

Industrial Control

Problems and challenges in LLM-based intelligent traffic motion prediction under cyberattacks

Junyi Yang¹, Sheng Gao^{2,3,*}, Kai Rao^{2,3}, Yunkai Lyu², Hao Zhang^{1,3,**} , and Huaicheng Yan^{2,4}

¹ College of Electronics and Information Engineering, Tongji University, Shanghai 201801, China

² Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

³ Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai 201210, China

⁴ Faculty of Artificial Intelligence, Shanghai University of Electric Power, Shanghai 200090, China

Received: 18 December 2025 / Revised: 15 January 2026 / Accepted: 5 February 2026 / Published online: 25 April 2026

Citation Yang J, Gao S, Rao K, et al. Problems and challenges in LLM-based intelligent traffic motion prediction under cyberattacks. *Security and Safety* 2026; 5: 2026007. <https://doi.org/10.1051/sands/2026007>

1 Introduction

In traffic control applications such as adaptive signal control and cooperative adaptive cruise control, the traffic processes are typically modeled as dynamic systems, where traffic flow speed, density, and queue length are constrained by traffic flow conservation laws and vehicle dynamics models. Large language models (LLMs) are not based on first-principles physical models, but rather on pattern recognition and generative models based on large-scale data correlations, which fundamentally changes the previous paradigm. When used to predict the future behavior of traffic participants, such as vehicle lane changes or pedestrian crossings, it essentially acts as a highly complex, nonlinear, time-varying, and uninterpretable state predictor or behavior generator module embedded in the control loop. From a control perspective, this brings about a fundamental transformation, the system's prediction component no longer corresponds to an interpretable dynamic model and is difficult to analyze using traditional control theory. Its output is probabilistic natural language or discrete action labels, rather than continuous state variables with clear physical meaning. This requires us to re-examine aspects such as attack-detection-defend and control system. Recent studies have begun exploring LLMs and multimodal LLMs as high-level reasoning modules for autonomous driving, including behavior planning and motion prediction. Representative efforts investigate LLM-guided behavioral decision interfaces in simulators and modular driving stacks, as well as LLM-powered context extraction to enhance motion prediction performance [1]. These developments suggest an emerging trend of embedding semantic reasoning modules into the perception–prediction–planning pipeline [2]. In contrast to application-driven works that focus primarily on accuracy improvements, the present paper centers on the control-theoretic implications under cyberattacks, emphasizing how semantic-level vulnerabilities can propagate into estimation, control design, and closed-loop stability and safety. Figure 1 shows the overall framework diagram of intelligent traffic motion prediction based on LLMs under cyberattacks.

2 The core problems and challenges in attack, detection, and defense

2.1 Cyberattack strategies: Targeted penetration against the prediction-control loop

- (1) Control perspective: Traditional cyberattacks against control systems, such as false data injection and denial-of-service (DoS) attacks, primarily compromise the integrity or availability of state

* Corresponding author (email: sheng-gao@ecust.edu.cn)

** Corresponding author (email: zhang-hao@tongji.edu.cn)

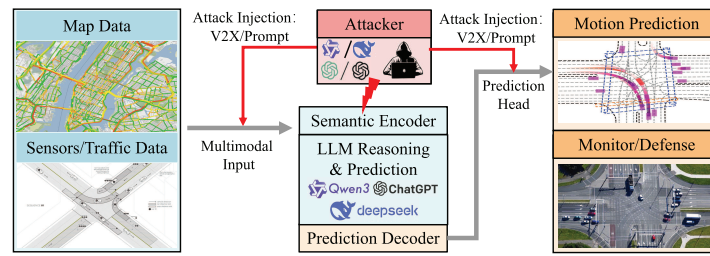


Figure 1. Conceptual architecture of LLM-based intelligent traffic motion prediction under cyberattacks

measurements [3, 4]. Their effects can be detected through observer residuals or state estimation consistency [5, 6]. Attacker typically assume that the disturbances have certain statistical properties or energy constraints.

- (2) Attack problems introduced by LLMs: LLMs introduce novel attack vectors based on semantic logic and cognitive biases. These attacks primarily include: (a) injecting carefully crafted, semantically coherent but misleading scenario descriptions into the LLM’s context window via V2X communication or roadside infrastructure; (b) for multimodal inputs such as images, point clouds, and text, the adversarial examples can be designed that are not obvious in a single modality but generate semantic ambiguity or errors after multimodal fusion; (c) exploiting the vulnerability of LLMs to understanding temporal relationships by designing attack sequences that unfold gradually over time. This type of attack aims to induce a slow drift in the prediction model parameters, making it extremely difficult to detect using detectors based on instantaneous anomalies.
- (3) LLMs vs. conventional deep predictors: While adversarial attacks on CNN-based perception and RNN/LSTM-based trajectory predictors are well studied, LLM-based modules introduce qualitatively different vulnerabilities. First, LLMs expose a prompt/context interface, enabling direct manipulation via text-level or semantic injection without perturbing raw sensor signals. Second, LLM reasoning can couple multiple heterogeneous cues across longer temporal horizons, which increases the attack surface for inducing consistent but wrong high-level hypotheses. Third, failures often occur as mode-level semantic errors that may not create immediate numerical residual anomalies, making them harder to detect with traditional physics-based consistency checks.

