Sentiment Matters: An Analysis of 200 Human-SAV Interactions

Lirui Guo[1], Michael G. Burke[2], and Wynita M. Griggs[1,2]

***Abstract*— Shared Autonomous Vehicles (SAVs) are likely to become an important part of the transportation system, making effective human–SAV interactions an important area of research. This paper introduces a dataset of 200 human-SAV interactions to further this area of study. We present an open-source human–SAV conversational dataset, comprising both textual data (e.g., 2,136 human–SAV exchanges) and empirical data (e.g., post-interaction survey results on a range of psychological factors). The dataset's utility is demonstrated through two benchmark case studies: First, using random forest modeling and chord diagrams, we identify key predictors of SAV acceptance and perceived service quality, highlighting the critical influence of response sentiment polarity (i.e., perceived positivity). Second, we benchmark the performance of an LLM-based sentiment analysis tool against the traditional lexicon-based TextBlob method. Results indicate that even simple zero-shot LLM prompts more closely align with user-reported sentiment, though limitations remain. This study provides novel insights for designing conversational SAV interfaces and establishes a foundation for further exploration into advanced sentiment modeling, adaptive user interactions, and multimodal conversational systems.**

I. Introduction

The rapid advancement of autonomous vehicle (AV) technology is transforming urban mobility, with shared autonomous vehicles (SAVs) anticipated to play a central role in future transportation systems [1], [2]. As SAVs transition from controlled experimental settings to real-world deployment, understanding the dynamics of human interaction with these vehicles and identifying key predictors of user acceptance becomes increasingly critical [3], [4]. In the context of high autonomy levels (SAE Levels 4/5), where vehicles are operated without the involvement of human drivers, successful adoption is influenced by a combination of advanced technical capabilities and effective human-centric interactions driven by seamless communication interfaces [5]. In particular, conversational interactions may serve as a key interface for trip requests, status updates, and in-ride assistance.

However, current research has largely overlooked the role of conversational dialogue in shaping psychological aspects such as psychological ownership, perceived service quality, and overall user acceptance. Prior studies on human–SAV interactions have primarily utilized structured surveys, driving simulations (e.g., Wizard-of-Oz protocols, motion-based simulators, and virtual reality environments), and assessments focusing on vehicle interior and exterior design [3], [6]–[8]. Although valuable, these methods typically focus on static or predefined scenarios, offering limited insight into the real-time dynamic conversational interactions users will experience in fully autonomous vehicles.

Recent attempts to bridge this gap have begun assembling conversational interaction datasets. For instance, [9] developed a multimodal Wizard-of-Oz dataset consisting of 30 hours of in-cabin interactions (10,590 utterances) collected from 30 participants during a scavenger-hunt activity, integrating speech, audio, and visual data to enhance intent recognition within vehicle cabins. Similarly, [10] introduced the doScenes dataset, pairing 1,000 nuScenes driving clips with natural-language instructions to enable voice-driven AV planning. However, neither of these datasets specifically addresses unconstrained conversational exchanges between users and SAV agents, such as analyzing how users perceive the sentiment or appropriateness of SAV responses to their requests.

Meanwhile, large language models (LLMs) like ChatGPT, have emerged as powerful tools capable of generating human-like conversational experiences across diverse domains, including healthcare, education, and intelligent vehicles [11]–[14]. Using LLMs for both conversational data collection and analysis enables granular measurements of user sentiment like polarity and subjectivity directly from dialogue [15].

In this study, we present a novel dataset comprising conversational interactions between 50 participants and four different SAE Level 5 SAV agents, simulated using GPT-3.5 turbo with varying prompting strategies. Our dataset includes:

- **Conversational Textual Data**: 2,136 request-response exchanges, accompanied by open-ended and interview insights.
- **Empirical Survey Data**: Structured post-interaction responses capturing user perceptions of a range of psychological factors, and self-reported sentiment measures (Polarity and Subjectivity) toward SAV responses.

To illustrate the utility of our dataset, we provide two benchmark case studies:

1) **Case Study 1 — SAV Acceptance Analysis** (Empirical Survey Data): We apply the Machine-Learning-Chord-Diagram framework proposed by [4], employing random forest modeling to analyze and visualize the most influential item-level

*This work was supported by an Australian Government Research Training Program (RTP) Scholarship.

[1]Department of Civil and Environmental Engineering, Monash University, Clayton, VIC 3800, Australia.

[2]Department of Electrical and Computer Systems Engineering, Monash University, Clayton, VIC 3800, Australia.

Email: {Lirui.Guo, Michael.G.Burke, Wynita.Griggs}@monash.edu

predictors of service quality and intention to use SAVs.

2) **Case Study 2 — Sentiment Analysis** (Conversational Textual Data): We use an LLM-based sentiment classifier on individual conversational exchanges, extracting metrics (minimum, maximum, mean, median) for polarity and subjectivity. We then benchmark performance against a traditional lexicon-based sentiment analysis tool (TextBlob) and self-reported sentiment scores.

The remainder of this paper is organized as follows. Section II details our dataset creation process and methodology. Section III presents the first case study, analyzing acceptance predictors derived from survey data. Section IV evaluates LLM-based sentiment analysis performance in the second case study. Finally, Section V concludes with a discussion of study limitations and recommendations for future research.

