# comBOT: an ensemble combination model combining results from SMART-tiny-CLSFT with a Cognitive Behavior Model

**Christian Rössert** **Johannes Drever** **Lukas Brostek**

cogniBIT GmbH
Agnes-Pockels-Bogen 1
80992 München, Germany
info@cognibit.de

June 10, 2025

**Abstract**

Transformer-based trajectory predictors such as *SMART(-tiny)* and its fine-tuned variant *SMART-tiny-CLSFT* reproduce well represented routine traffic with high realism, yet they degrade in sparsely represented high-speed highway scenes. In the Waymo Open Dataset, fewer than one percent of validation and test scenarios involve speed limits above 65mph; in precisely these situations both transformers drift beyond lane boundaries and even leave the roadway. We therefore pair them with *cogniBOT* - a mechanistic cognitive architecture model - in a simple scene-based ensemble dubbed *comBOT*. Whenever all actors in a scene occupy lanes faster than 65mph, comBOT delegates the rollout to cogniBOT; otherwise it relies on SMART. This hybrid reduces off-road and collision errors, illustrating how mechanistic cognitive models can systematically correct the longtail problem in large data-driven models.

## 1 Introduction

A currently popular approach to model human behavior in traffic is to implement purely data-driven models using machine learning techniques, such as transformer networks. These methods can reproduce traffic behavior in a realistic way, however, due to the complexity of human behavior in traffic, a very large amount of data is required to train the networks. While large amounts of traffic data for standard driving situations are available, it is questionable whether sufficiently large amounts of data will be available for all driving scenarios, especially critical driving situations such as accidents and near-accidents, as these kinds of situations are extremely rare. However, valid simulation of critical driving situations is of the utmost relevance for safeguarding autonomous vehicles.

To understand these data-driven models better we analyzed two prominent examples of transformer models used in the Waymo Open Sim Agents Challenge, SMART (Wu et al. 2024) and SMART-tiny-CLSFT (Zhang et al. 2024), which combines SMART with a CAT-K rollout fine-tuning strategy. We compared both models to our mechanistic cogniBOT cognitive architecture model which implements general principles of human sensorimotor information processing tuned to fit behavioral data (Brostek et al. 2024).

Unexpectedly, we found that both SMART models showed large lane overshoots and offroad driving (Figure 1, B) for highway scenarios with speed limits of 65 mph and larger. These scenarios are rather sparsely represented in the dataset with about only 1% in the validation and testing dataset. Our cogniBOT model, on the contrary, is working with an intrinsic representation of lane boundaries and capable of keeping the lane also at higher speed (Figure 1, A).

Within the 'comBOT' project we tested whether the combination of a SMART model with our cogniBOT could improve the overall metrics performance in the Sim Agents Challenge.

