Interviewing Silicon Experts: A Persona-Based LLM Interview Pipeline in Automated Driving

YING FANG, Eindhoven University of Technology, The Netherlands
PAVLO BAZILINSKYY, Eindhoven University of Technology, The Netherlands
MARIEKE MARTENS, Eindhoven University of Technology, The Netherlands

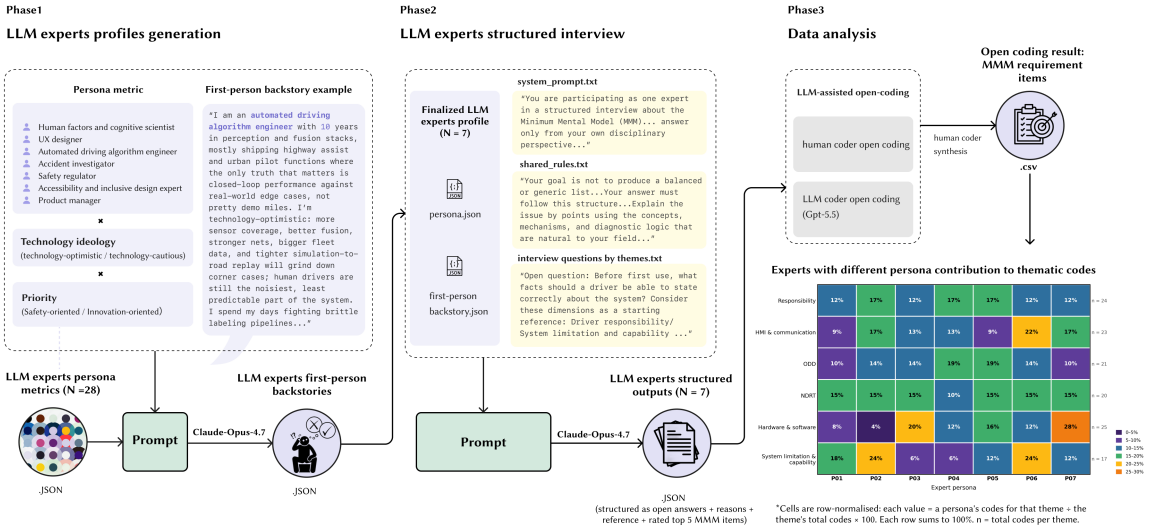

Fig. 1. A modular pipeline: LLM-expert interview deriving knowledge items for "*drivers' minimum mental model*".

Human expert interviews are valuable but often limited by slow recruitment, scarce expertise, and network-based sampling, especially for abstract human-automation topics requiring multidisciplinary input. This paper presents a persona-based LLM expert-interview pipeline for structuring expert knowledge before human involvement. We constructed 28 discipline-specific personas and purposively selected seven differentiated "silicon experts" across human factors, engineering, regulation, accessibility, and UX design. Using drivers' minimum mental model (MMM) as a demonstration topic, each persona completed a structured interview protocol and returned JSON-formatted responses. The pipeline generated 58 candidate MMM items, producing concrete, traceable, and auditable outputs. Results showed both convergence and persona-specific divergence: shared priorities included driver responsibility, operational design domain boundaries, and permitted non-driving-related tasks, while divergent contributions highlighted OTA updates, sensor limitations, and warning-channel perceivability. The pipeline is positioned as a preliminary scoping tool for preparing subsequent human expert validation.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Large language models, expert interview, persona, automated driving

Authors' Contact Information: Ying Fang, Eindhoven University of Technology, Eindhoven, The Netherlands, y.fang1@tue.nl; Pavlo Bazilinskyy, Eindhoven University of Technology, Eindhoven, The Netherlands, p.bazilinskyy@tue.nl; Marieke Martens, Eindhoven University of Technology, Eindhoven, The Netherlands, m.h.martens@tue.nl.