II. Dataset & Methodology

This section describes the design of the user study, the data collection approach, the structure of the resulting dataset, and details regarding data accessibility.

A. User Study Design and Data Collection

The user study was developed to investigate how different conversational styles and prompting strategies influence user perceptions, specifically regarding Psychological Ownership (PO), Anthropomorphism (A), Quality of Service (QoS), Disclosure Tendency (DT), Perceived Enjoyment (PE), Behavioral Intention (BI), and the sentiment (Polarity and Subjectivity) associated with SAV responses. Each participant interacted with four simulated SAE Level 5 SAV agents, powered by OpenAI's `gpt-3.5-turbo` model. The SAV user interface (UI) was built in Python using the `gradio` package and integrated with the OpenAI API [16], [17]. The SAV agents were differentiated by distinct prompting strategies as follows:

- **SAV 1 (Standard/Control):** Provided baseline functionality such as navigation assistance, climate control adjustments, and media management, without additional personalization.
- **SAV 2 (Standard + Psychological Ownership):** Enhanced interactions designed explicitly to foster users' sense of ownership over the vehicle.
- **SAV 3 (Standard + Anthropomorphism):** Participants selected a preferred anthropomorphic personality (e.g., friendly, sassy, or cool), facilitating personalized and human-like interactions.
- **SAV 4 (Combined PO + Anthropomorphism):** Integrated both psychological ownership and anthropomorphic strategies, aiming to reinforce personalization and ownership engagement simultaneously.

All SAV agents were explicitly instructed to comply with Australian national traffic laws and specific regulations applicable within the State of Victoria, aiming for realistic and safe responses throughout the interactions. The exact wording of the prompts is provided in the dataset documentation.

Data collection was conducted with ethical approval from the Monash University Human Research Ethics Committee (MUHREC)[1]. The user studies were held from April 30, 2024, to July 2, 2024. A total of 50 participants were recruited from an Australian university community and external networks. The study employed a within-subjects experimental design, meaning each participant interacted with all four SAV agents. To minimize potential ordering effects or bias due to interaction sequence, the order in which participants experienced each SAV agent was randomized.

Descriptive statistics, theoretical background and literature on the psychological factors investigated, a detailed overview of the prompt design, statistical comparisons of the effects of psychological ownership and anthropomorphic strategies on user experience, and qualitative analysis of participants' feedback on perceived psychological ownership are provided in [18]. The findings indicated that SAV4, which combined both strategies, was perceived as more human-like and elicited more positive user responses overall. In this paper, we present the complete study design, release the full de-identified conversational dataset, and analyze the role of sentiment in human–SAV interactions through two benchmark case studies.

The system architecture for conversational interactions is illustrated in Fig. 1.

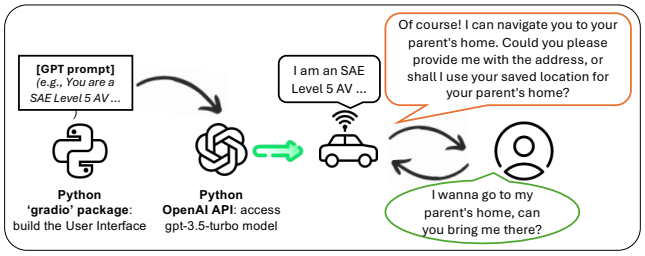


Fig. 1: System architecture of the simulated SAV agent.

Prior to their interactions, participants completed a pre-interaction survey capturing their prior experiences with AI systems and their preferred anthropomorphic personality style for the SAV (Cool & sophisticated; Engaging & friendly; or Sassy & tired). Preferences expressed in this survey informed the configuration of SAV3 and SAV4 during interactions.

A set of 15 in-vehicle interaction commands adapted from those commonly used in existing vehicles such as the Tesla Model Y [19] was provided to participants as example requests. These commands covered key vehicle functions, including climate control, vehicle settings, navigation, media controls, and contact management. A complete list of these commands is available in the dataset documentation. Participants were also encouraged to interact with the SAVs freely and make requests in their own words, as they would naturally do in real-life scenarios.

[1]Project ID: 40485

After each SAV interaction, participants completed a structured post-interaction survey designed to measure perceptions across the selected psychological factors. Additionally, participants were briefly interviewed to capture qualitative feedback on their experiences, particularly perceptions of psychological ownership relative to previous SAV interactions. All conversational exchanges were recorded and stored.

At the end of the entire interaction sequence (all four SAVs), participants completed a final survey collecting demographic information and responded to open-ended questions exploring features that contribute to psychological ownership, perceptions about sharing personal information and other additional insights or feedback.

The complete survey instrument used in this study can be found in the released dataset folder.

Overall, the collected data comprised:

- 2,136 conversational exchanges (request–response pairs) between users and SAV agents;
- 200 structured post-interaction survey responses (four responses per participant);
- 50 sets of open-ended qualitative statements or interview transcripts (one per participant).

Collected data was anonymized to protect participant privacy and analyzed inline with the approved ethics protocol. Even after broad recoding, several demographic records remained unique (k = 1), and so the dataset failed k-anonymity and we removed the demographic section from the public release [20]. The demographic overview is available in [18].

B. Data Structure and Availability

The fully de-identified dataset and accompanying documentation will be publicly accessible via the Monash Bridges repository, released under a CC BY 4.0 license. All files are provided in widely used accessible formats (.xlsx, .pdf, .pptx), facilitating easy reuse and analysis by the research community.

