Predictability in Autonomous Driving Systems

Félix Martí-Pérez

Universitat Politècnica de València
fmarper@upv.edu.es

Background

Autonomous driving (AD) systems are increasingly relying on artificial intelligence to handle the complex tasks of perception, prediction and planning. Traditional AD architectures typically employ a modular ensemble of black-box models (Sahoo and Varadarajan 2025), one for object detection, another for trajectory forecasting of vehicles and pedestrians, and a separate planning module that determines the vehicle's actions. Although each component remains opaque, this separable design helps engineers localise and diagnose failures.

Large language models (LLMs) have demonstrated an emergent "world model" purely from text. Vision-language models (VLMs) build on that success by adding visual inputs, which adds a richer, multimodal representation of the environment. A growing trend in AD research is to leverage VLMs' context-aware capabilities in end-to-end systems, replacing the traditional modular stack (Hwang et al. 2024; Goff et al. 2025).

Recent work from Waymo has shown that VLMs obey scaling laws remarkably similar to those of LLMs, as model size, data volume, and compute scale up, performance improves predictably (Baniodeh et al. 2025). By analogy with the leaps from GPT-2 to GPT-3 and GPT-4, we may witness exponential gains in AD capabilities over the next few years or even months.

However, as LLMs have repeatedly proven, capabilities are not in line with robustness nor in accordance with difficulty (Zhou et al. 2024). Even the most capable LLMs, able to match human performance on coding and mathematics benchmarks, can catastrophically fail on simple arithmetic tasks. Thus, scepticism arises in the deployment of AD systems, especially since current deployment relies largely on manufacturer-reported metrics. Independent, transparent evaluation is therefore critical.

To address these concerns and better understand the behaviour of such systems, researchers have developed several complementary approaches to evaluate the predictability and trustworthiness of AI. Three key approaches can be found in AI predictability: Uncertainty Estimation, which enables AI systems to quantify their confidence in each output or decision; Explainable AI (XAI), build techniques that make the decision-making process of AI systems transparent and interpretable to humans; capability-oriented evaluation, where AI systems are evaluated based on underlying capabilities and task dimensions.

We pursue the direction of capability-oriented evaluation. Recent work in this field has presented the notion of assessors (Burnell et al. 2022). An assessor is a lightweight meta-model that predicts a primary model's performance on each instance by leveraging task-specific meta-features. Mainly applied to forecast language model behaviour (Schellaert, Martínez-Plumed, and Hernández-Orallo 2025; Zhou et al. 2025), we extend assessors to vision-based AD systems, enabling proactive failure prediction, deeper analysis and more reliable deployment.

Central Question

The topic of the thesis aims to study the *evaluation and predictability in autonomous driving systems.*

In collaboration with Renault, we aim to develop both capability-oriented evaluation and reliable predictability methods to increase safety and trust in new, end-to-end autonomous driving systems.

Preliminary Work

Building on the recent success of assessor models in the large-language domain, we propose the first effort to bring this methodology to vision, in particular to the task of traffic-sign detection.

Our approach anticipates failures of an object-detection network before it processes each camera frame. We begin with a simple, well-controlled experimental setup and will iteratively scale to full-complexity, state-of-the-art detection systems.

At its core, our task is this: for each instance I (a single camera frame of the road), and for a fixed detection system S (a pre-trained network that outputs bounding boxes, class scores and confidences), predict the validity indicator $V(I, S)$. Here, V is the ground-truth quality metric, which in our case is the frame's mean Intersection-over-Union (IoU) between S's output and the labelled sign. Formally, we train a lightweight meta-model M so that $M(\phi(I), S) \approx V(I, S)$ where $\phi(I)$ is a feature vector summarising all available context for instance I.

To build $\phi(I)$, we combine external sensors and scene proxies with internal network signals. For now, we are utilising the following features:

- Spatio-temporal context: GPS coordinates, timestamp, and vehicle motion (speed, yaw rate, roll).
- Camera metadata: calibration parameters to estimate sign scale and off-axis views.
- Environmental proxies: weather tags (rain, fog, visibility), image-quality scores (blur via Laplacian variance, contrast, brightness), and road-surface roughness.
- Network-intrinsic signals: raw class-confidence.
- Scene embedding: a low-dimensional representation of the image.

By combining spatio-temporal context, image-based quality proxies and the detector's internal confidence, our predict-before-you-detect assessor aims to flag challenging frames. This novel vision-domain application of assessor methodology promises to improve the robustness of traffic-sign recognition systems and allows for a more comprehensive evaluation of the pitfalls of these systems.

Planned Work

An immediate extension to the work we are doing is to shift to object-level assessment. In this regime, our meta-model will have as input features tied to each detected bounding box. By focusing on individual instances rather than whole images, we expect to capture fine-grained failure modes of specific objects at the frame level.

Future research plans to integrate our assessor paradigm into end-to-end Vision-Language Models (VLMs) designed for autonomous driving. This integration promises a unified, multi-modal assessor capable of proactively flagging uncertain or high-risk outputs in real time.

Beyond methodological extensions, we aim to pursue three complementary research directions. First, we will perform a scaling-law analysis for VLM safety: by systematically varying model parameter count, training-data volume, and compute budget, we intend to quantify how these factors correlate with safety-critical metrics. Fitting log-linear curves to our empirical results will allow us to project reliability improvements for future, larger-scale VLM releases.