2026. Manuscript submitted to ACM

1 Introduction

Large language models (LLM) have evolved from simple chatbots to serious research instruments [3, 33]. Recent studies support researchers in hypothesis generation [36], enabling multi-stage agentic research workflows [19, 33], and acting as peer-review assistants [13, 37]. Within this wider shift, one methodological strand explores how LLM can be designed to mimic human responses, potentially augmenting research approaches that traditionally rely on human participants. This work has evolved along two main empirical trajectories: first, quantitative studies that use LLM to replicate public opinions in structured survey formats [1, 5, 14, 29]; second, qualitative studies that use LLM to conduct semi-structured interviews that produce richly contextualised detailed narratives [2, 15, 35].

Developing personas for LLM has recently gained traction as a promising method to elicit human-like responses [20, 27]. Extending the UX design tradition of using user archetypes to understand needs, behaviours, and constraints [26], a persona is used to steer LLMs to generate outputs more aligned with specific social and psychological standpoints, each tied to particular human traits [24]. For example, Argyle et al. coined the term "algorithmic fidelity" to capture how well an LLM can mirror subgroup-specific patterns in human attitudes and behaviours, demonstrating that conditioning GPT-3 with detailed first-person sociodemographic backstories produced results that closely align with its human benchmarks [5]. Similarly, De Winter et al. generated text-based personas by prompting GPT-4 to produce concise profiles that include age, gender, occupation, and brief descriptive traits, indicating that constructing personas for LLM can support preliminary questionnaire piloting prior to recruiting real participants [14].

Expert interviews are a standard method for investigating emerging topics [11], with semi-structured interviews used in human-automation interaction research [23, 32] and Delphi methods supporting the iterative synthesis of expert opinions into consensus [8, 25, 34]. However, these methods face key challenges. First, assembling experts takes considerable time [22]. Second, sampling often relies on the researcher's own professional network, producing overly homogeneous panels [21]. Third, experts can become fatigued during long or abstract interviews and rely more on quick experience-based judgements, resulting in shorter and less detailed qualitative responses [28]. These conventional constraints become particularly challenging when studying complicated topics such as **minimal mental models (MMM)**, which captures the requisite understanding that a driver must possess to interact safely with the automated system [12]. Although prior studies have examined drivers' mental models in relation to specific SAE levels or systems features [6, 7, 30, 31], it does not always translate into a concrete, cross-SAE-level catalogue of minimum driver knowledge. Thus, for an abstract and multidisciplinary topic such as MMM, researchers still need to turn broad concepts into probeable candidate items that can be discussed, challenged, and refined in expert interviews.

Recent research underscores the potential of persona-conditioned LLM to produce knowledge tailored to specific disciplines [5, 24]. If such personas can stand in for experts, they provide a rapid preparatory tool for expert interviews. It can help researchers generate tentative arguments on complex topics such as MMM and bring multidicsiplinary perspectives [15]. Since LLM outputs are sensitive to the prompts they receive and can be influenced by them [24], the main challenge in incorporating an LLM-based approach into expert interviews does not lie in querying the LLM itself, but in devising a procedure that ensures the responses are reproducible, comparable, and transparent enough for critical evaluation.

1.1 Aim of study

This study introduces and demonstrates an iterative, time- and resource-efficient pipeline to support early-stage qualitative exploration in the domain of automated driving. We use MMM as an example domain with three objectives.

Persona_01.JSON


```
{
  "persona_id": "P01",
  "discipline": "Human factors
& cognitive scientist",
  "ideological_orientation":
"technology-cautious",
  "priority": "safety-
oriented",
  "years_of_experience": 10
}
```


