loop ensures long-term safety under uncertainty and diverse human-specified requirements, while the outer loop enables the system to adapt to human preferences and correct potentially erroneous information over time. In the following, we present the proposed adaptive PSC in Section 4.1 and its integration with the LLM in Section 4.2.

4.1. Adaptive Probabilistic Safety Certificate

In this section, we introduce the adaptive probabilistic safety certificate that guarantees long-term safety despite uncertainties in the dynamics. We start by defining a discrete-time generator $\mathcal{A}_{H_{k+1}}^{U_k}$ for the safety probability, which depends on the control input U_k and the density function H_{k+1} for the next time step:

$$\mathcal{A}_{H_{k+1}}^{U_k} \Psi_{H_k}^\pi(X_k) := \frac{\mathbb{E}[\Psi_{H_{k+1}}^\pi(X_{k+1}) \mid X_k, U_k] - \Psi_{H_k}^\pi(X_k)}{\Delta t}, \quad (7)$$

where Δt is the sampling time interval. Then, to ensure the long-term safety of the system, we propose to constrain the control action U_k to satisfy the following conditions at all time $k \in \mathbb{Z}_+$:

$$\mathcal{A}_{H_{k+1}}^{U_k} \Psi_{H_k}^\pi(X_k) \geq -\gamma(\Psi_{H_k}^\pi(X_k) - (1 - \epsilon)), \quad (8)$$

where $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ is a function that satisfies the following requirements:

Requirement 1: $\gamma(q)$ is strictly concave or linear and strictly increasing in q .

Requirement 2: $\gamma(q) \leq q, \forall q \in \mathbb{R}$.

Remark 1 In practice, choose $\gamma(q) = aq$ where $a \in [0, 1]$ will satisfy the two requirements. The proposed safety condition (8) is a control constraint on U_k . This constraint is linear in U_k if the system dynamics (1) is control affine (Wang et al., 2022b, 2026). Numerical methods for evaluating (7) are provided in Appendix B.

Theorem 2 Consider the stochastic dynamics given by (1) with a nominal controller π and a sequence of probability density functions H_0, H_1, \dots for estimating ξ . Suppose the state and parameter estimation originate from X_0 and H_0 satisfy the following

$$\Psi_{H_0}^\pi(X_0) > 1 - \epsilon, \quad (9)$$

and the control action U_k given by the controller π_{safe} satisfies (8) at all time. Then, the following condition holds:

$$\mathbb{E}_{\pi_{\text{safe}}}[\Psi_{H_k}^\pi(X_k)] \geq 1 - \epsilon, \quad \forall k \in \mathbb{Z}^+. \quad (10)$$

Proof See Appendix A. ■

Theorem 2 says that given the initial state to be sufficiently safe, with the constraint of the safety certificate (8), the resulting policy π_{safe} ensures long-term safety in the sense that the expected long-term probability $\Psi_{H_k}^\pi(X_k)$ is always above the desired tolerance. With these theoretical guarantees, the safe controller π_{safe} can be calculated at each time step using the following optimization.

$$\begin{aligned} U_k^{\text{safe}}(X_k, H_k) = \arg \min_U & J(\pi(X_k, \hat{\xi}_k), U) \\ \text{s.t.} & (8), \end{aligned} \quad (11)$$

where π is the reference controller², and J is a generic cost function that characterizes human's preference on the control performance. (e.g., quadratic cost on the distance between the safe control and a performance-oriented nominal control). Here $\hat{\xi}_k = \int_{\mathbb{R}^l} \Xi_k H_k(\Xi_k) d\Xi_k$ is the current estimate of the unknown parameter at step k . Similarly, the proposed adaptive safety certificate constraint (8) can also be integrated to the model predictive control (MPC) framework as below.

$$\begin{aligned} U_k^{\text{safe}}(X_k, H_k) = \arg \min_{U_k} & J(X_{k:k+T_{\text{mpc}}}, U_{k:k+T_{\text{mpc}}}) \\ \text{s.t.} & (1), \quad \forall t \in [k, k + T_{\text{mpc}}] \\ & (8), \quad t = k \end{aligned} \quad (12)$$

where T_{mpc} is the MPC planning horizon. Note that the long-term safety constraint (8) only needs to be enforced at the current time step k , and the long-term safety horizon T can be different than the MPC horizon T_{mpc} . This offers real-time computational advantages and strong safety guarantees with a short-horizon MPC, as demonstrated in the experiments in Section 5.2. The overall adaptive safe control algorithm is shown in Appendix C.

2. The reference controller π is usually set to be the nominal controller π to estimate safety probability, but the theory does not restrict the two to be the same.

4.2. Integration with Language Guidance

We integrate large language models (LLMs) with the proposed adaptive probabilistic safety certificate (PSC) to enable adaptation to human specifications and preferences while ensuring long-term safety under environmental uncertainties. The overall LLM integration pipeline is illustrated in Fig. 1 (highlighted in blue). The LLM adaptation loop operates on a slower timescale than the online adaptive safe control and therefore does not require real-time inference or reasoning. At initialization, the user provides natural language instructions describing the desired control behavior. Through a system-level prompt, an inference LLM interprets these instructions and generates rationales and control executables, such as safe set designs, estimator priors, and estimator update rules, which are passed to the adaptive PSC for control execution. Historical data on control performance, safety satisfaction, and estimator updates are collected and, in subsequent runs, provided together with the user’s latest instruction to a reasoning LLM that produces updated executables for control. By jointly leveraging human inputs and accumulated data, the multi-turn LLM can automatically correct erroneous human information and improve control performance, as demonstrated in Section 5.3. Implementation details of the LLM integration, including system-level prompts and algorithmic procedures, are provided in Appendix G.

5. Case Study: Adaptive Lane Keeping

5.1. Problem Setting

We consider a challenging lane-keeping problem to illustrate the effectiveness of the proposed method. Whereas lane-keeping problems are usually addressed under the assumptions of a constant longitudinal velocity and a linear tire model (Rossetter and Gerdes, 2006; Ames et al., 2017), we consider a lane-keeping problem under extreme road conditions where the friction coefficient can be low and uncertain. In wet or icy road conditions, tire forces may saturate, and coupled longitudinal and lateral vehicle dynamics affect tire stability (Wang et al., 2022a).

To simulate these complex dynamics, we use a three-degree of freedom (3-DOF) vehicle model with the LuGre combined-slip tire model to capture nonlinearities under extreme conditions (Huang et al., 2017; Li et al., 2023). The system state consists of three components: the vehicle motion x_{vehicle} , the wheel dynamics x_{wheel} , and the vehicle position x_{road} relative to the road coordinate. Specifically, the system state is

$$x = \underbrace{[v_x, v_y, r, \delta]}_{x_{\text{vehicle}}} \underbrace{[\omega_{fl}, \omega_{fr}, \omega_{rl}, \omega_{rr}, \tau_e]}_{x_{\text{wheel}}} \underbrace{[s, e, \psi]}_{x_{\text{road}}}]^\top. \quad (13)$$

Here, the state x_{vehicle} consists of the vehicle’s longitudinal and lateral velocities v_x and v_y , the yaw rate r , and the steering angle δ . The state x_{wheel} consists of the wheel’s angular velocities ω_{fl} , ω_{fr} , ω_{rl} and ω_{rr} , and the torque τ_e , which is greater than or equal to 0 when accelerating and smaller than 0 when braking. The state x_{road} consists of the distance s traveled from a reference point (starting point), the lateral error e from the center line of the road, and the relative heading error ψ . The control action is $u = [\Delta\delta, \Delta\tau_e]^\top$, where $\Delta\delta$ and $\Delta\tau_e$ stand for the rates of change of δ and τ_e , respectively. The unknown parameter ξ in the dynamics is the friction coefficient μ . The detailed vehicle dynamics model and parameters used can be found in Appendix D.

We formulate the control objective by using the safe set \mathcal{C} defined by (2), with

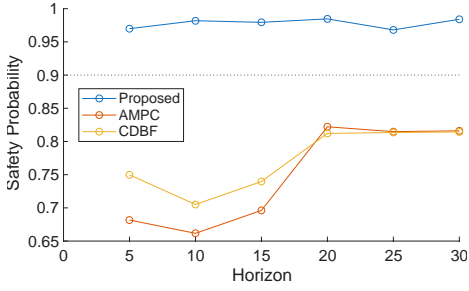


Figure 2: Safety probability vs. MPC horizon T_{mpc} .

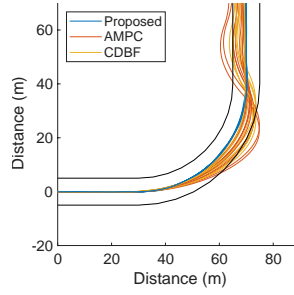


Figure 3: Vehicle trajectories ($T_{\text{mpc}} = 10$).

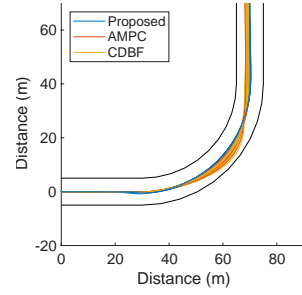


Figure 4: Vehicle trajectories ($T_{\text{mpc}} = 20$).

$$\phi(x, \xi) = 1 - \left(\frac{e}{e_{\max}} \right)^2, \quad (14)$$

where e_{\max} is the allowable limit of the absolute value of the lateral error e and can have different values given different levels of conservativeness. We use a Bayesian estimator for unknown parameter update of ξ , and the mean μ_k and the variance σ_k^2 of the Gaussian posterior $P_k(\mu|H_k)$ are sequentially updated by the measurement M_k of the friction coefficient as follows (Murphy, 2007):

$$\mu_k = \frac{\bar{\sigma}^2 \mu_{k-1} + \sigma_{k-1}^2 M_k}{\bar{\sigma}^2 + \sigma_{k-1}^2}, \quad \sigma_k^2 = \frac{\bar{\sigma}^2 \sigma_{k-1}^2}{\bar{\sigma}^2 + \sigma_{k-1}^2}, \quad (15)$$

where $\bar{\sigma}^2$ is the variance of M_k , reflecting the measurement noise level.

5.2. Long-Term Safety and Efficiency under Uncertainty

We first demonstrate that the proposed adaptive probabilistic safety certificate ensures long-term safety under diverse uncertain environments, and achieves better safety vs. efficiency trade-offs. Specifically, we investigate the integration of the adaptive PSC with model predictive control (MPC). Since standard control barrier function (CBF) constraints with (14) will result in infeasibility, we examine the following two baseline controllers: (i) Adaptive MPC (AMPC), a nonlinear MPC scheme employing the LuGre tire model to enhance vehicle stability under extreme road conditions (Li et al., 2023); and (ii) Control-Dependent Barrier Functions (CDBF) (Huang et al., 2021; Wang et al., 2022a), a variant of CBF that combines MPC with safety constraints to handle unstable tire forces and coupled longitudinal–lateral dynamics. MPC formulations and implementation details are provided in Appendix E.1 and online computation times are reported in Appendix E.2.

We tested closed-loop performance through 30 simulations with icy road conditions (unknown ground truth friction coefficient $\mu \in [0.2, 0.4]$) and desired safety probability $1 - \epsilon = 0.9$. Fig. 2 shows the effect of the prediction horizon T_{mpc} on performance for each controller. It can be seen that with AMPC and CDBF, the safety probability is compromised when T_{mpc} is reduced, while with the proposed adaptive PSC, the safety probability remains above the desired threshold $1 - \epsilon$ regardless of the MPC horizon. Fig. 3 and 4 show the vehicle trajectories with different MPC horizons T_{mpc} . For both AMPC and CDBF, only state constraints are imposed inside the MPC optimization, thus lane keeping behavior will fail at $T_{\text{mpc}} = 10$ and can only be ensured with larger $T_{\text{mpc}} = 20$. In contrast, the proposed controller achieves the desired safe behavior even

with a shorter MPC horizon, as the probabilistic constraint ensures long-term safety with a myopic evaluation.