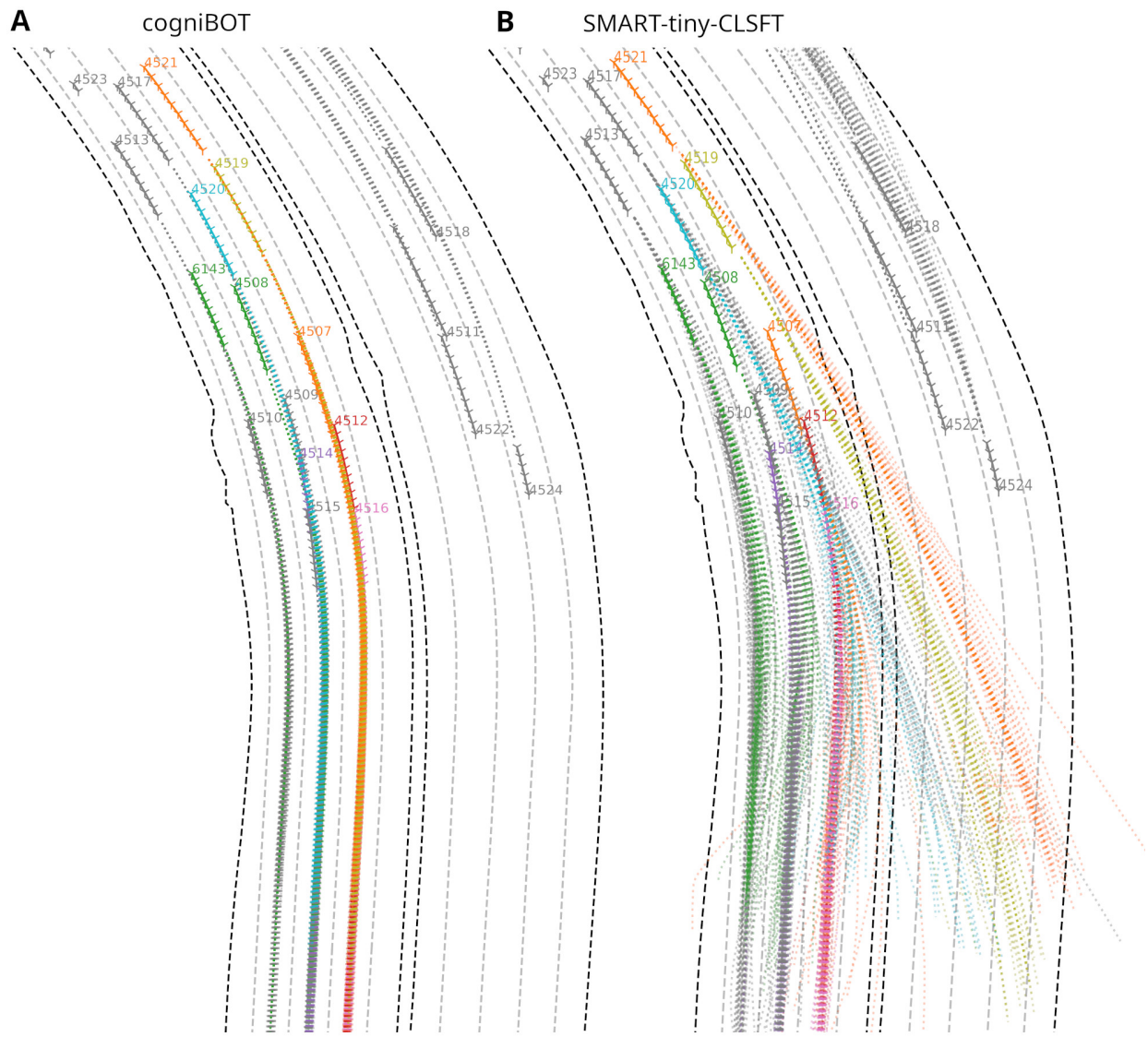


Figure 1: Comparison between trajectories of cogniBOT (A) and SMART-tiny-CLSFT (B) for scenario a428e7a7676c2b52. Dashed gray and black lines represent lane and road edges, respectively. Stars represent the positions of sim agents during the 1 second history. Dotted lines represent the 32 8-second rollouts. Colored lines represent sim agents that are directly evaluated, gray agents are not directly evaluated.

## 2 Methods

Both SMART implementations are publicly available. For SMART we trained the weights ourselves for 13 epochs. For SMART-tiny-CLSFT a version of the weights (dated 05.02.25), graciously provided by Zhang et al. 2024, was used.

In both cases the ensemble model 'comBOT' was constructed in the following way: If all agents in a scene were placed on lanes with speed limits larger than 29 m/s it was classified as a highway scenario and the cogniBOT model was used for the whole scenario. In all other scenarios SMART-tiny-CLSFT was used. This was done for testing and validation datasets. We also combined cogniBOT with the classic SMART model for the validation set only (identified as comBOT (SMART)).