1) Data Structure: The dataset comprises the following:

- **Data_human-SAV_interaction.xlsx:** Contains 2,136 conversational exchanges (request–response pairs) between users and SAV agents.
- **Data_Survey1-4.xlsx:** Structured responses from post-interaction surveys (four per participant), capturing psychological ownership, anthropomorphism, quality of service, disclosure tendency, perceived enjoyment, behavioral intention, and self-reported sentiment measures (polarity and subjectivity) regarding SAV responses.
- **Data_survey_open&end.xlsx:** Pre-interaction survey data assessing participants' prior experience with AI systems, their preferred anthropomorphic SAV personalities (e.g., friendly, cool), and responses to the final survey completed at the end of the study.
- **Data_interview_and_openendedQ.xlsx:** Transcribed qualitative data from participant interviews and responses to open-ended survey questions, capturing nuanced user feedback and perceptions.
- **Prompts.pptx:** Prompts used for developing each of the four SAV agents.
- **Sample_user_input.pdf:** Representative examples of user commands, illustrating navigation requests, comfort adjustments, and emotional prompts.
- **Survey.pdf:** Full survey instrument utilized during the study.

2) Access and Reuse: The dataset will be hosted for open and persistent access.

- **URL:** https://doi.org/10.26180/29486447
- **DOI:** 10.26180/29486447
- **License:** CC BY 4.0 (Creative Commons Attribution)

III. Case Study 1: Item Importance in Predicting SAV Acceptance

A. Motivation

Identifying specific aspects of user experience that most significantly influence SAV acceptance is crucial for designing interfaces and interaction strategies that foster user satisfaction and sustained use. Prior research has examined various psychological factors influencing public acceptance of SAVs and human-computer interaction more broadly, including Psychological Ownership (PO) [2], [21], Anthropomorphism (A) [22], Disclosure Tendency (DT) [23], Quality of Service (QoS) [2], Perceived Enjoyment (PE) [24], and sentiment (Polarity and Subjectivity) [25].

Most previous studies have applied traditional statistical methods, such as Structural Equation Modeling (SEM), relying primarily on aggregate scores to explore these relationships. Recent research, however, has demonstrated the effectiveness of machine learning methods, such as Neural Networks and Random Forests, in predicting user acceptance of AVs [4], [26]. To overcome limitations associated with traditional SEM, particularly its inability to identify item-level predictors, [4] introduced a predictive modeling framework combined with chord diagrams, enabling researchers to visualize the relative importance of specific questionnaire items. This approach uncovers fine-grained insights into drivers of various psychological factors and behavioral intentions, providing actionable information for SAV interface designers and service providers.

B. Prediction and visualization framework

In this study, we applied the predictive modeling and visualization framework proposed by [4] to our post-interaction survey dataset. The framework predicts each psychological factor by averaging its corresponding survey items using Random Forest models. For example, the Quality of Service factor is computed from items QoS1–QoS3, which measure user's perceived SAV service quality (QoS1), communication pleasantness (QoS2), and recommendation likelihood (QoS3). Full item descriptions are provided in the dataset documentation. The relative importance of each item in predicting these factor-level scores was visualized through chord diagrams. An example of how to interpret these diagrams is given in Fig. 2.

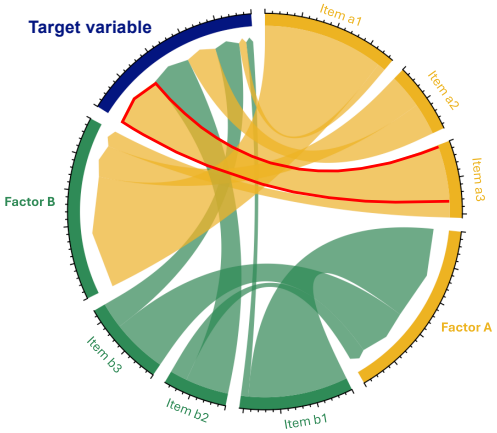


Fig. 2: Example chord diagram illustrating item-level importance in predicting a target variable. Two latent factors are shown, each comprising multiple items used as predictors. The width of each arc represents the relative importance of the item, normalized such that the total importance across all predictors sums to 100%. In this example, Item A3 emerges as the most influential predictor among the six items.

Comparisons across the four SAVs reveal variations in user perceptions attributable to different prompting strategies. The resulting relative item importance is shown visually in Fig. 3.

C. Psychological ownership prompt effect - SAV1 vs. SAV2

When predicting Behavioral Intention to use SAVs, the relative importance of the factor items for SAV1 and SAV2 is visualized in Fig. 3a and 3b, respectively. In SAV1, the most influential item is QoS3 (recommend to others, with a relative importance of 32.2%), followed by QoS1 (SAV service, 11.5%) and DT2 (tastes in music, 10.5%). In contrast, SAV2 demonstrates a different pattern, with Polarity emerging as the most influential item (22.4%), followed by QoS1 (SAV service, 22.0 %) and QoS3 (recommend to others, 14.7%). These findings suggest that integrating psychological ownership prompts increases the influence of the Polarity of SAV responses, indicating that users pay more attention to the sentiment of the SAV responses. In both groups, Polarity plays a key role in predicting Quality of Service (QoS), with relative importance of 44.7% and 57.9% respectively, suggesting that users' perceptions of SAV responses significantly shape their overall evaluation of the service.