Persona_01_backstory.JSON


```
{

  "first_person_backstory": "I am a human factors and cognitive scientist, and for the last 10
years I\u2019ve worked where automated driving actually breaks: at the handoff between machine
competence and human attention. My baseline is simple. If a driver cannot maintain calibrated
trust, recover from an out-of-the-loop state, and execute a safe takeover within verified timing
bounds across realistic ODD edge cases, it is not ready to ship. I do not care how pretty the
perception stack demo looks. In safety-critical systems, black-box behavior is a liability, not
a feature. Every week I argue with teams who want to wave away mode confusion, vigilance
decrement, inattentional blindness, and takeover lag because the disengagement charts look
decent in dry daylight. That is not validation. That is wishful thinking with a test fleet. I
push for human oversight, driver monitoring that actually measures state not just gaze,
conservative HMI, fault-tolerant fallback, and evidence that stands up under ISO 26262 and SOTIF
scrutiny. If we cannot explain failure and bound it, we do not deploy."

}
```


Fig. 2. Example of persona conditioning: converting a structured expert profile into a first-person professional backstory.

First, we use LLM persona profile design to systematically construct transparent and traceable expert standpoints to counter the network-bounded homogeneity in conventional expert sampling. Second, we introduce an interview protocol and a structured output format to translate a broad human–automation topic into concrete candidate items. Third, we explore the pipeline as an initial scoping tool to flag overlaps, conflicts, and gaps ahead of expert interviews, thereby guiding later interviews towards validation, refinement, and deeper exploration.

2 Method

We developed a three-phase pipeline (Figure 1) via scripted API calls in Python 3.14 scripts for generating the artefacts. We used Claude-Opus-4.7 in Phases 1 and 2 because, at the time of data collection, it performed well on step-by-step reasoning and document-heavy benchmarks, matching our need for long-context persona conditioning and structured qualitative outputs [4]. Data were gathered between 14 and 24 April 2026.

2.1 Phase 1: LLM-expert profiles generation

First, we constructed a candidate persona metric, following Argyle et al.'s silicon-sampling approach, which combines demographic factors with ideological, attitudinal, behavioural, and sociocultural traits to elicit subgroup-specific response patterns [5]. We identified key professional disciplines relevant in automated driving, including human factors, UX design, systems engineering, accident investigation, regulation, accessibility, and product management (Figure 1). Using these as standpoints, we created a heterogeneous persona matrix and combined it with two technological ideologies (technology-optimistic or technology-cautious) and two priorities (safety-focused or innovation-focused), yielding 28 distinct persona profiles ($7 \times 2 \times 2 = 28$), represented as 28 JSON files. Since the goal of expert recruiting is not to achieve full statistical coverage of the entire persona matrix, but rather to assemble a diverse and credible panel [5, 8]. Thus, we purposively selected seven personas for the interview stage to cover the representative standpoints while maximising contrast in technology ideology and priority. The final set was selected so that each persona's ideology and priorities reflected a plausible professional viewpoint, such as a technology-cautious, safety-focused human factors specialist and an innovation-oriented product manager. Our goal was not to stereotype professions, but to create a balanced, varied set of perspectives to enable efficient pipeline testing.

Second, we elaborated each chosen persona into a first-person backstory (Figure 2). This proven method helps make assumptions explicit and captures domain-specific reasoning patterns when designing silicon samples [5]. The API call for Claude-Opus-4.7 was made a total of seven times, once for each persona. Each call included three components: a

```
{
  "expert_id": "P01",
  "open_answer": {
    "points": [
      {
        "point": "The driver must know which non-driving related tasks are permitted, which are prohibited, and on what basis
that permission can be revoked in real time.",
        "reasoning": [
          NDRT permissibility is the most misunderstood facet of automated driving in the public, and it directly controls out-
of-the-loop depth. A driver who believes 'I can watch video' under a system that actually requires eyes-on-road will have
catastrophic takeover performance.",
        "scenarios": [
          "In L3 within ODD, certain NDRTs (reading, phone use) may be allowed, but sleeping, leaving the driver's seat, or
consuming alcohol are not, because the driver remains the fallback on transition demand.",
        "examples": [
          "Mercedes Drive Pilot (L3) explicitly permits some secondary tasks within its ODD but prohibits sleep and requires
readiness to resume within the transition demand window.",
        "references_or_legal_basis": [
          "UNECE R157 — driver availability requirements in L3.",
    "points": [
      {
        ...
      }
```


