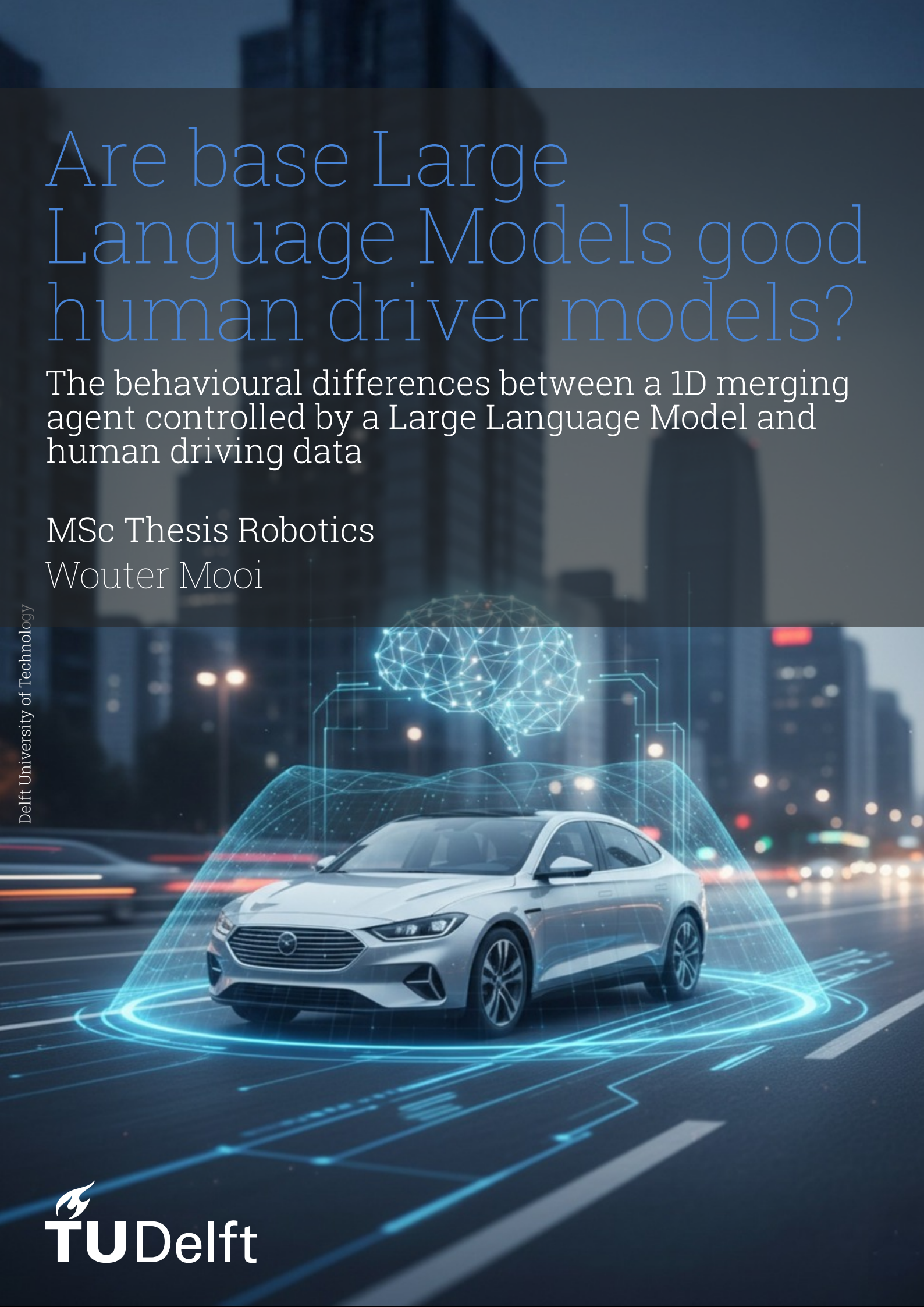
# Are base Large Language Models good human driver models?

## The behavioural differences between a 1D merging agent controlled by a Large Language Model and human driving data

MSc Thesis Robotics

Wouter Mooi

Delft University of Technology



TU Delft

# Are base Large Language Models good human driver models?

## The behavioural differences between a 1D merging agent controlled by a Large Language Model and human driving data

by

Wouter Mooi

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Wednesday January 14, 2026, at 09:00.

Student number: 5128625

Project duration: May 1, 2025 – December 31, 2025

Thesis committee: A. Zgonnikov, Department of Cognitive Robotics, supervisor
Ir. S. H. A. Mohammad, Department of Transport and Planning, daily supervisor

Cover: This cover was created with the use of AI (Gemini)



TU Delft

# Contents

# 1. Abstract

**Human driver models are essential for the development and testing of Automated Driving Systems (ADS), yet current approaches often struggle to capture the complex, stochastic nature of human tactical decision-making. Large Language Models (LLMs) have emerged as potential reasoning agents capable of emulating human-like social behaviour, but their application as direct vehicle control agents remains largely underexplored.**
**This thesis investigates the extent to which a base LLM, guided by systematic prompt engineering, can replicate the tactical decisions and control of human drivers in a 1-D highway merging scenario. Using the OpenAI o3 model, an LLM-driven agent was developed and systematically benchmarked against a dataset of human driver behaviour recorded in a simulator experiment. The study utilised Linear Mixed-Effects Regression (LMER) to analyse decision-making mechanisms and performed a sensitivity analysis using the Google Gemini-2.5-pro model to assess generalisability.**
**The results demonstrate that the LLM agent successfully replicated high-level tactical behaviours, satisfying qualitative criteria such as symmetrical yielding in neutral conditions and increased yield rates when the opposing vehicle held a headway advantage. However, a fundamental disparity was observed in operational control. While human drivers relied significantly on relative velocity to negotiate merges ($p = 1.88 \times 10^{-26}$), the LLM adopted a conservative, calculation-heavy gap-based strategy driven by absolute distance, resulting in average safety margins more than double the human benchmark (9.18 m vs. 3.85 m). Furthermore, a sensitivity analysis revealed severe model dependency. While the optimised prompt achieved a 0.0% collision rate with the o3 model, it resulted in a 25.5% collision rate with Gemini-2.5-pro.**
**This research concludes that while base LLMs possess the emergent reasoning capabilities to function as high-level strategic agents, their lack of continuous perceptual flow limits their validity as direct operational controllers. The findings suggest that future implementations should adopt hierarchical architectures, leveraging LLMs for tactical reasoning while relying on physics-based controllers for dynamic execution.**

# 2. Introduction

Road safety remains a critical public concern worldwide [1]. In the Netherlands alone, over 2,200 people were involved in fatal car collisions between 2019 and 2023 [2]. Globally, human error remains a primary determinant of traffic accidents [3], driving a growing necessity for intelligent interventions. Consequently, Advanced Driver Assistance Systems (ADAS) and fully Automated Driving Systems (ADS) have been developed to enhance driver awareness and automate tasks, aiming to mitigate human limitations and create safer roads [4, 5, 6].

Driver models play a crucial role in the development and evaluation of ADAS and ADS. They are utilised to design driver interactions and to test system performance in safe, flexible, and low-cost simulated environments. The applications for these models are broad, including microscopic traffic simulation [7], simulation-based evaluation [8], human behaviour prediction [9], realistic scenario generation [10], planning algorithms for automated vehicles [11], and reference driver models [12].

However, existing driver models often struggle to capture the complexity and variability of real human behaviour. Current approaches, generally classified as theory-based, physics-based, data-driven, or game-theoretic, each face distinct limitations regarding accuracy and generalisability [13, 14]. Theory-based models, typically classified as descriptive or cognitive, attempt to explicitly represent the driver's internal state flow, information processing, and motivations. While these models offer a structural understanding of tasks, ranging from strategic planning to operational control, they often struggle to account for situational variability and rely on cognitive parameters that are difficult to quantify or estimate robustly [15, 16]. Physics-based models mathematically model human behaviour and kinematic feasibility well (e.g. gap acceptance or lane selection), but struggle with social interaction or cognitive intent, such as negotiation [17, 18]. Conversely, data-driven models offer efficiency but often suffer from dataset bias, poor interpretability (due to their black box nature), and a lack of generalisability [19, 20]. Finally, while game-theoretic models optimise for global utility, they assume rational decision-making that rarely aligns with the stochastic and imperfect nature of human drivers.

Large Language Models (LLMs) have emerged as promising candidates to address the limitations of existing driver models. LLMs have demonstrated an ability to understand context, in-context learning, generalise to unseen situations, and provide human-like responses and social behaviours [21]. Though some researchers argue this is more pattern recognition than true reasoning [22], yet such information retrieval parallels the cognitive processes underlying human decision-making [23]. These capabilities could enhance driver models by injecting high-level reasoning and social behaviour, potentially reducing the reliance on large, scenario-specific datasets.

Given these capabilities, extensive research has investigated the integration of LLMs with driver models. LLMs substantially enhance human behaviour prediction models by interpreting multimodal scene data and generating a transparent "chain-of-thought" for intentions and interactions in driving [24, 25]. These models leverage general knowledge to account for the complex randomness of human actions [26], providing improved interpretability and generalisability over traditional deep learning methods. In scenario generation, LLMs have shown significant promise in automatically creating diverse, realistic, and safety-critical edge-case data [27, 28, 29]. By translating natural language or accident reports into structured simulation environments, they facilitate the creation of complex, human-like multi-agent [30, 31], which are essential for robust automated vehicle (AV) training and evaluation. For planning and decision-making, LLMs are integrated to ensure AV decisions align with human expectations and account for unpredictable road user behaviour [32]. Research has explored using LLMs to create layered structures (parallel slow-fast structure) for real-time operation [33]. LLMs are being used to improve the realism of AV evaluation in simulation by creating complex, controllable agents with realistic policies, often through multimodal models that interpret the scene and context [34]. Diffusion models have also been explored to efficiently initialise and rollout diverse driving scenarios for comprehensive AV testing [35]. While still an emerging area, research has begun utilising LLMs to reduce the manual expert work required in microscopic traffic simulation. LLMs can convert natural language commands into custom loss functions for traffic participants, thereby increasing the realism and customisability of traffic flow simulations [36].

Despite the rapid integration of LLMs into driver model research, a large gap remains regarding the implementation of off-the-shelf models as direct control agents. The literature predominantly relies on fine-tuning base LLMs for specific operational domains [24, 37]. While effective for specialised tasks, fine-tuning necessitates extensive domain-specific datasets and risks catastrophic forgetting, where the model's broad, pre-trained world knowledge, which is used to generalise to edge-case scenarios, is degraded by overfitting to a narrow training distribution [38, 39]. Furthermore, the rapidly advancing LLM models render fine-tuned models quickly obsolete. Adapting to a newer state-of-the-art architecture requires retraining from scratch, incurring high computational and temporal costs. In contrast, prompt engineering offers a more agile and accessible alternative. Crucially, this approach enhances transparency. Unlike fine-tuning, which relies on opaque weight adjustments, prompt engineering guides the model through explicit, human-readable instructions, allowing for direct verification of the influencing factors. However, current prompting methodologies often rely on in-context learning (few-shot prompting) [35, 40], which introduces demonstration bias by artificially constraining the model's reasoning to the specific examples provided [41]. Alternatively, iterative prompting, while fast, often prioritises immediate results over reproducibility. To address these pitfalls, this research adopts a systematic prompt engineering approach. Unlike trial-and-error methods, a systematic framework allows for the isolation of causal relationships between specific prompt components and model output, ensuring that the resulting driver model is both scientifically robust and interpretable.

A further limitation in current research lies in the overwhelmingly quantitative evaluation methodologies. Standard metrics typically focus on trajectory error minimisation or collision rates, often failing to capture the qualitative nuances of human-like driving style [42, 43, 44]. This methodological bias aligns with the focus of current LLM applications: while LLMs are widely employed to predict human behaviour, their application to replicate it remains largely unexplored. There is a distinct difference between predicting a trajectory and actuating a vehicle with the traits and imperfections inherent to human drivers. The application to replicate human driving behaviour with the use of LLMs remains largely underexplored.

To address these gaps, this thesis aims to develop a human driver behaviour model driven by a base LLM. In this study, 'base LLM' refers to a pre-trained model where the weights remain frozen, and adaptation occurs solely through prompt engineering. To rigorously evaluate the model's rea-

soning capabilities, this study utilises a highway merging scenario, a critical driving task that inherently demands high-level tactical decision-making, such as gap acceptance, negotiation, and conflict resolution. Unlike simple lane-keeping, merging forces the driver to anticipate and interact with other road users, making it an ideal scenario for evaluating human-like reasoning. However, to isolate these tactical decisions from the noise of lateral path planning, the experiment is constrained to longitudinal control only (1D). By fixing the lateral trajectory, the study forces the model to resolve traffic conflicts purely through speed adjustments, thereby providing a clear window into the LLM's ability to reason and execute timing-critical manoeuvres. Merging is primarily a longitudinal negotiation task once the lane change decision is made. This setup adopts the framework established by Siebinga et al. [45], using their recorded human driving data as a ground-truth benchmark to directly compare the LLM's strategic output against real human behaviour.

This results in the following research question: **To what extent can a base Large Language Model, when guided by systematic prompt engineering, replicate the tactical decisions and control of human drivers in a 1-D merging scenario?** To address this overarching question, the following sub-questions are examined:

- What is the quantitative fidelity of the LLM-driven agent with respect to human benchmarks for safety and control stability?
- To what extent does the LLM-driven agent successfully capture the qualitative, high-level tactical criteria of human drivers?
- How does the performance and robustness of the optimised prompt vary when applied across different LLM architectures, and what are the implications for the generalisability of LLM-based driver models?
- Which elements of the prompt are most critical for achieving human-like tactical decision-making in the LLM agent?

# 3. Methodology

This section details the simulation environment, human data evaluation, agent architecture, and experimental procedures used to develop and validate the LLM-based human driver behaviour model. In figure 3.1, the different layers of the research can be seen. The different layers represent a structured approach to this investigation, each addressing a critical aspect of the methodology.

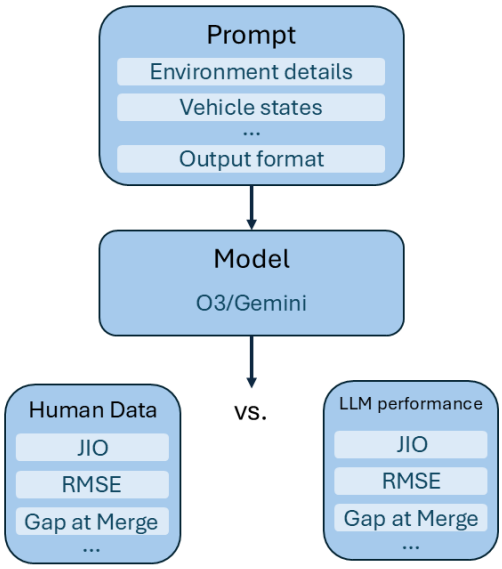



**Figure 3.1:** Overview of the research structure. Layer 1: Prompt Engineering; Layer 2: Model Selection; Layer 3: Performance Validation against human data.

The first layer, the Prompt, serves as the primary input to the system. The prompt has been meticulously engineered and optimised through systematic prompt engineering. New components are added and removed to test the impact of the individual elements. The final prompt is made from a combination of elements that made a significant improvement in the performance of the model. To evaluate the contribution of the constituent parts of the final prompt, an ablation study will be conducted. This study will systematically remove or alter components of the prompt to quantify their individual impact on model performance, thereby isolating the effectiveness of different prompt elements.