Similarly, when predicting user's Disclosure Tendency, A5 (feel compassion, 34.4%) emerged as the most important predictor for SAV1, followed by PO3 (interaction enhances my PO, 14.2%). While for SAV2, QoS2 (communication is pleasant, 11.4%) and A5 (feel compassion, 10.1%) are the most influential items. This reveals that after integrating psychological ownership prompts, although the beliefs of SAV can feel compassion is still important, the pleasantness of communication becomes a more significant factor.

Another notable difference arises when predicting Psychological Ownership. In SAV1, the most influential items are A2 (personality, 10.0%), A3 (preferences and mood, 9.9%), and DT3 (feelings and emotions, 9.1%). In contrast, for SAV2, the most influential items are PE1 (fun to interact, 12.2%), A3 (preferences and mood, 11.1%), and A2 (personality, 11.0%). These results indicate that while anthropomorphism-related items, particularly beliefs about the SAV having its own personality and preferences, are important in predicting psychological ownership for both SAV groups, other factors differ. In SAV1, the willingness to disclose personal feelings and emotions (DT3) plays a more important role, whereas in SAV2, the enjoyment of interaction (PE1) becomes a more critical factor.

D. Anthropomorphism prompt effect - SAV1 vs. SAV3

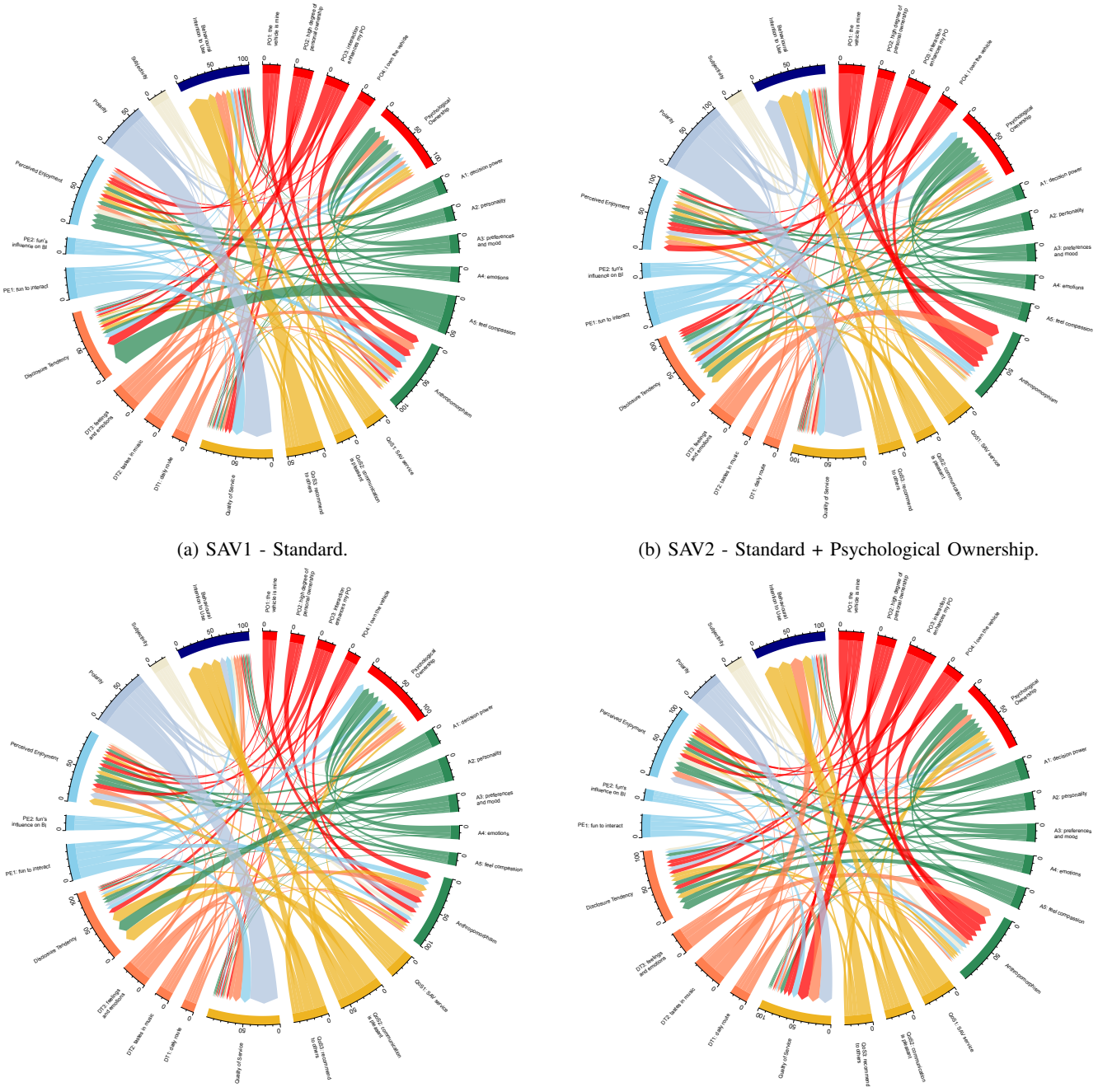
When comparing the relative importance of factor items for SAV1 and SAV3 (Fig. 3c), diverse patterns emerge. In predicting Behavioral Intention to use SAVs, the most influential items for SAV3 are QoS1 (SAV service, 20.5%), QoS2 (communication is pleasant, 17.3%), and QoS3 (recommend to others, 14.1%). Compared to SAV1, participants appear to place greater emphasis on whether communication with the SAV is pleasant. Similar to the psychological ownership prompt effect, the Polarity of SAV responses remains the most important predictor of QoS in both groups, with a relative importance of 44.7% in SAV1 and 43.0% in SAV3.