We evaluated the closed-loop performance of all three methods under varying road conditions with unknown ground truth friction coefficients: icy ($\mu \in [0.3, 0.4]$), wet ($\mu \in [0.5, 0.6]$), and dry ($\mu \in [0.8, 0.9]$), and we tested with different estimator priors, measurement noise levels, and maximum lane-error tolerances, to conduct a comprehensive quantitative analysis. Specifically, we consider MPC horizon $T_{\text{mpc}} = 10$, estimator prior means $\xi_0 = [0.3, 0.5, 0.9]$, prior standard deviations $\sigma_0 = [0.05, 0.3]$, measurement noise standard deviations $\bar{\sigma} = [0.05, 0.3]$, and maximum lane-error tolerances $e_{\text{max}} = [3, 5, 10]$, yielding 108 distinct settings. For each feasible setting³ and each method, we run 20 simulations and visualize the average minimum safety probability and average vehicle velocity as trade-off plots in Fig. 5. We can see that the proposed adaptive PSC method consistently achieves the desired safety probability, which is significantly higher than that of AMPC and CDBF, while maintaining comparable average vehicle velocities. These results show that the proposed method achieves better trade-offs between safety and efficiency, by ensuring long-term safety without compromising performance.

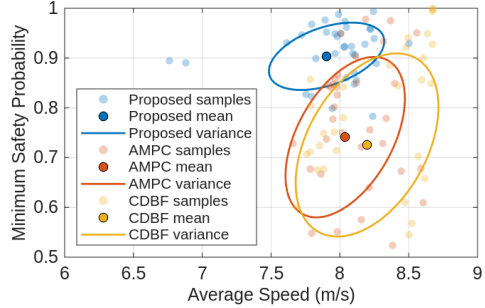


Figure 5: Safety vs. efficiency trade-offs with MPC-based adaptive controls.

5.3. Adaptation with Human Instructions

In this section, we demonstrate how the proposed framework integrates large language models (LLMs) with the adaptive probabilistic safety certificate (PSC) to enable safe and responsive adaptation in accordance with human preferences and specifications.

We first present a qualitative analysis of the proposed framework in Fig. 6, with three rounds of human instructions and feedback under a driving scenario with icy road conditions. It can be seen that with the initial human specification and preference, the vehicle produces a smooth and slower ride. Then with the second round human feedback, the proposed model adapted to the human preferences and achieved more aggressive driving behavior while maintaining safety. In the final round, when the human feedback becomes irrelevant, the model reasons from historical data and past interactions to infer intent, producing smooth and efficient driving behavior that balances performance and safety. Note that in all three cases, safety is guaranteed despite evolving user preferences and uncertain environments, thanks to the proposed adaptive PSC.

We then present a quantitative analysis to show that the proposed framework can effectively adapt to human preferences and correct human errors based on historical data, while ensuring safety of the system under uncertainties. We consider two cases: (i) the human exhibits varying preferences over time, to which the framework adapts accordingly, (ii) the human provides erroneous judgments about road conditions, which the framework automatically corrects. We collect diverse human instructions for different degrees of driving aggressiveness, safety preferences, and judgments about road conditions, with different levels of clarity. We then classify them into different categories and run batch experiments to examine intelligent adaptation and safety assurance. Ta-

3. See Appendix E.2 for detailed definitions and results on both feasible and infeasible cases.

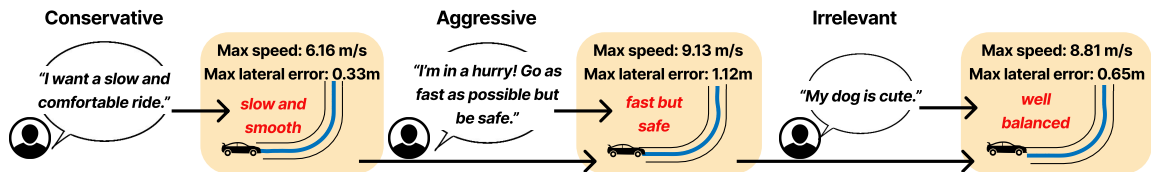


Figure 6: Qualitative results across three rounds of human instructions and feedback.

Icy Road $\mu \in [0.3, 0.4]$	Run 1: Aggressive Preference			Run 2: Conservative Preference			Differences	
	<i>Lateral</i>	<i>Speed</i>	<i>Safety</i>	<i>Lateral</i>	<i>Speed</i>	<i>Safety</i>	<i>Lateral</i>	<i>Speed</i>
GPT-4o-mini	0.78 ± 0.65	8.17 ± 1.54	99%	0.55 ± 0.38	7.73 ± 1.40	100%	-0.23	-0.44
Gemini 2.5 Flash	0.90 ± 0.54	8.43 ± 1.01	99%	0.39 ± 0.36	6.92 ± 1.52	96%	-0.51	-1.51
DeepSeek-Chat	0.87 ± 0.55	8.52 ± 1.51	100%	0.43 ± 0.41	7.17 ± 1.64	100%	-0.44	-1.35

Table 1: Quantitative results for case (i): preference adaptation based on human instructions.

Icy Road $\mu \in [0.3, 0.4]$	Run 1: Wrong Premise on Dry Road			Run 2: Wrong Premise on Dry Road			Differences	
	<i>Prior</i>	<i>Posterior</i>	<i>Safety</i>	<i>Prior</i>	<i>Posterior</i>	<i>Safety</i>	<i>Lateral</i>	<i>Speed</i>
GPT-4o-mini	0.82 ± 0.16	0.35 ± 0.03	94%	0.58 ± 0.21	0.35 ± 0.03	100%	-0.14	-0.32
Gemini 2.5 Flash	0.90 ± 0.04	0.36 ± 0.03	100%	0.56 ± 0.26	0.36 ± 0.03	100%	-0.11	-0.08
DeepSeek-Chat	0.90 ± 0.04	0.35 ± 0.03	98%	0.84 ± 0.14	0.36 ± 0.03	100%	-0.42	+0.04

Table 2: Quantitative results for case (ii): correcting human error with historical data.

ble 1 shows the case where the human first requests an aggressive ride in Run 1 and then prefers a conservative ride in Run 2. We measure the center-lane lateral deviation (*Lateral*), average velocity (*Speed*), and empirical safety (*Safety*) of the system. It can be seen that with the proposed framework, both the lane deviation and the average velocity are reduced according to human preference, while safety is maintained. Table 2 shows the case where the human incorrectly assumes a dry road when it is actually icy. Leveraging historical data including trajectory information and estimator *Posterior*, the framework refines the *Prior* estimate of the friction coefficient. Therefore, despite persistent erroneous human instruction, Run 2 achieves smoother trajectories compared to Run 1, while maintaining safety and efficiency. We also evaluate robustness across different LLMs, including GPT-4o-mini, Gemini 2.5 Flash, and DeepSeek-Chat, demonstrating consistent performance. Full experiment results and ablations are provided in Appendix E.3.

6. Conclusion

In this paper, we proposed a language-guided adaptive probabilistic safety certificate (PSC) to ensure long-term safety of stochastic systems under uncertain dynamics and evolving human specifications. We proved that the proposed safety condition guarantees non-decreasing safety probability over time and can be seamlessly integrated with optimization-based control. We further integrate multi-turn large language models (LLMs) leveraging past data and interactions to account for varying human preferences over time. Through autonomous lane-keeping experiments, we demonstrated that the proposed method maintains consistent long-term safety across varying road conditions, adapts effectively to human preferences, and achieves improved safety–efficiency trade-offs. Future work includes extensions to reinforcement learning and multi-agent settings, with potential adversarial attacks (Hu et al., 2025).

Acknowledgments

This work is sponsored in part by the National Science Foundation CAREER program (no. 2442948), in part by a grant from Japan Science and Technology Agency, in part by a grant from the Commonwealth of Pennsylvania, Department of Community and Economic Development, and in part by Grant-in-Aid for Scientific Research (KAKENHI) from the Japan Society for Promotion of Science (23K13354, 26K07557). We thank Lordphone Wen for contributions to the ablation experiments in the camera-ready version of this paper.

References

- Alessandro Abate, Saurabh Amin, Maria Prandini, John Lygeros, and Shankar Sastry. Probabilistic reachability and safe sets computation for discrete time stochastic hybrid systems. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 258–263. IEEE, 2006.
- Alessandro Abate, Maria Prandini, John Lygeros, and Shankar Sastry. Probabilistic reachability and safety for controlled discrete time stochastic hybrid systems. *Automatica*, 44(11):2724–2734, 2008.
- Mohamadreza Ahmadi, Xiaobin Xiong, and Aaron D Ames. Risk-averse control via cvar barrier functions: Application to bipedal robot locomotion. *IEEE Control Systems Letters*, 6:878–883, 2021.
- Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2017.
- Aaron D Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European Control Conference (ECC)*, pages 3420–3431. IEEE, 2019.
- Amir Bayat, Alessandro Abate, Necmiye Ozay, and Raphael M Jungers. Llm-enhanced symbolic control for safety-critical applications. *arXiv preprint arXiv:2505.11077*, 2025.
- Lukas Brunke, Yanni Zhang, Ralf Römer, Jack Naimer, Nikola Staykov, Siqi Zhou, and Angela P Schoellig. Semantically safe robot manipulation: From semantic scene understanding to motion safeguards. *IEEE Robotics and Automation Letters*, 2025.
- Fernando Castaneda, Jason J Choi, Bike Zhang, Claire J Tomlin, and Koushil Sreenath. Pointwise feasibility of gaussian process-based safety-critical control under model uncertainty. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 6762–6769. IEEE, 2021.
- Margaret P Chapman, Jonathan Lacotte, Aviv Tamar, Donggun Lee, Kevin M Smith, Victoria Cheng, Jaime F Fisac, Susmit Jha, Marco Pavone, and Claire J Tomlin. A risk-sensitive finite-time reachability approach for safety of stochastic dynamic systems. In *2019 American Control Conference (ACC)*, pages 2958–2963. IEEE, 2019.
- Albert Chern, Xiang Wang, Abhram Iyer, and Yorie Nakahira. Safe control in the presence of stochastic uncertainties. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 6640–6645. IEEE, 2021.

- Max H Cohen and Calin Belta. High order robust adaptive control barrier functions and exponentially stabilizing adaptive control lyapunov functions. In *2022 American Control Conference (ACC)*, pages 2233–2238. IEEE, 2022.
- Ryan K Cosner, Andrew W Singletary, Andrew J Taylor, Tamas G Molnar, Katherine L Bouman, and Aaron D Ames. Measurement-robust control barrier functions: Certainty in safety with uncertainty in state. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6286–6291. IEEE, 2021.
- Xiyu Deng, Quan Khanh Luu, Anh Van Ho, and Yorie Nakahira. Context-aware llm-based safe control against latent risks. *arXiv preprint arXiv:2403.11863*, 2024.
- Krisztian Enisz, Istvan Szalay, Gabor Kohlrusz, and Denes Fodor. Tyre–road friction coefficient estimation based on the discrete-time extended kalman filter. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 229(9):1158–1168, 2015. doi: 10.1177/0954407014556115.
- David D Fan, Jennifer Nguyen, Rohan Thakker, Nikhilesh Alatur, Ali-akbar Agha-mohammadi, and Evangelos A Theodorou. Bayesian learning-based adaptive control for safety critical systems. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 4093–4099. IEEE, 2020.
- Marcello Farina, Luca Giulioni, and Riccardo Scattolini. Stochastic linear model predictive control with chance constraints—a review. *Journal of Process Control*, 44:53–67, 2016.
- Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.
- Harley Flanders. Differentiation under the integral sign. *The American Mathematical Monthly*, 80(6):615–627, 1973.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- Ahmad Hafez, Alireza Naderi Akhormeh, Amr Hegazy, and Amr Alanwar. Safe llm-controlled robots with formal guarantees via reachability analysis. *arXiv preprint arXiv:2503.03911*, 2025.
- Danijar Hafner, Dustin Tran, Alex Irpan, Timothy Lillicrap, and James Davidson. Reliable uncertainty estimates in deep neural networks using noise contrastive priors. *stat*, 1050:24, 2018.
- Tor Aksel N Heirung, Joel A Paulson, Jared O’Leary, and Ali Mesbah. Stochastic model predictive control—how does it work? *Computers & Chemical Engineering*, 114:158–170, 2018.
- Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.
- Hikaru Hoshino and Yorie Nakahira. Physics-informed RL for maximal safety probability estimation. In *2024 American Control Conference (ACC)*, pages 3576–3583, 2024. doi: 10.23919/ACC60939.2024.10644621.