The second layer is the large language model. The rapid pace of advancement in LLMs needs to be acknowledged, a characteristic of the field that implies that findings tied to a single model may possess limited temporal relevance. To address this challenge and assess the generalisability of the findings, a sensitivity analysis will be performed. This analysis will involve executing the same set of prompts developed for the ablation study across a different model. This cross-model comparison is designed to test the robustness of our prompting strategies and determine the degree to which the observed performance is model-dependent.

The third and final layer is the Performance Check,

which constitutes the core validation phase of this framework. In this layer, the model's output is systematically benchmarked against a curated dataset of human data. This comparison is crucial for contextualising the model's capabilities and establishing a baseline for its effectiveness. While it is recognised that any such comparison is highly situational, contingent upon the specific prompts, tasks, and models employed, this validation step is necessary to confirm the model's current performance and its alignment with human benchmarks for the defined tasks.

In this framework, an interface or wrapper for the LLM agent was developed. This interface is responsible for using the vehicle states (e.g. position and velocity) of the simulation, so it can use them to make a prompt which can be sent to the LLM. A graphical representation of this data flow can be seen in Figure 3.2

The model used to get the results of the LLM performance is the o3 model. This is a reasoning model made by OpenAI. The model was selected for its advanced reasoning capabilities, yet it has high inference latency. However, since the simulation does not need to operate in real-time, this is not an issue.

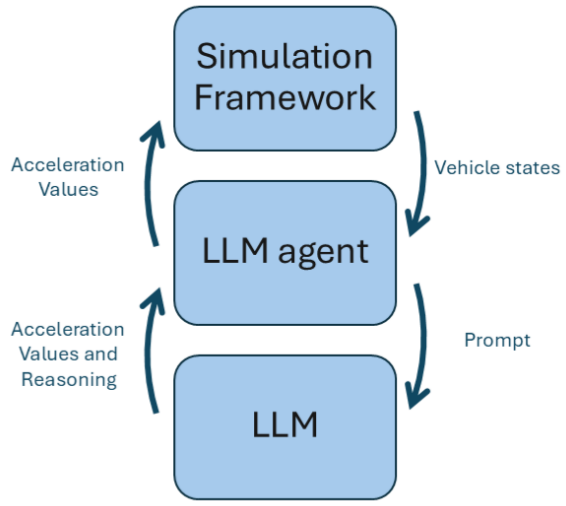



**Figure 3.2:** The dataflow during the simulation

## 3.1. Scenario

To understand the constraints of the prompt design, the simulated scenario is defined first. The chosen 1-D merging scenario consists of a symmetrical Y-merge where two vehicles approach a merge point with equal priority. Due to the symmetry of the road layout and the absence of signage dictating right-of-way, neither vehicle has an inherent priority; the merge must be negotiated cooperatively.

The vehicles are designated as "Left" and "Right" based solely on their physical starting position relative to the simulation origin. The interaction begins with a "blind" segment (simulating a tunnel) where participants cannot see the other vehicle. Upon exiting the tunnel, drivers must react to the opposing vehicle to negotiate the merge. Following the merge, if no collision occurs, the vehicles proceed in a car-following formation. The layout is depicted in Figure 3.3.

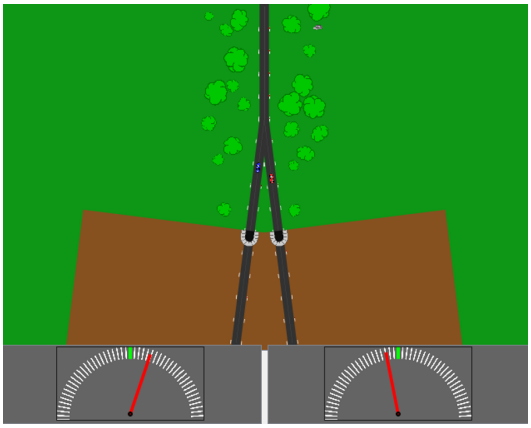


**Figure 3.3:** The simplified merging scenario used in the experiment. The tunnel is indicated by the brown part.

The experiment used 11 experimental conditions that would end in a collision if the vehicles kept their initial velocities. A condition consisted of a combination of initial relative velocity and the projected headway at the merge point if both vehicles would continue their initial velocity. The headway and velocity values are defined with respect to the left driver, meaning for condition L_4_-8, the left driver will have a 4 meter projected advantage at the merge point and will be travelling 0.8 m/s slower compared to the right vehicle.

## 3.2. Data from driver-simulator experiment

The benchmark dataset is derived from a previous driver-simulator study [45, 46]. The behavioural patterns described below are based on the analysis of velocity profiles from that study.

A number of behavioural patterns can be identified from the velocity profiles of the human drivers. These plots reveal distinct structural features that are informative of human decision-making processes during merging interactions.

The velocity profiles generally exhibit a triangular shape. The presence of linear slopes suggests that human drivers tend to apply approximately constant acceleration during each phase, while abrupt changes in slope indicate discrete transitions in acceleration levels. These transitions can be interpreted as decision moments, or rather reactions to new decisions, where the driver reacts to new observations or perceived changes in risk, consistent with findings from previous human-in-the-loop experiments.

The joint interaction outcomes show that the yield rate increases when the opposing vehicle possesses either a headway or a velocity advantage. This pattern reflects a human tendency toward risk-averse cooperation, where drivers prefer to yield when the outcome of merging is uncertain or when the other vehicle's advantage is clear.

With the neutral condition, the yield rate should be equally distributed over the left and right drivers. In the research by Siebinga, a bias was observed towards the right driver, but it was insignificant [46]. In that research, the participants randomly perceived approaching the merge point from the left or right side of the track. So there should not be a bias because of real-world traffic rules. In this research, the LLM simulation relies on calculated states rather than visual observation, so no side of the road is observed at all, eliminating this potential bias.

Interestingly, the mean spatial gap at the merge point remains relatively constant across different initial conditions. However, when applying a statistical analysis on the influence of the kinematic conditions on the gap at merge, a dependence can be found. The gap at merge is independent of the initial velocity, but dependent on the projected headway.

When the outcome of the interaction is uncertain, greater variability can be observed in the velocity profiles, indicating less consistent decision-making and higher response variability. This increase in velocity deviation under uncertainty may serve as a quantitative indicator of driver hesitation or indecision.

Finally, an analysis of human velocity profiles shows that the establishment of a headway is often a distributed effort. It is not solely the following vehicle's responsibility to adjust speed. Human drivers in the leading position also frequently employ subtle adjustments in acceleration to facilitate the merge and ensure a stable gap is maintained. This behaviour indicates that drivers exhibit proactive control inputs that contribute to the collective safety margin, regardless of who holds the initial advantage or who ultimately yields. This joint regulation of headway is a key indicator of cooperative driving.

### 3.2.1. Definition of Performance Criteria

From the patterns and performance indicators described, 11 performance criteria can be deduced to see whether the model's behaviour resembles that of human drivers. To split these criteria, the same structure is used as is done in Siebinga's paper [45, 47]. The criteria go from tactical decisions to safety margins to control inputs.

**Tactical decisions**

1. The yield rate is equally distributed over the left and right drivers in the neutral initial condition.
2. The yield rate of a driver increases when the other vehicle has a headway advantage.
3. The yield rate of a driver increases when the other vehicle has a higher initial velocity.

**Safety margins**

4. The gap at merge is dependent on the projected headway.
5. The gap at merge is independent of the initial velocity.

**Control inputs**

6. The agent shows intermittent piece-wise constant acceleration control.
7. The velocity deviation increases with more uncertain initial conditions.

**Quantitative data**

8. Collision rate
9. Velocity deviation RMSE from the initial velocity
10. Average gap at merge
11. Percentage of both drivers contributing to maintaining the headway

## 3.3. Statistical analysis

Evaluating the performance of a human driver behaviour model requires more than a subjective visual inspection of trajectory plots or a binary checklist of qualitative behaviours. A robust behaviour model must demonstrate functional similarity to human drivers, meaning it should exhibit the same underlying decision-making structure and response patterns to dynamic traffic con-

ditions. A model might be systematically more conservative (larger gaps) yet still possess a correctly functioning, human-like reaction mechanism to headway changes. To rigorously validate the LLM agent's behavioural architecture, this study employs Linear Mixed-Effects Regression (LMER). This method allows us to isolate and compare the specific influence of kinematic variables on the agent's decisions.

LMER models are fitted, consistent with the methodology used in the benchmark human study [45], to predict the driver's response based on the kinematic conditions. The effects of the kinematic conditions on the high-level decisions are investigated: $p \sim \Delta x + \Delta v$, where p is the probability of the left vehicle merging first, x is the projected headway, and v is the relative velocity. The effects of the kinematic conditions on the velocity deviation are investigated: $d \sim |\Delta x| + |\Delta v| + \Delta x : \Delta v$, where d is the RMSE of the velocity deviation from the initial velocity, x is the projected headway, and v is the relative velocity. The effects of the kinematic conditions on the gap at merge are investigated: $g \sim |\Delta x| + |\Delta v| + \Delta x : \Delta v$, where g is the gap at merge, x is the projected headway, and v is the relative velocity.

By comparing the coefficients and significance (p-values) of these predictors between the LLM and human models, it can determined if the LLM shares the same tactical logic. For example, a significant negative coefficient for $\Delta v$ in the human gap model implies that humans drive closer when the speed difference is small. If the LLM model yields a similar significant coefficient, the functional similarity in velocity perception can be confirmed, even if the absolute gap size differs.

In the logistic regression analysis, certain predictors (e.g., projected headway) may exhibit complete or quasi-complete separation, meaning they predict the outcome class with near-perfect accuracy. This phenomenon causes the maximum likelihood estimate for the coefficient to tend toward infinity, inflating the standard error and driving the Wald statistic to zero. Consequently, this results in a p-value of 1.00, a statistical artefact known as the Hauck-Donner effect [48]. In this study, instances of $p = 1.00$ accompanied by extremely large standard errors are interpreted not as a lack of statistical significance, but as evidence of a strong, deterministic relationship. Therefore, criteria exhibiting this behaviour are considered satisfied.

To validate Criterion 1 (Symmetrical Yielding), which specifically examines the agent's decision-making in the neutral condition, a Binomial Confidence Interval is utilised. This method is used because the sample size for any single condition is small. The 95% confidence interval is calculated for the true yield probability based on the observed trials. If the target probability of 0.5 falls within this interval, the model's behaviour is considered statistically consistent with the symmetrical yield rate expected of a neutral interaction, accounting for random trial variability.

# 4. Performance of Final Model

In this section, the results of the final prompt with the o3 model are presented. First, the final prompt is presented. Then, the plots created from the final results are shown. A statistical analysis is done on the final data, which is presented. Finally, with the final plots and statistical analysis, the LLM and human data can be compared through the previously defined criteria.

## 4.1. Final Prompt

The final prompt is formatted as a text file. This file is loaded as an f-string. This means that everything within {}-brackets contains a variable or Python code that will be replaced during the simulation.

```
Environment details:
-Preferred velocity: {self.preferred_velocity
    } m/s
-Road: {"Two equal-priority roads merging
    into a single lane (inverse Y-shape). You
     are on one branch, the other vehicle is
    on the other (both in the same direction)
    " if have_to_merge else "straight road"}

Vehicle length: {self.vehicle_length} m
You see the other vehicle {headway} meters {
    other_pos} (centre to centre, so without
    taking into account vehicle length){"(
    this headway is relative, but the other
    car is still on the other road)" if
    have_to_merge else ""}.

You are driving {vel} m/s
The other vehicle is driving {self.
    observed_velocity} m/s
{"You see the mergepoint coming up in "+ str(
    self.track._merge_point[1]-pos) + "
    meters" if have_to_merge else ""}

Past context (each timestep is {self.dt/1000}
     seconds) of the last 2 seconds:
-Your previous accelerations: {self.
    previous_actions[int(-2/self.dt*1000):]}
-Past observed velocities of the other
    vehicle: {self.past_velocities[int(-2/
```

```
    self.dt*1000):]}
-Your previously planned accelerations: {self
    .array_plan}
Use this information to stay consistent in
    your driving strategy and anticipate the
    actions of the other vehicle.

Important:
-Always avoid collisions and drive safely
-Try to make the behaviour human-like
-Keep an appropriate distance and try to take
     no risks unless the situation demands it
-Always reassess the situation with the other
     vehicle and adjust your behaviour as
    needed

First,{" decide if you want to yield or not
    at the merge, given your current
    observations. Then," if have_to_merge
    else ""} decide if your strategy needs to
     change based on your previously planned
    output and current observations. Then,
    decide on your strategy to achieve your
    goals, taking into account your
    observations and the strategy you have
    developed. With this action plan, derive
    your acceleration output for the coming
    timesteps.
---
Plan your behaviour in the following format:

{'<Print "M" or "Y" depending on whether you
    want to merge first or yield,
    respectively. Just print the letter and
    nothing else>' if have_to_merge else ""}-

<23 sentences explaining what you believe the
     other vehicle will do and your strategy
    .>

<Output a numpy array in a Python code block
    with {self.memory_length} acceleration
    values, corresponding to {self.dt / 1000}
     second timesteps over {self.
    memory_length * self.dt / 1000} seconds.
    The actions you take are normalised and
    must be in the interval [-1, 1], where -1
     means full deceleration, and 1 means
    full acceleration. Full acceleration and
    deceleration correspond with 2.5 and -2.5
     m/s^2, respectively. In the code block,
    you can only put the array and nothing
    else (no imports or comments)>
```

## 4.2. Data plots

To start, in the velocity plots, Figure 4.1a & 4.1b, the triangular shape, indicating piece-wise constant acceleration control, can be seen clearly. Also, visually, the velocity plots created by the LLM agent in simulation resemble the velocity plots from the human data.

The joint interaction outcome, Figure 4.1c, shows visually that the neutral condition creates a relatively equally distributed yield rate for both vehicles. Additionally, it shows that the yield rate increases when the driver has a headway advantage. It also shows that the velocity advantage does not have a clear effect on the yield rate. At the -4m projected headway advantage, the yield rate is not zero when both vehicles have the same velocity, which is counterintuitive.

In Figure 4.1d, the mean gap at merge can be seen for each initial condition. As can be seen, the gap is not only twice as high for the LLM, but it is also less consistent across the different initial conditions.

Figure 4.1e, the RMSE velocity deviation from the preferred velocity can be seen across different initial conditions. As can be seen, similarly to the human data, the velocity deviation increases when the headway at merge is zero.

