Evaluating Generative In-Vehicle Assistants: A Research Proposal for Mapping Claims to Human Outcomes

Document type: Research proposal and preliminary evidence map.
Status: Research proposal prepared for upload; not submitted.
Author: Yu Zhang
Affiliation: Independent Researcher
License: Creative Commons Attribution 4.0 International (CC BY 4.0). Copyright © 2026 Yu Zhang. This license applies to this proposal's original text; third-party materials retain their respective rights.

Abstract

Generative in-vehicle assistants are evaluated using heterogeneous measures, ranging from factual correctness and conversational preference to workload, sleepiness and driving performance. These measures support different claims. We propose a scoping review of how evaluation evidence supports claims about generative driver assistance, with particular attention to affect-aware support and the distinction between recognizing a state and improving it. An exploratory arXiv-led search has produced a candidate ledger and a provisional evidence map. This proposal specifies review questions, eligibility rules, extraction fields, quality considerations and a planned claim-to-measurement framework. Retrieval coverage and screening remain incomplete, and the framework has not been validated. No new driving experiments or intervention effects are reported. Planned outputs include an auditable review corpus, a measurement taxonomy, an evidence-gap map and evaluation recommendations bounded by participant role and study setting.

1. Motivation and related work

Existing work already studies automotive affective interfaces, driver monitoring, in-cabin AI and human–vehicle interaction [12–16]. This project therefore does not claim that connecting driver state to assistance is itself new. Its proposed contribution is a structured account of what evidence is needed to support different claims for generative in-vehicle systems, and which conclusions transfer across evaluation settings.

Expert-grounded dialogue evaluation addresses content and task quality [1, 2]. Proactivity and timing studies examine interaction design [3–5]. Simulator, test-track and on-road investigations examine selected human and driving-related outcomes [6–8]. Synthetic multimodal generation and response-style preference studies contribute different evidence [9, 10]. Historical work provides comparators for state regulation, user preferences and embodiment [11, 17, 18]. These examples motivate careful synthesis; they do not establish an absence of prior human validation.

2. Research questions

1. Which functions and claimed benefits of generative in-vehicle assistants are evaluated?
2. Which reference sources, measurements and experimental designs support those claims?
3. How do participant role, input provenance, exposure duration and setting limit interpretation?
4. Which links between content quality, interaction experience, human state and driving outcomes need further evidence?

3. Proposed scope and eligibility

The core corpus will include original studies of automotive assistants that generate context-dependent language or agentic responses. Relevant functions include conversation, explanation, clarification, proactive communication and affect-aware support. Papers must describe an evaluation with extractable methods and outcomes. Model-based evaluations will remain eligible but will be distinguished from human studies.

Fixed-message interfaces, historical affective interventions and passenger studies may enter explicitly labeled comparator strata. Generic driving perception, trajectory prediction and emotion classification without an assistance or evaluation connection will normally be excluded. Reviews will support background and citation

chasing, rather than count as original studies. Preprints and peer-reviewed papers will be identified separately. Publication versions and shared datasets or participant cohorts will be reconciled before counting independent evidence.

4. Preliminary work and its limits

The exploratory search used nine arXiv query families: 265 returned records represented 230 unique identifiers, and review seeds produced a ledger of 233 candidate records. Eighteen original-study publications were provisionally coded: eight generative core studies, five historical or non-generative comparators, and five evaluation-method comparators. These are publication counts, not independent trial counts.

Twenty-five PDFs were acquired for targeted inspection. However, 152 records remain without formal screening; seven original full-text decisions and 14 priority candidates remain pending. Retrieval from ACM, IEEE and broader metadata services was blocked by the environment. One selected review was available only at abstract level [24]. Extraction was performed with AI assistance and has not received independent human verification. These counts describe preparation for the proposed review, not completed systematic coverage. The protocol was developed after exploratory searching and is not preregistered.

5. Planned methods

5.1 Retrieval and selection

We will finalize a dated protocol before the next retrieval round, retaining the exploratory search as a separate phase. Searches will combine automotive setting terms, conversational or generative model terms, and evaluation or human-outcome terms. When accessible, arXiv will be complemented by ACM Digital Library, IEEE Xplore and multidisciplinary discovery services. Exact queries, dates, result limits and access failures will be recorded. Backward and forward citation searches will target the nearest reviews and core studies.

A human researcher will review all proposed inclusions and their supporting source passages. Borderline cases and exclusions will receive recorded reasons. If a second reviewer is available, independent screening of a specified subset and disagreement resolution will be documented. Otherwise, the final review will explicitly disclose single-reviewer selection and will not report inter-rater reliability. Unavailable full text will be logged rather than treated as a negative finding.

5.2 Extraction and appraisal

For each study we will extract: publication identity and version; assistant function; generative versus fixed-message behavior; participant role; data provenance; study setting; recruitment and exclusions; comparator; exposure duration; outcome definitions; measurement source; uncertainty reporting; source-page locators; and conflicts of interest where disclosed.