In predicting Disclosure Tendency, the most influential items in SAV3 are QoS2 (communication is pleasant, 19.0%), A2 (personality, 17.7%), and QoS3 (recommend to others, 16.6%), which differ from the most important predictors in SAV1 (A5, PO3, PE1). This shift suggests that, after integrating anthropomorphism prompts, the QoS items, particularly the pleasantness of communication and the likelihood of recommending the SAV to others, become more critical in predicting users' willingness to disclose personal information and emotions.

Surprisingly, when predicting Psychological Ownership, the most influential items in SAV3 are similar to but distinct from those in SAV2. For SAV3, the most important predictors are PE1 (fun to interact, 10.7%), A4 (emotions, 10.2%), and A1 (decision power, 9.5%). Compared to the control group (i.e., SAV1, where A2, A3, and DT3 are the most important predictors), these results indicate that, following the integration of anthropomorphism prompts, factors such as the enjoyment of interaction and beliefs about the SAV having its own emotions and decision-making power play a more significant role in predicting Psychological Ownership.

E. Combined prompt effect - SAV4

The combined group, SAV4 (Fig. 3d), integrates both PO and A prompts. When predicting Behavioral Intention to use SAVs, the most influential items are QoS1 (SAV service, 23.0%), QoS3 (recommend to others, 19.5%), and DT2 (tastes in music, 18.9%). These results are similar to SAV1 but show a higher relative importance for QoS1 and DT2, and a lower relative importance for QoS3. This suggests that while the quality of service and the likelihood of recommending the SAV remain important predictors across



(a) SAV1 - Standard.

(b) SAV2 - Standard + Psychological Ownership.

(c) SAV3 - Standard + Anthropomorphism.

(d) SAV4 - Standard + PO + A

Fig. 3: Chord Diagrams Showing Relative Importance of Predicting Target Factors. The arrows (chords) connect the predictor items (start nodes) and the target variables (end nodes). The width of the chords indicates the relative importance of each predictor item. The sum of the relative importance of each target variable (i.e., the overall questions) is 100%.

all four SAVs, the integration of both prompts enhances the influence of users' willingness to disclose their music preferences, unlike the focus on Polarity seen in SAV2 and SAV3.

When predicting Psychological Ownership, although Anthropomorphism items still play a significant role, the willingness to disclose daily routes (DT1, 8.7%) emerged as the third most important predictor. Interestingly, when predicting QoS, while the Polarity of SAV responses remains the most influential item (20.1%), its relative importance is lower than in SAV1, SAV2, and SAV3. This suggests that while the sentiment of SAV responses continues to be a critical factor, the integration of psychological ownership and anthropomorphism prompts shifts importance toward other factors, such as willingness to disclose music preferences (DT2, 16.9%) and the perception of interaction-enhanced psychological ownership (PO3, 16.6%).

Notably, the integration of psychological ownership

prompts (in SAV2 and SAV4) increases the overall relative importance of the Psychological Ownership factor in predicting other target variables. The relative importance is higher in SAV2 (116.1%) and SAV4 (126.2%) compared to SAV1 (105.7%) and SAV3 (82.4%). This finding suggests that psychological ownership prompts reinforce the influence of psychological ownership on users' perceptions of other factors, such as quality of service and anthropomorphism.

IV. Case Study 2: Using LLMs as Sentiment Analysis Tools

A. Motivation

Case Study 1 showed that conversational **sentiment polarity** is the strongest item-level predictor of perceived Quality of Service (QoS) across all four SAVs, and that QoS emerged as the most critical factor predicting Behavioral Intention (BI) to use SAVs. Even when aggregating importance at the factor-level, polarity remained notably influential, especially for SAV1, SAV2, and SAV3. These results point to a practical next step: *if polarity drives user perceptions, can we measure it automatically?* Specifically, can we reliably derive polarity and subjectivity measures (i.e., sentiment analysis metrics) directly from conversational text to predict user ratings? Addressing this question is the focus of Case Study 2, where we benchmark two sentiment-analysis approaches and evaluate how well their outputs track participants' self-reported sentiment.

Sentiment analysis integrates natural language processing (NLP), computational linguistics, and text analytics to extract and interpret emotional tone from textual data [27]. Traditional sentiment tools, such as TextBlob, adopt a lexicon-based method suitable for sentence-level evaluations. TextBlob offers straightforward interfaces for various NLP tasks – including sentiment scoring (polarity and subjectivity), part-of-speech tagging, and noun phrase extraction [28]. Polarity scores range from –1 (extremely negative) to 1 (extremely positive), while subjectivity ranges from 0 (highly objective) to 1 (highly subjective). With recent advancements in LLMs, researchers have begun exploring their effectiveness in sentiment analysis tasks. For example, [15] demonstrated that zero-shot LLM approaches, such as GPT-3.5 and GPT-4, can match or exceed traditional transfer learning methods across various sentiment benchmark datasets.