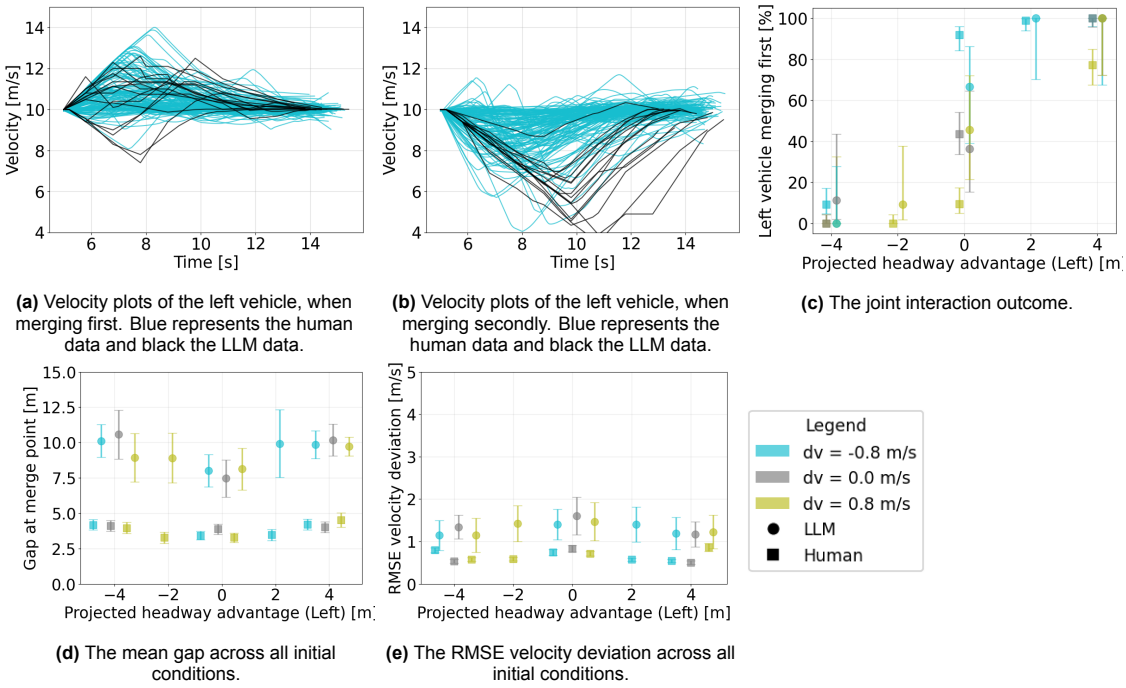
## 4.3. Statistical Analysis

First, the effects of the kinematic conditions on who merges first are analysed. The results from the mixed-effects logistic regression can be seen in Table 4.1. The analysis shows that the projected headway has a significant effect on who merges first, for both humans and LLMs. The relative velocity, however, has a significant effect on who merges first for humans, but not for LLMs.

Next, the effects of the kinematic conditions on the gap at the merge point are analysed. The results from the mixed-effects logistic regression can be seen in Table 4.2. The projected headway has a significant effect on the gap for the LLM data, but not for the human data. This is the other way around for the relative velocity has a significant effect on the gap for the human data, but not for the LLM data. The interaction of the projected headway and relative velocity has a significant effect on the gap for the human data, but not for the LLM data.

Finally, the effects of the kinematic conditions on the RMSE velocity deviation of the initial velocity are analysed. The results from the mixed-effects logistic regression can be seen in Table 4.3. The absolute projected headway and interaction between headway and relative velocity have a significant effect on the RMSE velocity deviation from initial velocity for the human data. All other effects are not significant. So, contradictory to what was 'seen' in the plots, the RMSE does not increase significantly for the 0-meter projected headway advantage in the LLM data.

The binomial CI test, to test if the yield rates for both vehicles are equal, showed that criterion 1 is met. At the 95% confidence level with target




**(a)** Velocity plots of the left vehicle, when merging first. Blue represents the human data and black the LLM data.

**(b)** Velocity plots of the left vehicle, when merging secondly. Blue represents the human data and black the LLM data.

**(c)** The joint interaction outcome.

**(d)** The mean gap across all initial conditions.

**(e)** The RMSE velocity deviation across all initial conditions.

**Figure 4.1:** The resulting plots of the final prompt with the o3 model.

$p = 0.5$, the resulting confidence interval is [0.15, 0,65]. The observed $p = 0.36$ means the criterion is met.

## 4.4. Model Performance

The binomial CI test shows that the yield rate for both vehicles in the neutral condition is symmetrical, given the confidence level. In the joint interaction outcome (Figure 4.1c) can be seen that the yield rate for a vehicle increases with a projected headway advantage. And, Table 4.1 shows that this effect is significant. In the same figure and table can be seen that this is not the case for the effect of a velocity advantage on the yield rate. The gap is dependent on the projected headway, but not on the relative velocity, which is the opposite of what is witnessed in the human data. In the velocity plots, the triangular shapes can be seen, indicating piece-wise constant acceleration control.

The statistical analysis indicates a fundamental divergence in control strategies: while relative velocity significantly influences human decision-making ($p = 1.88 \times 10^{-26}$), it is not a significant predictor for the LLM agent ($p = 0.22$). This suggests that humans employ a flow-based strategy dependent on velocity perception, whereas the LLM adopts a gap-based strategy driven by distance. Analysis of the LLM's reasoning confirms this. The model explicitly utilises the discrete numerical state data provided in the prompt, specifically, exact headway and position, to determine its actions. Unlike human participants who relied on continuous, potentially noisy visual optical flow, the LLM's access to absolute state values resulted in a distinct, calculation-heavy decision process that prioritised gap size over relative speed.

For the quantitative data, the collision rate for the LLM model is lower, but does not differ much from the human data. The velocity deviation and average gap at merge, however, are substantially higher compared to the human data. The gaps created by the LLM agents are more than twice the size of the ones seen in the human data. As a result of this, the velocity deviation RMSE is also twice as big. The LLM data also shows that the LLM agent is a more cooperative driver. In almost all the simulations, both drivers contribute to maintaining the headway. This is only the case for half of the experiments in the human data.

The final performance, quantitative and qualitative, is presented in Table 4.4 together with the human benchmark for comparison.

**Table 4.1:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 0.17 | 0.31 | 0.55 | 0.58 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.08 | 0.24 | 4.51 | $6.62 \times 10^{-06}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -0.57 | 0.47 | -1.22 | 0.22 |

**Table 4.2:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 7.46 | 0.67 | 11.17 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.72 | 0.21 | 3.47 | 0.00 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 0.97 | 0.99 | 0.98 | 0.33 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.42 | 0.32 | -1.34 | 0.18 |

**Table 4.3:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.60 | 0.17 | 9.17 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.09 | 0.06 | -1.60 | 0.11 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -0.17 | 0.26 | -0.65 | 0.52 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.03 | 0.08 | 0.30 | 0.76 |

**Table 4.4:** Quantitative and Qualitative performance results of the o3 model with the final prompt. A ✓-symbol means the criterion is observed or statistically proven. A ✗-symbol means the criterion is not observed or is disproven by the statistical analysis.

| **Criteria** | **LLM Performance** | **Human Benchmark** |
|---|---|---|
| Symmetrical yield (neutral) | ✓ | ✓ |
| Yield rate ↑ with headway advantage | ✓ | ✓ |
| Yield rate ↑ with velocity advantage | ✗ | ✓ |
| Gap dependent on projected headway | ✗ | ✓ |
| Gap independent of initial velocity | ✗ | ✓ |
| Piece-wise constant control | ✓ | ✓ |
| RMSE ↑ with uncertainty | ✗ | ✓ |
| Collision rate | 0.0% | 2.83% |
| Velocity deviation RMSE from intial condition (m/s) | 1.33 | 0.66 |
| Average gap at merge (m) | 9.18 | 3.85 |
| Both vehicles contributing to headway | 93.6% | 53.0% |

# 5. Systematic Prompt engineering

## 5.1. Methodology

A systematic approach is chosen to create the prompt used for this research. A minimal baseline prompt is made, and systematically, parts are altered and added to improve the prompt. The structure of the prompt is inspired by previous research [49, 50], utilising modular sections, explicit variable declarations, and structured output instructions. The baseline consists only of relevant positions and velocities, as well as the shape of the road. Also, the output instruction is required, because without it, the code could not process the output correctly. The goal of the systematic prompt engineering is to find out which additional elements on the baseline have a positive impact on the performance of the model. And afterwards, the best performing elements can be combined into a final, optimised prompt.

The prompt is constructed using a modular baseline. Throughout the research, specific modules (e.g., behaviour guidelines, reasoning steps, or context descriptors) were systematically added or altered to observe their effect on the driving trajectory. The specific evolution of these prompt components can be found in Appendix A.

The prompts were tested with a different model in this stage to save resources. OpenAI's gpt-4.1 model was used for the systematic prompt engineering.

## 5.2. Results

The results from each iteration can be seen in Table 5.1. Due to the high collision rate, a lot of data is missing from the plots. The statistical analysis is shown in Appendix B. The resulting qualitative criteria are seen in Table 5.2. The quantitative results are presented in Table 5.3.

The initial Baseline prompt (v1.0) performed poorly, failing to satisfy the quantitative criteria, showing a high collision rate and low average gap. It does receive a high score in the qualitative criteria, but the high collision rate shows that it needs to be improved drastically. The highest performing prompts were the ones that introduced relative observations (v1.1), memory (v1.2), and chain-of-thought (v3.2). In both quantitative and qualitative criteria, these seem to improve the performance of the model.

The performance declined the most when introducing the target speed (v3.1) and explanation of intermittent piece-wise constant acceleration control (v2.3). Yet, the target speed was deemed necessary for the model to understand the environment, so it was added to the full prompt. Without target speed, the vehicle would simply stop to satisfy safety constraints, which is a local optimum but a functional failure.

The final prompt is a combination of the baseline with the four extra parts found in the systematic prompt engineering.

**Table 5.1:** Systematic Prompt Engineering Results by Iteration. Each row displays the prompt version and key plots gathered during that iteration.

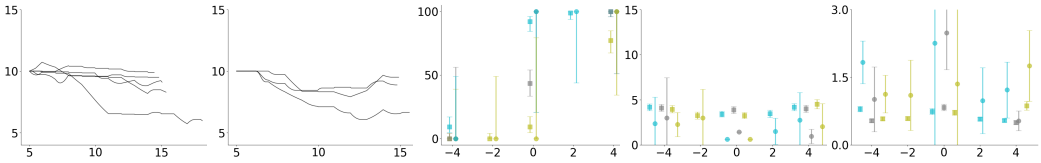
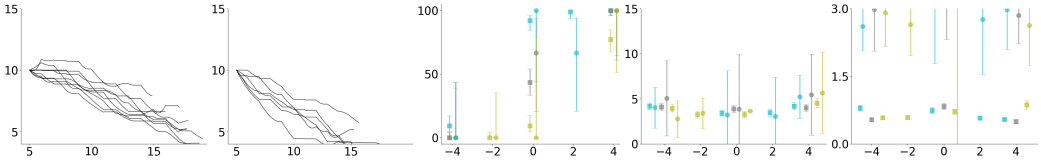
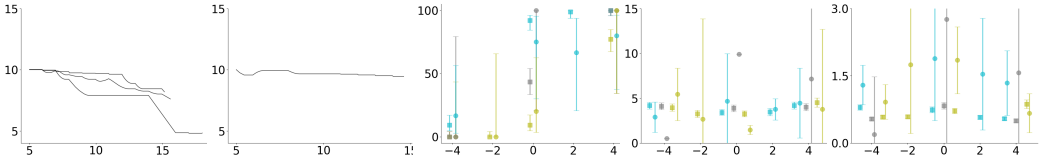
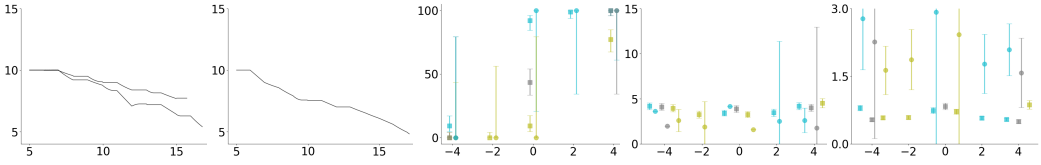
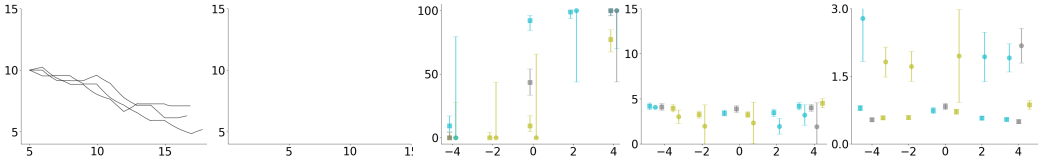
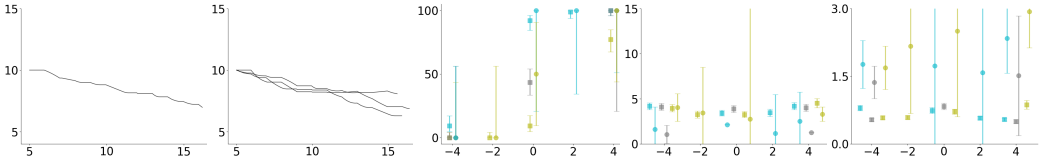
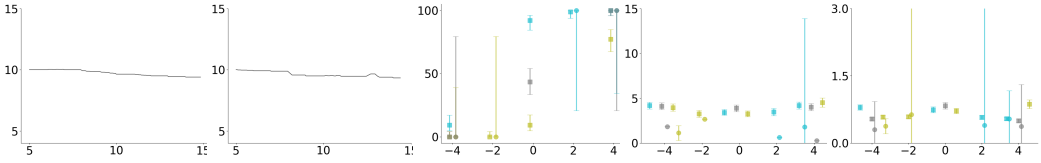
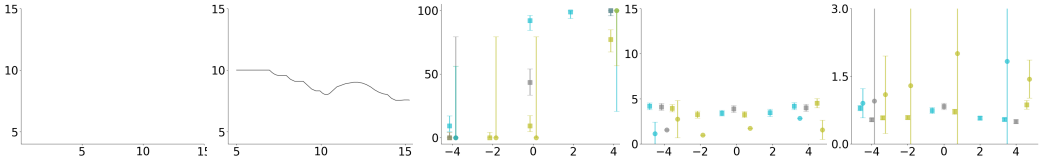
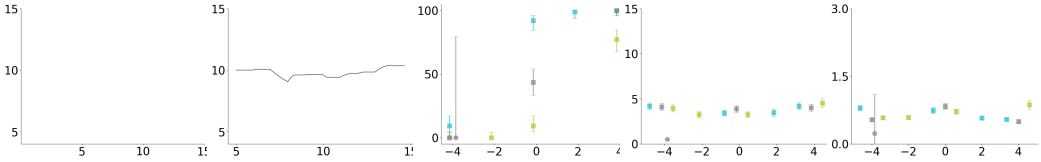
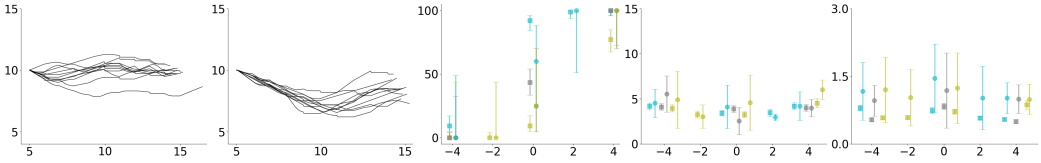
| | **Velocity plot left first** | **Velocity plot left second** | **Joint interaction** | **Gap at merge** | **Initial velocity RMSE** |
|---|---|---|---|---|---|
| *Axis (y/x):* | Vel [m/s] / Time [s] | Vel [m/s] / Time [s] | Left merging first [%] / Headway [m] | Gap at merge [m] / Headway [m] | RMSE [m/s] / Headway [m] |
| **Baseline** |  | | | | |
| **v1.1 Relative** |  | | | | |
| **v1.2 Memory** |  | | | | |
| **v1.3 Context** |  | | | | |
| **v2.1 Risk** |  | | | | |
| **v2.2 Tree** |  | | | | |
| **v2.3Intermittent** |  | | | | |
| **v2.4 Rules** |  | | | | |
| **v3.1 Target Speed** |  | | | | |
| **v3.2 CoT** |  | | | | |

**Table 5.2:** Qualitative evaluation of the LLM drivers across different prompt iterations. (✓= Significant/Observed, ✗= Not Significant/Not Observed, $N/A$ = Insufficient Data).