Appraisal will examine construct validity, comparator suitability, sample accounting, allocation and blinding where applicable, missing data, ecological validity and correspondence between claims and measurements. Model-judge scores, simulated tools, preference ratings, self-reported state, physiological indicators and driving behavior will be coded separately. Inconsistent denominators will be retained for verification rather than silently resolved.

5.3 Synthesis

We propose four descriptive layers: (1) content and task correctness, (2) interaction quality, (3) human state, and (4) driving and task consequences. Each evidence unit will connect assistant function, reference source, measured endpoint, supported claim and transfer boundary. The layers are not a validated scale or a ranking of study quality.

For example, benchmark-based safety or task assessments [20–23] will not be interpreted as demonstrated reductions in real-driver workload or crashes. Passenger timing results will remain distinct from active-driver

distraction evidence. Fatigue, stress, trust and workload will not be collapsed into a single affect outcome. Quantitative pooling is not planned unless genuinely comparable designs, endpoints and study independence are established.

6. Expected contributions and falsifiable expectations

The intended outputs are a traceable evidence map, a comparison with the closest existing reviews, a reusable extraction schema and practical reporting recommendations. Novelty remains a question to establish through that comparison.

Three expectations will guide extraction without being presented as findings:

- Evidence at one measurement layer may not support claims at another; we will document explicit claim-to-measurement matches and mismatches with source passages.
- Study role and setting may explain important differences in reported assistant benefits; we will stratify findings before considering cross-setting conclusions.
- Timing, confirmation and uncertainty handling may be evaluated less consistently than answer quality; the completed corpus may support or contradict this expectation.

We will report contradictory evidence and revise the proposed framework if the included literature does not support it. No numerical prevalence or superiority claim will be made from the current selected sample.

7. Work plan, resources and stopping conditions

The work proceeds through protocol finalization, broader retrieval, completed screening, human-verified extraction, synthesis, and author review. Phase completion will depend on auditable deliverables rather than an assumed publication deadline. Search and document processing can run on CPU; no GPU training or driver experiment is required for the review.

Any additional paid model analysis will require an explicit run budget and bounded requests. No experiments, participant recruitment, data collection or public posting will run automatically. A later human-subject intervention study would require a separate protocol, appropriate ethics review and explicit resource approval.

8. Transparency and limitations

This proposal and its preliminary corpus were developed with AI assistance for retrieval, coding, synthesis and drafting. Human authors remain responsible for source verification, interpretation and any public submission. The submitting researcher confirmed the author name Yu Zhang and affiliation Independent Researcher. The current arXiv-led sample may underrepresent automotive HCI and human-factors venues. The completed project may need a narrower scope if broader retrieval or sufficient human verification cannot be obtained.

This document reports no new experimental findings, does not claim a completed systematic review, and does not establish clinical or driving-safety benefits for any assistant. Supporting search logs, evidence tables and source identifiers are retained in the repository. Author and license details have been confirmed; platform-specific submission requirements remain to be verified.

References

[1] Ken E. Friedl, Abbas Goher Khan, Soumya Ranjan Sahoo, Md Rashad Al Hasan Rony, Jana Germies, Christian Süß. **InCA: Rethinking In-Car Conversational System Assessment Leveraging Large Language Models**. arXiv:2311.07469v2. https://arxiv.org/abs/2311.07469v2

[2] Rafael Giebisch, Ken E. Friedl, Lev Sorokin, Andrea Stocco. **Automated Factual Benchmarking for In-Car Conversational Systems using Large Language Models**. arXiv:2504.01248v1. https://arxiv.org/abs/2504.01248v1

[3] Huifang Du, Xuejing Feng, Jun Ma, Meng Wang, Shiyu Tao, Yijie Zhong, Yuan-Fang Li, Haofen Wang. **Towards Proactive Interactions for In-Vehicle Conversational Assistants Utilizing Large Language Models**. arXiv:2403.09135v1. https://arxiv.org/abs/2403.09135v1

[4] Josh Susak, Yifu Liu, Pascal Jansen, Mark Colley. **ProVoice: Designing Proactive Functionality for In-Vehicle Conversational Assistants using Multi-Objective Bayesian Optimization to Enhance Driver Experience**. arXiv:2601.19421v1. https://arxiv.org/abs/2601.19421v1

[5] Kaiser Hamid, Peihang Li, Nade Liang. **When2Talk: When Should a Proactive In-Car Agent Talk?**. arXiv:2609.12503v1. https://arxiv.org/abs/2609.12503v1

[6] Yeana Lee Bond, Mungyeong Choe, Baker Kasim Hasan, Arsh Siddiqui, Myounghoon Jeon. **ChatGPT on the Road: Leveraging Large Language Model-Powered In-vehicle Conversational Agents for Safer and More Enjoyable Driving Experience**. arXiv:2508.08101v1. https://arxiv.org/abs/2508.08101v1

[7] Lewis Cockram, Yueteng Yu, Jorge Pardo, Xiaomeng Li, Andry Rakotonirainy, Jonny Kuo, Sebastien Demmel, Mike Lenné, Ronald Schroeter. **Small Talk, Big Impact? LLM-based Conversational Agents to Mitigate Passive Fatigue in Conditional Automated Driving**. arXiv:2510.25421v2. https://arxiv.org/abs/2510.25421v2