Fig. 3. Shortened example of the structured JSON output used to store persona-based expert interview responses.

system prompt outlining the task, a JSON file for the persona, and a .txt file containing a sample first-person backstory. Every backstory starts with "I am," adopts a practitioner-style tone, and uses role-specific technical terminology with a clearly articulated professional position. For example, technology-optimistic experts assume that sensor fusion and large-scale data will ultimately handle all edge cases, whereas technology-cautious experts focus on risks arising from automation fragility. Similarly, innovation-focused experts emphasise rapid technical iteration and exploring future use cases, while safety-focused experts foreground zero-harm principles and adherence to regulation. These LLM experts ($N = 7$) then participated independently in structured interviews conducted in Section 2.2.

2.2 Phase 2: LLM-expert interview

First, we based the questions in the interview of *MMM requirements* on the validated frameworks interpreting mental model through [31]: *responsibility, HMI and communication, operational design domain (ODD), non-driving-related tasks (NDRT), system limitation and capability, hardware and software* [10]. The pre-defined themes offered a theoretically grounded structure for the interview protocol.

We queried each LLM expert once via an API call for Claude-Opus-4.7, using structured prompts to elicit responses strictly from their disciplinary standpoint. The prompts were assembled from four elements, with independent .txt and .json files: (1) the persona profile and first-person backstory; (2) a system prompt positioning the model as one expert in a structured interview; (3) a prompt of interview rules asking the persona to explain why each issue was proposed from its professional perspective and avoid fabricated references; (4) a theme-specific question asking for minimum knowledge drivers' must have. Each response was then constrained to a fixed output schema: a knowledge "point" beginning with "The driver must know...", followed by reasoning, scenarios, examples, and references where applicable (Figure 3). We therefore obtained an output format in which each response was represented as a JSON object that included explicit knowledge items with respective structural explanations.

2.3 Phase 3: data analysis

We analysed the seven JSON files as outputs from structured LLM-expert interviews, which followed six predefined themes (Section 2.2). The coding was carried out by the first author, informed by the Framework Method for qualitative data analysis [17]. The predefined interview themes formed the initial analytical structure, and within each theme,