The cogniBOT model integrates the strengths of rule-based and purely data-driven techniques to model road user behavior. Our core idea is to embed established knowledge of human information processing within a cognitive architecture. This architecture is composed of distinct yet interconnected models for perceptive, cognitive, and motor sub-processes. By explicitly modeling these known sensorimotor and cognitive stages, we drastically reduce the

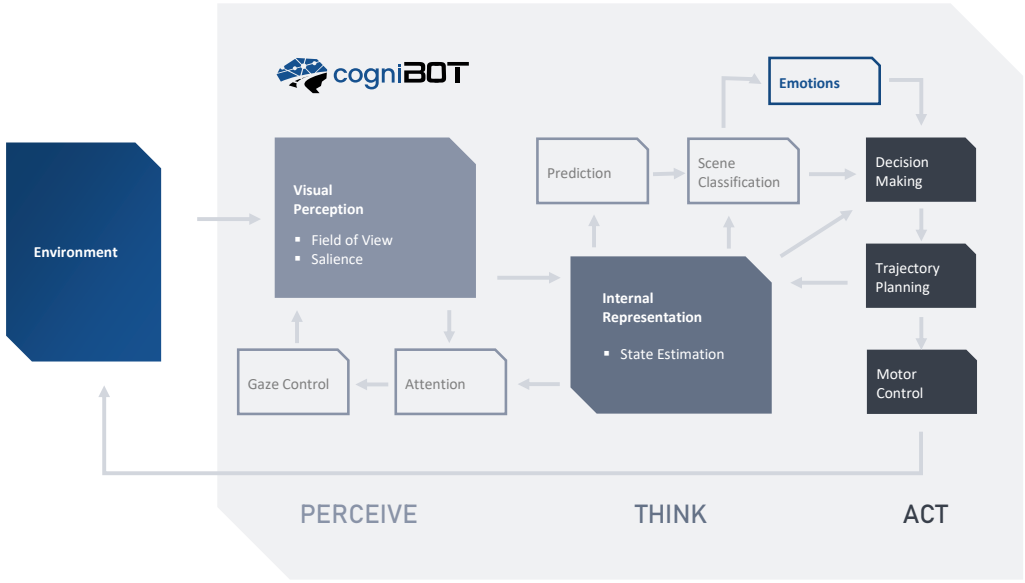


Figure 2: The cogniBOT system architecture

number of free model parameters. This, in turn, lessens the dependency on large training datasets and enables our models to generalize from common traffic situations to rare and critical events. Our behavior models are built upon this neuro-cognitive framework, which explicitly simulates the sequential sub-processes transforming sensory input into situation-aware actions. As depicted in Figure 2, these processes are organized into three primary functional groups.

The first functional group, visual perception, simulates the intake of environmental information by incorporating key limitations of human vision. For example, the agent has a restricted field of view, which it compensates for with simulated eye movements. These eye movements are governed by a sophisticated attention mechanism that balances top-down influences (e.g., the driver's current intent) and bottom-up signals (e.g., the sudden appearance of a vehicle in the periphery). To realistically model the inherent inaccuracies of human perception, we introduce stochasticity. For instance, noise is added to processed signals to simulate the natural decrease in visual acuity in peripheral vision, ensuring the model captures a range of perceptual successes and failures.

Following perception, the cognitive processing modules interpret this information. The agent first constructs an internal world model from the perceptual data - a representation containing only the objects and information it has actually perceived and recognized. Explicitly modeling this stage allows us to capture characteristic human misjudgments, such as the common overestimation of distance to vehicles in a rear-view mirror.

Drawing upon this internal model, the agent anticipates the future development of the traffic situation, predicting the behavior of all other road users from its unique assessment. Another stage in cognitive processing where stochastic variation and 'errors' frequently occur are cognitive judgments and predictions of all kinds, such as when judging the suitability of a gap when performing a lane change maneuver.