[8] Chris Monk, Allegra Ayala, Christine S. P. Yu, Gregory M. Fitch, Dara Gruber. **Visual and Cognitive Demands of a Large Language Model-Powered In-vehicle Conversational Agent**. arXiv:2601.15034v1. https://arxiv.org/abs/2601.15034v1

[9] Chang Liu, Dalai Mengke, Hanbo Zhou, Jia Hu, Peter Mihajlik, Tamas Sziranyi. **Multimodal Drivers' Emotion Recognition and Safety-Oriented Intervention for Intelligent Transportation Systems**. arXiv:2608.06378v1. https://arxiv.org/abs/2608.06378v1

[10] Harine Choi, Eun Hak Lee, Zhengzhong Tu. **Vision-Language Assistant for Emotional Reactions to Risky Driving**. arXiv:2607.16181v1. https://arxiv.org/abs/2607.16181v1

[11] A. J. Bequet, C. Jallais, J. Quick, D. Ndiaye, A. R. Hidalgo-Munoz. **Road to Serenity: Individual Variations in the Efficacy of Unobtrusive Respiratory Guidance for Driving Stress Regulation**. arXiv:2406.09777v1. https://arxiv.org/abs/2406.09777v1

[12] Michael Braun, Florian Weber, Florian Alt. **Affective Automotive User Interfaces – Reviewing the State of Emotion Regulation in the Car**. arXiv:2003.13731v2. https://arxiv.org/abs/2003.13731v2 The retrieved PDF identifies the published version in ACM Computing Surveys 54(7), Article 137 (2021), DOI: 10.1145/3460938.

[13] Anaïs Halin, Jacques G. Verly, Marc Van Droogenbroeck. **Survey and synthesis of state of the art in driver monitoring**. arXiv:2110.00472v1. https://arxiv.org/abs/2110.00472v1

[14] Yao Rong, Chao Han, Christian Hellert, Antje Loyal, Enkelejda Kasneci. **Artificial Intelligence Methods in In-Cabin Use Cases: A Survey**. arXiv:2101.02082v1. https://arxiv.org/abs/2101.02082v1

[15] Abu Jafar Md Muzahid, Xiaopeng Zhao, Zhenbo Wang. **Survey on Human-Vehicle Interactions and AI Collaboration for Optimal Decision-Making in Automated Driving**. arXiv:2412.08005v1. https://arxiv.org/abs/2412.08005v1

[16] Tima M. Yekta, Julius Schöning. **Keep Calm and Relax – HMI for Autonomous Vehicles**. arXiv:2408.09046v1. https://arxiv.org/abs/2408.09046v1

[17] Michael Braun, Jingyi Li, Florian Weber, Bastian Pfleging, Andreas Butz, Florian Alt. **What If Your Car Would Care? Exploring Use Cases For Affective Automotive User Interfaces**. arXiv:2004.02481v1. https://arxiv.org/abs/2004.02481v1

[18] Yuanchao Li, Lachlan Urquhart, Nihan Karatas, Shun Shao, Hiroshi Ishiguro, Xun Shen. **Beyond Voice Assistants: Exploring Advantages and Risks of an In-Car Social Robot in Real Driving Scenarios**. arXiv:2402.11853v2. https://arxiv.org/abs/2402.11853v2

[19] Johannes Kirmayr, Raphael Wennmacher, Khanh Huynh, Lukas Stappen, Elisabeth André, Florian Alt. **What Are You Doing? Effects of Intermediate Feedback from Agentic LLM In-Car Assistants During Multi-Step Processing**. arXiv:2602.15569v2. https://arxiv.org/abs/2602.15569v2

[20] Keito Inoshita. **Safety-Aware Evaluation of LLM-Generated Driver Intervention Messages through Multi-Task Risk Fusion**. arXiv:2606.22706v1. https://arxiv.org/abs/2606.22706v1

[21] Vaishnav Negi, Lev Sorokin, Soroosh Tayebi Arasteh, Andrea Stocco. **Automated Evaluation of Multi-Turn Dialogues in In-Car Conversational Assistants**. arXiv:2609.35812v1. https://arxiv.org/abs/2609.35812v1

[22] Johannes Kirmayr, Lukas Stappen, Elisabeth André. **CAR-bench: Evaluating the Consistency and Limit-Awareness of LLM Agents under Real-World Uncertainty**. arXiv:2601.22027v1. https://arxiv.org/abs/2601.22027v1

[23] Diba Afroze, Xingli Zhang, Yazhou Tu, Xiali Hei. **From Intent to Action: Benchmarking LLM Safety in Vehicle Voice Command Authorization**. arXiv:2609.19630v1. https://arxiv.org/abs/2609.19630v1

[24] Bilal Khana, Waseem Shariff, Rory Coyne, Muhammad Ali Farooq, Peter Corcoran. **From Review to Design: Ethical Multimodal Driver Monitoring Systems for Risk Mitigation, Incident Response, and Accountability in Automated Vehicles**. arXiv:2605.06439v1. https://arxiv.org/abs/2605.06439v1