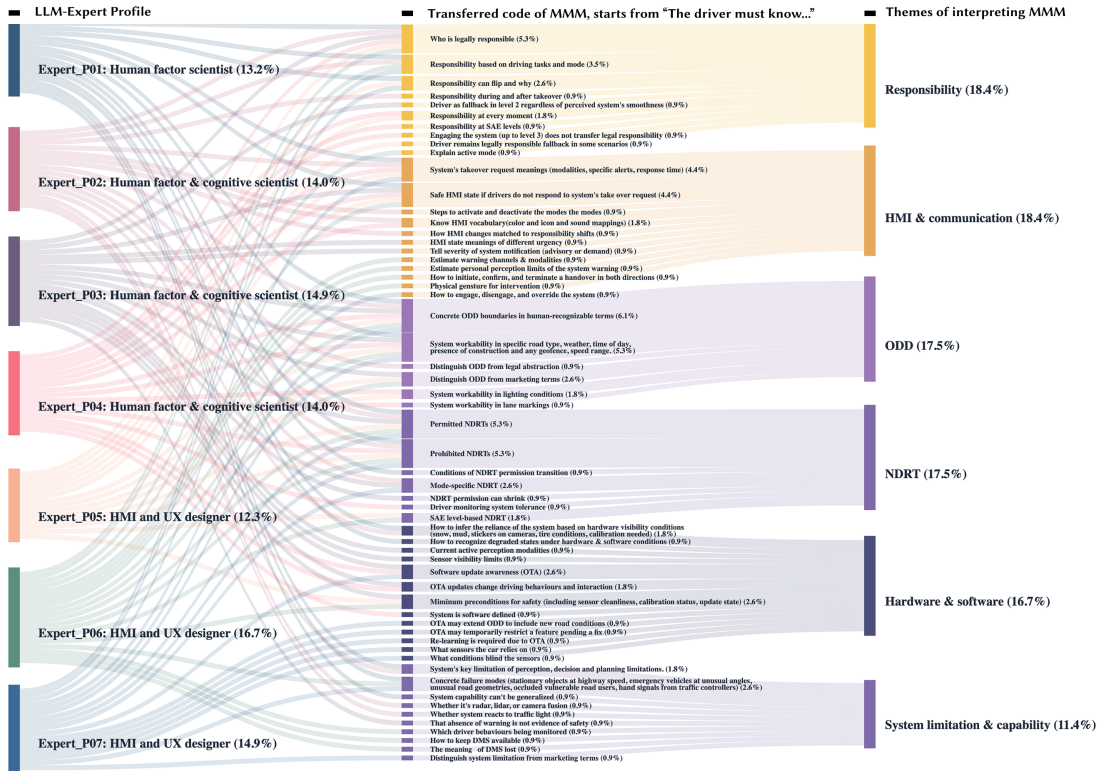

Fig. 4. Sankey diagram of visualising the connections between LLM-expert profiles, pre-defined themes, and codes, formulated as statements beginning with "*The driver must know...*"

inductive open coding was used to identify recurring concepts. Each "point" (a structured statement of driver knowledge within a theme) in the JSON files was decomposed into one or more codes. The accompanying reasoning, scenarios, examples, and references were used to support interpretation during the coding process. ChatGPT-5.5 acted as an auxiliary second coder and was kept distinct from Claude-Opus-4.7, which had generated the interview responses. It independently coded the materials in the same manner as the first author and returned the results as JSON files, which were then used to support comparison checking. The first author then reviewed, compared, and merged the codes within each theme. The coding results were combined into a single CSV file, with each row corresponding to one coded unit for data analysis.

3 Results

After merging the seven LLM expert outputs, the first author segmented the generated statements into point-level codes and removed duplicates, resulting in 58 distinct code terms across pre-defined six themes (Figure 4). Overall, the result shows three main characteristics. First, coverage across themes was relatively balanced, with **System limitation and capability** accounting for the smallest share of coded items (11.4%), **Responsibility** and **HMI and communication** the largest (18.4%). Second, the distribution of themes varied across LLM-expert profiles, suggesting that each persona contributed from its own professional standpoint. For example, P07, the product manager, reported

most code contributions were in **Hardware and software** (28%), while P06, the inclusive designer, contributed most codes to **System Limitation and capability** (22%) (Figure 1). P04, the accident investigator and P05, the safety regulator, contributed most strongly to **ODD** (19%). Third, the generated outputs transformed the overarching themes into concrete, rather than general, knowledge items. For example, **Responsibility** was explicitly interpreted by LLM through items such as "*who is legally responsible*", "*how responsibility can shift*". **HMI and communication** was translated into knowledge about "*takeover request meanings, warning channels and modalities*". Similarly, **System limitation and capability** was expressed through "*specific failure cases such as stationary objects at highway speed, occluded vulnerable road users, unusual road geometries, and hand signals from traffic controllers*".

The results further indicate consensus and discrepancy across the output of the LLM-experts. On the one hand, codes repeated in all the 7 LLM experts, for instance, "*who is legally responsible*", "*concrete ODD boundaries in human-recognisable terms*", "*system workability in specific road type, weather, time of day, presence of construction and any geofence, speed range.*", and "*The driver must differentiate permitted NDRTs and prohibited NDRTs*". On the other hand, the discrepancies were not direct contradictions. Instead, they showed that different personas set different boundaries for what should count as "minimum" knowledge. For example, P03, the algorithm engineer and P07, the product manager, expanded **Hardware & software**, including "*sensor obstruction and calibration after windshield replacement, tyre condition*". P07 highlighted "*Driver must be aware of OTA updates*", and "*OTA can change system behaviour, HMI signals...*". P06, the inclusive designer, brought the only accessibility focus, emphasising that drivers must know both the "*system's warning channels*" and "*their own ability to perceive these channels reliably*". Therefore, the observed discrepancies are better interpreted as complementary expansions rather than mutually exclusive disagreements.