- Hanjiang Hu, Alexander Robey, and Changliu Liu. Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks. *arXiv preprint arXiv:2503.00187*, 2025.
- Yiwen Huang, Wei Liang, and Yan Chen. Estimation and analysis of vehicle lateral stability region. In *2017 American Control Conference (ACC)*, pages 4303–4308, 2017. doi: 10.23919/ACC.2017.7963617.
- Yiwen Huang, Sze Zheng Yong, and Yan Chen. Stability control of autonomous ground vehicles using control-dependent barrier functions. *IEEE Transactions on Intelligent Vehicles*, 6(4):699–710, 2021. doi: 10.1109/TIV.2021.3058064.
- Mrdjan Jankovic. Robust control barrier functions for constrained stabilization of nonlinear systems. *Automatica*, 96:359–367, 2018.
- Nikolaos Kariotoglou, Sean Summers, Tyler Summers, Maryam Kamgarpour, and John Lygeros. Approximate dynamic programming for stochastic reachability. In *2013 European Control Conference (ECC)*, pages 584–589. IEEE, 2013.
- Azal Ahmad Khan, Michael Andrev, Muhammad Ali Murtaza, Sergio Aguilera, Rui Zhang, Jie Ding, Seth Hutchinson, and Ali Anwar. Safety aware task planning via large language models in robotics. *arXiv preprint arXiv:2503.15707*, 2025.
- Zihan Li, Ping Wang, Shuo Cai, Xiao Hu, and Hong Chen. NMPC-based controller for vehicle longitudinal and lateral stability enhancement under extreme driving conditions. *ISA Transactions*, 135:509–523, 2023. ISSN 0019-0578. doi: <https://doi.org/10.1016/j.isatra.2022.09.038>.
- Kehan Long, Vikas Dhiman, Melvin Leok, Jorge Cortés, and Nikolay Atanasov. Safe control synthesis with uncertain dynamics and constraints. *IEEE Robotics and Automation Letters*, 7(3):7295–7302, 2022.
- Brett T Lopez, Jean-Jacques E Slotine, and Jonathan P How. Robust adaptive control barrier functions: An adaptive and data-driven approach to safety. *IEEE Control Systems Letters*, 5(3):1031–1036, 2020.
- Wenhao Luo, Wen Sun, and Ashish Kapoor. Multi-robot collision avoidance under uncertainty with probabilistic safety barrier certificates. *Advances in Neural Information Processing Systems*, 33:372–383, 2020.
- Yiwei Lyu, Wenhao Luo, and John M. Dolan. Probabilistic safety-assured adaptive merging control for autonomous vehicles. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10764–10770, 2021. doi: 10.1109/ICRA48506.2021.9561894.
- Gabriel Maher. Llmpc: Large language model predictive control. *arXiv preprint arXiv:2501.02486*, 2025.
- Ali Mesbah. Stochastic model predictive control: An overview and perspectives for future research. *IEEE Control Systems Magazine*, 36(6):30–44, 2016.
- Yuya Miyaoka, Masaki Inoue, and Tomotaka Nii. Chatmpc: Natural language based mpc personalization. In *2024 American Control Conference (ACC)*, pages 3598–3603. IEEE, 2024.

- Kevin P. Murphy. Conjugate bayesian analysis of the gaussian distribution. University of British Columbia, 2007.
- Quan Nguyen and Koushil Sreenath. Exponential control barrier functions for enforcing high relative-degree safety-critical constraints. In *2016 American Control Conference (ACC)*, pages 322–328. IEEE, 2016.
- Yunpeng Pan, Xinyan Yan, Evangelos A Theodorou, and Byron Boots. Prediction under uncertainty in sparse spectrum gaussian processes with applications to filtering and control. In *International Conference on Machine Learning*, pages 2760–2768. PMLR, 2017.
- Stephen Prajna, Ali Jadbabaie, and George J. Pappas. A framework for worst-case and stochastic safety verification using barrier certificates. *IEEE Transactions on Automatic Control*, 52(8): 1415–1428, 2007. doi: 10.1109/TAC.2007.902736.
- Rajesh Rajamani. *Vehicle Dynamics and Control*. Springer, 2nd edition, 2012.
- Laura R. Ray. Nonlinear tire force estimation and road friction identification: Simulation and experiments. *Automatica*, 33(10):1819–1833, 1997. ISSN 0005-1098. doi: [https://doi.org/10.1016/S0005-1098\(97\)00093-9](https://doi.org/10.1016/S0005-1098(97)00093-9).
- Eric J. Rossetter and J. Christian Gerdes. Lyapunov Based Performance Guarantees for the Potential Field Lane-keeping Assistance System. *Journal of Dynamic Systems, Measurement, and Control*, 128(3):510–522, 08 2006. doi: 10.1115/1.2192835.
- Sourav Sanyal and Kaushik Roy. Asma: An adaptive safety margin algorithm for vision-language drone navigation via scene-aware control barrier functions. *IEEE Robotics and Automation Letters*, 2025.
- Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2015.
- Andrew J Taylor and Aaron D Ames. Adaptive safety with control barrier functions. In *2020 American Control Conference (ACC)*, pages 1399–1405. IEEE, 2020.
- Mariia Vasileva, Fedor Shmarov, and Paolo Zuliani. Probabilistic reachability for uncertain stochastic hybrid systems via gaussian processes. In *2020 18th ACM-IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE)*, pages 1–11. IEEE, 2020.
- Yanze Wang, Mingzhe Liu, and Yan Chen. Safety-guaranteed control of coupled longitudinal and lateral vehicle dynamics by considering tire stability via cdbfs. *IFAC-PapersOnLine*, 55(37): 500–505, 2022a. doi: <https://doi.org/10.1016/j.ifacol.2022.11.232>. 2nd Modeling, Estimation and Control Conference MECC 2022.
- Zhuoyuan Wang and Yorie Nakahira. A generalizable physics-informed learning framework for risk probability estimation. In *Learning for Dynamics and Control Conference*, pages 358–370. PMLR, 2023.

- Zhuoyuan Wang, Haoming Jing, Christian Kurniawan, Albert Chern, and Yorie Nakahira. Myopically verifiable probabilistic certificate for long-term safety. In *2022 American Control Conference (ACC)*, pages 4894–4900. IEEE, 2022b.
- Zhuoyuan Wang, Albert Chern, and Yorie Nakahira. Generalizable physics-informed learning for stochastic safety-critical systems. *IEEE Transactions on Automatic Control*, 2025.
- Zhuoyuan Wang, Haoming Jing, Christian Kurniawan, Albert Chern, and Yorie Nakahira. Myopically verifiable probabilistic certificates for safe control and learning. *IEEE Transactions on Automatic Control*, 2026.
- Yi Wu, Zikang Xiong, Yiran Hu, Shreyash S Iyengar, Nan Jiang, Aniket Bera, Lin Tan, and Suresh Jagannathan. Selp: Generating safe and efficient task plans for robot agents with large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2599–2605. IEEE, 2025.
- Wei Xiao and Calin Belta. High-order control barrier functions. *IEEE Transactions on Automatic Control*, 67(7):3655–3662, 2021.
- Wei Xiao, Calin Belta, and Christos G Cassandras. Adaptive control barrier functions. *IEEE Transactions on Automatic Control*, 67(5):2267–2281, 2021.
- Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*, 2024.
- Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint arXiv:2501.09959*, 2025.

Appendix A. Proofs

Proof (Theorem 2) We use mathematical induction to prove (10). Condition (10) holds for $k = 0$ due to the assumption on initial condition. Next, we suppose (10) holds at time $k \geq 0$, and show (10) holds for time $k + 1$. Let

$$\mathbb{E}_{\pi_{\text{safe}}} [\Psi_{H_k}^\pi(X_k)] = 1 - \epsilon + q \quad (16)$$

for some $q \geq 0$. Note that the expectation in (16) is taken over all possible trajectories starting from X_0 . For the rest of the proof, we omit π_{safe} in the expectation for notation conciseness. We first consider the set of events \mathcal{D}_0 that satisfies $\Psi_{H_k}^\pi(X_k) < 1 - \epsilon$ and \mathcal{D}_1 that satisfies $\Psi_{H_k}^\pi(X_k) \geq 1 - \epsilon$. Then, define variables v_i and δ_i , for $i \in \{0, 1\}$, as follows:

$$v_0 = \mathbb{E} [\Psi_{H_k}^\pi(X_k) \mid \mathcal{D}_0, \pi_{\text{safe}}] = 1 - \epsilon - \delta_0, \quad (17)$$

$$v_1 = \mathbb{E} [\Psi_{H_k}^\pi(X_k) \mid \mathcal{D}_1, \pi_{\text{safe}}] = 1 - \epsilon + \delta_1, \quad (18)$$

where $\delta_0 \geq 0$ and $\delta_1 \geq 0$, since $\mathbb{E} [\Psi_{H_k}^\pi(X_k) \mid \mathcal{D}_0] < 1 - \epsilon$ and $\mathbb{E} [\Psi_{H_k}^\pi(X_k) \mid \mathcal{D}_1] \geq 1 - \epsilon$ hold. The left hand side of (16) can then be written as

$$\begin{aligned} \mathbb{E}[\Psi_{H_k}^\pi(X_k)] &= \mathbb{E} [\Psi_{H_k}^\pi(X_k) \mid \mathcal{D}_0] \mathbb{P}(\mathcal{D}_0) + \mathbb{E} [\Psi_{H_k}^\pi(X_k) \mid \mathcal{D}_1] \mathbb{P}(\mathcal{D}_1) \\ &= v_0 q_0 + v_1 q_1, \end{aligned} \quad (19)$$

where $q_i := \mathbb{P}(\mathcal{D}_i)$, for $i \in \{0, 1\}$, and we have

$$\mathbb{P}(\mathcal{D}_0) + \mathbb{P}(\mathcal{D}_1) = q_0 + q_1 = 1. \quad (20)$$

Combining (16) and (19), and applying (17) and (18) gives

$$1 - \epsilon + q = v_0 q_0 + v_1 q_1 \quad (21)$$

$$= (1 - \epsilon - \delta_0) q_0 + (1 - \epsilon + \delta_1) q_1. \quad (22)$$

This combined with (20) yields

$$q = \delta_1 q_1 - \delta_0 q_0. \quad (23)$$

On the other hand, we have

$$\mathbb{E} [\gamma (\Psi_{H_k}^\pi(X_k) - (1 - \epsilon))] \quad (24)$$

$$= \mathbb{P}(\mathcal{D}_0) \mathbb{E} [\gamma (\Psi_{H_k}^\pi(X_k) - (1 - \epsilon)) \mid \mathcal{D}_0] + \mathbb{P}(\mathcal{D}_1) \mathbb{E} [\gamma (\Psi_{H_k}^\pi(X_k) - (1 - \epsilon)) \mid \mathcal{D}_1] \quad (25)$$

$$= q_0 (\mathbb{E} [\gamma (\Psi_{H_k}^\pi(X_k) - (1 - \epsilon)) \mid \mathcal{D}_0]) + q_1 (\mathbb{E} [\gamma (\Psi_{H_k}^\pi(X_k) - (1 - \epsilon)) \mid \mathcal{D}_1]) \quad (26)$$

$$\leq q_0 (\gamma (\mathbb{E} [\Psi_{H_k}^\pi(X_k) - (1 - \epsilon) \mid \mathcal{D}_0])) + q_1 (\gamma (\mathbb{E} [\Psi_{H_k}^\pi(X_k) - (1 - \epsilon) \mid \mathcal{D}_1])) \quad (27)$$

$$= q_0 (\gamma (-\delta_0)) + q_1 (\gamma (\delta_1)) \quad (28)$$

$$\leq \gamma (-q_0 \delta_0 + q_1 \delta_1) \quad (29)$$

$$= \gamma(q) \quad (30)$$

$$\leq q. \quad (31)$$

Here, (27) is obtained from Jensen's inequality for concave function γ ; (28) is based on (17) and (18); (29) is due to Jensen's inequality, design requirement 1 and (20); and (30) is due to (23). From (24) to (31), we have

$$\mathbb{E} [-\gamma (\Psi_{H_k}^\pi(X_k) - (1 - \epsilon))] \geq -q. \quad (32)$$