In this case study, we evaluate two sentiment detection strategies on conversational data collected during human–SAV interactions: a traditional lexicon-based tool (TextBlob) and an LLM-based sentiment analysis method. We subsequently assess the alignment between these automated sentiment scores and the sentiment perceptions participants reported in post-interaction surveys.

B. Experiment design and evaluation metrics

We used the `TextBlob` package (version 0.19.0) in Python. For the LLM-based sentiment analysis, we employed `gpt-4o-mini` using a zero-shot prompt, as shown below.

Both sentiment analysis methods were applied independently to every SAV response within each participant's conversation. For instance, if Participant P01 completed 15 request-response exchanges with SAV1, both methods produced 15 sentiment evaluations (each containing polarity and subjectivity scores). Aggregate statistics (minimum, maximum, mean, median, and mode) were then computed from these evaluations to create representative sentiment measures for each participant–SAV interaction.

Zero-Shot LLM Sentiment-Analysis Prompt

```
You are an advanced Sentiment Analysis
Model for evaluating responses from a
Shared Autonomous Vehicle (SAV). For
each SAV response, analyze its explicit
and implied sentiment, and then output
two scores:
   1) Polarity Score: A number from
      −1 (extremely negative) to 1
      (extremely positive).
   2) Subjectivity Score: A number from
      0 (highly objective) to 1 (highly
      subjective).
Instructions:
Evaluate its overall emotional tone and
degree of personal bias.
Return only a valid JSON object
with the keys "Polarity Score" and
"Subjectivity Score" - no additional
text.
Example JSON structure: "Polarity
Score": <number>, "Subjectivity Score":
<number> .
```

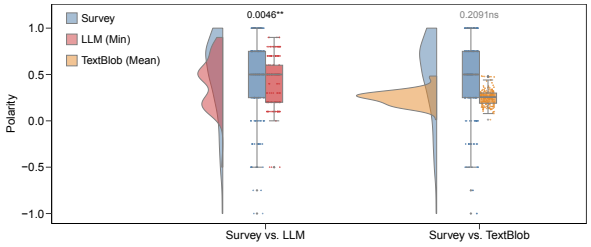
The non-parametric Spearman's rank correlation coefficient was employed to assess the relationship between the aggregated sentiment metrics and the corresponding survey-based sentiment ratings provided by participants [29]. Correlation outcomes are summarized in Table I. The sentiment features with the highest correlation from each method were visualized against survey-based distributions in density plots (Figs. 4a and 4b).

TABLE I: Spearman Correlation (r) Between Computed Sentiment Features and Survey-Based Ratings. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

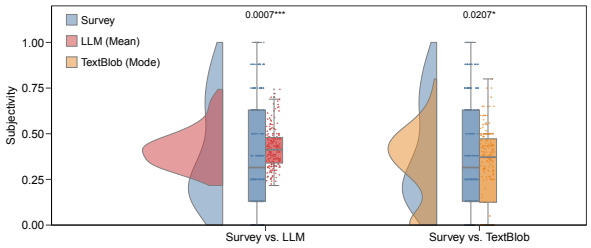
Method	r_{polarity}	$r_{\text{subjectivity}}$
`llm_min`	**0.199****	0.010
`llm_mean`	**0.182***	**0.237****
`textblob_mean`	0.089	0.032
`textblob_min`	0.089	0.006
`llm_median`	0.075	**0.217****
`textblob_median`	0.060	-0.056
`textblob_max`	0.021	0.076
`textblob_mode`	0.002	**0.163***
`llm_mode`	-0.001	0.131
`llm_max`	-0.089	**0.145***

C. Findings and Implications

Our results indicate that the LLM-based minimum polarity measure (`llm_min`) achieved the strongest correlation with



(a) Polarity.



(b) Subjectivity.

Fig. 4: Raincloud plots comparing survey scores to the highest-correlating sentiment features (among min, max, mean, median, mode) from LLM and TextBlob. Each combines density, boxplot, and jittered points, with Spearman p-values annotated. Kernel density estimates were computed using Scott's rule for bandwidth selection.

users' self-reported polarity ($r = 0.199$, $p = 0.005$) (Table I). This result is visually reinforced by the density plot in Fig. 4a, which shows that the LLM polarity distribution more closely matches the survey-based distribution than TextBlob's distribution. Similarly, for subjectivity, the LLM-derived mean score (`llm_mean`) had the highest correlation with survey-based subjectivity ratings ($r = 0.237$, $p = 0.001$) (Table I), outperforming TextBlob's best-performing measure (`textblob_mode`, $r = 0.163$, $p = 0.021$). In both sentiment dimensions, the LLM-based method consistently outperformed TextBlob at statistically significant levels $p < 0.05$.

Despite modest correlation magnitudes ($r \approx 0.2$), these results offer several meaningful insights:

1) **Holistic user evaluation**: Users appear to rate sentiment based on the overall conversational experience, encompassing factors beyond individual responses. Modest correlation values imply that text-only sentiment metrics alone are insufficient for capturing the complete user experience, which likely includes context, multimodal cues, and interpersonal interaction nuances.
2) **Sensitivity to extreme sentiment**: The prominence of the minimum polarity score indicates users are likely affected by the most negative ("worst-case") exchanges during the interactions.
3) **LLM efficacy**: Even with a simple zero-shot prompt, the LLM sentiment method significantly outperformed TextBlob, demonstrating strong potential for conversational sentiment analysis.

