Whose Personae? Synthetic Persona Experiments in LLM Research and Pathways to Transparency

Jan Batzner[W, C, M], **Volker Stocker**[W, B], **Bingjun Tang**[C], **Anusha Natarajan**[C], **Qinhao Chen**[C], **Stefan Schmid**[W, B], **Gjergji Kasneci**[M]

[W]Weizenbaum Institute [C]Columbia University [B]Technical University Berlin
[M]Munich Center for Machine Learning & Technical University Munich

Abstract

Synthetic personae experiments have become a prominent method in Large Language Model alignment research, yet the representativeness and ecological validity of these personae vary considerably between studies. Through a review of 63 peer-reviewed studies published between 2023 and 2025 in leading NLP and AI venues, we reveal a critical gap: task and population of interest are often underspecified in persona-based experiments, despite personalization being fundamentally dependent on these criteria. Our analysis shows substantial differences in user representation, with most studies focusing on limited sociodemographic attributes and only 35% discussing the representativeness of their LLM personae. Based on our findings, we introduce a persona transparency checklist that emphasizes representative sampling, explicit grounding in empirical data, and enhanced ecological validity. Our work provides both a comprehensive assessment of current practices and practical guidelines to improve the rigor and ecological validity of persona-based evaluations in language model alignment research.

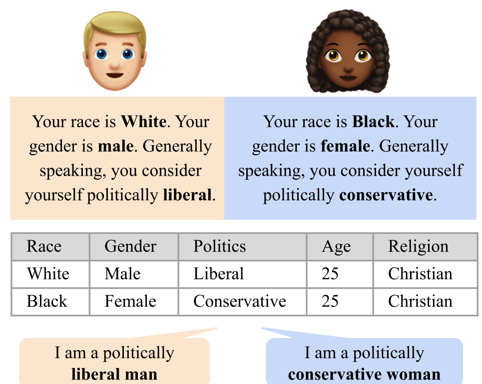

Figure 1: Differences in Synthetic Persona Construction. Demonstrated on an adapted example from Hu and Collier (2024), a study included in our review corpus.

Introduction

Large Language Models (LLMs) have rapidly proliferated across domains, yet ensuring their beneficial alignment with diverse users' preferences and values has become increasingly challenging (Weidinger et al. 2024). As heterogeneous user groups, organizations, and cultures interact with the same underlying models (Sorensen et al. 2024), LLM alignment is evolving beyond enforcing universal predefined values toward more "personalized alignment" approaches (Kirk et al. 2024a). These customization needs become particularly critical as systems are deployed in high-stakes environments, from healthcare consultation to educational contexts, where researchers have adopted synthetic personae as a methodological approach to evaluate and improve LLM performance across diverse user populations (Hu and Collier 2024; Gupta et al. 2023). For instance, while persona-based alignment can be used to communicate medical documents in a personalized language (Mullick et al. 2024), misaligned chatbots could be offensive in response to their assigned persona or user characteristics (Khan et al. 2024).

Correspondance: jan.batzner@weizenbaum-institut.de
Accepted for Publication at AAAI/ACM AIES 2025.
Presented at NeurIPS 2025 LLM-Eval Workshop.

Synthetic personae are constructed profiles using sociodemographic attributes, values, and behavioral traits. These can reflect real-world users or "imaginary people" (An et al. 2018), ranging from sociodemographic statements like "I am a woman. I have 2 kids" (Wan et al. 2023) to preferences such as "I enjoy teaching things to children" (Chen et al. 2025) or "I love to go to Disney World" (Kane and Schubert 2023). As LLMs are increasingly shaping our information ecosystems and used as decision support tools (Benary et al. 2023), persona-based evaluations have become an essential practice. Personae assigned through prompt instructions offer versatile applications, including in-context personalization, developing more engaging AI companions, and model evaluations.

Designing representative personae for real-world applications requires defining both the *task* and the *population of interest*. Unclear task boundaries can lead to overgeneralized claims and evaluations, a scenario Raji et al. (2021) refer to as the "everything and the whole wide world" benchmark problem. Therefore, unified dataset diversity scores might miss the essential specifications of *task* and *target population*. Talat et al. (2022) describe how attempts to aggregate diverse human judgments into unified models can be problematic, particularly when "the average view is implicitly identified with moral correctness" while obscuring whose perspectives are actually represented. While recent work has evaluated LLM benchmark quality and proposed more representative alternatives (Raji et al. 2021; Reuel et al. 2024; Kirk et al. 2024b), a comprehensive assessment of synthetic personae in LLM research remains a critical gap. In this paper, we address this shortcoming and make the following contributions:

1. **Literature Review:** We evaluate 63 papers published in leading NLP and AI venues between 2023 and 2025 that use synthetic personae, analyzing sociodemographic representation and methodological practices.
2. **Ecological Validity Assessment:** We find poor ecological validity in current LLM persona experiments, failing to reflect real-world demographics, user interactions, and domain datasets.[1]
3. **Pathways to Transparency:** We synthesize our findings into concrete guidelines and present a checklist for developing synthetic personae in LLM research.