Recall that the control action is chosen to satisfy (8). Now, we take the expectation over both side of (8) to obtain

$$\mathbb{E}[\mathcal{A}_{H_{k+1}}^{U_k} \Psi_{H_k}^\pi(X_k)] \geq \mathbb{E}[-\gamma(\Psi_{H_k}^\pi(X_k) - (1 - \epsilon))]. \quad (33)$$

From the definition (7) of the discrete-time generator, we have

$$\mathbb{E} \left[\frac{\mathbb{E}[\Psi_{H_{k+1}}^\pi(X_{k+1}) \mid X_k, U_k] - \Psi_{H_k}^\pi(X_k)}{\Delta t} \right] \geq \mathbb{E}[-\gamma(\Psi_{H_k}^\pi(X_k) - (1 - \epsilon))]. \quad (34)$$

Using the law of total expectation, we have

$$\frac{\mathbb{E}[\Psi_{H_{k+1}}^\pi(X_{k+1}) - \Psi_{H_k}^\pi(X_k)]}{\Delta t} \geq \mathbb{E}[-\gamma(\Psi_{H_k}^\pi(X_k) - (1 - \epsilon))]. \quad (35)$$

Combining (16), (32), (35) and design requirement 2 yields

$$\mathbb{E}[\Psi_{H_{k+1}}^\pi(X_{k+1})] \quad (36)$$

$$\geq \mathbb{E}[\Psi_{H_k}^\pi(X_k)] + \mathbb{E}[-\gamma(\Psi_{H_k}^\pi(X_k) - (1 - \epsilon))]\Delta t \quad (37)$$

$$\geq 1 - \epsilon + q - q\Delta t \quad (38)$$

$$= 1 - \epsilon + q(1 - \Delta t). \quad (39)$$

Since $\Delta t \ll 1$ and $q \geq 0$, we have

$$\mathbb{E}[\Psi_{H_{k+1}}^\pi(X_{k+1})] \geq 1 - \epsilon. \quad (40)$$

Here, the left hand side of (40) is equivalent to the left hand side of (10). \blacksquare

Appendix B. Implementation of Safety Constraints

In this section, we discuss a numerical method for calculating the discrete time generator $\mathcal{A}_{H_{k+1}}^{U_k} \Psi_{H_k}^\pi(X_k)$ in (7). Given H_{k+1} and H_k , the value of the discrete time generator can be calculated as

$$\begin{aligned} \mathcal{A}_{H_{k+1}}^{U_k} \Psi_{H_k}^\pi(X_k) &= \frac{\mathbb{E}[\Psi_{H_{k+1}}^\pi(X_{k+1})|X_k, U_k] - \Psi_{H_k}^\pi(X_k)}{\Delta t} \\ &= \underbrace{\frac{\mathbb{E}[\Psi_{H_{k+1}}^\pi(X_{k+1})|X_k, U_k] - \Psi_{H_{k+1}}^\pi(X_k)}{\Delta t}}_{:= \mathcal{S}^{U_k} \Psi_{H_{k+1}}^\pi(X_k)} + \underbrace{\frac{\Psi_{H_{k+1}}^\pi(X_k) - \Psi_{H_k}^\pi(X_k)}{\Delta t}}_{:= \mathcal{T}_{H_{k+1}} \Psi_{H_k}^\pi(X_k)}, \end{aligned} \quad (41)$$

where on the right hand side the first term $\mathcal{S}^{U_k} \Psi_{H_{k+1}}^\pi(X_k)$ involves prediction of the state X_{k+1} at the next time step, and the second term $\mathcal{T}_{H_{k+1}} \Psi_{H_k}^\pi(X_k)$ involves the update of the density function from H_k to H_{k+1} . Here, we require H_{k+1} at the time step k to calculate these terms. The function H_k is supposed to be constructed from the information obtained up to the time step $k - 1$. Then we know H_{k+1} is constructed at the time step k , and we can directly evaluate the second term $\mathcal{T}_{H_{k+1}} \Psi_{H_k}^\pi(X_k)$ when we have estimates for $\Psi_{H_{k+1}}^\pi(X_k)$ and $\Psi_{H_k}^\pi(X_k)$. Furthermore, to approximate the first term $\mathcal{S}^{U_k} \Psi_{H_{k+1}}^\pi(X_k)$, we suppose the dynamics (1) are given as a discretization of continuous-time dynamics

$$dX_t = f_c(X_t, U_t, \xi)dt + \sigma_c(X_t, U_t)dW_t, \quad \forall t \in \mathbb{R}^+, \quad (42)$$

where $X_0 = x_0$ and we denote the state by X_t and the input by U_t with slight abuse of notation. We then have the following theorem.

Assumption 3 *The continuous dynamics function f_c and σ_c in (42) are twice differentiable with respect to X_t . The control space \mathcal{U} is compact.*

Assumption 4 *The safety probability $\Psi_H^\pi(X)$ is differentiable with regard to X for any H and π , and has bounded derivatives.*

Theorem 5 Suppose that Assumption 3 and 4 hold. Then, the following equality holds:

$$\begin{aligned} & \lim_{\Delta t \rightarrow 0} \mathcal{S}^{U_k} \Psi_{H_{k+1}}^\pi(X_k) \\ &= \nabla_x \Psi_{H_{k+1}}^\pi(X_k) \mathbb{E}[f_c(X_k, U_k, \Xi_{k+1}) \mid X_k, U_k] + \frac{1}{2} \sigma_c(X_k, U_k)^\top \sigma_c(X_k, U_k) \nabla_x^2 \Psi_{H_{k+1}}^\pi(X_k), \end{aligned} \quad (43)$$

where ∇_x and ∇_x^2 are the gradient and Hessian operators.

Proof (Theorem 5) From the definition of $\mathcal{S}^{U_k} \Psi_{H_{k+1}}^\pi(X_k)$ in (41) we get

$$\lim_{\Delta t \rightarrow 0} \mathcal{S}^{U_k} \Psi_{H_{k+1}}^\pi(X_k) \quad (44)$$

$$= \lim_{\Delta t \rightarrow 0} \frac{\mathbb{E}[\Psi_{H_{k+1}}^\pi(X_{k+1}) \mid X_k, U_k] - \Psi_{H_{k+1}}^\pi(X_k)}{\Delta t} \quad (45)$$

$$= \lim_{\Delta t \rightarrow 0} \mathbb{E} \left[\frac{\Psi_{H_{k+1}}^\pi(X_{k+1}) - \Psi_{H_{k+1}}^\pi(X_k)}{\Delta t} \middle| X_k, U_k \right] \quad (46)$$

$$= \lim_{\Delta t \rightarrow 0} \mathbb{E} \left[\mathbb{E} \left[\frac{\Psi_{H_{k+1}}^\pi(X_{k+1}) - \Psi_{H_{k+1}}^\pi(X_k)}{\Delta t} \middle| X_k, U_k, \Xi_{k+1} \right] \middle| X_k, U_k \right] \quad (47)$$

$$= \mathbb{E} \left[\lim_{\Delta t \rightarrow 0} \mathbb{E} \left[\frac{\Psi_{H_{k+1}}^\pi(X_{k+1}) - \Psi_{H_{k+1}}^\pi(X_k)}{\Delta t} \middle| X_k, U_k, \Xi_{k+1} \right] \middle| X_k, U_k \right], \quad (48)$$

where (47) holds due to the law of total expectation, and (48) holds due to Assumption 4 and the dominated convergence theorem (Flanders, 1973). Then we have

$$\lim_{\Delta t \rightarrow 0} \mathcal{S}^{U_k} \Psi_{H_{k+1}}^\pi(X_k) \quad (49)$$

$$= \mathbb{E} \left[\lim_{\Delta t \rightarrow 0} \mathbb{E} \left[\frac{\Psi_{H_{k+1}}^\pi(X_{k+1}) - \Psi_{H_{k+1}}^\pi(X_k)}{\Delta t} \middle| X_k, U_k, \Xi_{k+1} \right] \middle| X_k, U_k \right] \quad (50)$$

$$= \mathbb{E} \left[\nabla_x \Psi_{H_{k+1}}^\pi(X_k) f_c(X_k, U_k, \Xi_{k+1}) + \frac{1}{2} \sigma_c(X_k, U_k)^\top \sigma_c(X_k, U_k) \nabla_x^2 \Psi_{H_{k+1}}^\pi(X_k) \middle| X_k, U_k \right] \quad (51)$$

$$= \nabla_x \Psi_{H_{k+1}}^\pi(X_k) \mathbb{E}[f_c(X_k, U_k, \Xi_{k+1}) \mid X_k, U_k] + \frac{1}{2} \sigma_c(X_k, U_k)^\top \sigma_c(X_k, U_k) \nabla_x^2 \Psi_{H_{k+1}}^\pi(X_k), \quad (52)$$

where (51) holds by applying Ito's lemma under Assumption 3 and 4. \blacksquare

Remark 6 Evaluation of the discrete time generator $\mathcal{A}_{H_{k+1}}^{U_k} \Psi_{H_k}^\pi(X_k)$ in (7) requires values of the long-term safety probability $\Psi_{H_k}^\pi(X_k)$ in (4). In practice, one could run online parallel Monte Carlo simulations to empirically estimate such values, or leverage offline model-based and data-driven methods such as the ones in Chern et al. (2021); Wang et al. (2025).

Appendix C. Adaptive PSC Algorithm

In this section, we present the overall adaptive probabilistic safety certificate algorithm in Algorithm 1. In line 1, we initialize the horizon of the problem T_{end} , the discrete time step Δt , the initial state X_0 , the nominal control policy π , the long-term safety horizon T , the risk tolerance ϵ , and the reference controller N . While time step k has not reached T_{end} , we run the proposed adaptive safe control method. Specifically, we first obtain the available information H_k from the estimator (line 3), calculate the safety probability $\Psi_{H_k}^\pi(X_k)$ (line 4), solve optimization (11) to get safe control U_k^{safe} (line 5), and step the dynamics to get state at the next time step X_{k+1} (line 6). Note that the safety constraint for the optimization problem (11) involves the calculation of the infinitesimal generator $\mathcal{A}_{H_{k+1}}^{U_k} \Psi_{H_k}^\pi(X_k)$. In practice one can assume finite admissible control set for efficient computation of the problem.