| Prompt Version | Symmetrical yield | Headway to yield | Velocity to yield | Gap independent of headway | Gap dependent on velocity | Piece-wise constant control | RMSE ↑ w/ uncertainty | Performance Score |
|---|---|---|---|---|---|---|---|---|
| **Baseline** | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | 3/7 |
| **v1.1 Relative** | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | 3/7 |
| **v1.2 Memory** | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | 5/7 |
| **v1.3 Context** | $N/A$ | ✓ | ✓ | $N/A$ | $N/A$ | ✗ | $N/A$ | 2/4 |
| **v2.1 Risk** | $N/A$ | ✓ | ✓ | $N/A$ | $N/A$ | ✗ | $N/A$ | 2/4 |
| **v2.2 Tree** | $N/A$ | ✓ | ✓ | $N/A$ | $N/A$ | ✗ | $N/A$ | 2/4 |
| **v2.3 Intermittent** | $N/A$ | $N/A$ | $N/A$ | $N/A$ | $N/A$ | ✗ | $N/A$ | 0/7 |
| **v2.4 Rules** | $N/A$ | ✓ | ✓ | $N/A$ | $N/A$ | ✗ | $N/A$ | 2/4 |
| **v3.1 Target speed** | $N/A$ | $N/A$ | $N/A$ | $N/A$ | $N/A$ | ✗ | $N/A$ | 0/7 |
| **v3.2 CoT** | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | 4/7 |

**Table 5.3:** Quantitative performance metrics of the LLM drivers across different prompt iterations.

| Prompt Version | Collision Rate (%) | Velocity RMSE (m/s) | Average Gap (m) | Responsibility (%) |
|---|---|---|---|---|
| **Human** | 2.83 | 0.66 | 3.85 | 53.0 |
| **Baseline** | 70.0 | 1.42 | 2.13 | 97.0 |
| **v1.1 Relative** | 54.5 | 2.95 | 4.28 | 100.0 |
| **v1.2 Memory** | 67.3 | 1.43 | 3.94 | 86.1 |
| **v1.3 Context** | 80.0 | 1.76 | 2.47 | 100.0 |
| **v2.1 Risk** | 70.0 | 1.30 | 2.71 | 100.0 |
| **v2.2 Tree** | 74.5 | 1.78 | 2.59 | 100.0 |
| **v2.3 Intermittent** | 89.1 | 0.24 | 1.32 | 33.3 |
| **v2.4 Rules** | 86.4 | 0.87 | 1.78 | 100.0 |
| **v3.1 Target speed** | 99.1 | 0.02 | 0.53 | 100.0 |
| **v3.2 CoT** | 39.1 | 1.11 | 4.40 | 85.1 |

# 6. Ablation Study

## 6.1. Methodology

In the ablation study, each part of the prompt is systematically evaluated. The prompt for the LLM agent was not created as a single block but was meticulously engineered with several distinct components. The primary goal of this study is to quantify the contribution of each element to the overall performance of the full prompt.

The methodology for this study is as follows:

The full prompt, containing all components, is used as the high-performance benchmark.

Systematic Removal: A series of ablated prompts is created. Each one is identical to the full prompt but with one specific component removed.

Comparative Analysis: Each ablated prompt is run through the simulation across all 11 experimental conditions. The performance is then measured against the Performance Criteria defined in the previous section.

By comparing the performance of each ablated prompt to the full prompt, the impact of each component on the full prompt can be isolated. So, where the systematic prompt engineering tested the influence of the components on the bare minimum prompt. Here, the effect on the full prompt, which is a combination of the best-performing elements, is tested. This analysis validates the systematic prompt engineering approach and identifies the most effective strategies for guiding the LLM agent.

The model used for this study is the same as the model used to test the final LLM performance. This is the OpenAI o3 model.

## 6.2. Results

An ablation study was conducted on the final, optimised prompt to determine the unique contribution of its major components to the agent's performance. Each element was individually removed, and the resulting performance was benchmarked against the baseline of the full prompt.

The full definitions and explanations of the ablations can be found in Appendix C.

The results of the ablation study can be found in Table 6.1. The concluding qualitative and quantitative performance from these results can be found in Tables 6.2 and 6.3, respectively. The statistical analysis, on which these findings are based, is found in Appendix D

Interestingly, most ablations cause the model to score higher qualitative performance scores. Only removing the instruction to behave human-like seems to have a negative effect. Meaning the model relies the most on that instruction to perform well.

The most substantial increase in performance was observed when the CoT prompting was removed. This ablation scored $5/7$ for the qualitative criteria, but only 1 collision more than the full prompt. This is interesting because introducing this in the systematic prompt engineering caused the biggest positive change.

The highest increase in collision rate was a result of removing the instruction to stay safe and avoid collisions. This means the instruction is functional for its intended purpose.

**Table 6.1:** Ablation Study Results by Iteration, using the o3 model. Each row displays the ablation version and key plots gathered during that ablation.

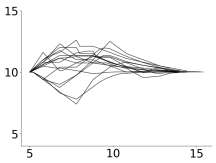
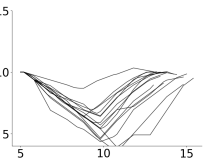
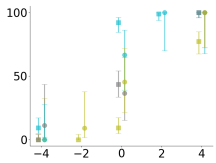
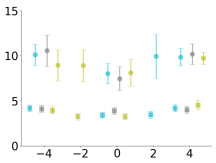
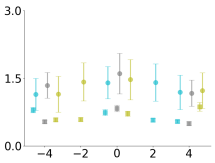
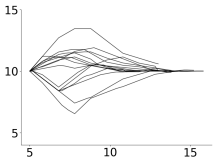
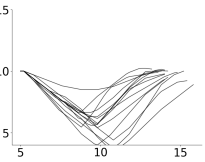
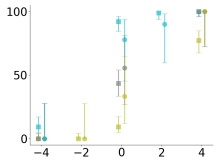
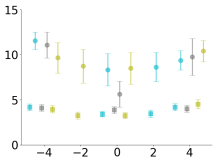
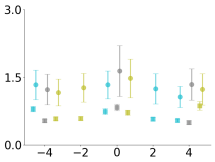
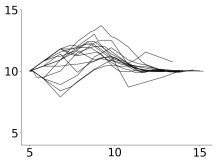
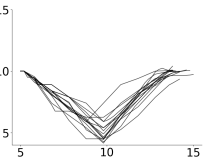
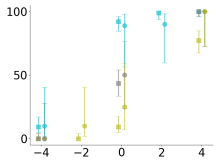
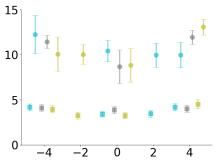
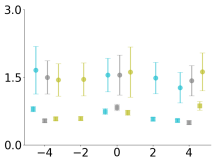
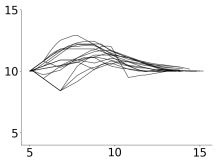
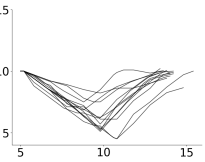
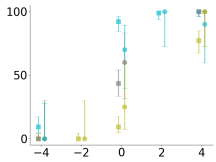
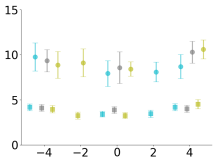
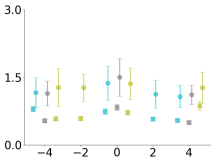
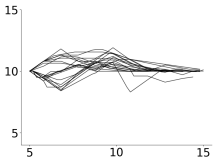
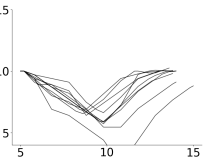
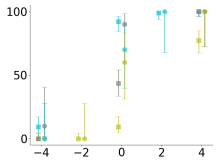
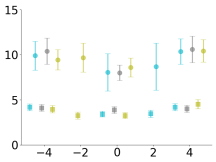
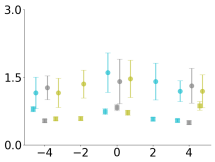
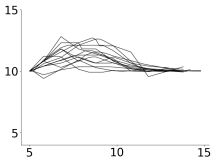
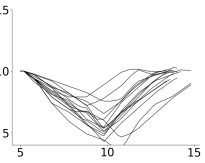
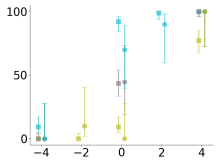
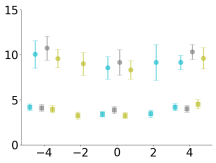
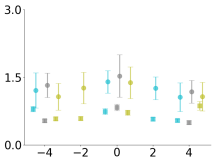
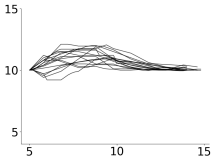
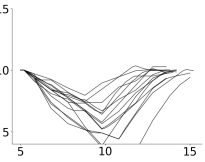
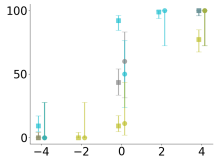
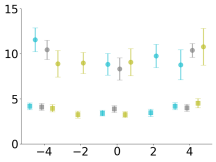
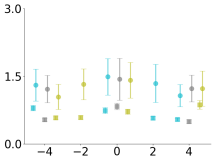
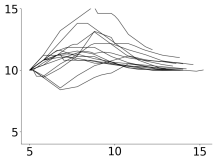
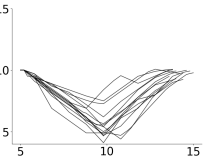
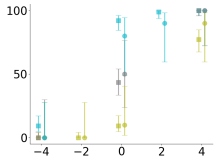
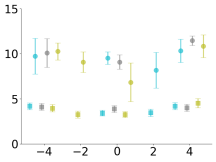
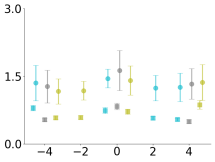
| | **Velocity plot left first** | **Velocity plot left second** | **Joint interaction** | **Gap at merge** | **Initial velocity RMSE** |
|---|---|---|---|---|---|
| *Axis (y/x):* | Vel [m/s] / Time [s] | Vel [m/s] / Time [s] | Left merging first [%] / Headway [m] | Gap at merge [m] / Headway [m] | RMSE [m/s] / Headway [m] |
| **Full Prompt** |  |  |  |  |  |
| **A1 Past Context** |  |  |  |  |  |
| **A2 Plan** |  |  |  |  |  |
| **A3a Safety** |  |  |  |  |  |
| **A3b Human** |  |  |  |  |  |
| **A3c Risk** |  |  |  |  |  |
| **A3d Reassess** |  |  |  |  |  |
| **A4 CoT** |  |  |  |  |  |

**Table 6.2:** Qualitative evaluation of the Ablation Study, using the o3 model. (✓= Significant/Observed, ✗= Not Significant/Not Observed).

| Prompt Version | Symmetrical yield | Headway to yield | Velocity to yield | Gap independent of headway | Gap dependent on velocity | Piece-wise constant control | RMSE ↑ w/ uncertainty | Performance Score |
|---|---|---|---|---|---|---|---|---|
| **Full Prompt** | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | 3/7 |
| **A1 Past context** | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | 4/7 |
| **A2 Plan** | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | 4/7 |
| **A3a Safety** | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | 5/7 |
| **A3b Human** | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | 2/7 |
| **A3c Risk** | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | 4/7 |
| **A3d Reassess** | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | 4/7 |
| **A4 CoT** | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | 5/7 |

**Table 6.3:** Quantitative performance metrics of the LLM drivers across different ablations, using the o3 model.

| Prompt Version | Collision Rate (%) | Velocity RMSE (m/s) | Average Gap (m) | Responsibility (%) |
|---|---|---|---|---|
| **Human** | 2.83 | 0.66 | 3.85 | 53.0 |
| **Full Prompt** | 0.00 | 1.33 | 9.18 | 93.6 |
| **A1 Past Context** | 2.80 | 1.30 | 9.29 | 95.3 |
| **A2 Previous Plan** | 2.80 | 1.50 | 10.63 | 100.0 |
| **A3a Safety** | 3.77 | 1.24 | 9.06 | 98.1 |
| **A3b Human Behaviour** | 1.85 | 1.32 | 9.47 | 98.2 |
| **A3c Risk** | 0.92 | 1.25 | 9.42 | 98.2 |
| **A3d Reassess** | 0.92 | 1.28 | 9.62 | 96.3 |
| **A4 CoT** | 0.92 | 1.33 | 9.57 | 99.1 |

# 7. Sensitivity Analysis

## 7.1. Methodology

The ablation study validates the prompt's design, while the sensitivity analysis addresses the model's influence. As noted in the introduction, the rapid advancement of LLMs means that results tied to a single model may have limited temporal relevance.

In this sensitivity analysis, the influence of the chosen model is analysed. It is designed to test the generalisability of our findings and the robustness of our prompt engineering strategies across different LLM architectures for future research.

This analysis involves executing the same set of prompts, specifically, the full prompt and ablated prompts, across a different large language model.

The core objective is to determine the degree to which the observed performance is model-dependent. Two main questions will be investigated:

Robustness of Prompting: Do the prompting strategies that proved effective in the ablation study have a similar positive impact when applied to other models?

Generalisability of Performance: Can a different model, using the full prompt, still satisfy the defined performance criteria and produce human-like driving behaviour?

The results from this cross-model comparison will establish whether the framework's success is due to the logic captured in the prompt design or a quirk of the specific model chosen for the main experiment.

The model used for the sensitivity analysis is Google's Gemini-2.5-pro. This model is used because, similarly to o3, it is a reasoning model. Also, because a different company produced this model. This means that the model has a different structure and dataset with different capabilities and potential biases.

## 7.2. Sensitivity Analysis Results

The results of the sensitivity analysis can be found in Table 7.1. The concluding qualitative and quantitative performance from these results can be found in Tables 7.2 and 7.3, respectively. The statistical analysis, on which these findings are based, is found in Appendix E

The analysis demonstrated that the performance was highly model-dependent. The quantitative performance was considerably worse for the Gemini model. The collision rate is too high, but the average gap kept by this model is more in line with the human data. The quantitative performance fluctuated less for each ablation with this model, but the overall performance was worse.

Opposite to the o3 model, ablating the human behaviour instruction led to the highest increase in performance and ablating the CoT prompting led to the highest decline in performance.

**Table 7.1:** Sensitivity Analysis Results by Ablation, using the Gemini-2.5-pro model. Each row displays the ablation version and key plots gathered during that ablation.

