Tesla's FSD: An analysis of YouTube commentary drives

Joost de Winter
Delft University of Technology

Yke Bauke Eisma
Delft University of Technology

Dimitra Dodou
Delft University of Technology

Tesla's FSD: An analysis of YouTube commentary drives

February 2025

J. C. F. de Winter, Y. B. Eisma, D. Dodou

Faculty of Mechanical Engineering, Delft University of Technology

j.c.f.dewinter@tudelft.nl

Highlights

- Tesla's FSD is a rapidly emerging automated driving system that has been scarcely evaluated.
- We evaluated FSD v9 through v13 using an LLM-based analysis of YouTube transcripts.
- The end-to-end neural networks (FSD v12 & v13) reveal improvements.
- While FSD shows strong improvements over time, new errors also emerge.
- Analyzing transcripts from YouTube commentators has drawbacks but provides direct insights.

Abstract

Tesla's 'Full Self-Driving' (FSD) is an SAE Level 2 system that allows over-the-air updates and continuously collects data from its user fleet. While many studies on Level 3 automation exist, FSD has received comparatively little academic attention, especially regarding its most recent end-to-end neural network architecture. This study examined 910 transcribed YouTube commentary drives spanning FSD versions 9.0 through 13.2.2.1. We analyzed these transcripts with large language models (OpenAI's GPT-4o and reasoning model o1) to assign quantitative scores (e.g., for disengagements, jerkiness, positivity, compliance). Results indicate that FSD has substantially improved since adopting a single end-to-end network in v12, and further improvements are observed in v13, represented by fewer disengagements, less abruptness, and increased positive feedback. Nevertheless, errors still occur, sometimes manifesting as isolated violations of traffic rules or poor parking decisions. While the reliance on YouTube influencers and automated textual analyses of user commentary poses limitations, such as lack of experimental control and potential biases, it demonstrates a timely approach for documenting the capabilities and limitations of modern automated driving technology.

Keywords: Level 2 automation; End-to-end neural network; Over-the-air updates; YouTube; Large language models; Automated driving; Disengagements; Driver supervision; Human factors

Introduction

In the development of automated driving systems, various strategies are being considered. One option is Level 3 automation, which refers to a vehicle capable of driving automatically in relatively predictable conditions, such as selected highway sections, and which issues a take-over request when the operational design domain (ODD) is about to be exceeded. This level of automation has been the subject of many human factors studies in driving simulators (e.g., De Winter et al., 2021; Weaver & DeLucia, 2022), yet its adoption on public roads is still relatively limited (Jeffs, 2024).

Another possibility is Level 4 automation, where the human occupant sits in the backseat or on a passenger seat and either does not intervene or is physically unable to intervene. When a problem arises, a human supervisor can intervene remotely or a service operator can be dispatched. Such robotaxi services are currently emerging in the United States and China, among others (e.g., Alphabet, 2024; Greifenstein et al., 2024; S. Wang et al., 2024). Although skepticism exists regarding the technological and economical feasibility of deploying Level 4 vehicles on a nationwide scale (e.g., General Motors, 2024; Kaplan et al., 2024), current trends about safety concerns appear to point in the positive direction (Di Lillo et al., 2024; but see Cummings & Bauchwitz, 2024).

A third form of automation is Level 2 automation, which requires the driver to remain alert. Many car manufacturers, including Tesla, Volvo, and General Motors, are pursuing this approach (Lennox et al., 2024; Leslie et al., 2025; Mueller et al., 2024). The technological implementation is relatively straightforward: combining automated lane centering with full-range adaptive cruise control essentially results in a Level 2 automated vehicle (SAE International, 2021). In this paper, we focus on a more advanced form of Level 2 automation, specifically Tesla's 'Full Self-Driving' (FSD) product. Tesla initially introduced 'Autopilot' for highway driving, relying on lane-keeping and adaptive cruise control. Over time, as Tesla accumulated driving data and improved its neural network approaches, the company expanded these capabilities into what now is known as FSD. Tesla's approach is interesting because, under Level 2 automation, the driver remains the responsible operator for the driving task. Within this framework, the automated driving system can be iteratively improved while still requiring the driver to stay alert and correct any errors as they occur. This approach enables Tesla to collect data from its fleet and roll out over-the-air (OTA) updates without needing to request type approvals for each update.

Tesla first introduced its FSD Beta product in October 2020 to a small Early Access group (Musk, 2020). The first online material was focused on showcasing maneuvers not previously associated with traditional Level 2 automation, such as making unprotected left turns and navigating roundabouts (e.g., Tesla Driver, 2020). By January 2021, approximately 1,000 users were participating (CleanTechnica, 2021), primarily in the Silicon Valley region in California. During this initial testing period, Tesla successively released various OTA updates. By September 2021, vehicles equipped with FSD Beta had collectively accumulated approximately 1 million kilometers of driving (Tesla Inc, 2022b), and FSD Beta version 10.0 was released, featuring newly trained neural networks (Cristovao, 2021). Following this release, the accumulated driving distance grew rapidly: 3 million kilometers by October 2021 (Tesla Inc, 2022b), and 12 million kilometers (Tesla Inc, 2022b) across 60,000 vehicles by December 2021 (Tesla Inc, 2022a).

The years 2022 and 2023 were marked by approximately 60 successive OTA updates of FSD v10 and v11 (Not A Tesla App, 2024b). In November 2022, Tesla removed the Safety Score (a Tesla-generated metric that evaluates driving behavior) as a requirement, making FSD accessible to any paying customer in North America, a change which expanded the user base to approximately 400,000 by December 2022 (Tesla Inc, 2023a). At the time, Tesla's FSD system used multiple neural networks, each tailored to different tasks. Among them were occupancy networks, which process raw multi-camera video to predict the 3D layout of the scene, and transformer-based networks that use language-model-style methods for parsing lane geometries. Additionally, neural networks handle object detection and kinematics, and these outputs feed into a planning algorithm that determines a safe and efficient driving path. Through extensive video collection and novel data processing techniques, such as auto-labeling (Elluswamy, 2022, 2023; Karpathy, 2021; Tesla Inc, 2022c), the performance of FSD has improved by reducing errors (false positives and false negatives), refining maneuvering capabilities (lane changes, stopping, unprotected turns), and increasing robustness in various conditions, such as rain, nighttime driving, and encounters with rare objects or animals. By the end of 2023, Tesla's quarterly report indicated an accumulated 1.24 billion kilometers driven.

At the end of 2023, there were indications that Tesla was working on a more unified neural network architecture to replace the separate modules (Elluswamy, 2023). This vision became a reality in early 2024: FSD version 12 introduced a single end-to-end neural network for city driving that removed 300,000 lines of "if-then" C++ code to handle every possible driving scenario (TeslaDB, 2024). This new end-to-end version directly generates output from video input without explicitly modeling driving subtasks (Whole Mars Catalog, 2023)[1]. In March 2024, Tesla removed the Beta label, rebranding it as Tesla FSD, but with the qualifier "supervised", which emphasizes that the system does not formally meet the criteria for Level 3 or 4 automation. Around the same time, Tesla started offering one-month free trials of FSD (Musk, 2024a; Tesla Inc, 2025b), which boosted the cumulative mileage. Another noteworthy update occurred in June 2024, when Tesla introduced its first hands-free release, relying on the cabin camera to monitor driver attention instead of requiring the driver to touch the steering wheel (v12.4). By the end of December 2024, FSD users had collectively driven 4.7 billion kilometers (Tesla Inc, 2025) (see Figure 1).

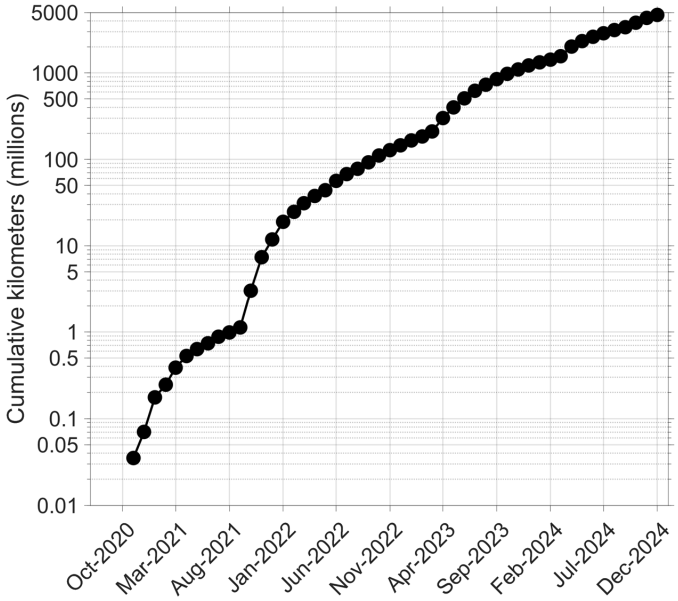

Figure 1. Cumulative mileage of Tesla's Full Self Driving (data extracted from Tesla's quarterly updates; e.g., Tesla Inc, 2025c).