Second, we would like to incorporate the publicly available Waymo disengagement reports [1] to bring real-world failure cases into our training and evaluation pipelines. It is essential to have rare but safety-critical scenarios, calibrate our meta-models to these corner cases and conduct root-cause analyses of mispredictions.

Third, we intend to adapt recent demand–ability evaluation frameworks, originally developed for LLMs, to the autonomous-driving domain. Rather than treating all frames equally, we want to define a set of continuous instance-difficulty scales (e.g. degree of occlusion, lighting contrast, weather severity, complexity of scene) and corresponding rubrics that quantify the perceptual and contextual demands placed on any detection model. This aims to align instance demands with system abilities to provide high explanatory power, revealing, for example, whether a model's failure rate is driven by poor handling of low-contrast signs or heavy rain conditions, enabling out-of-distribution generalisation by predicting performance on novel combinations of demands.

These efforts hope to carry our assessor framework from a proof-of-concept frame-level predictor to a comprehensive, safety-aware system suitable for next-generation VLM-based autonomous driving architectures. By advancing object-level prediction, multi-modal integration, scaling-law understanding, real-world failure incorporation and diversity-driven evaluation, we aim to deliver reliability guarantees that help address safety concerns from both governments and users.

Significance

In summary, our work aims to make the following contributions:

- First adaptation of assessor methodology to object-detection, enabling proactive failure prediction before any frame is processed.
- Extension to object-level assessment and integration with end-to-end Vision-Language Models AD systems.
- Scaling-law analysis ties model size, data volume, and compute budget to safety-critical metrics, enabling projections for future systems.
- Incorporation of Waymo disengagement reports brings rare, real-world corner cases into training and evaluation, improving root-cause understanding of failures.
- Demand–ability profiling applies continuous instance-difficulty scales to quantify how weather, occlusion, or scene complexity drives performance, enhancing out-of-distribution generalisation.

Given these contributions, we hope this research helps:

- Researchers gain a holistic view of the model's reliability under diverse, corner-case conditions.
- Industry to safely deploy autonomous-driving systems and be aware of the limitations and bottlenecks of their system.
- Regulators can establish clear, evidence-based certifications and try to adapt safety regulations to the development of autonomous-driving systems.

Acknowledgements

This work is currently being financed by the Renault Group. We gratefully acknowledge their support and collaboration.

References

Baniodeh, M.; Goel, K.; Ettinger, S.; Fuertes, C.; Seff, A.; Shen, T.; Gulino, C.; Yang, C.; Jerfel, G.; Choe, D.; Wang, R.; Kallem, V.; Casas, S.; Al-Rfou, R.; Sapp, B.; and Anguelov, D. 2025. Scaling Laws of Motion Forecasting and Planning – A Technical Report. arXiv:2506.08228.

[1] https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/disengagement-reports/

Burnell, R.; Burden, J.; Rutar, D.; Voudouris, K.; Cheke, L. G.; and Hernández-Orallo, J. 2022. Not a Number: Identifying Instance Features for Capability-Oriented Evaluation. In *International Joint Conference on Artificial Intelligence*.

Goff, M.; Hogan, G.; Hotz, G.; du Parc Locmaria, A.; Raczy, K.; Schäfer, H.; Shihadeh, A.; Zhang, W.; and Yousfi, Y. 2025. Learning to Drive from a World Model. arXiv:2504.19077.

Hwang, J.-J.; Xu, R.; Lin, H.; Hung, W.-C.; Ji, J.; Choi, K.; Huang, D.; He, T.; Covington, P.; Sapp, B.; Zhou, Y.; Guo, J.; Anguelov, D.; and Tan, M. 2024. EMMA: End-to-End Multimodal Model for Autonomous Driving. arXiv:2410.23262.

Sahoo, L. K.; and Varadarajan, V. 2025. Deep learning for autonomous driving systems: technological innovations, strategic implementations, and business implications - a comprehensive review. *Complex Engineering Systems*, 5(1).

Schellaert, W.; Martínez-Plumed, F.; and Hernández-Orallo, J. 2025. Analysing the Predictability of Language Model Performance. *ACM Trans. Intell. Syst. Technol.*, 16(2).

Zhou, L.; Pacchiardi, L.; Martínez-Plumed, F.; Collins, K. M.; Moros-Daval, Y.; Zhang, S.; Zhao, Q.; Huang, Y.; Sun, L.; Prunty, J. E.; Li, Z.; Sánchez-García, P.; Chen, K. J.; Casares, P. A. M.; Zu, J.; Burden, J.; Mehrbakhsh, B.; Stillwell, D.; Cebrian, M.; Wang, J.; Henderson, P.; Wu, S. T.; Kyllonen, P. C.; Cheke, L.; Xie, X.; and Hernández-Orallo, J. 2025. General Scales Unlock AI Evaluation with Explanatory and Predictive Power. arXiv:2503.06378.

Zhou, L.; Schellaert, W.; Martínez-Plumed, F.; Moros-Daval, Y.; Ferri, C.; and Hernández-Orallo, J. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032): 61–68.