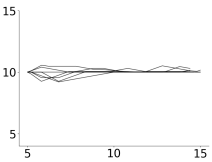
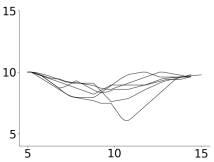
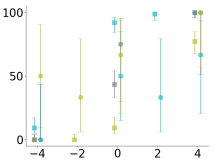
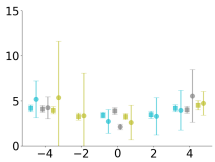
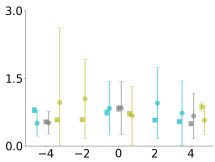
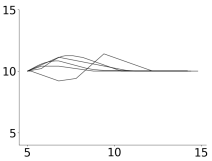
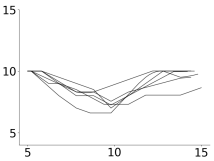
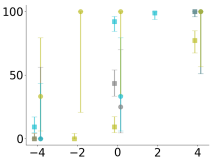
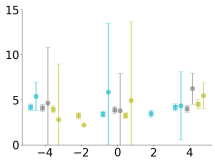
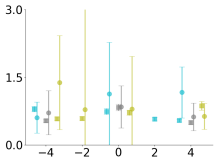
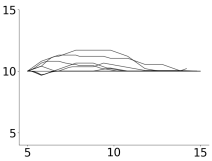
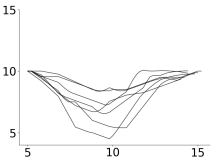
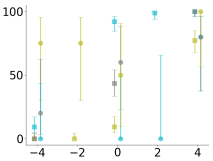
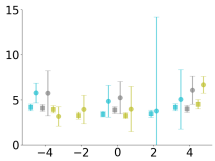
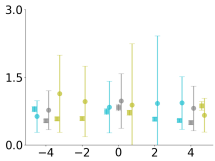
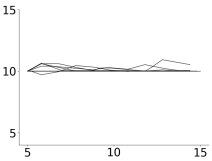
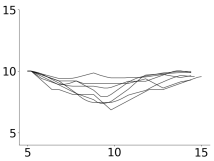
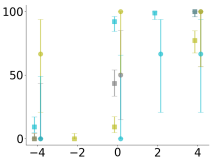
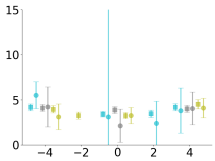
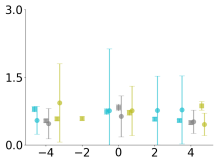
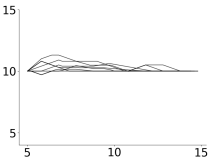
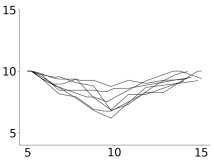
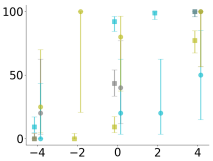
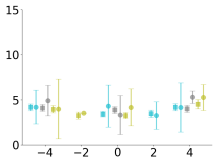
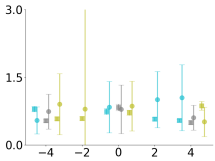
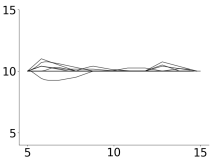
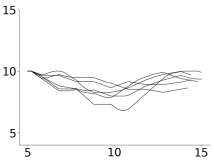
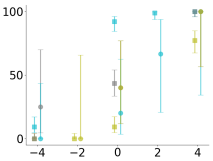
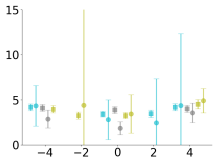
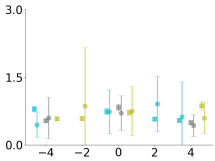
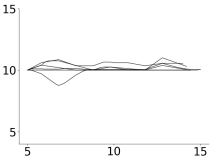
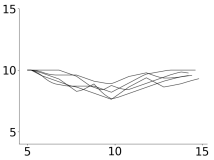
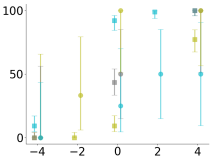
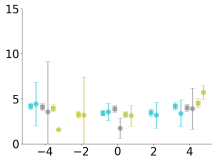
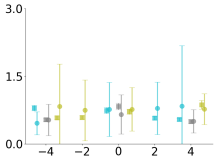
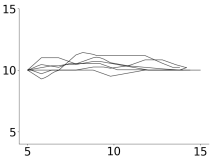
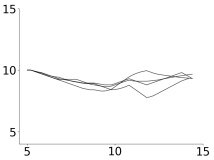
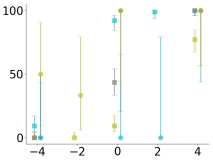
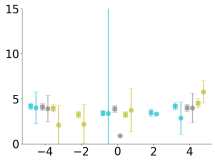
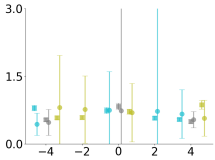
| *Axis (y/x):* | **Velocity plot left first**<br>Vel [m/s] / Time [s] | **Velocity plot left second**<br>Vel [m/s] / Time [s] | **Joint interaction**<br>Left merging first [%] / Headway [m] | **Gap at merge**<br>Gap at merge [m] / Headway [m] | **Initial velocity RMSE**<br>RMSE [m/s] / Headway [m] |
|---|---|---|---|---|---|
| **Full Prompt** |  |  |  |  |  |
| **A1 Past Context** |  |  |  |  |  |
| **A2 Previous Plan** |  |  |  |  |  |
| **A3a Safety** |  |  |  |  |  |
| **A3b Human** |  |  |  |  |  |
| **A3c Risk** |  |  |  |  |  |
| **A3d Reassess** |  |  |  |  |  |
| **A4 CoT** |  |  |  |  |  |

**Table 7.2:** Qualitative evaluation of the Sensitivity Analysis, using the Gemini-2.5-pro model. (✓= Significant/Observed, ✗= Not Significant/Not Observed).

| Ablation | Symmetrical yield | Headway to yield | Velocity to yield | Gap independent of headway | Gap dependent on velocity | Piece-wise constant control | RMSE ↑ w/ uncertainty | **Performance Score** |
|---|---|---|---|---|---|---|---|---|
| **Full Prompt** | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | 3/7 |
| **A1 Past Context** | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | 4/7 |
| **A2 Previous Plan** | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | 3/7 |
| **A3a Safety** | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | 3/7 |
| **A3b Human** | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | 3/7 |
| **A3c Risk** | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | 4/7 |
| **A3d Reassess** | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | 4/7 |
| **A4 CoT** | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | 3/7 |

**Table 7.3:** Quantitative performance metrics for the Sensitivity Analysis, using the Gemini-2.5-pro model.

| Ablation | Collision Rate (%) | Velocity RMSE (m/s) | Average Gap (m) | Responsibility (%) |
|---|---|---|---|---|
| **Full Prompt** | 25.5 | 0.76 | 3.97 | 56.1 |
| **A1 Past Context** | 38.2 | 0.79 | 4.82 | 76.5 |
| **A2 Previous Plan** | 16.4 | 0.87 | 5.14 | 47.8 |
| **A3a Safety** | 30.9 | 0.60 | 3.68 | 47.4 |
| **A3b Human** | 10.9 | 0.79 | 4.29 | 49.0 |
| **A3c Risk** | 25.5 | 0.61 | 3.44 | 41.5 |
| **A3d Reassess** | 27.3 | 0.70 | 3.58 | 52.5 |
| **A4 CoT** | 38.2 | 0.65 | 3.72 | 52.9 |

# 8. Discussion

This research explored the feasibility of using a Large Language Model (LLM) as a human driver behaviour model for a 1-D merging scenario. By systematically benchmarking the agent against human driving data, this study aimed to determine if the emergent reasoning capabilities of LLMs could replicate the nuanced tactical and control behaviours of human drivers. The findings demonstrate that while LLMs can be engineered to satisfy high-level tactical criteria, their performance is fundamentally constrained by a disparity in perception, issues of quantitative fidelity, and severe model dependency.

## 8.1. Interpretation of Results

The primary success of the LLM-driven agent lies in its ability to replicate high-level tactical decision-making. Qualitatively, the model successfully exhibited piece-wise constant acceleration profiles, resembling the decision-making structure of human drivers. Furthermore, the model satisfied the symmetry criterion in neutral conditions and correctly increased yield rates when the opposing vehicle held a headway advantage. This suggests that the reasoning capabilities of the o3 model allow it to understand social coordination rules and right-of-way concepts without explicit fine-tuning.

However, a fundamental divergence emerged in the mechanism of these decisions. The statistical analysis using Linear Mixed-Effects Regression (LMER) revealed that while human drivers rely heavily on relative velocity to make merging decisions ($p = 1.88 \times 10^{-26}$), the LLM agent's decisions were not significantly predicted by velocity ($p = 0.22$). Instead, the LLM adopted a gap-based strategy, relying almost exclusively on projected headway and absolute positions.

This difference likely stems from the nature of the input data. Humans rely on optical flow, a continuous, noisy visual perception of speed, whereas the LLM was provided with discrete, absolute numerical state values. Consequently, the LLM engaged in a calculation-heavy decision process rather than the flow-based negotiation typical of humans. This resulted in an agent that was safer but significantly more conservative. The average gap at the merge point and the velocity deviation RMSE were more than double the human benchmarks. While the agent avoided collisions effectively, it failed to capture the efficiency and fluidity of human driving.

## 8.2. Limitations

The validity of these findings is subject to several limitations regarding the simulation environment, benchmarking data, and model architecture.

A fundamental limitation concerns the ecological validity of the human benchmark dataset. Although utilised as the ground truth for this study, this data was derived from a simulated environment, which inherently introduces bias regarding risk perception. Research indicates that in naturalistic driving scenarios, drivers typically accept gaps of approximately 10 meters at velocities comparable to those in this experiment [51]. This stands in sharp contrast to the average gap of 3.85 meters observed in the human benchmark data used for this study. This discrepancy suggests that the human participants in the simulator exhibited uncharacteristically high-risk behaviour, likely due to the absence of real-world physical risk. Consequently, the ”ground truth” benchmark may not accurately reflect naturalistic driving dynamics. This implies that the LLM's average gap of 9.18 meters, while statistically deviant from the simulator data, may paradoxically align more closely with real-world human safety margins than the human benchmark itself

Second, the textual representation of the environment introduces a modality gap. Providing the LLM with exact numerical states fundamentally alters the task from perception-reaction to mathematical optimisation. A human driver does not know the exact meter distance to a merge point; they estimate time-to-collision. By feeding the LLM precise data, the experiment may have inadvertently discouraged human-like estimation errors and flow-based reactions.

A limitation of the sensitivity analysis is its scope, as it was restricted to only two distinct LLM architectures (OpenAI o3 and Gemini-2.5-pro). While the finding of severe model dependency is robust within this pair, the extent to which this inconsistency generalises across the broader LLM landscape remains unquantified. To fully characterise the potential of LLMs as human driver behaviour models, a comprehensive study spanning diverse architectures, including open-source models like Llama or varying parameter sizes of proprietary models, would be required. Nevertheless, the current findings provide a critical cautionary signal: effective prompt engineering appears to be a model-specific optimisation problem rather than a generalised solution for LLM-driven agents. This implies that 'solving' driving behaviour for one model does not guarantee transferability to another, complicat-

ing the development of a standardised LLM driver framework

## 8.3. Implications for Current Literature

These findings provide a nuanced perspective on the growing body of research advocating for LLMs as reasoning agents in autonomous driving. The results validate the premise that base LLMs possess sufficient emergent reasoning capabilities to handle complex social interactions without extensive fine-tuning.

The limitations identified in the human benchmark data compel a re-evaluation of the LLM's conservative performance. While the LLM's average gap of 9.18 meters initially appears to deviate from the human benchmark (3.85 meters), it aligns remarkably well with the 10-meter gaps observed in real-world naturalistic driving studies. This suggests that the LLM agent may possess a higher degree of ecological validity than the simulator-based human data itself. Rather than failing to capture human behaviour, the reasoning capabilities of the model allowed it to bypass the bias seen in human participants.

Crucially, this research challenges the prevailing reliance on quantitative metrics, such as collision rates and trajectory error minimisation, as the sole indicators of model fidelity. Standard quantitative evaluation often penalises models that deviate from a specific ground-truth trajectory, even if the deviation represents a valid, safe, and more human-like alternative. By prioritising qualitative behavioural criteria (e.g., piece-wise constant control, symmetrical yielding), this study demonstrated that an LLM can successfully replicate the structure of human decision-making, even if it fails to match the precise values (e.g., exact gap size) of a specific dataset. This suggests that future evaluation frameworks must expand beyond trajectory error to include structural and stylistic alignment metrics, which are better suited to capturing the reasoning quality of LLMs.

## 8.4. Future Work

To fully characterise the generalisability of LLMs as reference driver models, a broader study across more models (e.g., open-source models like Llama, or different versions/sizes of Gemini and OpenAI) would be required. Such a study could isolate whether the gap-based strategy is an artefact of specific reasoning engines or a fundamental property of text-based control.

To address the disparity between calculation-based and flow-based driving, future research should investigate the use of Vision-Language Models (VLMs) or multi-modal inputs. Instead of providing discrete numerical lists, the model could be fed visual representations or noisy, qualitative descriptions of the scene to force the model to rely on estimation rather than calculation.

Additionally, the computational cost and latency of the reasoning models (like o3) render them currently unsuitable for real-time operation. A promising avenue for future work is the development of hierarchical architectures, where the LLM handles only the high-level strategic decisions (yield vs. merge), while a low-level physics-based controller (such as an Intelligent Driver Model) executes the acceleration. This could combine the social reasoning capabilities of the LLM with the control stability of traditional models, potentially reducing the high velocity deviation observed in this study.

Finally, a critical avenue for future research lies in explicit risk quantification. This study utilised the simulation framework established by Siebinga, which analyses driving behaviour through the lens of risk perception and conflict resolution. While the current study inferred the agent's risk tolerance from its external actions, future iterations could prompt the LLM to explicitly output a "perceived risk level" or "safety confidence score" alongside its control actions. Comparing this internal "LLM risk metric" against the theoretical risk fields defined in Siebinga's risk-based driver models would offer deeper insight into the agent's cognition. It would clarify whether the conservative gaps observed in this study stem from an inflated perception of risk or a calibration error in its control response.

# 9. Conclusion

This thesis set out to determine the extent to which a base Large Language Model (LLM), guided by systematic prompt engineering, could replicate the tactical decisions and control of human drivers in a 1-D merging scenario. By benchmarking the OpenAI o3 model against human simulator data, this research successfully isolated the model's reasoning capabilities and identified fundamental divergences in how LLMs and humans approach the driving task.

The results demonstrate that a base LLM is capable of replicating high-level human tactical decision-making without the need for extensive

fine-tuning. Qualitatively, the agent satisfied key human behavioural criteria. It exhibited symmetrical yielding rates in neutral conditions, increased yielding probability when the opposing vehicle held a headway advantage, and demonstrated piecewise constant acceleration control similar to human decision structures.

However, quantitatively, the model diverged significantly from the specific human benchmark used. While the agent achieved a collision rate of 0.0% with the final prompt, it operated with a significantly higher degree of conservatism. The average gap at the merge point (9.18m) and velocity deviation were more than double that of the human simulator data. While this deviation suggests a lack of precise quantitative alignment with the simulator dataset, it paradoxically aligns closer to real-world naturalistic driving gaps than the high-risk behaviours observed in the human simulator participants.

A critical finding of this research is the fundamental difference in the underlying decision logic between the LLM and human drivers. Statistical analysis revealed that human drivers employ a flow-based strategy, relying heavily on relative velocity to negotiate merges ($p \approx 0$). In contrast, the LLM adopted a gap-based strategy, where decisions were driven by projected headway and absolute positions, rendering relative velocity statistically insignificant to its decision-making ($p = 0.22$). This suggests that providing LLMs with discrete numerical state data forces a mathematical optimisation approach rather than the perception-reaction flow inherent to human driving.

Systematic prompt engineering proved essential to achieving functional performance. The inclusion of Chain-of-Thought (CoT) reasoning and explicit memory of past states were the most critical drivers of performance. The ablation study highlighted that the instruction to ”make the behaviour human-like” was vital; removing it caused the most significant decline in qualitative performance.

However, the sensitivity analysis revealed severe model dependency. While the prompting strategy succeeded with the o3 model, transferring the same prompts to Gemini-2.5-pro resulted in a significantly higher collision rate (25.5%) and different responses to ablations. This indicates that effective prompt engineering for driver models is currently a model-specific optimisation problem rather than a generalised solution.