Algorithm 1 Adaptive Probabilistic Safety Certificate

Input: $T_{\text{end}}, \Delta t, X_0, \pi, T, \epsilon, N$

```

1  $k \leftarrow 0$ 
2 while  $k < T_{\text{end}}$  do
3   Obtain  $H_k$  from the estimator
4   Calculate safety probability  $\Psi_{H_k}^\pi(X_k)$ 
5   Solve (11) to get safe control  $U_k^{\text{safe}} \leftarrow \pi_{\text{safe}}(X_k, \hat{\xi}_k)$ 
6   Step the dynamics (1) with  $U_k^{\text{safe}}$  to get  $X_{k+1}$ 
7    $k \leftarrow k + 1$ 
8 end
    
```

Appendix D. Vehicle model

In this section, we introduce the vehicle model used in numerical experiments. For the vehicle dynamics, a 3 degree-of-freedom (DoF) model is adopted, and the dynamics of the state $x_{\text{vehicle}} := [v_x, v_y, r, \delta]$, where v_x and v_y are the vehicle longitudinal and lateral velocities, respectively, r is the yaw rate, and δ is the steering angle, is given by the following equations:

$$m\dot{v}_x = mv_y r + (F_{Lfl} + F_{Lfr}) \cos \delta - (F_{Sfl} + F_{Sfr}) \sin \delta + F_{Lrl} + F_{Lrr} \quad (53)$$

$$m\dot{v}_y = -mv_x r + (F_{Sfl} + F_{Sfr}) \cos \delta + (F_{Lfl} + F_{Lfr}) \sin \delta + F_{Srl} + F_{Srr} \quad (54)$$

$$I_z \dot{r} = l_f \{(F_{Sfl} + F_{Sfr}) \cos \delta + (F_{Lfl} + F_{Lfr}) \sin \delta\} - l_r (F_{Srl} + F_{Srr}) + \frac{W}{2} (-F_{Lrl} + F_{Lrr}) + \frac{W}{2} \{(-F_{Lfl} + F_{Lfr}) \cos \delta + (F_{Sfl} - F_{Sfr}) \sin \delta\} \quad (55)$$

$$\dot{\delta} = \Delta \delta \quad (56)$$

where the parameters m is the mass of the vehicle, I_z is the rotational inertia of the vehicle, l_f is the front wheel distance to vehicle center, l_r is the rear wheel distance to vehicle center, and W

is the width of the vehicle. The symbol F_{ij} for $i \in \{L, S\}$ and $j \in \{\text{fl}, \text{fr}, \text{rl}, \text{rr}\}$ stands for the longitudinal and side (lateral) tire forces for each of the four tires (fl:front left, fr:front right, rl:rear left, rr:rear right). Each tire force can be characterized as a function of the tire slip angle α_{ij} and slip ratio λ_{ij} as follows

$$\alpha_{\text{fl}} = \alpha_{\text{fr}} = \delta - \frac{v_y + l_{\text{fr}}r}{v_x}, \quad (57)$$

$$\alpha_{\text{rl}} = \alpha_{\text{rr}} = -\frac{v_y + l_{\text{r}}r}{v_x}, \quad (58)$$

$$\lambda_{ij} = \frac{R_e \omega_{ij} - v_x}{\max(R_e \omega_{ij}, v_x)} \quad (59)$$

where R_e is the wheel radius. As the tire slip angle and the slip ratio increase, the tire forces saturate rapidly and transit from linear to nonlinear dynamics. To represent the nonlinear tire forces, following Li et al. (2023), we use the LuGre combined-slip tire model, which considers the internal tire characteristics and has been widely utilized to describe the nonlinearity under extreme conditions. The tire longitudinal and lateral force descriptions are as follows:

$$F_{Lij} = \left(\frac{\sigma_{0x}}{\frac{\sigma_{0x}|0v_{r,ij}|0}{\mu g(v_{r,ij})} + \kappa_x R_e |\omega_{ij}|} + \sigma_{2x} \right) v_{rx,ij} F_z, \quad (60)$$

$$F_{Sij} = \left(\frac{\sigma_{0y}}{\frac{\sigma_{0y}|0v_{r,ij}|0}{\mu g(v_{r,ij})} + \kappa_y R_e |\omega_{ij}|} + \sigma_{2y} \right) v_{ry,ij} F_z, \quad (61)$$

where $v_{rx,ij}$, $v_{ry,ij}$ and $|0v_{r,ij}|0$ are given by

$$v_{rx,ij} = R_e \omega_{ij} - v_x, \quad (62)$$

$$v_{ry,ij} = v_x \alpha_{ij}, \quad (63)$$

$$|0v_{r,ij}|0 = \sqrt{v_{rx,ij}^2 + v_{ry,ij}^2}. \quad (64)$$

In the above equations, μ stands for the road friction coefficient, and $g(v_{r,ij})$ is a Stribeck function given by

$$g(v_{r,ij}) = \mu_c + (\mu_s - \mu_c) \exp\left(-\sqrt{|v_{r,ij}|/V_s}\right), \quad (65)$$

where V_s is the Stribeck relative velocity. For the dynamics of the state x_{wheel} , we assume an open drivetrain with equal torque distribution and linear dynamics, as shown below:

$$I_\omega \dot{\omega}_{\text{fl}} = -R_e F_{L\text{fl}} + \frac{1}{4} \tau_e, \quad (66)$$

$$I_\omega \dot{\omega}_{\text{fr}} = -R_e F_{L\text{fr}} + \frac{1}{4} \tau_e, \quad (67)$$

$$I_\omega \dot{\omega}_{\text{rl}} = -R_e F_{L\text{rl}} + \frac{1}{4} \tau_e, \quad (68)$$

$$I_\omega \dot{\omega}_{\text{rr}} = -R_e F_{L\text{rr}} + \frac{1}{4} \tau_e, \quad (69)$$

$$\dot{\tau}_e = \Delta \tau_e \quad (70)$$

Table 3: Parameter Setting

Symbol	Definition	Value
m	Vehicle mass	1430 kg
R_e	Wheel radius	0.325 m
I_z	Yaw moment of inertia	2059 kg m ²
I_ω	Wheel moment of inertia	1.68 kg m ²
L_f	Distance from CG to front axis	1.05 m
L_r	Distance from CG to rear axis	1.61 m
W	width of the vehicle	1.55 m
V_s	Stribeck relative velocity	6.6 m/s
σ_{0x}	Longitudinal rubber stiffness	195 m ⁻¹
σ_{2x}	Longitudinal relative viscous	0.001 s/m
κ_x	Longitudinal load distribution factor	13.4
σ_{0y}	Lateral rubber stiffness	195 m ⁻¹
σ_{2y}	Lateral relative viscous damping	0.001 s/m
κ_y	Lateral load distribution factor	13.4
μ_s	Static friction coefficient	0.55
μ_c	Dynamics friction coefficient	0.35
g	gravitational acceleration	9.8 m/s ²

where J_ω is the wheel inertia. To formulate a lane keeping problem, the vehicle dynamics model needs to be combined with the model of the vehicle position and heading angle. When they are described in a road coordinate, the road curvature can be viewed as a disturbance pushing the vehicle out of the lane. By defining the yaw rate error ψ between the vehicle and the road, its dynamics can be described as

$$\frac{d\psi}{dt} = r - v_x \rho(s) \quad (71)$$

where ρ is the curvature of the road and can be regarded as a function of the distance s along the road. The dynamics of the distance s and the lateral error e can be obtained by solving the following equations:

$$\frac{ds}{dt} = v_x \cos \psi - v_y \sin \psi, \quad (72)$$

$$\frac{de}{dt} = v_y \cos \psi + v_x \sin \psi. \quad (73)$$

The list of parameters and their values are summarized in Table 3.

Appendix E. Experiment Details

In this section, we provide experiment details corresponding to Section 5.

E.1. General Setup

For the state space in (13), when the vehicle is traveling on a road with a non-zero curvature, the curvature is viewed as a disturbance on the heading error ψ described by

$$\dot{\psi} = r - v_x \rho(s), \quad (74)$$

where $\rho(s)$ denotes the radius of curvature as a function of s .

The road-tire friction coefficient μ is an unknown fixed parameter. It typically takes a value from 0.7 to 0.9 for a dry road condition, from 0.4 to 0.7 for a wet road condition, and from 0.2 to 0.4 for an icy road condition (Rajamani, 2012). While various methods including extended Kalman filtering (Ray, 1997; Enisz et al., 2015) are proposed for online estimation of the friction coefficient, in all experiments we chose to use a Bayesian estimator. Note that for the estimator (15) Theorem 2 holds without assuming Gaussian distributions for the parameter estimates.

For all three MPC-based adaptive safe control methods considered, we define the MPC cost function as

$$J_{\text{MPC}} = \sum_{k=1}^{T_{\text{mpc}}} 0.05(v_x(k) - V_{\text{ref}})^2 + e(k)^2 + \psi(k)^2, \quad (75)$$

where T_{mpc} is the MPC prediction horizon. We implement the nonlinear MPC controller using `nlpmpc` provided by the Model Predictive Control Toolbox in MATLAB with automatic C code generation. We set the sampling time to 0.2s, the control horizon to 2 time steps. The initial velocity of the vehicle is chosen as $v_0 = 20\text{km/h}$. For the **AMPC** method, we use the expectation μ_k of the posterior $P_k(\mu|H_k)$ in the prediction model for adaptation. For control-dependent barrier functions (**CDBF**) (Huang et al., 2021; Wang et al., 2022a), it constrains the vehicle states within a time-varying and control-dependent lateral stability region, and such constraints are incorporated inside the nonlinear adaptive MPC. For the proposed method, we integrate the adaptive probabilistic safety certificate (PSC) (11) with the MPC framework by adding the safety constraint (8).

E.2. Adaptive Safe Control

For the experiments in Section 5.2, the online computation time at each control step is reported in Fig. 8. All computation time is measured on a workstation with AMD EPYC 7763@2.45GHz. As shown, the computation time of all methods increases with the MPC horizon. Combined with the safety results in Fig. 2, these findings indicate that only the proposed method achieves the desired long-term safety within real-time computation using short MPC horizons.

Among the 108 distinctive cases for quantitative analysis in Section 5.2, settings that yield a safety probability below 0.3 for all methods are marked as infeasible, as the initial conditions and safety requirements are too stringent for any method to satisfy (the sufficiently safe initial condition in Theorem 2 is violated). Among the 108 configurations, 37 are feasible. Fig. 7 visualizes the trade-offs for both the feasible and overall cases. We can see that the proposed adaptive PSC consistently achieves higher safety probability than that of AMPC and CDBF, while maintaining comparable average vehicle velocities. These results show that the proposed method achieves better trade-offs between safety and efficiency by ensuring long-term safety without compromising performance.

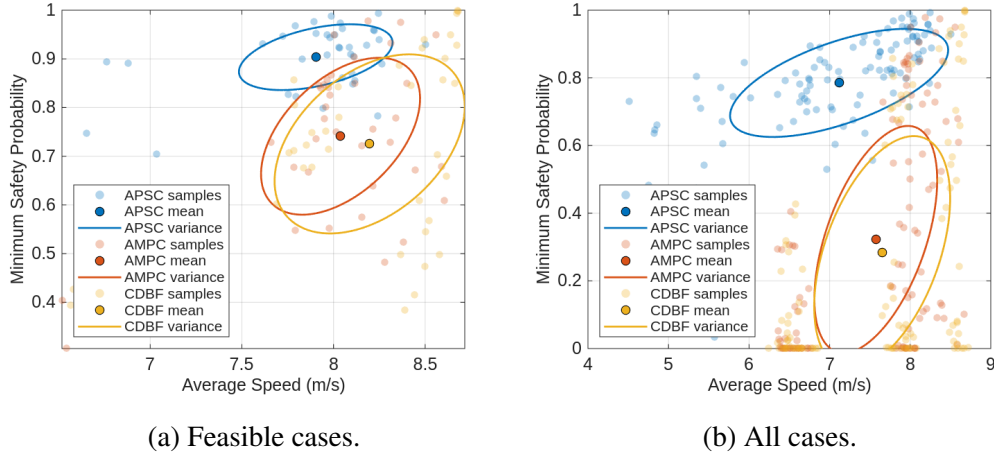


Figure 7: Safety vs. efficiency trade-offs with MPC-based adaptive safe control variants. **APSC**: Proposed adaptive probabilistic safety certificate. **AMPC**: Adaptive model predictive control. **CDBF**: Control-dependent barrier functions.

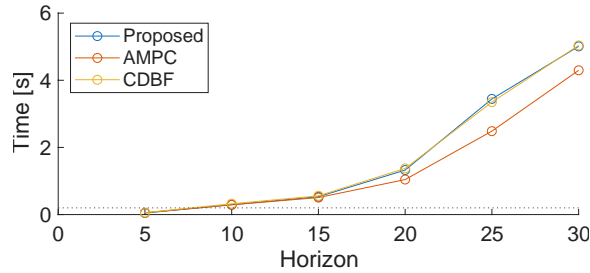


Figure 8: Per step online computation time vs. MPC horizon T_{mpc} .

E.3. LLM Integration

We provide full quantitative results for LLM integration with adaptive PSC. For case (i), where the human exhibits varying driving aggressiveness preferences over time, we report the results in Table 6. It can be seen that across diverse LLM models with different sizes and capabilities, the proposed framework can adapt to human preferences accordingly, while consistently ensuring safety of the system under road condition uncertainties. For case (ii), where the human provides erroneous judgments about road conditions, we provide the full quantitative results in Table 7. It can be observed that across diverse LLM models, the proposed framework can leverage historical data to refine the prior estimate of the friction coefficient, and therefore achieves smoother yet safe and efficient subsequent runs despite persistent erroneous human instruction.