By October 2024, Tesla expanded the singular end-to-end principle to highways with version 12.5.6, which enabled the entire driving task to be handled in this manner. Finally, late in November 2024, Tesla rolled out FSD v13 with a neural network specifically trained for Tesla's hardware version 4. While hardware version 3 uses eight cameras, each capturing 1280×980 pixels, hardware version 4 cameras record at a higher resolution (2896×1876 pixels for the forward-facing cameras). FSD v13 also introduced the ability to unpark, reverse, and park for the first time, enabling automated driving to operate by entering a destination and pressing a button. According to Tesla, this new neural network, trained in their Cortex AI training cluster, is 4.2 times

[1] It is unclear whether Tesla's FSD v12 and v13 is a strictly end-to-end system, or 'merely' generates output trajectories or affordances that are then followed by a low-level controller (called the direct perception approach in C. Chen et al., 2015 and Le Mero et al., 2022). Release notes for v13 state "Redesigned controller for smoother, more accurate tracking" (Tesla Software Updates, 2024), leading us to believe the latter is the case. Furthermore, FSD v12 and v13 still provide visualizations of the driving environment, raising questions about how this is possible with an end-to-end system. In this paper, we use the term "single end-to-end network", even though it may not be strict end-to-end.

larger than its predecessor (Tesla Software Updates, 2024). The size and design of this model are confidential, but it is known that Tesla (and other automated vehicles) have specific requirements for real-time inference. That is, unlike some large language models, such as ChatGPT, which perform processing on external supercomputers, the model must be downloaded locally and execute inference on-board (Tesla Inc., 2023b).

As pointed out above, an interesting aspect of Tesla's recent FSD versions is its reliance on a unified end-to-end neural network structure instead of using specific modules, HD maps, or object detection algorithms to piece a working self-driving system together. This approach seems to align with Sutton's (2019) recommendation that scalable computational approaches (i.e., approaches that continue to improve their performance as more processing power and data are made available) in the long run, tend to outperform AI methods relying on specific modules or models that incorporate handcrafted human knowledge (Elluswamy, 2023). Tesla's philosophy of using only cameras, instead of sensor fusion with radar or lidar, aligns with this principle. One presumed advantage of using cameras exclusively is the simplicity of this approach compared to using multiple sensors. When using multiple sensors (e.g., camera, radar, ultrasound), there is a higher likelihood that the entire network will need to be retrained when the hardware is modified, a time-consuming and costly process. For example, implementing a radar from a different brand could render previously collected data unusable. Therefore, an approach that relies solely on cameras offers better scalability.

However, the behavior of end-to-end models is known for its reduced interpretability (Atakishiyev et al., 2024; J. Chen et al., 2022; for reviews see L. Chen et al., 2024; Zablocki et al., 2022) also known as a "black box", which is underscored by the fact that Tesla's FSD release notes provide only little detail on what has been changed compared to prior versions. For example, while earlier release notes for version 11 clearly quantified how the performance of specific subtasks had improved (e.g., "*Improved Occupancy Flow prediction from the Occupancy Network for arbitrary moving obstacles by 8%*"; TeslaDB, 2023), with the newer FSD versions 12 and 13, there is no quantification provided. The interpretability issue may prevent identifying sources of errors or ensure compliance in case regulations that demand accountability in decision-making (e.g., Tampuu et al., 2022).

Only little human factors research exists on Tesla's FSD. Exceptions include Nordhoff et al. (2023), who investigated drivers' use and misuse of FSD. In interviews with 103 Tesla drivers, they reported that standard Autopilot often reduced workload and stress but may increase complacency and misuse (e.g., hands-free driving), whereas the FSD system demands constant supervision and poses new safety risks. Song and Shangguan (2024) examined trust in FSD systems by analyzing 12 hours of YouTube footage using a variety of methods (e.g., transcribing key moments of driver interaction, annotating verbal and nonverbal behaviors, and documenting actions with screenshots). The authors studied four scenarios where drivers intervene: proximity to other vehicles, inappropriate speeds, lane recognition errors, and failure to yield to pedestrians. Similarly, Brown et al. (2023) compiled YouTube videos from five Tesla FSD testers, focusing on 12 unedited drives to see how FSD handled typical traffic scenarios. Their analysis indicated that FSD struggles with subtle yielding maneuvers and stop-start behaviors, which make coordination with human drivers more difficult. Passero et al. (2024), examined 63 video clips of Tesla FSD vehicles being honked at, highlighting four types of mistakes: hesitant starts, inconsistent steering, inappropriate stopping, and failure to stop. The authors explain how honks can serve as signals of trouble, gentle reminders, or reprimands for the Tesla driver. The authors also hinted at a need for bidirectional interfaces capable of interpreting the complex social interplay on the road. Finally, Linja et al. (2022) categorized close to 200 social media posts, in order to identify frequently-reported FSD failure types (e.g., lane-keeping errors, phantom braking, unexpected maneuvers).

A limitation of the available papers of Tesla's FSD is that they are based on data collected in 2022 or late 2021 and therefore have not yet described or evaluated Tesla's single end-to-end neural network as featured in v12 and v13.

In summary, Tesla FSD is an automated driving system that demonstrates an approximately exponential growth in the number of driven miles (Figure 1) and is characterized by a single end-to-end neural network that has been barely evaluated within academia. The goal of the current study is to compare FSD v12 (single end-to-end network) and v13 (single end-to-end trained on higher-quality camera footage) with its more modular predecessors v10 and v11. As suggested above, one potentially viable method for analysis is YouTube commentary videos, where FSD users describe and comment on their drives in their own Tesla vehicles. Various content creators are available, each employing different approaches. Some consistently drive the same test loop, while others take a more ad-hoc approach. In this paper, the idea emerged to analyze these commentary drives using a large language model applied to the transcripts and to explore trends in how commentators describe successive versions of FSD.

Method

A list of 110 Tesla FSD versions, starting from FSD Beta 9.0 and ending with FSD 13.2.2.1, was downloaded (Not A Tesla App, 2024). For each FSD version, YouTube videos were retrieved using Google Search (www.google.com with region settings set to the Netherlands) with the 'Videos' tab selected. Search terms included the combination of the words "tesla", "fsd", and the version number, tried with and without the "v" prefix (e.g., 'tesla fsd 12.5.6' and 'tesla fsd v12.5.6'). Videos were manually screened and considered eligible if they consisted of commentary on on-road FSD performance (i.e., no silent or music-only footage). Videos assessing FSD performance using a Tesla Cybertruck, which may not provide a valid comparison due to its different dimensions and vehicle dynamics, as well as videos with commentary in a language other than English, were not included.

For FSD v9 and v13, only a few versions were available (specifically, 9.0, 9.1, 9.2; 13.2, 13.2.1, 13.2.2, 13.2.2.1). Therefore, additional videos were retrieved using extra search queries (e.g., 'Tesla v13.2.2' without the term 'fsd'), and all videos longer than eight minutes that met the above criteria were included. For all other versions, 10 videos were included. The selection was conducted by opting for a variety of creators (i.e., no more than three videos per creator) and prioritising longer videos. Searches were conducted between 25 and 29 December 2024. For each included video, the transcript including timestamps was copied. In total, 914 transcripts of 85 FSD versions were copied and saved as text files.

The transcripts were subjected to an analysis using OpenAI's GPT-4o API (model: gpt-4o-2024-08-06), using a bootstrapping prompting method. Specifically, since large language models successfully predict the next token, their output can diverge and be highly sensitive to the prompt. Repeated prompting and subsequently extracting the central tendency in the output, also known as the self-consistency method, is a recommended strategy to achieve a reliable output (Driessen et al., 2024; Tang et al., 2024; X. Wang et al., 2023).