These findings underscore the potential of LLM-based sentiment analysis as a powerful and efficient tool for evaluating conversational interactions. However, the modest correlation between LLM-derived sentiment scores and participants' self-reported ratings also reveals its current limitations. Surprisingly, the LLM-based tool did not accurately reflect users' perceived sentiment toward SAV responses. Two possible explanations emerge: First, as previously discussed, participants' self-ratings may have reflected their aggregate experience over the entire interaction, particularly influenced by notably negative exchanges such as the refusal of a request, rather than the sentiment of isolated SAV responses. Participants may disproportionately emphasize negative experiences or interactions that fail their expectations, a phenomenon in social psychology whereby negative experiences have a stronger impact on overall evaluations than equivalent positive ones [30], [31]. Second, LLMs may still lack the capacity to fully capture the nuanced sentiment embedded in complex, real-world dialogues. These limitations point to promising directions for future research, including the integration of richer contextual and multimodal cues, such as dialogue history, speech prosody, and facial expressions – to improve the accuracy and depth of sentiment modeling in human–SAV interactions.

The identified importance of sentiment polarity offers actionable insights for SAV interface designers. As demonstrated in Case Study 1, SAV response polarity plays an important role in influencing perceived service quality. This suggests that LLM-powered sentiment analysis tools could be effectively integrated into autonomous vehicle systems to enable real-time sentiment monitoring. When signs of user dissatisfaction or confusion are detected, the system could proactively respond to negative sentiment dips by deploying adaptive strategies to maintain a positive and stable user experience. These targeted design strategies, grounded in sentiment polarity insights, offer a promising pathway toward emotionally responsive and user-centered SAV interfaces.

V. Conclusions

This paper introduced an open-source human–SAV interactions dataset. The dataset features diverse prompting strategies designed to elicit psychological ownership and anthropomorphic engagement, and includes both structured survey responses and rich conversational textual data. It offers insights into key user perceptions such as psychological ownership, anthropomorphism, quality of service, disclosure tendency, perceived enjoyment, behavioral intention, and sentiment.

We demonstrated the utility of this dataset through two benchmark case studies. In Case Study 1, we applied Random Forest modeling to the survey data to examine item-level predictors of SAV acceptance, revealing how different conversational strategies influence user experience. In Case Study 2, we analyzed the conversational data using both a LLM-based sentiment analysis tool and TextBlob, comparing their alignment with user-reported sentiment ratings.

Several promising directions emerge for future work. First, this screen-based prototype could be replicated in realistic

driving environments to capture richer interaction cues. Second, future studies should recruit more diverse participants to improve generalization. Third, benchmarking additional LLM architectures and advanced sentiment analysis methods could address potential model biases and improve prediction accuracy. Lastly, randomizing or removing example commands could encourage more realistic and spontaneous user interactions. These extensions will further clarify sentiment's role in shaping user experience to inform SAV design.

We invite researchers to further explore this dataset in the context of human–SAV interaction. Future studies could, for example, investigate the predictive power of conversational text in estimating perceived quality of service, or develop more advanced prompting techniques to enhance the performance of LLM-based sentiment analysis, building upon the findings of Case Study 2.

REFERENCES

[1] K. Merfeld, M.-P. Wilhelms, S. Henkel, and K. Kreutzer, "Carsharing with shared autonomous vehicles: Uncovering drivers, barriers and future developments – A four-stage Delphi study," *Technological Forecasting and Social Change*, vol. 144, pp. 66–81, Jul. 2019.

[2] J. Dai, R. Li, Z. Liu, and S. Lin, "Impacts of the introduction of autonomous taxi on travel behaviors of the experienced user: Evidence from a one-year paid taxi service in Guangzhou, China," *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103311, Sep. 2021.

[3] T. Zhang, D. Tao, X. Qu, X. Zhang, R. Lin, and W. Zhang, "The roles of initial trust and perceived risk in public's acceptance of automated vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 98, pp. 207–220, Jan. 2019.

[4] L. Guo, M. G. Burke, and W. M. Griggs, "A new framework to predict and visualize technology acceptance: A case study of shared autonomous vehicles," *Technological Forecasting and Social Change*, vol. 212, Mar. 2025.

[5] P. A. M. Ruijten, J. M. B. Terken, and S. N. Chandramouli, "Enhancing Trust in Autonomous Vehicles through Intelligent User Interfaces That Mimic Human Behavior," *Multimodal Technologies and Interaction*, vol. 2, no. 4, p. 62, Dec. 2018.

[6] X. Zou, S. O'Hern, B. Ens, S. Coxon, P. Mater, R. Chow, M. Neylan, and H. L. Vu, "On-road virtual reality autonomous vehicle (VRAV) simulator: An empirical study on user experience," *Transportation Research Part C: Emerging Technologies*, vol. 126, p. 103090, May 2021.

[7] A. Zhang and P.-L. Patrick Rau, "Tools or peers? Impacts of anthropomorphism level and social role on emotional attachment and disclosure tendency towards intelligent agents," *Computers in Human Behavior*, vol. 138, p. 107415, Jan. 2023.