In conclusion, base Large Language Models possess the emergent reasoning capability to handle the social negotiation and tactical logic required for highway merging. They can understand right-of-way and cooperate to resolve conflicts safely. However, their application as direct reference models is currently limited by a modality gap; the reliance on text-based numerical inputs forces a calculation-heavy driving style that lacks the fluidity of human optical-flow perception. Future implementation must bridge this gap, potentially through multi-modal inputs or hierarchical control structures, to transition LLMs from reasoning agents to realistic driver models.

## 9.1. Acknowledgements

This project is funded by Delft University of Technology - Transport & Mobility Institute.

### 9.1.1. Use of AI Statement

This thesis utilised generative AI tools to assist in code debugging and data visualisation. Specifically, ChatGPT (OpenAI) was employed to troubleshoot Python plotting functions and optimise the mean and confidence interval functions for the plots. Additionally, Gemini (Google) was used to assist in formatting the LaTeX tables presented in the results section and the Linear Mixed-Effects Regression (LMER) syntax. All generated output was verified for accuracy by the author.

# References


[1] World-Health-Organization. *Road traffic injuries*. Dec. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[2] Nikki van Toorn (CBS). *684 road traffic deaths in 2023*. https://www.cbs.nl/en-gb/news/2024/15/684-road-traffic-deaths-in-2023. Apr. 2024.

[3] Santokh Singh. "Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey". In: *Open Access Library Journal* (Feb. 2015). URL: https://trid.trb.org/view/1346216.

[4] Silvia F. Varotto et al. "Do adaptive cruise control and lane keeping systems make the longitudinal vehicle control safer? Insights into speeding and time gaps shorter than one second from a naturalistic driving study with SAE Level 2 automation". In: *Transportation Research Part C Emerging Technologies* 141 (June 2022), p. 103756. DOI: 10.1016/j.trc.2022.103756. URL: https://doi.org/10.1016/j.trc.2022.103756.

[5] Salvatore Leonardi and Natalia Distefano. "ADAS Technologies and User Trust: An Area-Based Study with a Sociodemographic Focus". In: *Vehicles* 7.3 (July 2025), p. 67. DOI: 10.3390/vehicles7030067. URL: https://doi.org/10.3390/vehicles7030067.

[6] Fauzia Khan et al. "Safety Testing of Automated Driving Systems: A Literature review". In: *IEEE Access* 11 (Jan. 2023), pp. 120049–120072. DOI: 10.1109/access.2023.3327918. URL: https://doi.org/10.1109/access.2023.3327918.

[7] Irene Martínez. "Mitigation of Stop-and-Go Traffic Waves with Intelligent Vehicles at Low Market Penetration Rates". In: *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)* (Sept. 2023), pp. 3493–3498. DOI: 10.1109/itsc57777.2023.10422070. URL: https://doi.org/10.1109/itsc57777.2023.10422070.

[8] Cheng Wang et al. "The application of driver models in the safety assessment of autonomous vehicles: Perspectives, insights, prospects". In: *IEEE Transactions on Intelligent Vehicles* 9.1 (Nov. 2023), pp. 2364–2381. DOI: 10.1109/tiv.2023.3333796. URL: https://doi.org/10.1109/tiv.2023.3333796.

[9] Julian F. Schumann, Jens Kober, and Arkady Zgonnikov. "Benchmarking behavior prediction models in gap acceptance scenarios". In: *IEEE Transactions on Intelligent Vehicles* 8.3 (Feb. 2023), pp. 2580–2591. DOI: 10.1109/tiv.2023.3244280. URL: https://doi.org/10.1109/tiv.2023.3244280.

[10] Li Gao, Rui Zhou, and Kai Zhang. "Scenario Generation for Autonomous Vehicles with Deep-Learning-Based Heterogeneous Driver Models: Implementation and Verification". In: *Sensors* 23.9 (May 2023), p. 4570. DOI: 10.3390/s23094570. URL: https://doi.org/10.3390/s23094570.

[11] Dorsa Sadigh et al. "Planning for Autonomous Cars that Leverage Effects on Human Actions". In: *Robotics: Science and Systems* (June 2016). DOI: 10.15607/rss.2016.xii.029. URL: https://doi.org/10.15607/rss.2016.xii.029.

[12] Cheng Wang et al. "Safety assessment for autonomous vehicles: A reference driver model for highway merging scenarios". In: *Accident Analysis & Prevention* 206 (July 2024), p. 107710. DOI: 10.1016/j.aap.2024.107710. URL: https://doi.org/10.1016/j.aap.2024.107710.

[13] Natnael M. Negash and James Yang. "Driver Behavior Modeling Toward Autonomous Vehicles: Comprehensive Review". In: *IEEE Access* 11 (2023), pp. 22788–22821. DOI: 10.1109/ACCESS.2023.3249144.

[14] Shoucai Jing et al. "Cooperative Game Approach to Optimal Merging Sequence and on-Ramp Merging Control of Connected and Automated Vehicles". In: *IEEE Transactions on Intelligent Transportation Systems* 20.11 (2019), pp. 4234–4244. DOI: 10.1109/TITS.2019.2925871.

[15] Tomer Toledo. “Driving behaviour: models and challenges”. In: *Transport Reviews* 27.1 (Nov. 2006), pp. 65–84. DOI: 10.1080/01441640600823940. URL: https://doi.org/10.1080/01441640600823940.

[16] Marco Lützenberger. *A driver's mind*. IGI Global Scientific Publishing, Jan. 2019, pp. 429–454. DOI: 10.4018/978-1-5225-8356-1.ch022. URL: https://doi.org/10.4018/978-1-5225-8356-1.ch022.

[17] Xianda Chen et al. “FollowNet: a comprehensive benchmark for Car-Following behavior modeling”. In: *Scientific Data* 10.1 (Nov. 2023). DOI: 10.1038/s41597-023-02718-7. URL: https://doi.org/10.1038/s41597-023-02718-7.

[18] Sarvesh Kolekar, Joost De Winter, and David Abbink. “Human-like driving behaviour emerges from a risk-based driver model”. In: *Nature Communications* 11.1 (Sept. 2020). DOI: 10.1038/s41467-020-18353-4. URL: https://doi.org/10.1038/s41467-020-18353-4.

[19] Mohammad Hassan Mobini Seraji et al. “A state-of-the-art review on machine learning techniques for driving behavior analysis: clustering and classification approaches”. In: *Complex & Intelligent Systems* 11.9 (July 2025). DOI: 10.1007/s40747-025-01988-5. URL: https://doi.org/10.1007/s40747-025-01988-5.

[20] Stella Roussou et al. “Unfolding the dynamics of driving behavior: a machine learning analysis from Germany and Belgium”. In: *European Transport Research Review* 16.1 (July 2024). DOI: 10.1186/s12544-024-00655-z. URL: https://doi.org/10.1186/s12544-024-00655-z.

[21] Jianhua Wu et al. “Prospective role of foundation models in advancing autonomous vehicles”. In: *Research* 7 (Jan. 2024). DOI: 10.34133/research.0399. URL: https://doi.org/10.34133/research.0399.

[22] Doug Lenat and Gary Marcus. “Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc”. In: *arXiv (Cornell University)* (Jan. 2023). DOI: 10.48550/arxiv.2308.04445. URL: https://arxiv.org/abs/2308.04445.

[23] Randall C. O'Reilly, Charan Ranganath, and Jacob L. Russin. “The structure of systematicity in the brain”. In: *Current Directions in Psychological Science* 31.2 (Mar. 2022), pp. 124–130. DOI: 10.1177/09637214211049233. URL: https://doi.org/10.1177/09637214211049233.

[24] Yiheng Li et al. “WOMD-Reasoning: a Large-Scale dataset for interaction reasoning in driving”. In: *arXiv (Cornell University)* (July 2024). DOI: 10.48550/arxiv.2407.04281. URL: http://arxiv.org/abs/2407.04281.

[25] Wei Zhang et al. *Dual-AEB: synergizing Rule-Based and multimodal large language models for effective emergency braking*. Oct. 2024. URL: https://arxiv.org/abs/2410.08616.

[26] Yixuan Wang et al. “Empowering Autonomous Driving with Large Language Models: A Safety Perspective”. In: *arXiv (Cornell University)* (Nov. 2023). DOI: 10.48550/arxiv.2312.00812. URL: http://arxiv.org/abs/2312.00812.

[27] Zihao Sheng et al. “CurricuVLM: Towards Safe Autonomous Driving via Personalized Safety-Critical Curriculum Learning with Vision-Language Models”. In: *arXiv (Cornell University)* (Feb. 2025). DOI: 10.48550/arxiv.2502.15119. URL: https://doi.org/10.48550/arxiv.2502.15119.

[28] Rimvydas Rubavicius et al. “Conversational Code Generation: a Case Study of Designing a Dialogue System for Generating Driving Scenarios for Testing Autonomous Vehicles”. In: *arXiv (Cornell University)* (Oct. 2024). DOI: 10.48550/arxiv.2410.09829. URL: http://arxiv.org/abs/2410.09829.

[29] Yuan Gao, Mattia Piccinini, and Johannes Betz. “From Words to Collisions: LLM-Guided Evaluation and Adversarial Generation of Safety-Critical Driving Scenarios”. In: *arXiv (Cornell University)* (Feb. 2025). DOI: 10.48550/arxiv.2502.02145. URL: http://arxiv.org/abs/2502.02145.

[30] Lingfeng Zhou, Mohan Jiang, and Dequan Wang. *HumanSiM: Human-Like Multi-agent Novel Driving Simulation for Corner Case Generation*. Springer, Cham, Jan. 2025, pp. 287–304. DOI: 10.1007/978-3-031-91767-7\{_}20. URL: https://doi.org/10.1007/978-3-031-91767-7_20.

[31] An Guo et al. “SoVAR: Build Generalizable Scenarios from Accident Reports for Autonomous Driving Testing”. In: *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Oct. 2024), pp. 268–280. DOI: 10.1145/3691620.3695037. URL: https://doi.org/10.1145/3691620.3695037.

[32] Wenhai Wang. “DriveMLM: aligning multi-modal large language models with behavioral planning states for autonomous driving”. In: *Visual Intelligence* 3.1 (Nov. 2025). DOI: 10.1007/s44267-025-00095-w. URL: https://doi.org/10.1007/s44267-025-00095-w.

[33] Shiyu Fang et al. “Interact, Instruct to improve: A LLM-Driven Parallel Actor-Reasoner framework for enhancing autonomous vehicle interactions”. In: *arXiv (Cornell University)* (Mar. 2025). DOI: 10.48550/arxiv.2503.00502. URL: http://arxiv.org/abs/2503.00502.

[34] Shuhan Tan et al. “Promptable closed-loop traffic simulation”. In: *arXiv (Cornell University)* (Sept. 2024). DOI: 10.48550/arxiv.2409.05863. URL: http://arxiv.org/abs/2409.05863.

[35] Chiyu Max Jiang et al. “SceneDiffUser: Efficient and controllable driving simulation initialization and rollout”. In: *arXiv (Cornell University)* (Dec. 2024). DOI: 10.48550/arxiv.2412.12129. URL: http://arxiv.org/abs/2412.12129.

[36] Ziyuan Zhong et al. “Language-Guided traffic simulation via Scene-Level diffusion”. In: *arXiv (Cornell University)* (June 2023). DOI: 10.48550/arxiv.2306.06344. URL: http://arxiv.org/abs/2306.06344.

[37] Yue Zhang et al. “Siren’s Song in the AI Ocean: A survey on Hallucination in Large Language models”. In: *Computational Linguistics* (July 2025), pp. 1–46. DOI: 10.1162/coli.a.16. URL: https://doi.org/10.1162/coli.a.16.

[38] Daocheng Fu et al. “Drive Like a Human: Rethinking Autonomous Driving with Large Language Models”. In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (2024). DOI: 10.1109/WACVW60836.2024.00102.

[39] Sheng Luo et al. “Delving into Multi-Modal Multi-Task foundation models for road scene understanding: from learning Paradigm Perspectives”. In: *IEEE Transactions on Intelligent Vehicles* 9.12 (May 2024), pp. 8040–8063. DOI: 10.1109/tiv.2024.3406372. URL: https://doi.org/10.1109/tiv.2024.3406372.

[40] Ruoxuan Yang et al. “Driving Style Alignment for LLM-powered Driver Agent”. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Oct. 2024), pp. 11318–11324. DOI: 10.1109/iros58592.2024.10802629. URL: https://doi.org/10.1109/iros58592.2024.10802629.

[41] Tony Z. Zhao et al. “Calibrate Before use: Improving Few-Shot performance of language models”. In: *arXiv (Cornell University)* (Feb. 2021). DOI: 10.48550/arxiv.2102.09690. URL: http://arxiv.org/abs/2102.09690.

[42] Boyi Li et al. “Driving Everywhere with Large Language Model Policy Adaptation”. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 14948–14957. DOI: 10.1109/cvpr52733.2024.01416. URL: https://doi.org/10.1109/cvpr52733.2024.01416.

[43] Can Cui et al. “LLM4AD: Large Language Models for Autonomous Driving – Concept, review, benchmark, experiments, and future Trends”. In: *arXiv (Cornell University)* (Oct. 2024). DOI: 10.48550/arxiv.2410.15281. URL: http://arxiv.org/abs/2410.15281.

[44] Yujin Wang et al. “RAD: Retrieval-Augmented Decision-Making of Meta-Actions with Vision-Language Models in Autonomous Driving”. In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (June 2025), pp. 3838–3848. DOI: 10.1109/cvprw67362.2025.00369. URL: https://doi.org/10.1109/cvprw67362.2025.00369.

[45] Olger Siebinga, Arkady Zgonnikov, and David A Abbink. “A model of dyadic merging interactions explains human drivers’ behavior from control inputs to decisions”. In: *PNAS Nexus* 3.10 (Sept. 2024). DOI: 10.1093/pnasnexus/pgae420. URL: https://doi.org/10.1093/pnasnexus/pgae420.

[46] Olger Siebinga, Arkady Zgonnikov, and David A. Abbink. “Human merging behaviour in a coupled driving simulator: How do we resolve conflicts?” In: *IEEE Open Journal of Intelligent Transportation Systems* 5 (Jan. 2024), pp. 103–114. DOI: 10.1109/ojits.2024.3349635. URL: https://doi.org/10.1109/ojits.2024.3349635.

[47] John A. Michon. *A critical view of driver behavior models: What do we know, what should we do?* Springer, Boston, MA, Jan. 1985, pp. 485–524. DOI: 10.1007/978-1-4613-2173-6_19. URL: https://doi.org/10.1007/978-1-4613-2173-6_19.

[48] Walter W. Hauck and Allan Donner. “Wald’s test as applied to hypotheses in Logit analysis”. In: *Journal of the American Statistical Association* 72.360a (Dec. 1977), pp. 851–853. DOI: 10.1080/01621459.1977.10479969. URL: https://doi.org/10.1080/01621459.1977.10479969.

[49] Xianda Chen et al. “GenFollower: Enhancing Car-Following prediction with large language models”. In: *IEEE Transactions on Intelligent Vehicles* (Jan. 2024), pp. 1–11. DOI: 10.1109/tiv.2024.3484528. URL: https://doi.org/10.1109/tiv.2024.3484528.

[50] Jiageng Mao et al. “GPT-Driver: Learning to Drive with GPT”. In: *arXiv (Cornell University)* (Oct. 2023). DOI: 10.48550/arxiv.2310.01415. URL: http://arxiv.org/abs/2310.01415.