2.2 Anomaly detection: Deep monitoring from data statistics to cognitive logic

- (1) Control perspective: Traditional anomaly detection methods are based on system dynamic models [7, 8]. The core assumption is that under normal operation, the system behavior conforms to a known model or statistical distribution.
- (2) Detection problems for LLM-based prediction modules: Due to the propensity of LLMs to generate hallucinations, producing seemingly correct but factually incorrect logical inferences, anomaly detection must go beyond traditional numerical residual analysis and delve into semantic consistency and logical plausibility. A multi-layered detection framework is needed. (a) The semantic-level predictions from the LLM, for instance, a “lane-change intent” can be operationalized as an increase in expected lateral displacement rate and a reduction in time-to-lane-boundary, which can be checked against physical-model-based predictions. (b) Embedding lightweight probes within key layers of the LLM, such as the attention layers and feedforward network layers, to monitor their activation patterns. An offline analysis establishes a cognitive baseline pattern library for different typical scenarios such as normal driving, congestion, and emergency avoidance. During online operation, the similarity or deviation between the current activation pattern and the baseline patterns is calculated in real-time. (c) In a connected vehicle environment, cross-validation is performed using the prediction results of neighboring vehicles for the same traffic scenario or the same target vehicle. Prediction results or intermediate features are exchanged through secure V2V communication to form a local consensus.

2.3 Closed-loop stability and safety: Systemic risks under cyberattack

- (1) Control perspective: Stability and safety are the lifelines of control systems. Formal verification of reachability analysis and control barrier functions [9] are used to prove that the system state always remains within the safe set.
- (2) Challenges in closed-loop stability and security for LLM-based control systems under cyberattack: When LLMs are part of a closed-loop system, the stability and security of the entire system become almost impossible to formally analyze using existing control theory tools. The high-dimensional nonlinear structure of LLMs and their scale and complexity make a complete analysis of their input-output behavior infeasible. Cyberattacks can be viewed as time-varying, adaptive, and strong disturbances injected into the closed-loop. Analyzing the impact of such attacks, launched by attackers against LLMs at the semantic level, on the stability of the physical closed-loop system is an unresolved challenge.

3 The core problems and challenges in control systems

3.1 Modeling and representation: The gap between physical state and semantic understanding

- (1) Control perspective: In control theory, an accurate state-space representation is fundamental to designing effective controllers. State variables such as position, velocity, and acceleration are measurable and quantifiable [10].
- (2) Challenges in LLMs-based modeling under cyberattacks: Through data preprocessing, high-level semantic representations are extracted from raw sensor data such as image pixels and laser point clouds. Based on this semantic information, the LLM forms an understanding of the traffic scene. However, there is a gap in the mapping process from low-level physical signals to high-level semantic representations. There is a lack of a strict, reversible mapping relationship between the LLM's internal semantic representation and the precise numerical state required by the control system. Attackers can introduce slight perturbations to the input image of an adversarial sample, causing the LLM's semantic understanding to make fundamental errors, such as misidentifying stopping as driving. Since this error occurs at the semantic level, it may be difficult for subsequent modules to detect at the physical state level.

3.2 Controller design: Working in cooperate with a black-box predictor

- (1) Control perspective: Modern robust control methods [11] and model predictive control (MPC) [12] explicitly incorporate model uncertainty into the design, but require a structured description of the uncertainty, such as norm boundedness.
- (2) Challenges in LLM-based control systems under cyberattacks: As a prediction module, an LLM is essentially a highly complex black-box system, and the prediction errors or uncertainties it generates do not possess the structured characteristics required by control theory. This failure mode, such as completely ignoring a critical obstacle, is discrete, semantic, and unpredictable. Traditional robust control methods assume that disturbances are continuous and bounded, making it difficult to cover this mode-level failure of LLMs. Therefore, extreme scenarios must be explicitly considered in controller design, specifically how the system can maintain basic safety even when the LLM's predictions are severely inaccurate or completely wrong. This requires a new paradigm for resilient controller design, necessitating the introduction of redundancy, parallel traditional predictors in the architecture, or the design of fault-state switching logic based on safety boundaries.