4 Discussion

These findings indicate that the LLM-based pipeline offers a controlled, transparent and auditable approach for supporting early-stage interviews with human experts. First, the pipeline turns an abstract concept into a concrete, structured space: it produced 58 distinct items, each of which can be traced back to its persona and reasoning. Second, the pipeline produces differentiation across disciplines together with convergence on items the personas were never prompted to agree on [5]. These features are built into the pipeline in a way that one-shot prompts simply can't reach. They enable traceability, reproducibility, and cross-model auditing, turning the process into something closer to expert interviews or Delphi studies than to mere querying. Aligning with calls for greater transparency in automotive user research [16], we therefore frame the pipeline as a strategic preparation tool that structures and stress-tests the requirement space for the human validation detailed below in future work.

4.1 Limitation and future work

This study has several limitations that suggest directions for future work. First, the outputs may reflect training-data biases [9, 18], tending to more easily replicate viewpoints prevalent in academic publications, while regulations and tacit industry design practices may be inconsistently captured. Future work should therefore examine whether different model providers, prompt phrasings, and cross-model critique procedures between models produce stable or divergent thematic clusters. Second, LLM experts are not human experts: they lack tacit knowledge and direct accountability for safety-critical design decisions. The generated statements should therefore be used as starting points for a Delphi-style rating study, in which practising human experts validate their reliability, ambiguity, and completeness. Third, the structured interview pipeline should be further iterated toward a semi-structured format, in which initial outputs are fed back into the model to generate follow-up probes and clarify ambiguous points. In general, this pipeline should

be understood as a strategic tool to prepare and test early-stage materials to anchor broad and abstract topics in fast-evolving domains such as driver–automation interaction.

5 Supplementary material

Supplementary material including code, generated JSON responses, and the coding table is currently available via https://www.dropbox.com/scl/fo/tz6awwf3zyjqpjx70hjbb/ADE_C3yh2OvKzfuA0kqWe0I?rlkey=jhmezp3zq2y76yto7tjpsmcbo&st=q4x9ifrk&dl=0. A maintained version of the code is available at https://github.com/FayeFang-creator/llm-expert-interview.

References

[1] Md Shadab Alam and Pavlo Bazilinskyy. 2025. Cross or Nah? LLMs Get in the Mindset of a Pedestrian in front of Automated Car with an eHMI. In *Adjunct Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '25 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 19–34. doi:10.1145/3744335.3758477

[2] Aliya Amirova, Theodora Fteropoulli, Nafiso Ahmed, Martin R. Cowie, and Joel Z. Leibo. 2024. Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. *PLOS ONE* 19, 3 (March 2024), e0300024. doi:10.1371/journal.pone.0300024

[3] Yadagiri Annepaka and Partha Pakray. 2025. Large language models: a survey of their development, capabilities, and applications. *Knowledge and Information Systems* 67, 3 (March 2025), 2967–3022. doi:10.1007/s10115-024-02310-4

[4] Anthropic. 2026. Introducing claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7

[5] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (July 2023), 337–351. doi:10.1017/pan.2023.2

[6] Matthias Beggiato and Josef F. Krems. 2013. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour* 18 (May 2013), 47–57. doi:10.1016/j.trf.2012.12.006

[7] Matthias Beggiato, Marta Pereira, Tibor Petzoldt, and Josef F. Krems. 2015. Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F: Traffic Psychology and Behaviour* 35 (Nov. 2015), 75–84. doi:10.1016/j.trf.2015.10.005

[8] Daniel Beiderbeck, Nicolas Frevel, Heiko A. Von Der Gracht, Sascha L. Schmidt, and Vera M. Schweitzer. 2021. Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. *MethodsX* 8 (2021), 101401. doi:10.1016/j.mex.2021.101401