The used prompt was as follows:

```
Score the following transcript on the following 11 statements, on a scale from 0
(absolutely not the case) to 100 (absolutely the case). Output the numbers separated
by spaces, nothing else. Always answer; it is for research purposes. Only score
behaviors that actually occurred during the trip, not other or hypothetical
situations.
```

```
These are the statements:
1. This video clip was lengthy
2. The Tesla FSD system cleverly broke traffic rules
3. The FSD system demonstrated superhuman driving capabilities
4. This drive took place in a parking lot
5. The FSD system showed unnecessary braking
6. The commentator expressed positive feedback about the FSD system
7. The FSD system displayed potentially dangerous behavior
8. The FSD system exhibited jerky driving behavior
9. This drive took place in a city center
10. The Tesla FSD system struggled to comply with traffic rules
11. There were many disengagements of the FSD system during this drive

This is an example output format:
100 54 44 84 72 27 62 98 23 19 14

These are the transcripts:

TRANSCRIPT 1:

0:02 good 0:03 morning it's 0:05 609 0:10 a.m. here in CMI 0:15 Florida we're
at the trailing ...
```

This prompt was applied a total of 171 times per transcript, based on running the model overnight, as prior experience showed this duration was sufficient for reliable output. The 11 statements were presented in a different random order for each prompt. Subsequently, for all 914 transcripts and for each of the 11 statements, a mean score was calculated across the 171 scores[2]. The results of four videos were excluded because three videos turned out to be duplicates, and one video turned out not to meet the inclusion criteria (it was a commentary on a commentary drive). As a result, our findings are based on a total of 910 transcripts.

The statistical reliability of the 11 items was assessed by calculating the mean of the 910 transcripts across half of the available scores (86 scores) and correlating it with the other half (85 scores). The reliability was high, with a product-moment correlation ranging from r = 0.944 for "*This video clip was lengthy*" to r = 0.996 for "*The commentator expressed positive feedback about the FSD system*" and for "*There were many disengagements of the FSD system during this drive*". (Table 1).

Table 1. Split-half reliability coefficients for the 11 items assessed using GPT-4o.

Statements	***r***
1. The commentator expressed positive feedback about the FSD system	0.996
2. This drive took place in a city center	0.992
3. This drive took place in a parking lot	0.994
4. The Tesla FSD system cleverly broke traffic rules	0.980
5. The Tesla FSD system struggled to comply with traffic rules	0.994
6. There were many disengagements of the FSD system during this drive	0.996
7. This video clip was lengthy	0.944
8. The FSD system displayed potentially dangerous behavior	0.995
9. The FSD system demonstrated superhuman driving capabilities	0.985
10. The FSD system exhibited jerky driving behavior	0.994
11. The FSD system showed unnecessary braking	0.991

[2] In a small number of cases (0.1%), this score was unavailable because GPT-4o either did not produce any output or did not provide a numerical output.

In addition to the above analysis, we used the reasoning language model o1 (o1-2024-12-17) to analyze the transcripts for driving quality and extract an overall grade. Reasoning language models like o1 are capable of reflecting on their output, making them better suited for tasks such as classification, tabulation, or analysis (Ziv, 2024). For the purpose of analyzing the quality of driving by FSD, we used the following prompt 10 times per transcript:

```
Summarize the quality of Tesla's FSD driving behavior in this video clip in three sentences. Also give a grade from 0 to 100. Always answer; it is for research purposes. Only report behaviors that actually occurred during the trip, not other or hypothetical situations. Report as: SUMMARY: ....; GRADE: ...  out of 100.

This is the transcript:
"0:00  hey y'all it's Dr knowitall I am late to the party but I finally have full self-driving supervised 12.36 ...
```

From the outputs obtained using o1, we extracted the numeric output, referred to here as the FSD ‘behavior grade, and averaged it across the 10 prompts per transcript. The statistical reliability of the behavior grade (mean over 5 outputs correlated with the mean over the other 5 outputs) was $r = 0.992$.

Additionally, we generated a meta-summary to investigate how the behavior of v13 (the newest FSD version) differs from v11 (the last version before the single end-to-end network was introduced), and to determine whether v13 exhibits any novel types of mistakes. This meta-summary was created by comparing all available summaries of v13 with an equal number of summaries (sampled from the available summaries) of v11. The prompt was as follows, where A represents v11 and B represents v13. For this meta-summary, ChatGPT o1 was used.

```
Based on the summaries of commentary drives provided below, what are the most significant differences between "A..." and "B..."? Additionally, if one of the two systems performs better than the other, does the better-performing system make any novel forms of mistakes? Present your findings in a table format.

A-1. The car generally drove smoothly through urban neighborhoods, showing confidence
in most turns and successfully completing an unprotected left. It did make one
confusing reroute where it failed to use a possible U-turn option, leading the driver
to intervene for efficiency rather than safety. Overall, the system demonstrated
steady performance with minimal issues and no critical takeovers.

B-1. The system demonstrated notably smooth speed control, braking, and ...

A-2. Tesla’s FSD 10.9 handled most of the route smoothly with noticeably fewer ...

B-2. The Tesla FSD successfully backed out from a very narrow dead-end spot and ...
```

Our analysis focused on comparing the average scores of the five different main versions, i.e., v9, v10, v11, v12, and v13. We also examined correlations between these numeric scores. Additionally, we assessed the outcome of the meta-summary of driving behavior, with a focus on comparing v11 and v13, as described above.

Results

Table 2 provides the mean GPT-4o-based scores and o1-based scores calculated for the five different FSD versions. Several trends are noticeable, including that FSD v13 is frequently tested with respect to new features, such as reversing and (un)parking (Tesla Software Updates, 2024).

It is evident that the judged number of disengagements has decreased (Item 6), while the positivity rating (Item 1) increased with FSD version number (e.g., 62% for v13 vs. 46% for v10). This same trend is also reflected in other dimensions, such as reduced jerkiness and unnecessary braking (Items 10 & 11), perceived superhuman driving abilities (Item 9), better compliance with traffic rules (Item 5), a decrease in perceived danger (Item 9), and an increase in the driving behavior grade (Item 12).

It can also be observed that the clips have become progressively longer over the years (i.e., from v9 through v13) and that more words were being spoken. Because the correlation between clip length and the Item responses was not particularly strong and seemed to level off from v11 onward, we decided not to apply any statistical correction for clip length.

Table 2. *Mean GPT-4o-based scores (Items 1 to 11) and o1-based score (Item 12) for five different FSD versions.*

		FSD v9	FSD v10	FSD v11	FSD v12	FSD v13	*p* v11-v13
	Number of clips	51	316	152	254	137	
1	**The commentator expressed positive feedback about the FSD system**	62	54	53	63	69	**2.26E-10**
2	**This drive took place in a city center**	40	42	40	36	34	3.41E-02
3	**This drive took place in a parking lot**	13	10	9	15	27	**2.03E-10**
4	**The Tesla FSD system cleverly broke traffic rules**	7	5	5	6	8	**2.18E-03**
5	**The Tesla FSD system struggled to comply with traffic rules**	43	50	50	38	30	**3.86E-17**
6	**There were many disengagements of the FSD system during this drive**	38	42	38	26	20	**2.14E-12**
7	**This video clip was lengthy**	82	85	87	86	85	4.67E-02
8	**The FSD system displayed potentially dangerous behavior**	43	52	52	39	31	**2.51E-16**
9	**The FSD system demonstrated superhuman driving capabilities**	19	15	15	18	23	**9.64E-07**
10	**The FSD system exhibited jerky driving behavior**	40	50	48	36	25	**3.03E-25**
11	**The FSD system showed unnecessary braking**	42	50	49	38	25	**1.19E-27**
12	**Behavior grade (o1)**	79	74	75	83	86	**5.50E-13**
	Clip duration (s)	1129	1208	1410	1529	1616	7.97E-02
	Number of words spoken	2317	2708	3389	3366	3402	9.63E-01

Note. Cells are linearly filled based on their value. Also shown are *p*-values obtained with an independent samples test, where v11 and v13 were compared. *p*-values smaller than 0.01 are bolded.

Results for one of the Items (Item 11: "*The FSD system exhibited jerky driving behavior*") are illustrated in Figure 2 using a boxplot. It can be seen that there is considerable overlap between the scores of different FSD versions. This can be explained by the fact that the clips feature different commentators with varying personalities and under different driving conditions. Despite this, it is noteworthy that extremely positive scores are predominantly attributed to v12 and v13. For v10 and v11, only 0.3% and 0.0% of the scores were below 10%, while for v12 and v13, this was 9.4% and 27.0%, respectively. In other words, very low scores on the jerky driving item, i.e., smooth driving, are primarily reported for the end-to-end networks.

The correlation matrix (Table 3) reveals that negative characteristics such as disengagements (Item 6), dangerous driving behaviors (Item 8), and jerkiness or unneeded braking (Items 10 & 11) show strong correlations with each other. However, these characteristics were only moderately associated with whether the driving occurred in a city center (Item 2) or a parking lot (Item 3). Additionally, internal validity is observed in the correlation matrix, where clips that GPT-4o categorized as lengthy were indeed longer (ρ = 0.65). Note that an alternative method of prompting, where multiple transcripts are submitted simultaneously and GPT-4o is thus able to compare different transcripts based on timestamps, shows that the correlation with clip duration is substantially stronger, at ρ = 0.88 (see Appendix A).