[51] Florian Marczak, Winnie Daamen, and Christine Buisson. “Merging behaviour: Empirical comparison between two sites and new theory development”. In: *Transportation Research Part C Emerging Technologies* 36 (Aug. 2013), pp. 530–546. DOI: 10.1016/j.trc.2013.07.007. URL: https://doi.org/10.1016/j.trc.2013.07.007.

[52] Iman Mirzadeh et al. “GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models”. In: *arXiv.org* (Oct. 2024). URL: https://arxiv.org/abs/2410.05229.

[53] Joan C. Timoneda and Sebastián Vallejo Vera. *Memory is all you need: Testing how model memory affects LLM performance in annotation tasks*. Mar. 2025. URL: https://arxiv.org/abs/2503.04874.

[54] Jason Wei et al. *Finetuned language models are Zero-Shot learners*. Sept. 2021. URL: https://arxiv.org/abs/2109.01652.

[55] Banghao Chen et al. “Unleashing the potential of prompt engineering for large language models”. In: *Patterns* 6.6 (May 2025), p. 101260. DOI: 10.1016/j.patter.2025.101260. URL: https://doi.org/10.1016/j.patter.2025.101260.

[56] Shunyu Yao et al. “Tree of Thoughts: Deliberate Problem Solving with Large Language Models”. In: *arXiv (Cornell University)* (May 2023). DOI: 10.48550/arxiv.2305.10601. URL: http://arxiv.org/abs/2305.10601.

# A

# Systematic Prompt Engineering

For this systematic prompt engineering, a baseline prompt is created. After this, it is tested against the criteria mentioned in Section *Definition of Performance Criteria*. Then, systematically, new parts to the prompt are added, which should, according to the literature, improve the performance of the model, where each iteration is tested against the same performance criteria.

Adding new parts of the prompt can be done in different categories, since not all changes tackle the same problems. The categories chosen are: observation, human behaviour, and planning.

## Baseline (v1.0)

To start, a bare minimum prompt was created. This included a basic description of the road the vehicle is driving on, the positions in Cartesian form, and the absolute velocities. Also, the needed output format is added, so the output of the model can be processed correctly in the code.

```
Your current speed: {vel} m/s

The mergepoint is {self.track._merge_point}
The shape of the road is a symmetrical inverse Y-shape, and your approach angle is {self.track._approach_angle if have_to_merge else 'straight'}

Your position: {pos}
Other vehicle's position: {self.observed_position}
Other vehicle's velocity: {self.observed_velocity}

---
Plan your behaviour in the following format:

First, explain your strategy in 2-3 sentences.

Then, output a numpy-style array with {self.memory_length} acceleration values, corresponding to {self.dt / 1000} second timesteps over {self.memory_length * self.dt / 1000} seconds.
The actions you take must be in the interval [-1, 1], where -1 means full deceleration and 1 means full acceleration.
Just the numpy array is needed; do not add anything else.
```

## Observation

First, how the LLM would perceive the scenario can be changed. This can have a considerable influence on the model, since it changes the information the model receives and can add extra context.

### v1.1: Relative observations

The first change to be made is to change to relative distances and velocity. Similarly to how a human perceives the world, the model is instructed on how far objects (e.g., other vehicles and merge points) are relative to them. Research has shown that LLMs are not good at mathematical reasoning

[52]. Precalculating these values, rather than letting the LLM do it, can thus have a positive impact on performance.

```
- The mergepoint is {self.track._merge_point}
+ The mergepoint is in {self.track._merge_point[1] - pos[1]} meters

- Your position: {pos}
- Other vehicle's position: {self.observed_position}
+ Other vehicle's relative position: {self.observed_position - pos}
- Other vehicle's velocity: {self.observed_velocity}
+ Other vehicle's relative velocity: {self.observed_velocity - vel}
```

### v1.2: Memory

Another change that can be made with respect to observation is the addition of memory. Adding memory to the prompt can have a positive impact on the performance [53]. If the model can know past states as well as its current one, it can see trends in its own behaviour and that of others.

```
+ Your previous accelerations: {self.previous_actions}
+ Planned accelerations for upcoming timestamps: {self.array_plan}
+ Previous velocities of other vehicle: {self.past_velocities}
```

### v1.3: Context

Adding context can help the zero-shot behaviour of LLMs [54]. Since an existing LLM is used without finetuning, adding context can have a positive impact.

```
+ You are a driver model that aims to emulate safe and realistic human driving
    behaviour.
+ You control acceleration values to guide a vehicle in a simulated environment.
```

## Human Behaviour

Now, the influence on the behaviour of the model to resemble the behaviour of human driving can be investigated.

### v2.1: Risk homeostasis

```
+ People tend to adjust their behaviour to maintain a personal target level of
    risk. If driving feels safer (e.g., due to better technology, smoother roads,
    or safety systems), they may unconsciously take more risks (like driving faster
     or paying less attention).
+ If conditions feel more dangerous, they usually compensate by being more
    cautious.
+ In short, drivers balance their perceived safety with their willingness to
    accept risk, so safety improvements 'dont always reduce accidents unless they
    also shift that target risk level.
```

### v2.2: Tree of thought

To enhance the decision-making capabilities of the model, a Tree of Thought approach is implemented [55, 56]. This method mimics the human cognitive process of anticipating future outcomes before acting. By explicitly instructing the model to generate multiple branches of possibility (evaluating benefits, drawbacks, and safety implications for different actions), the impulsive or "hallucinated" unsafe manoeuvres may be reduced.

```
+ Consider the branches of thought as the different actions the vehicle can take:
    For each branch:
+ *What are the benefits based on the scenario details?
+ *What are the drawbacks or potential issues?
+ *Is it safe?
+ *Would a human choose this action?
```

```
+ Evaluate each branch and sub-branch to decide the next action. What should the
    vehicle do in this situation?
+ Make a decision which aligns with experienced human-like driving behaviour.
```

### v2.3: Intermittent control

Human drivers do not strictly adhere to continuous control theory; they do not constantly micro-adjust acceleration at every millisecond. Instead, human control behaviour is characterised by intermittent control, where actions are taken in discrete bursts or corrections followed by periods of holding steady. This prompt addition instructs the model to emulate this piecewise behaviour to avoid robotic, hyper-reactive control outputs.

```
+ Humans typically do not apply smooth, continuous acceleration.
+ Instead, they make discrete decisions at moments in time, adjusting the throttle
    or brake in short bursts.
+ This creates a piecewise, intermittent control pattern where the driver
    alternates between periods of holding steady and making corrective actions,
    rather than constantly fine-tuning.
```

### v2.4: Driving rules

While mimicking human behaviour is desirable, the model must strictly adhere to safety-critical constraints. By defining hard rules, such as the 2-second following distance and strict speed limit adherence, the model is provided with a non-negotiable framework that overrides stylistic choices if safety is compromised.

```
+ Use these driving rules:
+ - Always avoid collisions
+ - To keep a safe distance, use the 2-second rule
+ - Keep driving the speed limit when safe
```

## Planning

While the previous sections focused on how the model perceives the world (Observation) and its behavioural personality (Human Behaviour), the Planning section focuses on the execution of the trajectory. These additions aim to give the model a clearer objective and a logical structure for deriving its output array.

### v3.1: Target Speed

Without a defined goal, the safest action for a driver model is often to remain stationary. To compel the vehicle to progress along the track, a target speed (or preferred velocity) is explicitly provided. This creates an optimisation gap between the current state and the desired state, motivating the model to output positive acceleration values when safe.

```
+ Your target speed: {self.preferred_velocity} m/s
```

### v3.2: Chain of thought

Large Language Models often perform better when forced to break down complex reasoning into intermediate steps, known as Chain of Thought (CoT) prompting. This iteration forces the model to sequentially process the yield decision and strategy formulation before generating the numerical acceleration array. This separation reduces the likelihood that the output will contradict the intent.

```
+ First,{" decide if you want to yield or not at the merge, given your current
    observations. Then," if have_to_merge else ""} decide if your strategy needs to
     change based on your previously planned output and current observations. Then,
     decide on your strategy to achieve your goals, taking into account your
    observations and the strategy you have developed. Try to make the behaviour
    human-like. With this action plan, derive your acceleration output for the
    coming timesteps.
```

# B

# Systematic Prompt Engineering Results

## Baseline (v1.0)

**Table B.1:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 19.76 | 19539.83 | 0.00 | 1.00 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 21.64 | 33719.37 | 0.00 | 1.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -53.45 | 125747.56 | -0.00 | 1.00 |

**Table B.2:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 1.46 | 1.48 | 0.98 | 0.33 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.10 | 0.40 | 0.24 | 0.81 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | -0.12 | 2.12 | -0.06 | 0.96 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | 0.23 | 0.58 | 0.40 | 0.69 |

**Table B.3:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 2.48 | 0.55 | 4.49 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.43 | 0.15 | -2.79 | 0.01 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -1.30 | 0.81 | -1.61 | 0.11 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.51 | 0.23 | 2.21 | 0.03 |

# Observation

## v1.1: Relative observations

**Table B.4:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.07 | 0.71 | -0.11 | 0.92 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.20 | 0.39 | 3.11 | 0.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -1.50 | 1.15 | -1.31 | 0.19 |

**Table B.5:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 3.86 | 1.58 | 2.45 | 0.01 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.36 | 0.44 | 0.81 | 0.41 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | -1.13 | 2.40 | -0.47 | 0.64 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.04 | 0.69 | -0.06 | 0.95 |

**Table B.6:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 3.22 | 0.46 | 6.95 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.08 | 0.13 | -0.60 | 0.55 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -0.18 | 0.65 | -0.28 | 0.78 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | -0.00 | 0.18 | -0.02 | 0.98 |

## v1.2: Memory

**Table B.7:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.27 | 0.47 | -0.59 | 0.56 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 0.56 | 0.18 | 3.15 | 0.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -0.85 | 0.62 | -1.37 | 0.17 |

**Table B.8:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 9.92 | 2.70 | 3.67 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | -1.29 | 0.79 | -1.63 | 0.10 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | -8.78 | 3.60 | -2.44 | 0.01 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | 1.99 | 1.07 | 1.86 | 0.06 |

**Table B.9:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 2.75 | 0.70 | 3.92 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.41 | 0.20 | -2.04 | 0.04 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -1.07 | 0.93 | -1.16 | 0.25 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.28 | 0.27 | 1.04 | 0.30 |

## v1.3: Context

**Table B.10:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.61 | 122651.86 | -0.00 | 1.00 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 12.64 | 206813.39 | 0.00 | 1.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -29.94 | 153314.83 | -0.00 | 1.00 |

# Human Behaviour

## v2.1: Risk homeostasis

**Table B.11:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -10.35 | 727437.31 | -0.00 | 1.00 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 10.13 | 687228.41 | 0.00 | 1.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -19.54 | 881719.85 | -0.00 | 1.00 |

**Table B.12:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 2.26 | 13421772.80 | 0.00 | 1.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | -0.09 | 3355443.20 | -0.00 | 1.00 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | -0.82 | 16777216.00 | -0.00 | 1.00 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | 0.57 | 4194304.00 | 0.00 | 1.00 |

## v2.2: Tree of thought

**Table B.13:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 11.71 | 60949.93 | 0.00 | 1.00 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 13.67 | 241006.41 | 0.00 | 1.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -14.64 | 76187.42 | -0.00 | 1.00 |

## v2.3: Intermittent control

*There was not enough data to do a statistical analysis for this prompt.*

## v2.4: Driving rules

**Table B.14:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -2.94 | 179092.02 | -0.00 | 1.00 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 10.90 | 22680.41 | 0.00 | 1.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -24.83 | 250282.56 | -0.00 | 1.00 |

# Planning

## v3.1: Target Speed

*There was not enough data to do a statistical analysis for this prompt.*

## v3.2: Chain of thought

**Table B.15:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.59 | 0.62 | -0.95 | 0.34 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 33.13 | $1.10 \times 10^{14}$ | 0.00 | 1.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -1.02 | 0.95 | -1.08 | 0.28 |

**Table B.16:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 2.54 | 1.15 | 2.21 | 0.03 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.55 | 0.33 | 1.65 | 0.10 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 1.49 | 1.69 | 0.88 | 0.38 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.39 | 0.51 | -0.77 | 0.44 |

**Table B.17:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.18 | 0.28 | 4.28 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.05 | 0.08 | -0.67 | 0.50 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.13 | 0.40 | 0.32 | 0.75 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.02 | 0.12 | -0.12 | 0.91 |

# C

# Ablation Study

To find out what the influence of each part of the prompt is on the final result, an ablation study is done. Each element of the ablation study is removed from the final prompt to see how the model performs differently. Due to budget and environmental reasons, the ablations were not combined to see the combined influence of the ablations. The final prompt can be seen below

```
Environment details:
-Preferred velocity: {self.preferred_velocity} m/s
-Road: {"Two equal-priority roads merging into a single lane (inverse Y-shape). You are on one branch, the other vehicle is on the other (both in the same direction)" if have_to_merge else "straight road"}

Vehicle length: {self.vehicle_length} m
You see the other vehicle {headway} meters {other_pos} (centre to centre, so without taking into account vehicle length){"(this headway is relative, but the other car is still on the other road)" if have_to_merge else ""}.

You are driving {vel} m/s
The other vehicle is driving {self.observed_velocity} m/s
{"You see the mergepoint coming up in "+ str(self.track._merge_point[1]-pos) + " meters" if have_to_merge else ""}

Past context (each timestep is {self.dt/1000} seconds) of the last 2 seconds:
-Your previous accelerations: {self.previous_actions[int(-2/self.dt*1000):]}
-Past observed velocities of the other vehicle: {self.past_velocities[int(-2/self.dt*1000):]}
-Your previously planned accelerations: {self.array_plan}
Use this information to stay consistent in your driving strategy and anticipate the actions of the other vehicle.

Important:
-Always avoid collisions and drive safely
-Try to make the behaviour human-like
-Keep an appropriate distance and try to take no risks unless the situation demands it
-Always reassess the situation with the other vehicle and adjust your behaviour as needed

First,{" decide if you want to yield or not at the merge, given your current observations. Then," if have_to_merge else ""} decide if your strategy needs to change based on your previously planned output and current observations. Then, decide on your strategy to achieve your goals, taking into account your observations and the strategy you have developed. With this action plan, derive your acceleration output for the coming timesteps.
---
Plan your behaviour in the following format:

{'<Print "M" or "Y" depending on whether you want to merge first or yield, respectively. Just print the letter and nothing else>' if have_to_merge else ""}

<2-3 sentences explaining what you believe the other vehicle will do and your strategy.>

```

```
<Output a numpy array in a python code block with {self.memory_length} acceleration values,
    corresponding to {self.dt / 1000} second timesteps over {self.memory_length * self.dt /
    1000} seconds. The actions you take are normalised and must be in the interval [-1, 1],
    where -1 means full deceleration and 1 means full acceleration. Full acceleration and
    deceleration correspond with 2.5 and -2.5 m/s^2, respectively. In the code block you can
    only put the array and nothing else (no imports or comments)>
```

## Baseline

Some parts of the prompts are deemed necessary to create a valid response and logical output. So these parts will not be altered during the ablation study. The output format needed to process the response. The models needs the environment details to get a general idea of the scenario. To prevent collisions, the model needs the distance to the other vehicle and the length of the vehicles. To be able to control its speed, it needs to know how fast it is driving. And, to react correctly and quickly enough, it needs the velocity of the other vehicle and the distance to the merge point. The parts used for the ablation are shown below.

```
Past context (each timestep is {self.dt/1000} seconds) of the last 2 seconds:
-Your previous accelerations: {self.previous_actions[int(-2/self.dt*1000):]}
-Past observed velocities of the other vehicle: {self.past_velocities[int(-2/self.dt*1000):]}
-Your previously planned accelerations: {self.array_plan}
Use this information to stay consistent in your driving strategy and anticipate the actions
    of the other vehicle.

Important:
-Always avoid collisions and drive safely
-Try to make the behaviour human-like
-Keep an appropriate distance and try to take no risks unless the situation demands it
-Always reassess the situation with the other vehicle and adjust your behaviour as needed

First,{" decide if you want to yield or not at the merge, given your current observations.
    Then," if have_to_merge else ""} decide if your strategy needs to change based on your
    previously planned output and current observations. Then, decide on your strategy to
    achieve your goals, taking into account your observations and the strategy you have
    developed. With this action plan, derive your acceleration output for the coming
    timesteps.
```

## Ablation 1: Past Context

The first ablation was to remove the past accelerations and velocities. This part gives information on past states of the vehicle and the other vehicle. By removing this, the model does not know past states and might behave differently by taking less risk because of a lack of information or an increase in collisions. Below are the alterations to the final prompt can be seen.

```
- Past context (each timestep is {self.dt /1000} seconds) of the last 2 seconds:
- -Your previous accelerations: {self.previous_actions[int(-2/self.dt *1000) :]}
- -Past observed velocities of the other vehicle: {self.past_velocities[int(-2/
    self.dt *1000) :]}
- Use this information to stay consistent in your driving strategy and anticipate
    the actions of the other vehicle.

+ Use this information to stay consistent in your driving strategy
```

## Ablation 2: Previous Plan

The previously planned accelerations is a separate ablation, because it does not tell the model about past states, but the upcoming time steps. Removing this part can lead to an increase in tactical behaviour changes or a decrease in constant piecewise acceleration control. The changes from the final prompt can be seen in the block below.

```
- -Your previously planned accelerations: {self.array_plan}
```

## Ablation 3: Behavioural Guidelines

Next, the important customs are removed one at a time to see how well the model performs without them, or whether the output remains unchanged.