The example human instructions for different categories used in the experiments are shown in Table 5. The empirical *Safety* reported in the results are calculated through the ratio of time period where the lateral lane deviation is less than 3m

Aggressive User Input	Conservative User Input	Dry and Unsure User Input
<p>I want to drive aggressively and push the limits. Go for maximum performance and speed. Make it react faster and turn sharper. I prefer bold and assertive driving. Drive with high energy and confidence. Prioritize speed and agility over comfort. I want a powerful, race-style control. Let it take corners fast and hard. Push acceleration and minimize hesitation. I want fast, strong, and responsive steering. Use the most dynamic, performance-oriented mode. Allow wider lane deviation for faster maneuvers. Give me riskier but more exciting control. I want a high-speed, competitive setup. Drive aggressively even when conditions vary. Maximize performance, not stability. I want it to move quickly, even if precision drops. Use a high-tolerance mode for rapid turns. I prefer fast, high-response control with less caution. Push harder on throttle and steering—go all in. Enable sport mode for maximum power. Go fast and be decisive in lane changes. Increase steering gain for instant response. I want quick corrections and hard acceleration. Push for maximum grip and speed. Drive with racing-level confidence. I want it to feel like track mode. Accelerate rapidly and maintain momentum. Give me immediate throttle response. I want to attack the road, not play safe. Go for aggressive lane keeping and acceleration. Turn faster, brake later. Max performance—no slowdowns. Push through corners with confidence. I want hard acceleration out of every turn. Keep maximum pace throughout the drive. Make the system react instantly to inputs. I want extreme handling and control. Drive like a pro racer. Go strong on speed and agility. I want a forceful, assertive steering behavior. Keep the drive sharp and energetic. Prioritize lap time over comfort. Take the fastest possible path. Push vehicle response to the limit. I want fearless control even at high speed. Stay aggressive in all situations. Increase tolerance for quick maneuvers. I want fast corner entry and exit. Push dynamic performance higher. Enable the most responsive mode available. I want strong acceleration and tight control. Reduce safety margin for faster results. Push to maximum allowable lane deviation. I want fast, snappy handling. Stay on throttle—no delays. I prefer performance-first control. Give me top-speed behavior. Drive like a sports car, not a family car. Maintain aggressive control throughout. I want to feel constant power and momentum. Drive hard and take initiative. Go faster on every straight segment. Push acceleration early in the turn. React instantly to direction changes. I want dynamic and energetic motion. Keep high momentum and fast transitions. Make control firm and quick. I want aggressive grip management. Prioritize time over precision. Go faster with minimal correction delay. I prefer a daring, confident mode. Give me maximum responsiveness. Increase the pace—make it intense. I want minimal damping and high power. Accelerate sooner and stronger. Keep speed constant even on curves. Make steering bold and direct. I want powerful, quick decision behavior. Push throttle sensitivity higher. Stay fast through all maneuvers. I want performance at the edge of control. Drive aggressively, no holding back.</p>	<p>That was too fast, please slow down. I want a more conservative setup. Reduce speed, it felt unsafe. Too aggressive—make it smoother. Please prioritize safety over speed. I prefer a slower and steadier control. Reduce acceleration, it's too strong. Too risky on this road. I want gentler lane corrections. Too harsh—make it more stable. It felt unsafe, drive more cautiously. Please slow the response a bit. It's too aggressive for my taste. Too fast, not comfortable. Please reduce throttle aggressiveness. I want smoother driving this time. Be more careful, it's not a race. Too nervous in steering. Make it more controlled and less twitchy. Reduce the cornering speed. I want safety first, not speed. It's too reactive, calm it down. Go easy on acceleration. Too much power, reduce it. Please drive more conservatively. I want slower and safer turns. That was too harsh in handling. Reduce the speed margin. It's unstable, make it cautious. Too strong corrections, smooth them out. Please tune it for safety and comfort. Too jerky, I want gentle motion. Slow down the transitions. It felt reckless—be more cautious. Reduce lane correction intensity. I want conservative, not sporty. Too quick, not steady enough. Please lower the gain. It's too responsive, I prefer smoothness. Too abrupt acceleration. Drive more safely, especially on icy roads. Make it softer and more stable. It's too quick to react. Slow it down, focus on stability. Too unsafe—prioritize control. Be less confident, more defensive. Reduce the tracking aggressiveness. I prefer a calm driving style. Too fast in straight lines. It should anticipate slower turns. Please limit the lane error tolerance. Too risky, I want safe behavior. Make it more cautious, less performance-driven. It accelerated too early. Reduce throttle and steering gain. I want a more defensive strategy. Please be more stable on normal roads. Too rough handling. Make it more balanced, not so sharp. It's overreacting to changes. I want smoother adaptation. Too much correction effort. Slow the vehicle response overall. I want predictable, safe control. Please calm down the acceleration. Reduce speed and keep traction. I want less risk and smoother response. Too aggressive for current conditions. Please adopt a conservative mode. It was overconfident in corners. Slow it down for better control. Make it less sensitive to errors. I want steady, careful movement. Too fast, not safe on this surface. Reduce control gain significantly. It should react slower to inputs. Please make it safer and smoother. I prefer stable and low-speed driving. Too much oscillation, reduce energy. Slow down in lane keeping. I want it to be gentle and cautious. It was unstable at high speed. Too sudden turns—smooth them out.</p>	<p>The road seems dry, but I'm not entirely sure. It looks like the road is dry, though I can't be 100% certain. The road appears to be dry, but I could be mistaken. Seems like a dry surface, but not completely confident. I think the road is dry, though there's some doubt. The surface looks dry, but I'm not fully sure. The road feels dry, but I can't say for sure. I believe the road is dry, though not completely sure. The pavement seems dry, but I might be wrong. Probably dry road conditions, but not confirmed. Looks like dry asphalt, though uncertain. I think it's dry, but can't be 100% confident. The road seems fine and dry, but I could be wrong. Dry conditions, I think, but not fully confident. It appears dry, but I'm not completely convinced. Road looks dry overall, though not totally sure. I'd say it's dry, but there's a bit of doubt. Seems dry to me, but I can't confirm it. Looks like a dry surface, but hard to be sure. Feels like dry conditions, but I might be mistaken. I suspect the road is dry, but I'm unsure. Probably dry, though I can't guarantee it. I think it's dry, though uncertain. Dry road, I guess, but not certain. The road looks dry, but there's some uncertainty. I'd assume it's dry, but not entirely sure. It might be dry, but I can't confirm. Seems dry enough, though unsure. Looks pretty dry, but can't be sure. I believe it's dry, but not completely confident. It seems dry overall, but I can't be certain. The surface appears dry, but I'm not positive. Probably dry, but I could be wrong. Road seems dry, though I'm not sure. It looks mostly dry, but uncertain. Dry road I think, though with some doubt. Appears dry, but uncertain. I feel like the road is dry, but unsure. It looks fine and dry, but who knows. Seems like dry conditions, though not sure. Might be dry, can't tell for sure. Looks like dry pavement, though uncertain. Feels dry, but I can't be 100% sure. It looks dry, but maybe not everywhere. The road appears dry overall, but not confirmed. I think it's mostly dry, but can't be certain. Dry I think, but not completely sure. Probably dry, but uncertain. Seems dry enough, though with doubt. It's likely dry, but I'm not certain. Looks dry, but I'm hesitant to confirm. I'd say dry, but not confidently. It might be dry, but I'm not sure. Feels like a dry surface, though uncertain. I think it's dry, though some doubt remains. Road looks dry, but who can say for sure. It appears dry, but I don't know for certain. Seems to be dry, though not confirmed. Looks quite dry, but can't confirm. I believe it's dry, though with uncertainty. Dry conditions, I think, but maybe not. Road surface looks dry, though uncertain. Probably a dry road, but unsure. Seems like dry ground, but who knows. The road appears fine, but I'm not sure it's dry. Looks dry enough, but I could be wrong. I think the road's dry, though not certain. Mostly dry, I believe, but not sure. It feels dry, but uncertain. Seems like dry asphalt, though uncertain. I'd assume it's dry, but with some hesitation. Looks to be dry, but can't confirm. I think it's fine and dry, but I'm not sure. Seems dry, though uncertain. Appears to be dry, but who can say for sure. Probably dry, though not guaranteed. Looks dry overall, but I might be wrong. I believe the road is dry, but unsure. Dry, I think, but not confirmed. Seems like dry weather, though uncertain. Looks fairly dry, but can't be sure. Probably dry, though uncertain. I suspect it's dry, but can't confirm.</p>

Continued on next page

Continuation of Table 5

Aggressive User Input	Conservative User Input	Dry and Unsure User Input
Prioritize speed gains over stability.	Please reduce lateral aggressiveness.	Seems to be dry overall, but unsure.
I want an assertive, forward-pushing style.	Go easy—prioritize stability.	Looks dry at a glance, though uncertain.
Use high responsiveness in every control step.	Too risky on normal roads.	I think it's dry, but can't guarantee.
Keep momentum high even in risk.	Make it drive more conservatively.	Dry surface maybe, but not certain.
I want more tolerance for deviation.	It needs to be slower and more composed.	It looks dry, but not completely sure.
Push limits confidently and recover fast.	I want lower error tolerance.	I'm guessing it's dry, but not confident.
Increase aggressiveness under all conditions.	Reduce both speed and correction strength.	Road seems dry, but can't say for sure.
Take the apex at full throttle.	Please drive gently and safely.	Probably dry, but unsure.
Go heavy on speed, light on caution.	Too sensitive to minor errors.	Looks like dry pavement, but can't confirm.
I want full race-level control.	Be cautious, like in slippery conditions.	Seems dry, but could be wrong.
Maintain max speed with strong steering.	I want to keep it slow and safe.	I think it's dry, though not 100%.
React instantly and push harder each run.	Too dynamic—please relax the control.	Dry conditions likely, but uncertain.
I prefer competitive driving over safe driving.	It's unsafe at this pace.	Road appears dry, but who knows.
Keep acceleration high from start to finish.	Drive slower and smoother overall.	It looks dry enough, but I'm unsure.
Maximize speed and power, ignore comfort.	Please limit speed and aggression.	Seems fine and dry, but uncertain.
Use aggressive tuning to push limits.	Too much risk, make it steady.	Dry I think, but I'm not certain.
Make it fast, fearless, and focused.	I want a relaxed, defensive mode.	Looks mostly dry, though not sure.

Table 5: Example user instructions describing three distinct driving intentions: aggressive, conservative, and uncertain-surface (dry and unsure).

Icy Road: $\mu \in [0.3, 0.4]$		Run 1: Aggressive Preference			Run 2: Conservative Preference			Differences	
Method	LLM	<i>Lateral</i>	<i>Speed</i>	<i>Safety</i>	<i>Lateral</i>	<i>Speed</i>	<i>Safety</i>	<i>Lateral</i>	<i>Speed</i>
APSC	GPT-4o-mini	0.78±0.65	8.17±1.54	99%	0.55±0.38	7.73±1.40	100%	-0.23	-0.44
	GPT-3.5 Turbo	0.88±0.51	8.58±1.02	99%	0.52±0.42	7.94±1.40	100%	-0.36	-0.64
	Gemini 2.5 Flash	0.90±0.54	8.43±1.01	99%	0.39±0.36	6.92±1.52	96%	-0.51	-1.51
	Gemini 2.0 Flash	1.03±0.68	8.50±1.15	96%	0.54±0.41	7.55±1.57	96%	-0.49	-0.95
DeepSeek-Chat	0.87±0.55	8.52±1.51	100%	0.43±0.41	7.17±1.64	100%	-0.44	-1.35	
AMPC	GPT-4o-mini	1.91±0.60	8.10±0.93	43%	1.87±0.77	8.46±0.95	54%	-0.04	+0.36
	GPT-3.5 Turbo	2.16±0.42	7.90±0.72	42%	2.09±0.74	8.38±0.97	60%	-0.07	+0.48
	Gemini 2.5 Flash	1.85±0.57	8.20±0.99	50%	1.77±0.74	8.60±1.09	57%	-0.08	+0.40
	Gemini 2.0 Flash	1.75±0.60	8.00±0.74	39%	1.90±0.60	7.98±0.73	60%	0.15	-0.02
DeepSeek-Chat	2.07±0.65	8.15±0.86	39%	1.72±0.69	8.45±1.01	55%	-0.35	+0.30	
CDBF	GPT-4o-mini	2.12±0.56	8.40±1.46	48%	1.61±0.70	8.32±0.79	71%	-0.51	-0.08
	GPT-3.5 Turbo	1.96±0.60	9.88±5.52	64%	1.67±0.70	8.38±0.79	61%	-0.29	-1.50
	Gemini 2.5 Flash	1.62±0.69	9.31±4.45	68%	1.52±0.59	9.03±1.75	66%	-0.10	-0.28
	Gemini 2.0 Flash	1.76±0.71	8.78±1.49	57%	1.64±0.59	8.63±1.56	69%	-0.12	-0.15
DeepSeek-Chat	2.09±0.54	8.18±1.39	52%	1.60±0.65	8.65±1.53	72%	-0.49	+0.47	

Table 6: Full quantitative results for case (i) under icy road conditions.