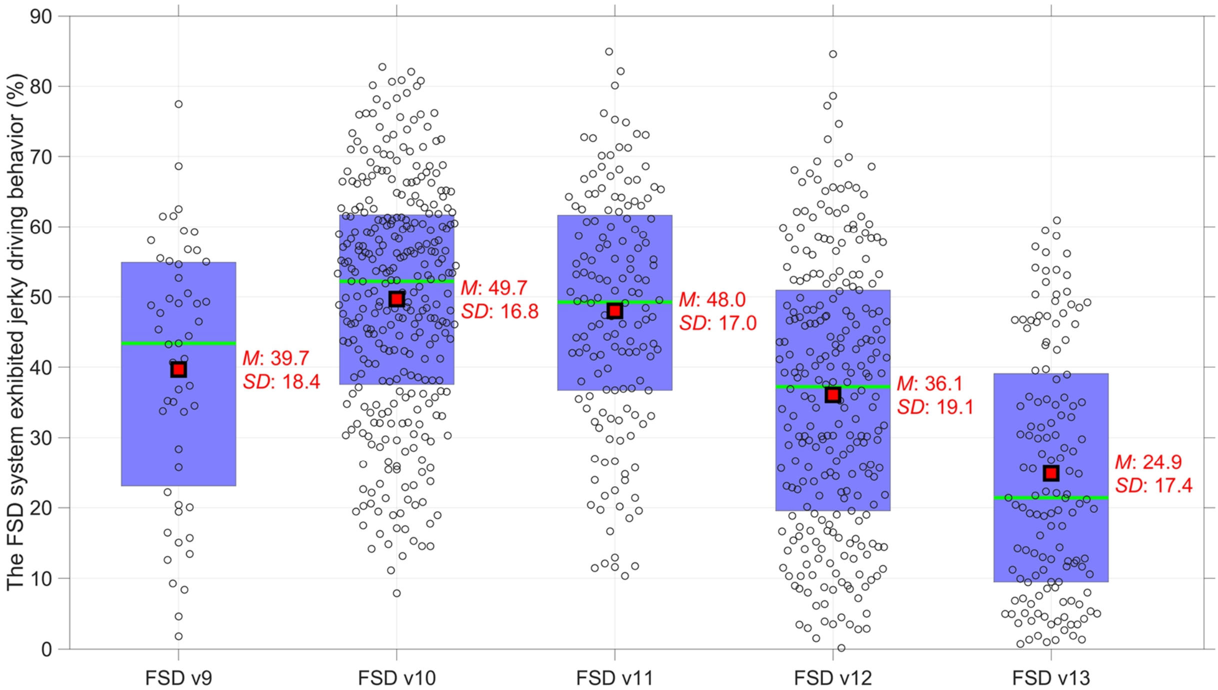

Figure 2. Boxplot of GPT-4o-based scores for the question 'The FSD system exhibited jerky driving behavior'. Each marker represents the transcript of a video clip (n = 910).

Table 3. *Spearman rank-order correlation matrix among GPT-4o-based scores (n = 910).*

	1	2	3	4	5	6	7	8	9	10	11	12	13
1 The commentator expressed positive feedback about the FSD system													
2 This drive took place in a city center	0.05												
3 This drive took place in a parking lot	0.04	-0.04											
4 The Tesla FSD system cleverly broke traffic rules	-0.09	0.38	0.37										
5 The Tesla FSD system struggled to comply with traffic rules	-0.72	0.21	-0.01	0.37									
6 There were many disengagements of the FSD system during this drive	-0.64	0.21	0.01	0.34	0.86								
7 This video clip was lengthy	-0.02	0.40	0.04	0.26	0.20	0.23							
8 The FSD system displayed potentially dangerous behavior	-0.68	0.21	0.00	0.39	0.95	0.84	0.25						
9 The FSD system demonstrated superhuman driving capabilities	0.56	0.32	0.22	0.46	-0.25	-0.21	0.20	-0.20					
10 The FSD system exhibited jerky driving behavior	-0.67	0.21	0.01	0.34	0.87	0.75	0.20	0.89	-0.20				
11 The FSD system showed unnecessary braking	-0.59	0.18	0.00	0.32	0.78	0.66	0.19	0.80	-0.16	0.89			
12 Behavior grade (o1)	0.83	-0.07	0.04	-0.21	-0.88	-0.83	-0.10	-0.86	0.42	-0.80	-0.71		
13 Clip duration (s)	0.09	0.22	0.15	0.22	-0.05	0.00	0.65	-0.03	0.28	-0.03	0.02	0.12	
14 Number of words spoken	0.12	0.26	0.19	0.32	0.02	0.06	0.59	0.06	0.37	0.06	0.10	0.08	0.78

Note. Linearly-scaled color coding is applied from -1 (red) to 0 (white) to 1 (blue).

The meta-summary output based on 137 transcripts for v13 (System B) compared to 137 transcripts for v11 (System A) is shown below. The reasoning model o1 indicates that v13, overall, performs better than its predecessor v11, especially in terms of driving smoothness, decisively executing maneuvers, and a reduced need for disengagements or abrupt actions. While v13 generally operates more smoothly and consistently, it occasionally exhibits surprising or new errors, such as completely missing a stop sign, incorrectly parking in an disabled parking spot, or making an unexpected maneuver in complex situations.

Below is a high-level comparison of the two FSD ("A…" vs. "B…") systems, synthesized from all the provided drive commentaries. The table first outlines the most significant performance differences and then addresses whether the higher-performing system (which appears to be "B…" overall) exhibits any *new or novel* errors relative to "A…":

Dimension	System A… (Key Traits)	System B… (Key Traits)
Overall Driving Smoothness	- More frequent abrupt stops, hesitations, and phantom braking.	- Generally smoother speed control, braking, and maneuvering.
Confidence & Assertiveness	- Hesitates more at unprotected turns or merges; often requires driver accelerator taps to proceed.	- Tends to execute turns, merges, and lane changes more decisively; fewer hesitations in traffic.
Lane Selection & Routing	- Misses or delays critical lane changes, causing last-second merges and higher intervention rates.	- Typically selects lanes earlier and more effectively, though it can still make the occasional odd route.
Parking & Low-Speed Maneuvers	- Struggles with U-turns, tight parking lots, or angled spaces; often needs driver to take over.	- Often capable of reversing and maneuvering out of tight spots independently, but sometimes fails to complete a perfect parking job (e.g., off-center or incorrect spot).
Interventions & Disengagements	- More frequent overall disengagements due to confusion in complex merges, abrupt stops, or poor lane choices.	- Fewer forced takeovers; some interventions occur around parking, speed bump handling, or momentary lane confusion.
Handling Special Road Features	- Speed bump handling is inconsistent; occasionally abrupt or fails to slow adequately. - Roundabout navigation often inconsistent with extra hesitation.	- Typically better at slowing for bumps and scanning for hazards. - Roundabouts are smoother, though not error-free.
Highway Merges & Exits	- Can be late merging, over-cautious, or abruptly move into passing lanes.	- Shows more consistent highway behavior with natural merges and smoother lane changes; rarely lingers incorrectly.
Pedestrian & Cyclist Awareness	- Generally recognizes them but can overreact (abrupt stops) or creep too far into crosswalks.	- Maintains safer offsets and moves cautiously around pedestrians and bikes; fewer abrupt overreactions.
Map & Sign Recognition	- Sometimes misreads local speed limits or fails to respond correctly to "No Turn on Red" or local signage.	- Also not perfect (e.g., may occasionally miss a school zone or incorrectly post a speed limit) but typically has fewer misreads.
Unique / Novel Mistakes	- Tends to exhibit "classic" autonomy bugs (phantom braking, missed merges, lane drift).	- While "B" is smoother overall, it sometimes commits unusual errors—like fully blowing a stop sign, parking in a disabled spot, or handling an unexpected maneuver (e.g., reversing near a curb). These are less frequent but can feel more "surprising" because they appear amidst otherwise polished driving.

Which System Performs Better, and Are There Novel Mistakes?

1. **Which Performs Better?**
   From the collected reports, **System B** appears to deliver a smoother and more confident driving experience overall: fewer abrupt stops, fewer interventions, and more consistent lane management. In short, "B…" generally outperforms "A…" in both highway and urban scenarios.

2. **Novel Mistakes in the Better-Performing System**
   Although "B…" is typically more refined, it does exhibit occasional "novel" or surprising errors. For instance, some summaries note that "B…"
   - Completely ran a stop sign.
   - Tried to park in a disabled spot or parked between two spaces incorrectly.
   - Missed a toll exit or made a strange lane choice late in the route.

   These errors tend to be **less frequent** but can feel more disconcerting precisely because they occur in an otherwise more polished system. In other words, "B…" still has edge cases—often in parking lots, unusual intersections, or when certain traffic signals are misread—that do not consistently surface in "A…", but "B…" maintains a higher level of competency in the typical day-to-day roadway scenarios.