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[10] Anika Boelhouwer, Arie P. van den Beukel, Mascha C. van der Voort, and Marieke H. Martens. 2019. Should I take over? Does system knowledge help drivers in making take-over decisions while driving a partially automated car? *Transportation Research Part F: Traffic Psychology and Behaviour* 60 (Jan. 2019), 669–684. doi:10.1016/j.trf.2018.11.016

[11] Alexander Bogner, Beate Littig, and Wolfgang Menz. 2009. Introduction: Expert Interviews — An Introduction to a New Methodological Debate. In *Interviewing Experts*, Alexander Bogner, Beate Littig, and Wolfgang Menz (Eds.). Palgrave Macmillan UK, London, 1–13. doi:10.1057/9780230244276_1

[12] Oliver Carsten and Marieke H. Martens. 2019. How can humans understand their automated cars? HMI principles, problems and solutions. *Cognition, Technology & Work* 21, 1 (Feb. 2019), 3–20. doi:10.1007/s10111-018-0484-0

[13] Sully F. Chen, Anton Alyakin, Andreas Seas, Eunice Yang, Joanne J. Choi, Jin Vivian Lee, Amelia L. Chen, Pranav I. Warman, Rochelle T. Bitolas, Robert J. Steele, Daniel A. Alber, and Eric K. Oermann. 2026. LLM-assisted systematic review of large language models in clinical medicine. *Nature Medicine* 32, 3 (March 2026), 1152–1159. doi:10.1038/s41591-026-04229-5

[14] Joost C.F. de Winter, Tom Driessen, and Dimitra Dodou. 2024. The use of ChatGPT for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences* 228 (Oct. 2024), 112729. doi:10.1016/j.paid.2024.112729

[15] Andreas Dengel, Rupert Gehrlein, David Fernes, Sebastian Görlich, Jonas Maurer, Hai Hoang Pham, Gabriel Großmann, and Niklas Dietrich genannt Eisermann. 2023. Qualitative Research Methods for Large Language Models: Conducting Semi-Structured Interviews with ChatGPT and BARD on Computer Science Education. *Informatics* 10, 4 (Oct. 2023). doi:10.3390/informatics10040078

[16] Patrick Ebel, Pavlo Bazilinskyy, Mark Colley, Courtney Michael Goodridge, Philipp Hock, Christian P. Janssen, Hauke Sandhaus, Aravinda Ramakrishnan Srinivasan, and Philipp Wintersberger. 2024. Changing Lanes Toward Open Science: Openness and Transparency in Automotive User Research. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Stanford CA USA, 94–105. doi:10.1145/3640792.3675730

[17] Nicola K. Gale, Gemma Heath, Elaine Cameron, Sabina Rashid, and Sabi Redwood. 2013. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology* 13, 1 (Sept. 2013), 117. doi:10.1186/1471-2288-13-117

[18] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (Sept. 2024), 1097–1179. doi:10.1162/coli_a_00524

[19] Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. 2026. A multi-agent system for automating scientific discovery. *Nature* (May 2026). doi:10.1038/s41586-026-10652-y

[20] Giulia Iadisernia and Carolina Camassa. 2025. Prompting for Policy: Forecasting Macroeconomic Scenarios with Synthetic LLM Personas. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25)*. Association for Computing Machinery, New York, NY, USA, 335–343. doi:10.1145/3768292.3770385

[21] Karina Lovell John Baker. 2006. How expert are the experts? An exploration of the concept of 'expert' within Delphi panel techniques. doi:10.7748/nr2006.10.14.1.59.c6010

[22] Sinead Keeney, Hugh P. McKenna, and Felicity Hasson. 2010. *The Delphi Technique in Nursing and Health Research*. John Wiley & Sons. Google-Books-ID: osrJ9Pz6xpAC.