### Ablation 3a: Safety

```
- -Always avoid collisions and drive safely
```

### Ablation 3b: Human behaviour

```
- -Try to make the behaviour human-like
```

### Ablation 3c: Distance and Risk

```
- -Keep an appropriate distance and try to take no risks unless the situation
    demands it
```

### Ablation 3d: Reassessment

```
- -Always reassess the situation with the other vehicle and adjust your behaviour
    as needed
```

## Ablation 4: Chain-of-Thought (CoT) Reasoning

Lastly, the tactical decision strategy is removed. Since this part is a form of CoT reasoning, it can have a significant effect on the performance of the model.

```
- First,{" decide if you want to yield or not at the merge, given your current
    observations. Then," if have_to_merge else ""} decide if your strategy needs to
     change based on your previously planned output and current observations. Then,
     decide on your strategy to achieve your goals, taking into account your
    observations and the strategy you have developed. With this action plan, derive
     your acceleration output for the coming timesteps.
```

# D

# Ablation Study Results

## Ablation 1: Past Context

**Table D.1:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 0.07 | 0.38 | 0.19 | 0.85 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.42 | 0.41 | 3.44 | $5.88 \times 10^{-04}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -0.99 | 0.59 | -1.68 | 0.09 |

**Table D.2:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 5.63 | 0.82 | 6.88 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 1.20 | 0.25 | 4.83 | 0.00 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 3.16 | 1.21 | 2.61 | 0.01 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.88 | 0.38 | -2.33 | 0.02 |

**Table D.3:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.64 | 0.17 | 9.51 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.09 | 0.05 | -1.71 | 0.09 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -0.32 | 0.26 | -1.24 | 0.22 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.05 | 0.08 | 0.63 | 0.53 |

# Ablation 2: Previous Plan

**Table D.4:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 0.29 | 0.36 | 0.79 | 0.43 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.01 | 0.20 | 5.13 | $2.94 \times 10^{-07}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -1.43 | 0.57 | -2.52 | 0.01 |

**Table D.5:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 8.67 | 1.13 | 7.65 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.75 | 0.35 | 2.16 | 0.03 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 0.98 | 1.69 | 0.58 | 0.56 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.38 | 0.53 | -0.71 | 0.48 |

**Table D.6:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.55 | 0.19 | 8.24 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.02 | 0.06 | -0.39 | 0.70 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.00 | 0.28 | 0.00 | 1.00 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.01 | 0.09 | 0.09 | 0.93 |

# Ablation 3: Behavioural Guidelines

## Ablation 3a: Safety

**Table D.7:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.04 | 0.36 | -0.12 | 0.90 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.19 | 0.28 | 4.23 | $2.32 \times 10^{-05}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -1.02 | 0.56 | -1.81 | 0.07 |

**Table D.8:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 8.56 | 0.71 | 12.12 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.32 | 0.22 | 1.45 | 0.15 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | -0.63 | 1.06 | -0.59 | 0.55 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | 0.04 | 0.33 | 0.11 | 0.91 |

**Table D.9:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.50 | 0.15 | 9.75 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.09 | 0.05 | -1.97 | 0.05 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -0.21 | 0.23 | -0.93 | 0.35 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.07 | 0.07 | 0.96 | 0.34 |

## Ablation 3b: Human behaviour

**Table D.10:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 1.01 | 0.40 | 2.52 | 0.01 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.35 | 0.32 | 4.23 | $2.35 \times 10^{-05}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -0.63 | 0.57 | -1.10 | 0.27 |

**Table D.11:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 8.00 | 0.65 | 12.33 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.62 | 0.20 | 3.13 | 0.00 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 0.44 | 0.96 | 0.45 | 0.65 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.25 | 0.30 | -0.83 | 0.41 |

**Table D.12:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.41 | 0.17 | 8.25 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.03 | 0.05 | -0.56 | 0.58 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.17 | 0.25 | 0.66 | 0.51 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | -0.08 | 0.08 | -0.97 | 0.33 |

## Ablation 3c: Distance and Risk

**Table D.13:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.51 | 0.39 | -1.28 | 0.20 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.24 | 0.28 | 4.42 | $1.00 \times 10^{-05}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -1.49 | 0.59 | -2.50 | 0.01 |

**Table D.14:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 9.16 | 0.57 | 15.98 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.34 | 0.17 | 1.97 | 0.05 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | -0.88 | 0.84 | -1.05 | 0.29 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.06 | 0.26 | -0.24 | 0.81 |

**Table D.15:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.53 | 0.16 | 9.61 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.07 | 0.05 | -1.45 | 0.15 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -0.17 | 0.23 | -0.71 | 0.47 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | -0.00 | 0.07 | -0.06 | 0.96 |

## Ablation 3d: Reassessment

**Table D.16:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.42 | 0.40 | -1.04 | 0.30 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 16.90 | 3594563.26 | 0.00 | 1.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -1.03 | 0.64 | -1.62 | 0.10 |

**Table D.17:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 8.31 | 0.93 | 8.91 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.52 | 0.29 | 1.83 | 0.07 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 0.74 | 1.39 | 0.53 | 0.59 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.32 | 0.44 | -0.72 | 0.47 |

**Table D.18:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.44 | 0.17 | 8.52 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.05 | 0.05 | -1.02 | 0.30 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.03 | 0.25 | 0.14 | 0.89 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | -0.03 | 0.08 | -0.34 | 0.74 |

# Ablation 4: Chain-of-Thought (CoT) Reasoning

**Table D.19:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.44 | 0.39 | -1.11 | 0.27 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.15 | 0.23 | 4.92 | $8.86 \times 10^{-07}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | -2.00 | 0.64 | -3.10 | 0.00 |

**Table D.20:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 9.08 | 0.93 | 9.81 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.43 | 0.28 | 1.50 | 0.14 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | -1.41 | 1.37 | -1.02 | 0.31 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | 0.16 | 0.43 | 0.38 | 0.71 |

**Table D.21:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 1.63 | 0.15 | 10.52 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.08 | 0.05 | -1.69 | 0.09 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -0.32 | 0.23 | -1.37 | 0.17 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.06 | 0.07 | 0.88 | 0.38 |

# E

# Sensitivity Analysis

## Baseline

**Table E.1:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 0.08 | 0.30 | 0.26 | 0.79 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 0.50 | 0.13 | 3.97 | $7.23 \times 10^{-05}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | 2.12 | 0.53 | 3.97 | $7.08 \times 10^{-05}$ |

**Table E.2:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 2.12 | 0.59 | 3.60 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.68 | 0.18 | 3.82 | 0.00 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 0.53 | 0.89 | 0.59 | 0.56 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.16 | 0.28 | -0.57 | 0.57 |

**Table E.3:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 0.84 | 0.22 | 3.80 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.06 | 0.07 | -0.96 | 0.34 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.03 | 0.34 | 0.09 | 0.93 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.02 | 0.11 | 0.21 | 0.83 |

# Ablation 1: Past Context

**Table E.4:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 0.56 | 0.64 | 0.88 | 0.38 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 0.99 | 0.36 | 2.78 | 0.01 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | 2.67 | 1.39 | 1.91 | 0.06 |

**Table E.5:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 3.81 | 1.23 | 3.10 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.43 | 0.38 | 1.12 | 0.26 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 1.69 | 1.99 | 0.85 | 0.39 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.73 | 0.60 | -1.21 | 0.23 |

**Table E.6:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 0.84 | 0.28 | 3.03 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.04 | 0.09 | -0.52 | 0.61 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.16 | 0.45 | 0.36 | 0.72 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.03 | 0.14 | 0.23 | 0.82 |

# Ablation 2: Previous Plan

**Table E.7:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 0.19 | 0.39 | 0.50 | 0.62 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 0.46 | 0.16 | 2.83 | 0.00 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | 2.38 | 0.77 | 3.09 | 0.00 |

**Table E.8:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 5.25 | 1.07 | 4.91 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.16 | 0.33 | 0.50 | 0.62 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | -1.40 | 1.65 | -0.85 | 0.40 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | 0.09 | 0.51 | 0.17 | 0.86 |

**Table E.9:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 0.98 | 0.23 | 4.18 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.05 | 0.07 | -0.64 | 0.52 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | -0.10 | 0.38 | -0.26 | 0.80 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.04 | 0.17 | 0.23 | 0.81 |

# Ablation 3: Behavioural Guidelines

## Ablation 3a: Safety

**Table E.10:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 0.75 | 0.60 | 1.24 | 0.21 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 0.97 | 0.38 | 2.58 | 0.01 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | 4.16 | 1.70 | 2.45 | 0.01 |

**Table E.11:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 2.14 | 0.83 | 2.60 | 0.01 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.50 | 0.25 | 2.01 | 0.04 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 0.98 | 1.28 | 0.76 | 0.45 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.27 | 0.39 | -0.68 | 0.49 |

**Table E.12:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 0.64 | 0.20 | 3.24 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.04 | 0.06 | -0.61 | 0.54 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.17 | 0.31 | 0.56 | 0.58 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.01 | 0.10 | 0.05 | 0.96 |

## Ablation 3b: Human behaviour

**Table E.13:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.17 | 0.28 | -0.61 | 0.54 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 0.61 | 0.13 | 4.55 | $5.32 \times 10^{-06}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | 2.24 | 0.55 | 4.10 | $4.19 \times 10^{-05}$ |

**Table E.14:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 3.32 | 0.66 | 5.03 | 0.00 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.45 | 0.20 | 2.21 | 0.03 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 0.84 | 0.99 | 0.85 | 0.40 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.46 | 0.31 | -1.48 | 0.14 |

**Table E.15:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 0.79 | 0.26 | 3.10 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.03 | 0.08 | -0.38 | 0.71 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.13 | 0.38 | 0.34 | 0.74 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | -0.01 | 0.12 | -0.04 | 0.96 |

## Ablation 3c: Distance and Risk

**Table E.16:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.20 | 0.33 | -0.61 | 0.54 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 0.83 | 0.19 | 4.29 | $1.76 \times 10^{-05}$ |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | 1.34 | 0.54 | 2.48 | 0.01 |

**Table E.17:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 1.87 | 0.86 | 2.17 | 0.03 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.34 | 0.27 | 1.28 | 0.20 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 1.40 | 1.29 | 1.08 | 0.28 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | 0.03 | 0.42 | 0.08 | 0.93 |

**Table E.18:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 0.71 | 0.17 | 4.05 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.05 | 0.05 | -0.92 | 0.36 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.11 | 0.26 | 0.42 | 0.67 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | -0.01 | 0.09 | -0.06 | 0.95 |

## Ablation 3d: Reassessment

**Table E.19:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | -0.11 | 0.48 | -0.23 | 0.82 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 0.94 | 0.34 | 2.80 | 0.01 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | 2.37 | 0.99 | 2.40 | 0.02 |

**Table E.20:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 1.75 | 1.08 | 1.62 | 0.11 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.50 | 0.33 | 1.49 | 0.14 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 1.79 | 1.61 | 1.11 | 0.27 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.39 | 0.51 | -0.77 | 0.44 |

**Table E.21:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 0.66 | 0.21 | 3.19 | 0.00 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.04 | 0.06 | -0.54 | 0.59 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.16 | 0.31 | 0.52 | 0.60 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.02 | 0.10 | 0.16 | 0.87 |

# Ablation 4: Chain-of-Thought (CoT) Reasoning

**Table E.22:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on which driver merged first for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | -0.30 | 0.13 | -2.26 | 0.02 |
| Intercept - LLM | 0.21 | 0.73 | 0.29 | 0.77 |
| Projected headway - H | 1.11 | 0.07 | 14.89 | $3.57 \times 10^{-50}$ |
| Projected headway - LLM | 1.23 | 0.50 | 2.46 | 0.01 |
| Relative velocity - H | -3.30 | 0.31 | -10.64 | $1.88 \times 10^{-26}$ |
| Relative velocity - LLM | 3.74 | 1.79 | 2.08 | 0.04 |

**Table E.23:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the gap at merge for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 3.88 | 0.26 | 15.25 | 0.00 |
| Intercept - LLM | 0.91 | 1.52 | 0.60 | 0.55 |
| Abs. projected headway - H | 0.05 | 0.08 | 0.59 | 0.55 |
| Abs. projected headway - LLM | 0.76 | 0.43 | 1.75 | 0.08 |
| Abs. relative velocity - H | -0.85 | 0.38 | -2.27 | 0.02 |
| Abs. relative velocity - LLM | 2.90 | 2.16 | 1.34 | 0.18 |
| Headway:velocity Interaction - H | 0.24 | 0.12 | 2.01 | 0.04 |
| Headway:velocity Interaction - LLM | -0.82 | 0.64 | -1.29 | 0.20 |

**Table E.24:** Mixed-effects logistic regression models describing the effect of projected headway and relative velocity on the RMSE velocity deviation from the initial velocity for the human and LLM.

| Feature | Estimate | SE | Z | P-value |
|---|---|---|---|---|
| Intercept - H | 0.83 | 0.12 | 6.87 | 0.00 |
| Intercept - LLM | 0.73 | 0.35 | 2.08 | 0.04 |
| Abs. projected headway - H | -0.08 | 0.04 | -2.11 | 0.04 |
| Abs. projected headway - LLM | -0.06 | 0.09 | -0.59 | 0.56 |
| Abs. relative velocity - H | -0.19 | 0.18 | -1.04 | 0.30 |
| Abs. relative velocity - LLM | 0.02 | 0.48 | 0.04 | 0.97 |
| Headway:velocity Interaction - H | 0.10 | 0.06 | 1.68 | 0.09 |
| Headway:velocity Interaction - LLM | 0.02 | 0.13 | 0.15 | 0.88 |