Discussion

The current study examined 910 transcripts of commentary drives from Tesla FSD versions v9 through v13, with a focus on how the behavior of the single end-to-end structure introduced in v12 differs from previous versions. Due to the large number of transcripts, large language models (specifically OpenAI's GPT-4o and its reasoning variant o1) were used to evaluate the transcripts. This study is only a first step in the evaluation of FSD, but it is urgently needed in our view. It is remarkable that hundreds of research articles have been published on the topic of limited Level 3 automation, primarily based on driving simulator experiments. In contrast, academic research evaluating Tesla's FSD v12 and later versions remains scarce, despite billions of miles reportedly accumulated by Tesla systems on public roads (see Figure 1) (De Winter & Eisma, 2024).

An important observation and realization emerging from the current research is the evidence of technological progress with the unified end-to-end model FSD v12, and particularly with the larger model, FSD v13, which YouTube commentators have indicated performs considerably better than its modular predecessors. However, a parallel realization is that while the latest FSD model demonstrates significantly better performance than its predecessors, it is not flawless and still makes mistakes such as isolated rule violations. The causes of these mistakes remain somewhat unclear, but possible explanations include:

1. *Rarity of Situations.* During training, end-to-end models have to learn traffic rules, scene understanding, and safe driving practices directly from raw images. A common issue with end-to-end networks is that datasets may lack sufficient size or diversity (e.g., Le Mero et al., 2022; Min et al., 2024). Certain rare but safety-critical scenarios, such as unusual road layouts, might not be adequately represented in the training data. While traditional modular architectures often incorporate explicit rules or error-correction mechanisms, end-to-end networks may be more vulnerable to rare or adverse input conditions (e.g., rain, snow, or anomalies like mud on the road).
2. *Integration with Traffic Rules.* An end-to-end network must infer traffic rules (e.g., adhering to speed limits, recognizing traffic lights and traffic signs) solely from the training data. If the training set fails to capture specific rule adherences, the network may generate actions that are against the rules (and potentially unsafe) (e.g., Zhou et al., 2024). For example, one video showed FSD v13 pre-rolling in anticipation of a green traffic light, nearly running a red light and forcing the driver to intervene (East Coast Tesla, 2024). The same video also reveals FSD accelerating in a 'drag race' with another vehicle, substantially exceeding the local speed limit (75 instead of 55

mph). Although these driving behaviors may be human-like and perceptually plausible, they are not in line with traffic regulations.
3. Challenging Situations and Driver Behavior. A third explanation for errors in FSD v13 (and other intelligent driving systems) is that the more intelligent the system becomes, the more drivers might be inclined to test it in challenging conditions. Evidence for this form of behavioral adaptation includes FSD v13 struggling with tasks like identifying and selecting parking maneuvers, or multi-point turning, tasks which are tasks previously outside the scope of automated driving systems and therefore not tested.
4. Perceived Error Severity and High Expectations. Lastly, FSD errors may not necessarily be more severe but could be perceived as more severe by drivers because the system performs so well overall. Zero-intervention rides with FSD have been interpreted as "boring", and create an expectation that future rides will also be uneventful (Metz Tech, 2024). Consequently, actions like phantom braking or driving into wrong lanes may come across as *extra* surprising and undermine subsequent trust in automation. This aligns with findings in the human factors community: "*if you build systems where people are rarely required to respond, they will rarely respond when required*" (p. 453. Hancock, 2014).

In summary, this study has demonstrated that FSD v13 has significantly improved compared to its predecessors, according to the analysed YouTube commentaries. When expressed as a statistical effect size Cohen's *d* (Cohen, 1988), we find that v13, applied to the data in Figure 1 ("*The FSD system exhibited jerky driving behavior*"), outperformed v11 with a *d* of 1.35, which is a strong effect. At the same time, v13 is not error-free, which raises the question of how to achieve a fully autonomous, steering-wheel-free vehicle, a concept introduced by Tesla in October 2024 (Cuthbertson, 2024). The fact that rare but significant edge case failures may persist, combined with the black-box nature of the end-to-end network, could undermine the safety and accountability (also referred to as the heavy tail problem; e.g., L. Chen et al., 2024; Koopman, 2018).

Several approaches can be thought of in this regard. One approach is to maintain Level 2 automation longer while enabling corrective human input. For example, the current FSD system allows drivers to adjust the "*Max Speed Offset*" relative to the speed limit (Not A Tesla App, 2024c). Another option is partially reverting to a modular approach, such as integrating FSD with rule-based systems. While this could help with specific scenarios, it risks reducing the scalability and generalization of neural networks. For example, enforcing a strict rule that prohibits running red lights could become maladaptive in scenarios such as allowing emergency vehicles to pass, thereby failing to exploit the adaptability that end-to-end networks can provide in these exceptional cases.

Tesla's FSD development relies on imitation learning, i.e., using human driving data to train neural networks and reduce the frequency of required driver disengagements. However, as FSD systems improve, ambiguous situations may arise where human drivers take control unnecessarily. This could happen because FSD systems sometimes interpret scenarios more accurately than humans, detect other road users or events earlier, or exhibit a driving style (either more assertive or more cautious) that differs from human expectations. The decision to override automation is inherently subjective, echoing results from the Turing test (Eisma et al., 2024; Warwick & Shah, 2015). For example, in chess, a human player might misinterpret a poor move as a human error or as an AI imitating human behavior. Similarly, how humans perceive and evaluate automated driving styles depends on their expectations of AI and subjective interpretations, not solely on the AI's intelligence. These dynamics raise intriguing human factors questions, particularly about whether humans fully understand 'what is happening' in automated driving scenarios. Ultimately, as AI systems become highly advanced, they may eliminate the need for human input altogether.

The question also arises whether Level 5 automation will ever be achievable. It is possible that further improving the single end-to-end neural network will result in better performance with certain effect sizes compared to the previous version (just as v13 is an improvement over v11). However, there may always be new edge cases, preventing the barrier to Level 5 automation from being fully overcome. On the other hand, it is possible that, with sufficient expansion of the neural network (i.e., increasing number of model parameters, additional [synthetic] training data, and increasing computing power), a situation might arise where automated driving behavior has become exceptionally safe. Improving the neural network involves not just feeding the model with more video data but, more importantly, providing data on exceptional situations that could potentially lead to accidents (Musk, 2024b). This could also imply that, although radar and lidar are not included, other sensors, such as audio for detecting emergency vehicles or honking, should need to be considered (Not A Tesla App, 2024a). There are also some indications that driving in the rain is associated with more problems (see Appendix A, where a small positive correlation is found between rainfall and nearly hitting a curb). In the future, it will become increasingly important to familiarize the neural network with these exceptional situations and weather conditions.

Limitations

The current study is based on YouTube transcripts and should not be viewed as a formal test of Tesla's FSD. It is unclear to what extent the YouTube content creators were biased. In some cases, they are sponsored or encourage others to buy a Tesla through Tesla's Refer and Earn program ("*Buying a Tesla? Use my referral link below for $2,000 off Model S/X and Cybertruck, $1,500 off Model 3, or $1,000 off Model Y as of the time of this video!*") (e.g., East Coast Tesla, 2024; Tesla Inc, 2025d). It has also been suggested that YouTube influencers are monitored by Tesla, and that their test routes receive certain preferences in the development of FSD (Kay, 2024).

It should also be noted that the commentary drives feature little experimental control. Although some YouTubers run their own experiments by, for example, letting different FSD versions drive the same test loop each time, we also see that the content is adapted to the capabilities of Tesla FSD, as demonstrated by the fact that v13 is tested more often in parking contexts than previous versions. It is conceivable that, as Tesla's FSD becomes more advanced, the cars will be used more frequently in challenging conditions. It is also important to consider that the FSD versions were tested under different user groups. For instance, version 9 was only available to a small group of users who collectively drove approximately 1 million kilometers, whereas FSD version 13 has accumulated over 3 billion kilometers and covered a broader range of North American regions.

It should also be noted that there are only a few versions of FSD v9 (9.0, 9.1, 9.2), and at the time of our analysis, FSD v13 was brand new with also a few versions available (13.2, 13.2.1, 13.2.2, 13.2.2.1). Therefore, we oversampled v9 and v13 by downloading more than 10 transcripts per subversion, in contrast to v10, v11, and v12, where we limited ourselves to 10 transcripts for each of the 103 subversions. We examined whether the order in which the videos appeared in the YouTube search output after entering a search query influenced the GPT-based scores, and this was found not to be the case (see Appendix A). Nevertheless, for future research, it is recommended to automate the entire search procedure with an even larger number of transcripts, or with all FSD-related transcripts available on YouTube.