These predictions then form the basis for the agent's decision-making, which is guided by a cost function that evaluates potential actions by weighing competing goals, such as maintaining speed, ensuring a safe distance, and minimizing accident risk. To account for the full spectrum of human driving behavior, the model also incorporates emotional states, which allows us to simulate a wide range of driver archetypes, from relaxed "cruisers" to aggressive "speeders."

The final functional group is motor control, which simulates the physical execution of the chosen action (e.g., steering, braking). This component is crucial for representing individual differences in motor performance, thereby capturing characteristic behavioral variations between, for example, older and younger drivers.

The implementation of these modules leverages a range of technical approaches from theoretical neuroscience, control engineering, and machine learning. For specific tasks like scene classification or trajectory prediction, we employ

compact neural networks (e.g., MLPs, LSTMs). In contrast to monolithic systems, we favor smaller models with approximately 1,000 parameters, which significantly reduces the complexity and allows them to be trained effectively on synthetically generated data.

Consequently, the number of free parameters in the complete cogniBOT model is on the order of 100, substantially lower than in purely data-driven approaches. The parameterization process leverages existing scientific knowledge: physiological parameters, such as the field of view size or eye movement dynamics, are determined from established findings in the academic literature, while behavioral parameters defining decision-making, such as preferred following distance (see Fig. 3), are calibrated using data from traffic science research.

It is crucial to note that while the baseline SMART models were trained on the training dataset, the cogniBOT model was not fitted to any data from the Waymo Open Dataset. Its parameters are based entirely on the external knowledge sources described above (Brostek et al. 2024).

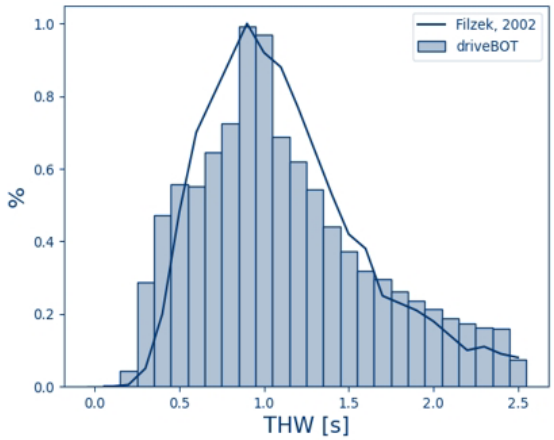


Figure 3: Distributions of time headway observed on German Autobahn (from Filzek 2003) and simulated by driveBOT v1.3 (the car driver model based on the cogniBOT system architecture)

## 3 Results and Discussion

In the following we compare the metrics of the comBOT and comBOT (SMART) models to the SMART-tiny-CLSFT, SMART and cogniBOT models. We start our comparison with the validation set (Table 1). Here we notice that both comBOT models show a slight increase in realism meta-metric over the respective SMART models. This increase is in both cases mostly accounted for by the improvement in offroad- and collision-likelihood. Both of these metrics are greatly affected by trajectories that disregard lane boundaries.

When comparing comBOT to the official testing set (Table 2) performance of the latest SMART-tiny-CLSFT model (Sim Agent Challenge Leaderboard, submission 11.04.2025), we see that there still is an improvement in offroad- and collision- likelihood, however this does not carry forward to the overall realism meta-metric. Since cogniBOT itself is a general model and not trained to the specific kinematics of the Waymo dataset, it shows a lower performance especially in these aspects.