3.3 Model scale and robustness

The scale of an LLM (e.g., parameter count and context length) may improve semantic understanding and few-shot generalization, which can indirectly benefit motion prediction in complex scenes. However, robustness to adversarial manipulation is not guaranteed to increase monotonically with scale, especially under adaptive attackers who optimize perturbations against the deployed model. Understanding how model scale interacts with semantic attack surfaces, confidence calibration, and downstream closed-loop safety remains an open problem that requires joint analysis from machine learning robustness and control-theoretic stability perspectives.

4 Conclusion

Introducing LLMs into intelligent transportation motion prediction under cyberattack environments is far from a simple technological replacement, it represents a paradigm shift in control. It highlights the fundamental differences in core principles between data-driven artificial intelligence and classic control engineering. While LLMs offer new possibilities for handling extremely complex and uncertain traffic environments, blindly deploying them without seriously addressing the issues of verifiability, robustness, and security from a control perspective will lead to significant systemic risks. Future research can focus on developing novel attack models that describe semantic-level attack dynamics; developing new hybrid dynamic system theories to unify the discrete semantic decision-making of LLMs with continuous vehicle dynamics within a single framework for modeling and analysis; researching hierarchical architectures where LLMs act as advisors and traditional verifiable controllers act as executors; studying the adversarial training, input sanitization, and real-time attack detection algorithms based on anomalies in internal activation patterns specifically tailored to the characteristics of LLMs; building comprehensive testing environments and benchmarks that integrate traffic scenarios, cyberattack vectors, and LLM modules to promote the development of safety-oriented verification tools.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62503174, 62433014, 62333005, 62403200, 62473133, 62533002), the China University Industry-Research Innovation Fund (Grant No. 2024IT032), the State Key Laboratory of Autonomous Intelligent Unmanned Systems (Grant No. ZZKF2025-1-24), the Shanghai International Science and Technology Cooperation Project (Grant No. 24510714000), the Shanghai Natural Science Foundation (Grant No. 24ZR1416200), and the Fundamental Research Funds for the Central Universities.

Author contribution statement

Junyi Yang proposed the main idea. Sheng Gao co-wrote the commentary. Kai Rao and Yunkai Lyu reviewed and revised the manuscript. Hao Zhang helped develop the framework and revise the paper. Huaicheng Yan supervised the work and suggested future directions.

References

- [1] Zheng X, Wu L, Yan Z, et al. Large language models powered context-aware motion prediction in autonomous driving. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024, 980–985.
- [2] Lan Z, Liu L, Fan B, et al. Traj-LLM: A new exploration for empowering trajectory prediction with pre-trained large language models. IEEE Trans Intell Veh 2025; **10**: 794–807.
- [3] Gao S, Zhang H, Wang Z, et al. Optimal injection attack strategy for cyber-physical systems: A dynamic feedback approach. Secur Saf 2022; **1**: 2022005.
- [4] Wang Z, Gao S, Zhang H, et al. Optimal DoS attack strategy for cyber-physical systems via energy allocation. IEEE Trans Control Netw Syst 2024; **11**: 2022–2032.
- [5] Huo JR and Li XJ. Stealthy switching attacks on sensors against state estimation in cyber-physical systems. Int J Robust Nonlinear Control 2023; **33**: 1169–1183.
- [6] Zhang H, He L, Wang Z, et al. The detection mechanism for false data injection attack via the ellipsoidal set-membership approach. Asian J Control 2023; **25**: 4853–4863.
- [7] Zhang Z, Deng R, Cheng P, et al. On feasibility of coordinated time-delay and false data injection attacks on cyber-physical systems. IEEE Internet Things J 2022; **9**: 8720–8736.
- [8] Ye D and Zhang TY. Summation detector for false data-injection attack in cyber-physical systems. IEEE Trans Cybern 2020; **50**: 2338–2345.
- [9] Du Z, Zhang H, Cui P, et al. Safe trajectory generation for nonholonomic multi-robot systems: A compensation-based MPC approach. IEEE Trans Autom Sci Eng 2025; **22**: 21831–21842.
- [10] Rao K, Yan H, Huang Z, et al. Swift pursuer: A topology-accelerated and robust approach for pursuing an evader in obstacle environments with state measurement uncertainty. IEEE Robot Autom Lett 2025; **10**: 3972–3979.
- [11] Francis BA. A Course in H_∞ Control Theory. Berlin, Heidelberg: Springer, 1987.
- [12] Berberich J, Köhler J, Müller MA, et al. Data-driven model predictive control with stability and robustness guarantees. IEEE Trans Autom Control 2021; **66**: 1702–1717.