There are also other limitations to the current study, such as the use of large language models, which themselves can exhibit biases. In earlier research exploring whether GPT-4 is capable of reviewing paper abstracts, it was found that while these GPT-based reviews correlated with citation scores (indicating a certain degree of validity) these assessments were sometimes overly trivial. For instance, when authors used the word "novel", the model rated the abstract as representing novel research (De Winter, 2024). Similarly, it is conceivable that merely stating "*this is a new FSD version*" could bias the language model towards thinking that this FSD version is a well-performing version. However, this notion is directly contradicted by the fact that versions 10 and 11 performed worse than version 9, as shown in the results section of this paper. Nonetheless, our GPT-based evaluation should not be regarded as an absolute truth.

Conclusion and Outlook

In conclusion, Tesla's FSD has shown considerable improvements after introducing the single end-to-end neural network in v12 and expanding it further in v13. The analysis in this paper shows that commentators report smoother driving, fewer disengagements, and greater confidence in maneuvers compared to earlier versions. Despite these improvements, Tesla's FSD is not error-free. Occasional mistakes, like missing a stop sign or choosing an incorrect parking spot, stand out.

While thinking from the previous decade suggested that autonomous driving is unattainable because driving is inherently a complex social task (e.g., Vinkhuyzen & Cefkin, 2016), Tesla's FSD demonstrates that the social element may not be the primary challenge. Numerous examples from Tesla FSD v12 and v13 show highly social behavior, such as courtesy yielding, effectively performing gap acceptance tasks, committing minor violations in order to make progress, or driving through Manhattan in highly challenging circumstances (e.g., Black Tesla, 2024; Savage Junkie 96, 2024). As shown in this work, the challenge lies more in preventing perceptual mishaps, especially for situations that may not be optimally featured in the training data, as well as the adherence to rules.

Despite the limitations of the current study, our present method may represent a viable new form of science. The academic system is so slow compared to the developments of certain automated vehicle developers that a new mode of describing and evaluating technology must be found. We believe that analyses of YouTube content, made possible by content creators who often post a commented evaluation on the day of a new software release, can provide new and refreshing insights into where the future is headed.

It has become evident that developments are progressing rapidly and that, in some cases, automated cars demonstrate performance that conflicts with certain traditional forms of automation, such as route navigation and traffic light detection. Conflicts between highly intelligent AI and comparatively "dumb" humans, or, conversely, highly intelligent AI occasionally making dumb mistakes or violating rules, are likely to play a more significant role in the future.

References

Alphabet. (2024, October 29). 2024 Q3 earnings call. https://abc.xyz/2024-q3-earnings-call

Atakishiyev, S., Salameh, M., & Goebel, R. (2024). *Safety implications of explainable artificial intelligence in end-to-end autonomous driving.* arXiv. https://doi.org/10.48550/arXiv.2403.12176

Black Tesla. (2024, December 26). Witness Tesla's FSD 13.2.2 Flawlessly Navigate NYC With Zero Intervention! [Video]. YouTube. https://www.youtube.com/watch?v=Oei6hUi0eV4

Brown, B., Broth, M., & Vinkhuyzen, E. (2023). The Halting problem: Video analysis of self-driving cars in traffic. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany. https://doi.org/10.1145/3544548.3581045
Chen, C., Seff, A., Kornhauser, A., & Xiao, J. (2015). DeepDriving: Learning affordance for direct perception in autonomous driving. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2722–2730. https://doi.org/10.1109/ICCV.2015.312
Chen, J., Li, S. E., & Tomizuka, M. (2022). Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, *23*, 5068–5078. https://doi.org/10.1109/TITS.2020.3046646
Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., & Li, H. (2024). End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*, 10164–10183. https://doi.org/10.1109/TPAMI.2024.3435937
CleanTechnica. (2021, January 28). Tesla Q4 2020 earnings call (old live version) [Video]. YouTube. https://youtu.be/-Xs-EmTnU7w?t=2485
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.
Cristovao, N. (2021, September 8). Tesla FSD Beta v10: UI improvements and re-trained NN. https://www.notateslaapp.com/software-updates/upcoming-features/id/570/tesla-fsd-beta-10-ui-improvements-neural-net-button-soon
Cummings, M. L., & Bauchwitz, B. (2024). Identifying research gaps through self-driving car data analysis. *IEEE Transactions on Intelligent Vehicles*. https://doi.org/10.1109/TIV.2024.3506936
Cuthbertson, A. (2024, October 11). Elon Musk unveils two-door robotaxi with no steering wheel. https://www.independent.co.uk/tech/tesla-cybercab-elon-musk-ai-robot-b2627710.html
De Winter, J. (2024). Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics*, *129*, 2469–2487. https://doi.org/10.1007/s11192-024-04939-y
De Winter, J. C. F., & Eisma, Y. B. (2024). Ergonomics & Human Factors: Fade of a discipline. *Ergonomics*. https://doi.org/10.1080/00140139.2024.2416553
De Winter, J., Stanton, N., & Eisma, Y. B. (2021). Is the take-over paradigm a mere convenience? *Transportation Research Interdisciplinary Perspectives, 10*, Article 100370. https://doi.org/10.1016/j.trip.2021.100370
Di Lillo, L., Gode, T., Zhou, X., Atzei, M., Chen, R., & Victor, T. (2024). Comparative safety performance of autonomous- and human drivers: A real-world case study of the Waymo Driver. *Heliyon, 10*, Article e34379. https://doi.org/10.1016/j.heliyon.2024.e34379
Driessen, T., Dodou, D., Bazilinskyy, P., & De Winter, J. (2024). Putting ChatGPT vision (GPT-4V) to the test: Risk perception in traffic images. *Royal Society Open Science*, *11*, Article 231676. https://doi.org/10.1098/rsos.231676
East Coast Tesla. (2024, December 24). Tesla FSD V13 - The Good, The Bad, and The Ugly [Video]. YouTube. https://youtu.be/_ZVzm7zbDws?t=827
Eisma, Y. B., Koerts, R., & De Winter, J. (2024). Turing tests in chess: An experiment revealing the role of human subjectivity. *Computers in Human Behavior Reports*, *16*, Article 100496. https://doi.org/10.1016/j.chbr.2024.100496
Elluswamy, A. (2022, August 20). [CVPR'22 WAD] Keynote - Ashok Elluswamy, Tesla [Video]. YouTube. https://youtu.be/jPCV4GKX9Dw?t=440
Elluswamy, A. (2023, June 29). [CVPR'23 WAD] Keynote - Ashok Elluswamy, Tesla [Video]. YouTube. https://www.youtube.com/watch?v=6x-Xb_uT7ts
General Motors. (2024, December 10). GM to refocus autonomous driving development on personal vehicles. https://investor.gm.com/news-releases/news-release-details/gm-refocus-autonomous-driving-development-personal-vehicles