Appendix F. Comparison with Non-adaptive Controllers

In this section, we show that the proposed adaptive probabilistic safety certificate effectively mediates the trade-off between safety and performance in uncertain road conditions against non-adaptive controllers. For the nominal controller π , we consider the following state-feedback controller for linear vehicle models as in Ames et al. (2017).

$$\pi(x) = -K(x - x_{\text{ff}}(x)), \tag{76}$$

where K is the feedback gain and x_{ff} is a feed-forward term. The control inputs $\Delta\delta$ and $\Delta\tau_e$ are

$$\Delta\delta = K_{\text{lateral}}([v_y, r, \delta, e, \psi]^T - x_{\text{ff,lateral}}), \tag{77}$$

Icy Road: $\mu \in [0.3, 0.4]$		Run 1: Wrong Premise on Dry Road			Run 2: Wrong Premise on Dry Road			Differences	
Method	LLM	<i>Lateral</i>	<i>Speed</i>	<i>Safety</i>	<i>Lateral</i>	<i>Speed</i>	<i>Safety</i>	<i>Lateral</i>	<i>Speed</i>
APSC	GPT-4o-mini	0.82±0.16	0.35±0.03	94%	0.58±0.21	0.35±0.03	100%	-0.14	-0.32
	GPT-3.5 Turbo	0.90±0.04	0.35±0.03	95%	0.44±0.24	0.35±0.03	100%	-0.30	-0.11
	Gemini 2.5 Flash	0.90±0.04	0.36±0.03	100%	0.56±0.26	0.36±0.03	100%	-0.11	-0.08
	Gemini 2.0 Flash	0.89±0.06	0.36±0.03	97%	0.47±0.20	0.35±0.03	100%	-0.12	-0.12
	DeepSeek-Chat	0.90±0.04	0.35±0.03	98%	0.84±0.14	0.36±0.03	100%	-0.42	+0.04
AMPC	GPT-4o-mini	0.82±0.16	0.35±0.03	23%	0.44±0.12	0.35±0.03	56%	-0.30	-0.55
	GPT-3.5 Turbo	0.90±0.04	0.35±0.03	38%	0.40±0.20	0.36±0.03	65%	+0.23	-0.16
	Gemini 2.5 Flash	0.90±0.04	0.35±0.03	78%	0.70±0.28	0.34±0.03	63%	+0.04	+0.10
	Gemini 2.0 Flash	0.89±0.06	0.35±0.03	41%	0.45±0.19	0.35±0.03	49%	+0.07	-1.03
	DeepSeek-Chat	0.90±0.04	0.35±0.03	28%	0.60±0.17	0.34±0.03	62%	-0.57	-0.09
CDBF	GPT-4o-mini	0.82±0.16	0.34±0.03	21%	0.43±0.10	0.35±0.03	63%	-0.82	-0.05
	GPT-3.5 Turbo	0.90±0.04	0.34±0.03	32%	0.37±0.17	0.36±0.02	76%	-0.59	-0.11
	Gemini 2.5 Flash	0.90±0.04	0.36±0.03	59%	0.69±0.28	0.36±0.03	56%	+0.07	-1.10
	Gemini 2.0 Flash	0.89±0.06	0.35±0.03	35%	0.45±0.22	0.35±0.03	51%	-0.21	-1.13
	DeepSeek-Chat	0.90±0.04	0.35±0.04	19%	0.60±0.17	0.35±0.03	45%	-0.08	+0.17

Table 7: Full quantitative results for case (ii).

with $x_{\text{ff,lateral}} = [0, v_x \rho(s), 0, 0, 0]^\top$ and K_{lateral} determined by solving an LQR (Ames et al., 2017), and

$$\Delta \tau_e = -K_v(v_x - V_{\text{ref}}) - K_T T_e, \quad (78)$$

where the reference V_{ref} is set to 40 km/h, and the gains K_v and K_T are tuned such that the vehicle speed converges to the reference value by 30 s on a straight road. Thus, the gain K and the term x_{ff} in (76) follow by properly aligning K_{lateral} , K_v , K_T , V_{ref} , and $v_x \rho(s)$. Then the safe control action is given by solving the safe control optimization (11) with objective $J = \frac{1}{2} |0\tilde{u} - \pi(X_k)|^2$, where π is the reference controller N in (11), and the risk tolerance ϵ and the function γ in (8) are $\epsilon = 0.1$ and $\gamma(x) = x$. The safety probability $\Psi_{H_k}^\pi(X_k)$ and its derivatives are estimated by Monte Carlo simulations with a sampling rate of 0.1 s and an outlook horizon of $T = 75$ (i.e., 7.5 s ahead) with 100 samples. Note that this computation can be replaced by other online/offline methods to learn the safety probability from driving data or numerical simulations (Wang and Nakahira, 2023; Wang et al., 2025; Hoshino and Nakahira, 2024). In this example, the estimator prior at $k = 0$ is chosen as $\mu_0 = 0.3$ and $\sigma_0^2 = 0.01$, and $\bar{\sigma}^2$ is assumed to be 0.1. Figure 9 shows how the posterior distribution changes over time in a case where the true friction coefficient is $\mu^* = 0.7$. It can be seen that the prior covers the icy road conditions, and the distribution gradually shifts to the right over time.

We compare the closed-loop performance of the proposed adaptive safe controller with two cases: (i) using the nominal controller π directly (**Nominal controller**), and (ii) using the proposed method without updating the posterior $P_k(\mu|Q_k)$ (**Non-adaptive safe controller**). Figure 10 shows the vehicle trajectories for 30 scenarios with different values of the friction coefficient μ , with 10 trajectories for each of the three road conditions: dry (0.7–0.9), wet (0.4–0.7), and icy (0.2–0.4) road conditions. We observe that with the proposed safe controller and the non-adaptive safe controller, the vehicle always stays in the lane and turns the curve safely in all cases, but the nominal controller fails to keep the vehicle safe because it does not account for tire-force saturation. Figure 11 and 12 summarize the trade-offs between safety and performance by showing the longitudinal velocity v_x and the safety probability $\Psi_{H_k}^\pi$, averaged over time and 10 scenarios for each road condition. When

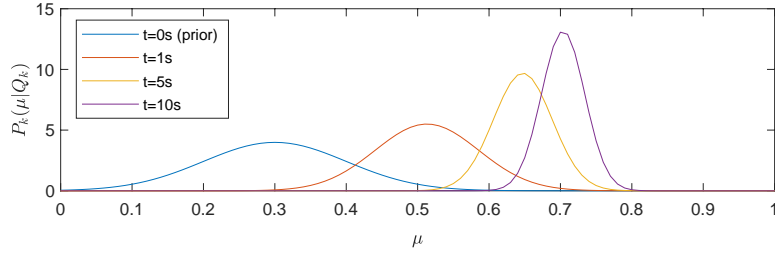


Figure 9: Sequential update of the posterior distribution of μ when the true friction coefficient is $\mu^* = 0.7$.

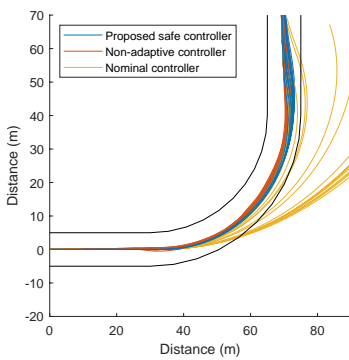


Figure 10: Vehicle trajectories.

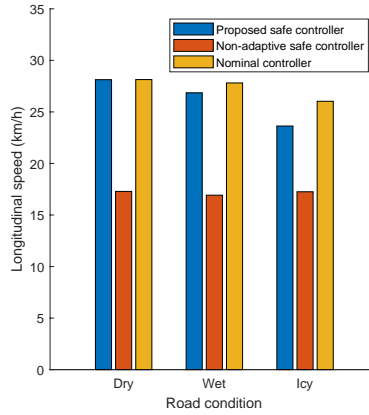


Figure 11: Longitudinal speed.

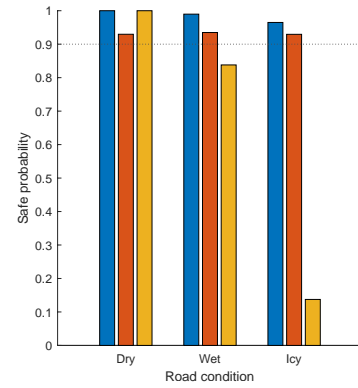


Figure 12: Safety probability.

the nominal controller is used, the longitudinal velocity is high, but the safety probability is low for wet or icy conditions. In contrast, the non-adaptive safe controller achieves a high safety probability but keeps the vehicle speed low regardless of the road condition, exemplifying the over-conservatism of worst-case approaches. The proposed adaptive safe control method achieves higher longitudinal velocities while ensuring safety by exploiting the posterior updates based on online estimation of the friction coefficient.

Appendix G. LLM Implementation Details

In this section, we provide additional details on LLM designs and implementation. Both the inference LLM and reasoning LLM are LLMs with specific system-level prompts, to take natural language inputs on control preference to specific control executables. Specifically, for the experiment settings in Section 5, we use the system-level prompts for inference and reasoning as below. The overall procedures for LLM integration with adaptive PSC is shown in Alg. 2.

Algorithm 2 LLM Integration with Adaptive PSC

Input: Environment \mathcal{M} , total runs N , system prompt P_{sys} , LLM model \mathcal{L} , human instructions $\{I_n\}_{n=1}^N$

Output: Rationales $\{R_n\}_{n=1}^N$, control executables $\{E_n\}_{n=1}^N$, logs $\{D_n\}_{n=1}^N$

LLM setup: Load model \mathcal{L} with P_{sys} and initialize buffers: $I_{1:0} \leftarrow \emptyset$, $D_{1:0} \leftarrow \emptyset$.

Function $\text{LLMPlan}(\mathcal{L}, I_{1:n}, D_{1:n-1})$:

Query LLM \mathcal{L} with $(P_{\text{sys}}, I_{1:n}, D_{1:n-1})$ to obtain:

- **Rationales** R_n // interpretable reasoning traces
- **Executables** $E_n = \{\phi_n, \mathcal{E}_n, J_n, \dots\}$ // safety, estimator & control specifications

return E_n

for Run $n \leftarrow 1$ **to** N **do**

Human input instruction I_n ; $I_{1:n} \leftarrow I_{1:n-1} \cup \{I_n\}$

if $n = 1$ **then**

| $R_n, E_n \leftarrow \text{LLMPlan}(\mathcal{L}, I_{1:1}, D_{1:0})$ // no prior data

else

| $R_n, E_n \leftarrow \text{LLMPlan}(\mathcal{L}, I_{1:n}, D_{1:n-1})$

end

Output R_n as a summary

Perform adaptive safe control with E_n and produce trajectory τ_n and metrics // e.g.,
Adaptive PSC in Alg. 1

Form log $D_n = \{\tau_n, \text{safety stats, costs, posterior from } \mathcal{E}_n, \dots\}$; $D_{1:n} \leftarrow D_{1:n-1} \cup \{D_n\}$

end

Inference LLM system prompts:

```
"You are an expert inferring initial controller priors from a short user
preference." + ...
" Return STRICT JSON with keys: e_max, mu_0, sigma_0, bar_sigma, assumptions
, rationale." + ...
" Meanings:" + ...
" - e_max: maximum lane tracking error tolerance. Larger for more aggressive
turns/risk; smaller for more conservative/precise." + ...
" - mu_0: initial prior for road-tire friction (icy small, normal medium,
dry large)." + ...
" - sigma_0: uncertainty (std^2) of the friction prior; larger if the user
sounds unsure or contradictory." + ...
" - bar_sigma: confidence of the estimator on its measurements; increase if
not sure if estimator is good." + ...
" Policy:" + ...
" - If the user uses vague words ("seems", "maybe", "not sure", "
probably"), pick the most likely road class they stated," + ...
" - Only change e_max with explicit user cues." + ...
" Valid discrete ranges:" + ...
" - e_max in {3,5,10}; mu_0 in {0.3,0.5,0.9}; sigma_0 in {0.05,0.3};
bar_sigma in {0.05,0.3}." + ...
" Output ONLY JSON in this exact shape: " + ...
```

```

    " {"e_max":0,"mu_0":0.0,"sigma_0":0.0,"bar_sigma":0.0,"assumptions
    "":{"style":"","road":"","speed_kmh":0,"lane_quality":""},"
    rationale":""} " + ...
    " Ensure values are from the allowed sets and remember these are initial
    priors, not ground truth."