[23] Stacey Li, Debargha Dey, Claudia Santacruz, and Wendy Ju. 2025. What Researchers Need from Driving Simulator Systems: A Thematic Analysis of Expert Interviews. In *Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '25)*. Association for Computing Machinery, New York, NY, USA, 55–68. doi:10.1145/3744333.3747836

[24] Yiren Liu, Pranav Sharma, Mehul Oswal, Haijun Xia, and Yun Huang. 2025. PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 506–534. doi:10.1145/3715336.3735789

[25] Ana Isabel Martins, Gonçalo Santinha, Ana Margarida Almeida, Óscar Ribeiro, Telmo Silva, Nelson Rocha, and Anabela G Silva. 2023. Consensus on the Terms and Procedures for Planning and Reporting a Usability Evaluation of Health-Related Digital Solutions: Delphi Study and a Resulting Checklist. *Journal of Medical Internet Research* 25 (June 2023), e44326. doi:10.2196/44326

[26] Tomasz Miaskiewicz and Kenneth A Kozar. 2011. Personas and user-centered design: How can personas benefit product design processes? *Design studies* 32, 5 (2011), 417–430.

[27] Suhong Moon. 2025. *Binding Large Language Models to Virtual Personas for Human Simulation*. Ph. D. Dissertation. UC Berkeley. https://escholarship.org/uc/item/8s32x9x7

[28] M. Granger Morgan. 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences* 111, 20 (May 2014), 7176–7184. doi:10.1073/pnas.1319946111

[29] Pat Pataranutaporn, Nattavudh Powdthavee, Chayapatr Archiwaranguprok, and Pattie Maes. 2025. Simulating human well-being with large language models: Systematic validation and misestimation across 64,000 individuals from 64 countries. *Proceedings of the National Academy of Sciences* 122, 48 (Dec. 2025), e2519394122. doi:10.1073/pnas.2519394122

[30] SAE International. 2021. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. https://www.sae.org/standards/j3016-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles

[31] Stephanie Seupke, Sarukan Segar, and Martin Baumann. 2025. Development of a Mental Model Questionnaire Framework: A Systematic Approach to Measuring Mental Models in Automated Driving. doi:10.2139/ssrn.5168351

[32] Wilbert Tabone, Joost C.F. de Winter, Claudia Ackermann, Jonas Bärgman, Martin Baumann, Shuchisnigdha Deb, Colleen Emmenegger, Azra Habibovic, Marjan Hagenzieker, P.A. Hancock, Riender Happee, Josef Krems, John D. Lee, Marieke H. Martens, Natasha Merat, Don Norman, Thomas B. Sheridan, and Neville A. Stanton. 2021. Vulnerable road users and the coming wave of automated vehicles: Expert perspectives. *Transportation Research Interdisciplinary Perspectives* 9 (March 2021), 100293. doi:10.1016/j.trip.2020.100293

[33] Dhruv Trehan and Paras Chopra. 2026. Why LLMs Aren't Scientists Yet: Lessons from Four Autonomous Research Attempts. doi:10.48550/arXiv.2601.03315 arXiv:2601.03315 [cs.LG].

[34] Norhanisha Yusof, Nor Laily Hashim, and Azham Hussain. 2022. A review of fuzzy Delphi method application in human-computer interaction studies. Kedah, Malaysia, 040026. doi:10.1063/5.0094417

[35] Jihong Zhang, Xinya Liang, Anqi Deng, Nicole Bonge, Lin Tan, Ling Zhang, and Nicole Zarrett. 2025. Leveraging Interview-Informed LLMs to Model Survey Responses: Comparative Insights from AI-Generated and Human Data. https://arxiv.org/abs/2505.21997v1

[36] Yunxiang Zhang, Muhammad Khalifa, Shitanshu Bhushan, Grant Murphy, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2025. MLRC-bench: Can language agents solve machine learning research challenges?. In *Advances in neural information processing systems*, D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (Eds.), Vol. 38. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2025/file/82c96f3c90741ef2c9b248e65d9b5db0-Paper-Datasets_and_Benchmarks_Track.pdf

[37] Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. *Information Fusion* 124 (Dec. 2025), 103332. doi:10.1016/j.inffus.2025.103332

Received 18 June 2026