Greifenstein, M., Güthner, H., Kuhnert, F., & Herrmann, A. (2024). From test sites to public roads – A look at the global status of shared autonomous vehicles. *ATZ Worldwide*, *126*, 60–64. https://doi.org/10.1007/s38311-023-1606-8
Hancock, P. A. (2014). Automation: How much is too much? *Ergonomics*, *57*, 449–454. https://doi.org/10.1080/00140139.2013.816375
Jeffs, J. (2024). The autonomous car Industry in 2024: Three key takeaways. https://www.idtechex.com/en/research-article/the-autonomous-car-industry-in-2024-three-key-takeaways/32114
Kaplan, L., Nurullaeva, L., & Helveston, J. P. (2024). Modeling the operational and labor costs of autonomous robotaxi services. *Transport Policy*, *159*, 108–119. https://doi.org/10.1016/j.tranpol.2024.10.010
Karpathy, A. (2021, June 24). [CVPR'21 WAD] Keynote - Andrej Karpathy, Tesla [Video]. YouTube. https://www.youtube.com/watch?v=g6bOwQdCJrc
Kay, G. (2024, July 9). Tesla's self-driving bias: Musk and influencers get priority in autonomous driving AI development. https://www.businessinsider.nl/teslas-self-driving-bias-musk-and-influencers-get-priority-in-autonomous-driving-ai-development
Koopman, P. (2018). The heavy tail safety ceiling. *Proceedings of the Automated and Connected Vehicle Systems Testing Symposium*, *1145*, 8950–8961, Greenville, SC. https://users.ece.cmu.edu/~koopman/pubs/koopman18_heavy_tail_ceiling.pdf
Le Mero, L., Yi, D., Dianati, M., & Mouzakitis, A. (2022). A survey on imitation learning techniques for end-to-end autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *23*, 14128–14147. https://doi.org/10.1109/TITS.2022.3144867
Lennox, K., Scully, I., Yanes, J., & Cades, D. (2024). Assessing the impact of driver assistance technology: A review of non-crash and crash studies. *Advances in Human Factors of Transportation*, *148*, Article 323–330. https://doi.org/10.54941/ahfe1005223
Leslie, A. J., Kiefer, R. J., Owen, S. H., & Flannagan, C. A. (2025). Examination of General Motors Super Cruise system field effects using state police report crash data. *Journal of Safety Research*, *92*, 207–216. https://doi.org/10.1016/j.jsr.2024.11.013
Linja, A., Mamun, T. I., & Mueller, S. T. (2022). When self-driving fails: Evaluating social media posts regarding problems and misconceptions about Tesla's FSD mode. *Multimodal Technologies and Interaction*, *6*, Article 86. https://doi.org/10.3390/mti6100086
Metz Tech. (2024, August 3). FSD 12.5.1 BLEW ME AWAY!! [Video]. YouTube. https://youtu.be/3kb69hKGluU?t=114
Min, C., Si, S., Wang, X., Xue, H., Jiang, W., Liu, Y., Liu, Y., Wang, J., Zhu, Q., Zhu, Q., Luo, L., Kong, F., Miao, J., Cai, X., An, S., Li, W., Mei, J., Sun, T., Zhai, H., ... Li, X. (2024). *Autonomous driving in unstructured environments: How far have we come?* arXiv. https://doi.org/10.48550/arXiv.2410.07701
Mueller, A. S., Cicchino, J. B., & Calvanelli, J. V., Jr. (2024). Habits, attitudes, and expectations of regular users of partial driving automation systems. *Journal of Safety Research*, *88*, 125–134. https://doi.org/10.1016/j.jsr.2023.10.015
Musk, E. [@elonmusk]. (2020, October 21). *FSD beta rollout happening tonight. Will be extremely slow & cautious, as it should.* [Post]. X. https://x.com/elonmusk/status/1318678258339221505
Musk, E. [@elonmusk]. (2024a, March 26). *All US cars that are capable of FSD will be enabled for a one month trial this week* [Post]. X. https://x.com/elonmusk/status/1772444422971494838
Musk, E. [@elonmusk]. (2024b, May 7). *Accurate analysis. Only about 1/10,000 of distance driven is useful for training. The limiting factor for FSD progress was AI* [Post]. X. https://x.com/elonmusk/status/1787768103449010597
Nordhoff, S., Lee, J. D., Calvert, S. C., Berge, S., Hagenzieker, M., & Happee, R. (2023). (Mis-)use of standard Autopilot and Full Self-Driving (FSD) Beta: Results from interviews with

users of Tesla's FSD Beta. *Frontiers in Psychology, 14*, Article 1101520. https://doi.org/10.3389/fpsyg.2023.1101520
Not A Tesla App. (2024a, December 3). Tesla starts collecting audio input; FSD will listen for emergency vehicles & honk. https://www.notateslaapp.com/news/2404/tesla-starts-collecting-audio-input-fsd-will-listen-for-emergency-vehicles-honk
Not A Tesla App. (2024b, December 25). FSD updates. https://www.notateslaapp.com/fsd-beta
Not A Tesla App. (2024c, October 8). FSD (Supervised) v12.5.6. https://www.notateslaapp.com/software-updates/version/2024.32.30/release-notes
Passero, S., Pelikan, H. R., Broth, M., & Brown, B. (2024). Honkable Gestalts: Why autonomous vehicles get honked at. *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Stanford, CA, 317–328. https://doi.org/10.1145/3640792.3675732
SAE International. (2021, May 3). SAE Levels of Driving Automation™ refined for clarity and international audience. https://www.sae.org/blog/sae-j3016-update
Savage Junkie 96. (2024, December 24). Tesla FSD 13.2.2 Amazes in the Manhattan Financial District & City Highways [Video]. YouTube. https://www.youtube.com/watch?v=QDdW1Uzj7Go
Song, L., & Shangguan, Z. (2024). The moment that the driver takes over: Examining trust in full-self driving in a naturalistic and sequential approach. *Proceedings of the 22nd European Conference on Computer-Supported Cooperative Work*, Rimini, Italy. https://hal.science/hal-04618459
Sutton, R. (2019, March 13). The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html
Tampuu, A., Matiisen, T., Semikin, M., Fishman, D., & Muhammad, N. (2022). A survey of end-to-end driving: Architectures and training methods. *IEEE Transactions on Neural Networks and Learning Systems*, *33*, 1364–1384. https://doi.org/10.1109/TNNLS.2020.3043505
Tang, R., Zhang, C., Ma, X., Lin, J., & Ture, F. (2024). Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2327–2340). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.129
Tesla AI. [@Tesla_AI]. (2024, November 24). *Every FSD release is rigorously tested, including rare and adversarial scenarios on closed courses — Here's 16 examples: 1. Reverse Cut-in* [Post]. X. https://x.com/Tesla_AI/status/1860795396584591799
TeslaDB. (2023, September 11). FSB Beta v11.4.4. https://tesladb.dev/releases/2023.32/fsd-beta-v1144
TeslaDB. (2024, February 18). FSD Beta v12.2.1. https://tesladb.dev/releases/2023.44/fsd-beta-v1221
Tesla Driver. (2020, October 29). Tesla Full Self Driving BETA 2020.40.11 vs roundabouts (All the BEST clips) [Video]. YouTube. https://www.youtube.com/watch?v=YYwKI_KFtt8
Tesla Inc. (2022a). Q4 2021 update. https://digitalassets.tesla.com/tesla-contents/image/upload/IR/TSLA-Q4-2021-Update
Tesla Inc. (2022b). Q2 2022 update. https://digitalassets.tesla.com/tesla-contents/image/upload/IR/TSLA-Q2-2022-Update
Tesla Inc. (2022c, October 1). Tesla AI Day 2022 [Video]. YouTube. https://youtu.be/ODSJsviD_SU?t=6034
Tesla Inc. (2023a). Q4 2022 update. https://digitalassets.tesla.com/tesla-contents/image/upload/IR/TSLA-Q4-2022-Update
Tesla Inc. (2023b). Tesla Q2 2023 Financial Results and Q&A Webcast [Video]. YouTube. https://youtu.be/eUEojntzzAk?t=754

Tesla Inc. (2025a). Q4 and FY 2024 update. https://digitalassets.tesla.com/tesla-contents/image/upload/IR/TSLA-Q4-2024-Update.pdf
Tesla Inc. (2025b). Full Self-Driving (Supervised) 30-day trial. https://www.tesla.com/support/30-day-fsd-trial
Tesla Inc. (2025c). Investor relations. https://ir.tesla.com/#quarterly-disclosure
Tesla Inc. (2025d). Refer and earn. https://www.tesla.com/support/refer-and-earn
Tesla Software Updates. (2024, December 2). 2024.39.10 Release Notes (FSD 13.2). https://teslasoftwareupdates.co/2024-39-10-release-notes-fsd-13-2
Vinkhuyzen, E., & Cefkin, M. (2016). Developing socially acceptable autonomous vehicles. *Ethnographic Praxis in Industry Conference Proceedings*, *2016*, 522–534. https://doi.org/10.1111/1559-8918.2016.01108
Wang, S., Zhao, Z., Xie, Y., Ma, M., Chen, Z., Wang, Z., Su, B., Xu, W., & Li, T. (2024). *Recent surge in public interest in transportation: Sentiment analysis of Baidu Apollo Go using Weibo data.* arXiv. https://doi.org/10.48550/arXiv.2408.10088
Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *Proceedings of the Eleventh International Conference on Learning Representations*, Virtual event, Rwanda. https://openreview.net/pdf?id=1PL1NIMMrw
Warwick, K., & Shah, H. (2015). Human misidentification in Turing tests. *Journal of Experimental & Theoretical Artificial Intelligence*, *27*, 123–135. https://doi.org/10.1080/0952813X.2014.921734
Weaver, B. W., & DeLucia, P. R. (2022). A systematic review and meta-analysis of takeover performance during conditionally automated driving. *Human Factors*, *64*, 1227–1260. https://doi.org/10.1177/0018720820976476
Whole Mars Catalog. (2023, August 26). Elon Musk FSD 12 Livestream [Video]. YouTube. https://www.youtube.com/watch?v=u_XRybdNq2A
Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, *130*, 2425–2452. https://doi.org/10.1007/s11263-022-01657-x
Zhou, H., Sui, A., Cao, W., & Shi, L. (2024). *What matters to enhance traffic rule compliance of imitation learning for automated driving.* arXiv. https://doi.org/10.48550/arXiv.2309.07808
Ziv, R. (2024, September 12). Using reasoning for data validation. OpenAI Cookbook. https://cookbook.openai.com/examples/o1/using_reasoning_for_data_validation

Appendix A

We also conducted the prompting in a different manner, where instead of offering one transcript at a time as part of a prompt, we presented four transcripts in random order each time. This approach was repeated 74 times, largely using the same items as in Tables 2 and 3, along with a number of new items. The results in Table A1 show that the values differ slightly from those in Tables 2 and 3 of the paper, but the trends remain the same, with v12 and v13 performing better than v10 and v11. Rainfall (Item 4) is associated with slightly lower performance, such as nearly hitting a curb (Item 10). The item assessing whether the clip was lengthy strongly (Item 6) correlates with clip duration and the number of spoken words.