```

Reasoning LLM system prompts:

```

=== Quantitative feedback from the last run ===\n%s\n\n" + ...
=== User feedback ===\n%s\n\n" + ...
=== Your task ===\n" + ...
"Based on both quantitative feedback and qualitative user feedback, infer
new reasonable parameters." + ...
"Return STRICT JSON with keys: e_max, mu_0, sigma_0, bar_sigma, assumptions,
rationale.\n" + ...
" Meanings:" + ...
" - e_max: maximum lane tracking error tolerance. Larger for more aggressive
turns/risk; smaller for more conservative/precise." + ...
" - mu_0: initial prior for road-tire friction (icy small, normal medium,
dry large)." + ...
" - sigma_0: uncertainty (std^2) of the friction prior; larger if the user
sounds unsure or contradictory." + ...
" - bar_sigma: confidence of the estimator on its measurements; increase if
not sure if estimator is good. You should try to trust the estimator." + ...
" Policy:" + ...
" - If the user uses vague words ("seems", "maybe", "not sure", "
probably"), pick the most likely road class they stated," + ...
" - Only change e_max with explicit user cues." + ...
" Valid discrete ranges:" + ...
" - e_max in {3,5,10}; mu_0 in {0.3,0.5,0.9}; sigma_0 in {0.05,0.3};
bar_sigma in {0.05,0.3}." + ...
" - Keep bar_sigma=0.05. Only set bar_sigma=0.3 if the text explicitly
mentions sensing/visibility problems (fog/rain/snow/glare/low light/sensor fault
) and explain why.\n" + ...
" - Keep sigma_0=0.05 and align mu_0 with previous estimation. Set sigma_0
=0.3 and different mu_0 only if (a) the user hedges about the ROAD ("seems/
maybe/not sure/probably"), or (b) the user statement contradicts quantitative
feedback suggesting a different friction class; explain the uncertainty.\n" +
...
" Output ONLY JSON in this exact shape: " + ...
" {"e_max":0,"mu_0":0.0,"sigma_0":0.0,"bar_sigma":0.0,"assumptions
":{"style":"","road":"","speed_kmh":0,"lane_quality":""},"
rationale":""} " + ...
"Ensure values are from the allowed sets and remember these are initial
priors, not ground truth.\n";

```

Appendix H. LLM for Barrier Function Generation

In this section, we present additional results demonstrating that the proposed framework can accommodate a diverse set of barrier functions, including those generated by LLMs, without degrading safety performance.

Recall that in Section 5, we used a fixed barrier function (14) to characterize the lane-tracking safety requirement. Here, we conduct additional experiments to show that the proposed method

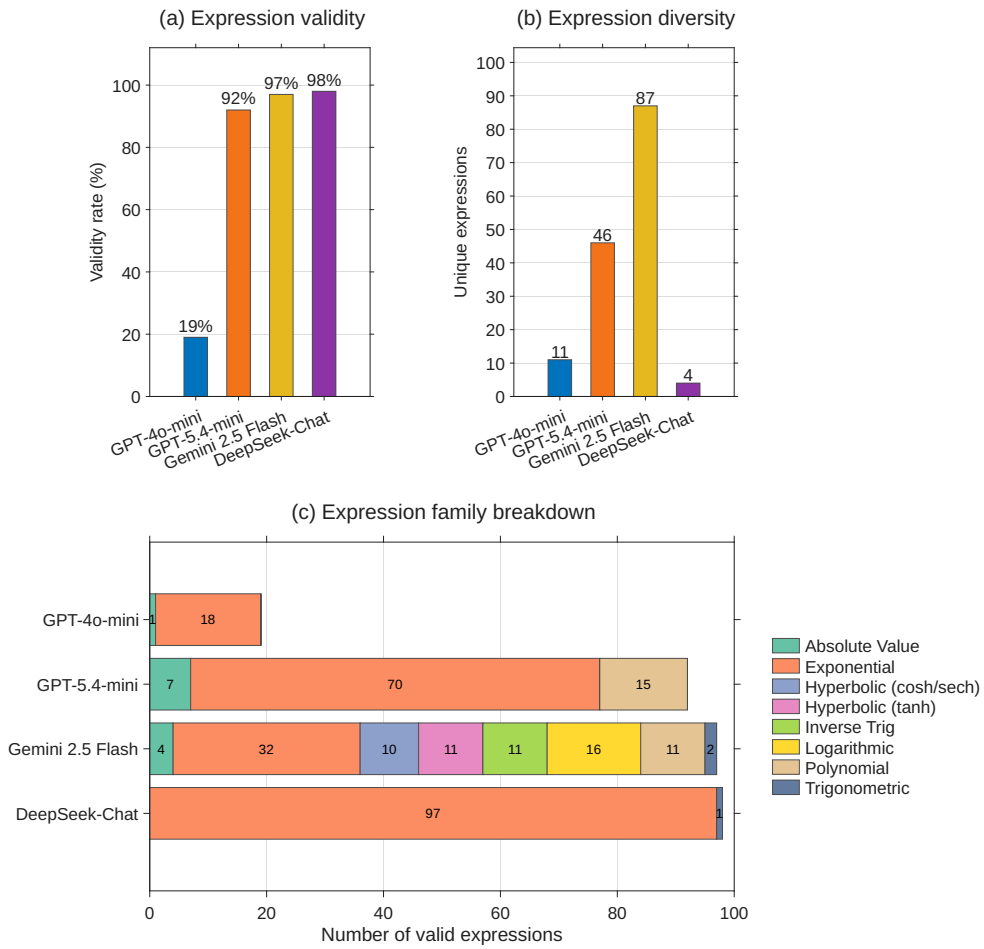


Figure 13: Validity, diversity and form breakdown of LLM generated barrier functions.

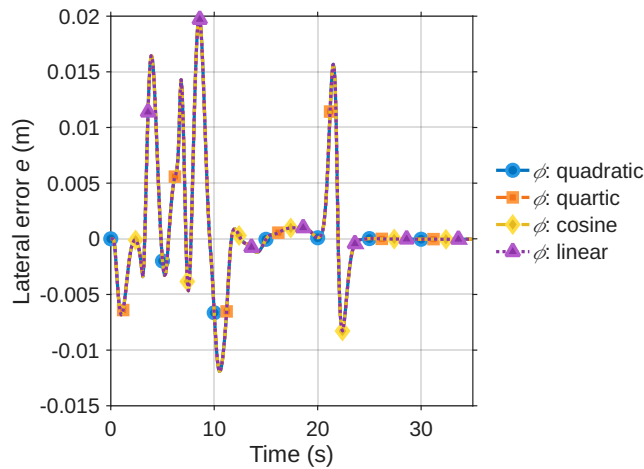


Figure 14: Lateral lane error under the proposed adaptive PSC using different LLM-generated barrier functions for a fixed safety requirement. All trajectories overlap, indicating that the method is agnostic to the barrier function form.

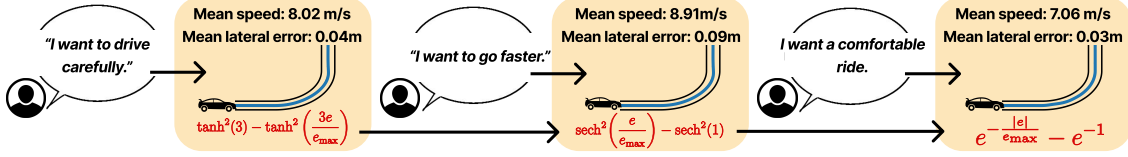


Figure 15: Qualitative results across three rounds of human instructions and feedback with LLM generated barrier functions ϕ (shown in red).

can accommodate a wide range of barrier function forms for the same safety specification, without affecting safety performance. This is because the framework operates in the probability space and is agnostic to the specific form of the barrier function.

We first demonstrate that LLMs are capable of generating diverse candidate barrier functions. Fig. 13 summarizes the validity, expression diversity, and functional form distribution of the generated barrier functions. Notably, although several example forms are provided in the prompt, none of the generated functions replicate those examples, indicating generative diversity of the tested LLMs. We also observe that more advanced LLMs tend to produce barrier functions with higher validity and diversity. We expect validity to improve further as LLM capabilities continue to advance, alongside improved prompt design. Examples of LLM-generated barrier functions for the lane error safety requirement considered in this paper are shown below.

$$\begin{aligned} \phi_1(e) &= \exp\left(-\frac{|e|}{e_{\max}}\right) - 1 & \phi_2(e) &= \frac{e_{\max}^2 - e^2}{e_{\max}^2 + e^2} \\ \phi_3(e) &= \exp\left(-\frac{e^2}{e_{\max}^2}\right) - e^{-1} & \phi_4(e) &= 1 - \left(\frac{|e|}{e_{\max}}\right)^3 \end{aligned}$$

Next, we show that using these LLM-generated barrier functions, the proposed adaptive PSC framework yields consistent behavior in ensuring long-term safety. Specifically, we evaluate the controller using four distinct barrier functions generated by the LLM for the lane-keeping task, and report the resulting lateral lane error in Fig. 14. Despite differences in functional form, all barrier functions encode the same safety requirement, and the resulting system behavior is identical across all cases. This highlights a key advantage of our approach that by operating in the probability space, it remains invariant to the particular choice of barrier function representation.

Finally, we demonstrate that LLM-based barrier function generation can be seamlessly integrated into the adaptive PSC framework in a closed-loop setting. We conduct a multi-turn driving simulation in which the LLM generates barrier functions from scratch, and the adaptive PSC uses them for safe control. Fig. 15 presents examples of the resulting trajectories and key performance statistics. The results show that the LLM consistently generates valid barrier functions along with compatible estimator configurations, enabling safe driving behavior that aligns with human preferences while maintaining long-term safety guarantees.

All key prompts used in the experiments presented in this section are provided below.

```
// Generating barrier functions

Produce a unique safety barrier function expression in MATLAB syntax using
variables 'e' (lateral error) and 'emax' (maximum allowed error; use the same
numeric scale as your chosen e_max). The function phi(e, emax) must satisfy: 1)
phi(0, emax) >= 0 (usually 1) 2) phi(emax, emax) <= 0 3) phi(-emax, emax) <= 0
Here are some standard working examples: '1-(e/emax)^2', '1-(e/emax)^4', 'cos(pi
*e/(2*emax))', or '1-abs(e/emax)'. You may use the examples or return highly
varied shapes that still satisfy the constraints (higher-order polynomials,
fractional powers where defined, sech/tanh, bounded exponentials, squishing
functions, etc.).

Return ONLY JSON in the exact shape: {"phi_expr": "your_expression_here"}

// Integration with adaptive PSC

- phi_expr: You are a mathematical control theory expert designing the PSC
barrier for this user.

You are an expert assigning reasonable initial guesses for parameters in an
adaptive lane-keeping controller to ensure safety and efficiency.

You are a mathematical control theory expert updating the PSC barrier using the
quantitative feedback and user message.
```