Even though the performance of comBOT is not able to outperform the results of SMART-tiny-CLSFT in the testing dataset, it clearly shows a possible disadvantage of purely data-driven models. When a specific scenario type is underrepresented in the dataset, this type of algorithm may produce highly unrealistic behavior for these scenarios. These kind of overtraining effects are most likely to occur not only in the high-speed scenarios focused on in this study, but must also be expected for other underrepresented types of traffic scenarios in the Waymo dataset,

The results of our comBOT study highlight how a hybrid approach can be used to advance the modeling of human behavior in traffic. Further research will analyze how purely data-driven methods can be combined with mechanistic models of cognitive processes to capture behavior in standard and underrepresented driving scenarios.

| Metric name | Validation | | | |
|---|---|---|---|---|
| | **SMART-tiny-CLSFT** (05.02.25) | **comBOT** | **SMART** (13 epochs) | **comBOT** (SMART) |
| Realism meta-metric | 0.7845 | **0.7847** | 0.7571 | **0.7574** |
| Linear Speed Likelihood | **0.3858** | 0.3840 | **0.3362** | 0.3351 |
| Linear Acceleration Likelihood | **0.4064** | 0.4054 | 0.2740 | **0.2749** |
| Angular Speed Likelihood | **0.5182** | 0.5177 | 0.4691 | **0.4692** |
| Angular Acceleration Likelihood | **0.6596** | 0.6587 | **0.6137** | 0.6133 |
| Distance To Nearest Object Likelihood | **0.3912** | 0.3911 | 0.3652 | 0.3652 |
| Collision Likelihood | 0.9702 | **0.9706** | 0.9534 | **0.9538** |
| Time To Collision Likelihood | **0.8323** | 0.8321 | **0.8217** | 0.8216 |
| Distance To Road Edge Likelihood | **0.6825** | 0.6823 | **0.6484** | 0.6482 |
| Offroad Likelihood | 0.9516 | **0.9528** | 0.9360 | **0.9371** |
| minADE | **1.3215** | 1.3388 | **1.7700** | 1.7812 |

Table 1: Metrics for validation

| Metric name | Testing | | |
|---|---|---|---|
| | **SMART-tiny-CLSFT** | **comBOT** | **cogniBOT** |
| Realism meta-metric | **0.7846** | 0.7837 | 0.7015 |
| Linear Speed Likelihood | **0.3868** | 0.3802 | 0.1939 |
| Linear Acceleration Likelihood | **0.4066** | 0.4033 | 0.2610 |
| Angular Speed Likelihood | **0.5203** | 0.5181 | 0.3539 |
| Angular Acceleration Likelihood | **0.6588** | 0.6580 | 0.5278 |
| Distance To Nearest Object Likelihood | **0.3922** | 0.3906 | 0.2806 |
| Collision Likelihood | 0.9702 | **0.9703** | 0.9401 |
| Time To Collision Likelihood | **0.8302** | 0.8297 | 0.7888 |
| Distance To Road Edge Likelihood | **0.6814** | 0.6789 | 0.4903 |
| Offroad Likelihood | 0.9524 | **0.9525** | 0.8801 |
| minADE | **1.3065** | 1.3687 | 4.2599 |

Table 2: Metrics for testing

## 4 Attribution

This work was made using the Waymo Open Dataset, provided by Waymo LLC under the Waymo Dataset License Agreement for Non-Commercial Use, available at waymo.com/open/terms. (*Waymo Open Dataset: An autonomous driving dataset* 2019-2025)

## References

Brostek, Lukas et al. (June 2024). *Achieving Realism in Traffic Simulations: Performance of a Cognitive Behavior Model on the Waymo Open Sim Agent Challenge*. DOI: 10.31219/osf.io/csf7b. URL: osf.io/csf7b.

Filzek, Björn (2003). "Abstandsverhalten auf Autobahnen-Fahrer und ACC im Vergleich". In: *FORTSCHRITT-BERICHTE VDI. REIHE 12, VERKEHRSTECHNIK/FAHRZEUGTECHNIK* 536.

*Waymo Open Dataset: An autonomous driving dataset* (2019-2025).

Wu, Wei et al. (May 2024). "SMART: Scalable multi-agent real-time motion generation via next-token prediction". In: *arXiv [cs.RO]*.

Zhang, Zhejun et al. (Dec. 2024). "Closed-loop supervised fine-tuning of tokenized traffic models". In: *arXiv [cs.LG]*.