[8] C. Wang, W. Ren, C. Xu, N. Zheng, C. Peng, and H. Tong, "Exploring the impact of conditionally automated driving vehicles transferring control to human drivers on the stability of heterogeneous traffic flow," *IEEE Transactions on Intelligent Vehicles*, pp. 1–17, 2024.

[9] E. Okur, S. H Kumar, S. Sahay, and L. Nachman, "Audio-visual understanding of passenger intents for in-cabin conversational agents," in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, A. Zadeh, L.-P. Morency, P. P. Liang, and S. Poria, Eds. Seattle, USA: Association for Computational Linguistics, Jul. 2020, pp. 55–59. [Online]. Available: https://aclanthology.org/2020.challengehml-1.7/

[10] P. Roy, S. Perisetla, S. Shriram, H. Krishnaswamy, A. Keskar, and R. Greer. doScenes: An autonomous driving dataset with natural language instruction for human interaction and vision-language navigation. [Online]. Available: http://arxiv.org/abs/2412.05893

[11] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940. [Online]. Available: https://www.nature.com/articles/s41591-023-02448-8

[12] L. Hua, N. Zheng, Y. Lu, L. Guo, and J. Xu, "Use of large language models in engineering education: A case study on infrastructure design report introductions," in *Proceedings of AAEE 2024*. [Online]. Available: https://easychair.org/publications/preprint/XpQv/download

[13] H. Du, S. Teng, H. Chen, J. Ma, X. Wang, C. Gou, B. Li, S. Ma, Q. Miao, X. Na, P. Ye, H. Zhang, G. Luo, and F.-Y. Wang, "Chat With ChatGPT on Intelligent Vehicles: An IEEE TIV Perspective," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 3, pp. 2020–2026, Mar. 2023, conference Name: IEEE Transactions on Intelligent Vehicles.

[14] M. Yuksekgonul, F. Bianchi, J. Boen, S. Liu, P. Lu, Z. Huang, C. Guestrin, and J. Zou, "Optimizing generative AI by backpropagating language model feedback," *Nature*, vol. 639, no. 8055, pp. 609–616. [Online]. Available: https://www.nature.com/articles/s41586-025-08661-4

[15] J. O. Krugmann and J. Hartmann, "Sentiment analysis in the age of generative AI," vol. 11, no. 1, p. 3. [Online]. Available: https://doi.org/10.1007/s40547-024-00143-4

[16] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ML models in the wild," Jun. 2019.

[17] OpenAI, "Openai python api library, version 0.27.7," 2024, accessed: Nov. 4, 2024. [Online]. Available: https://platform.openai.com/docs.

[18] L. Guo, M. G. Burke, and W. M. Griggs. Exploring human-SAV interaction using large language models: The impact of psychological ownership and anthropomorphism on user experience. [Online]. Available: http://arxiv.org/abs/2504.16548

[19] Tesla, Inc., "Model Y Owner's Manual," *Tesla*, Aug. 2023, [Online]. [Online]. Available: https://www.tesla.com/ownersmanual/modely/en_us/GUID-EA1715B0-A3A6-454E-995A-8AA2C3A32D44.html

[20] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 571–588, Oct. 2002.

[21] J. Lee, D. Lee, Y. Park, S. Lee, and T. Ha, "Autonomous vehicles can be shared, but a feeling of ownership is important: Examination of the influential factors for intention to use autonomous vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 411–422, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X19301895

[22] M. Wu, N. Wang, and K. F. Yuen, "Deep versus superficial anthropomorphism: Exploring their effects on human trust in shared autonomous vehicles," *Computers in Human Behavior*, vol. 141, p. 107614, Apr. 2023.

[23] X. Cheng, L. Su, X. R. Luo, J. Benitez, and S. Cai, "The good, the bad, and the ugly: Impact of analytics and artificial intelligence-enabled personal information collection on privacy and participation in ridesharing," *European Journal of Information Systems*, vol. 31, no. 3, pp. 339–363, May 2022.

[24] C. S. Song and Y.-K. Kim, "Predictors of consumers' willingness to share personal information with fashion sales robots," *Journal of Retailing and Consumer Services*, vol. 63, p. 102727, Nov. 2021.

[25] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022.

[26] M. L. Ahmed, R. Iqbal, C. Karyotis, V. Palade, and S. A. Amin, "Predicting the public adoption of connected and autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 1680–1688, 2022.

[27] V. Bonta, N. Kumaresh, and J. Naulegari, "A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," *Asian Journal of Computer Science and Technology*, vol. 8, pp. 1–6, Mar. 2019.

[28] S. Loria, "TextBlob: Simplified Text Processing," 2022. [Online]. Available: https://textblob.readthedocs.io/en/dev/

[29] Laerd Statistics. Spearman's rank-order correlation: A guide to when to use it, what it does, and what the assumptions are. Laerd Statistics. [Online]. Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

[30] P. Rozin and E. B. Royzman, "Negativity Bias, Negativity Dominance, and Contagion," *Personality and Social Psychology Review*, vol. 5, no. 4, pp. 296–320, Nov. 2001.

[31] Q. Xu, "The two sides of heuristics: How positive and negative bias influence chatbot acceptance," in *Artificial Intelligence in HCI*, H. Degen and S. Ntoa, Eds. Cham: Springer Nature Switzerland, 2025, pp. 336–349.