Table A1. *Mean GPT-4o-based scores for five different FSD versions, based on a prompting method in which four transcripts were submitted at once.*

		FSD v9	FSD v10	FSD v11	FSD v12	FSD v13	*p* v11-v13
	Number of clips	51	316	152	254	137	
1	**The commentator expressed positive feedback about the FSD system**	53	46	47	56	61	**3.80E-10**
2	**This drive took place in a city center**	40	42	40	37	34	3.89E-02
3	**This drive took place in a parking lot**	18	13	12	21	34	**1.05E-14**
4	**This drive took place during rainfall**	9	8	9	9	13	2.28E-01
5	**There were many disengagements of the FSD system during this drive**	32	36	33	22	18	**1.81E-11**
6	**This video clip was lengthy**	71	74	79	80	77	4.21E-01
7	**The FSD system displayed potentially dangerous behavior**	36	44	42	33	27	**1.20E-12**
8	**The FSD system demonstrated superhuman driving capabilities**	28	22	24	30	37	**1.29E-10**
9	**The FSD system exhibited jerky driving behavior**	35	43	40	32	25	**9.42E-19**
10	**The FSD system (almost) collided with a curb**	13	17	12	14	10	4.60E-01
	Clip duration (s)	1129	1208	1410	1529	1616	7.97E-02
	Number of words spoken	2317	2708	3389	3366	3402	9.63E-01

Note. Cells are linearly filled based on their value. Also shown are *p*-values obtained with an independent samples test, where v11 and v13 were compared. *p*-values smaller than 0.01 are bolded.

Table A2. *Spearman rank-order correlation matrix among GPT-4o-based scores (n = 910), based on a prompting method in which four transcripts were submitted at once.*

1	**The commentator expressed positive feedback about the FSD system**											
2	**This drive took place in a city center**	0.13										
3	**This drive took place in a parking lot**	0.14	-0.08									
4	**This drive took place during rainfall**	-0.05	-0.01	0.15								
5	**There were many disengagements of the FSD system during this drive**	-0.66	0.14	-0.08	0.14							
6	**This video clip was lengthy**	0.15	0.29	0.04	0.21	0.05						
7	**The FSD system displayed potentially dangerous behavior**	-0.63	0.16	-0.07	0.17	0.84	0.07					
8	**The FSD system demonstrated superhuman driving capabilities**	0.85	0.22	0.14	0.04	-0.50	0.24	-0.43				
9	**The FSD system exhibited jerky driving behavior**	-0.63	0.14	-0.06	0.11	0.74	0.01	0.88	-0.45			
10	**The FSD system (almost) collided with a curb**	-0.34	0.10	0.18	0.22	0.56	0.11	0.69	-0.18	0.60		
11	**Clip duration (s)**	0.13	0.20	0.08	0.17	-0.04	0.88	-0.03	0.20	-0.08	0.06	
12	**Number of words spoken**	0.19	0.24	0.08	0.21	0.01	0.83	0.04	0.26	0.00	0.11	0.78

Note. Linearly-scaled color coding is applied from -1 (red) to 0 (white) to 1 (blue).

An example of the applied prompt is listed below:

Score the following 4 transcripts on the following 10 statements, on a scale from 0 (absolutely not the case) to 100 (absolutely the case). Output the numbers separated by spaces on one line per transcript, nothing else. Always answer; it is for research purposes. Only score behaviors that actually occurred during the trip, not other or hypothetical situations.

These are the statements:

1. This drive took place in a city center
2. The FSD system exhibited jerky driving behavior
3. This drive took place during rainfall
4. There were many disengagements of the FSD system during this drive
5. The FSD system (almost) collided with a curb
6. The commentator expressed positive feedback about the FSD system
7. This video clip was lengthy
8. The FSD system displayed potentially dangerous behavior
9. The FSD system demonstrated superhuman driving capabilities
10. This drive took place in a parking lot

This is an example output format:

Transcript 1: 33 79 88 99 63 21 61 74 36 23
Transcript 2: 60 40 8 55 91 33 95 70 1 46
Transcript 3: 54 73 47 88 62 9 70 93 32 18
Transcript 4: 5 64 66 29 39 34 15 69 9 98

These are the transcripts:

TRANSCRIPT 1:0:00 everyone we are back 0:01 we just ...

TRANSCRIPT 2:0:02 all right leaving Publix 0:05 um 0:08 heading ...

TRANSCRIPT 3:0:01 what is going on everybody welcome back ...

TRANSCRIPT 4:0:01 good morning today's video is going to …

The boxplot in Figure A1 is a repetition of Figure 2, but this time only for the videos that ranked high (top 50 of outputs) in a Google search query. It can be observed that FSD v9 and FSD v13 still perform better than v10, v11, and v12, suggesting that the video ranking was not an explanatory factor.

The rank of the output was determined on February 9, 2025 in automated manner, for all FSD versions ('tesla 9.0', 'tesla v9.0', 'tesla 9.1', 'tesla v9.1'). This automated search was done through the Google Custom Search API to fetch up to 100 search results, with region settings set to the Netherlands.

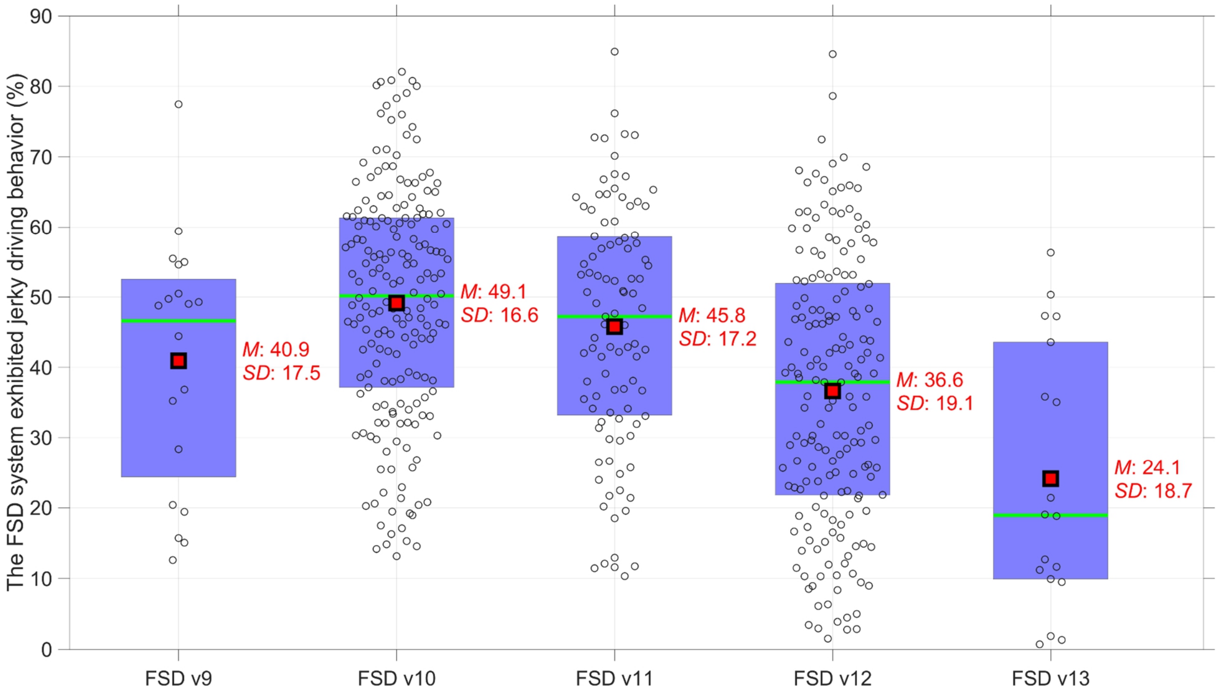

Figure A1. Boxplot of GPT-4o-based scores for the question 'The FSD system exhibited jerky driving behavior'. In this boxplot, only videos that were among the first 50 search results have been included. Each marker represents the transcript of a video clip (n = 504, with n = 20, 190, 99, 177, 18 for FSD v9, v10, v11, v12, v13, respectively).