Related Work

The use of personae in human-computer interaction literature predates LLMs, with researchers, product designers and marketers constructing personae since the 2000s to represent specific user types (Jung et al. 2017; Salminen et al. 2018). The user persona should enable companies to better identify the needs of their target users (Miaskiewicz and Kozar 2011). Early personae studies relied on surveys, interviews, and ethnographic studies but were constrained by small sample sizes, high costs, and temporal limitations (Zhang, Brown, and Shankar 2016). The availability of user data gathered through social media platforms allowed quantitative persona creation, leveraging computational methods on large-scale user data from online platforms to identify behavioral patterns across demographic groups (Salminen et al. 2020a; An, Kwak, and Jansen 2017). LLMs further enabled simulation studies with persona-based agents, allowing developers to examine scenarios where agents interact based on assigned personas and platform design (Park et al. 2022). However, researchers often did not assess whether these personae accurately capture the underlying user population they are intended to reflect or mimic (Salminen et al. 2020b). Critically, most persona creation research models "representative populations" rather than specific subgroups (Salminen et al. 2020a), a limitation mirrored in our LLM persona review (Table 3), where 43% (n=27) of the studies target undifferentiated general populations. The lack of representativeness assessment has therefore been a longstanding issue in personae research that warrants attention.

Checklists in AI Research Checklists have emerged as a critical tool for improving transparency, reproducibility, and methodological rigor in machine learning research (Gebru et al. 2021; Mitchell et al. 2019; Orr and Crawford 2024; Kapoor et al. 2024; Raji et al. 2021). They have only recently been formalized within the ML community as a response to identified reproducibility crises and systematic challenges in research quality assessment. The development of these checklists for ML-based research reflects a growing recognition that structured frameworks can help researchers address common pitfalls and improve transparency (Kapoor et al. 2024).

One early version of an AI checklists is the Model Cards project by Mitchell et al. (2019). They encouraged researchers to consider a model's target user group and how performance might vary across user characteristics. For example, facial recognition models exhibited different error rates based on skin color. Gebru et al. (2021)'s "Datasheets for Datasets" framework established a template for thorough dataset documentation, ranging from motivation to composition, preprocessing, use, distribution, and maintenance. They refer to datasheets for hardware components and advocate for more equal transparency in ML research. Other ML checklists have since emerged, including REFORMS for ML-based science (Kapoor et al. 2024), BetterBench for LLM benchmarks (Reuel et al. 2024), and guidelines for dataset curation (Orr and Crawford 2024; Zhao et al. 2024). Reuel et al. (2024)'s assessment of AI benchmarks revealed substantial quality differences among common benchmarking practices, identifying rigorous documentation standards. Similarly, REFORMS comprises 32 checklist questions across eight project steps of conducting and reporting a Machine Learning project, developed through expert consensus involving domain experts from various fields to ensure broad applicability.

In this paper, we create the Persona Transparency Checklist that builds upon the above practices, while addressing the unique challenges of LLM persona datasets. Building on previous checklist frameworks, our checklist emphasizes methodological transparency and reproducibility. However, we specifically focus on dimensions critical to persona-based evaluation: application domain, target population, data source, ecological validity, reproducibility, and generalizability. By situating our checklist within this broader tradition of ML evaluation frameworks, we contribute to ongoing efforts to enhance methodology standardization while addressing the specific needs of persona-based LLM research.

Method

Our study employs a structured literature review approach to map the landscape of synthetic personae studies in LLM research, identify key concepts and highlight knowledge gaps. We conducted a systematic search and screening process to identify relevant literature.

Eligibility Criteria We established the following inclusion criteria: (i) studies involving computational experiments with language models, excluding conceptual works; (ii) empirical evaluation of at least one pretrained large language model; (iii) publication as full papers, excluding abstracts, workshop papers, or work-in-progress submissions; and (iv) publication in high-impact AI and NLP venues that

[1] Ecological validity refers to the extent to which research experiments emulate and can be generalized to real-world settings and conditions (Schmuckler 2001).

influence research directions in conversational AI, specifically ICML, NeurIPS, ICLR, CHI, AAAI, FAccT, AIES, and conferences within the *ACL Anthology.

Search Strategy We conducted searches across the proceedings of the specified venues for papers published between January 2023 and April 2025. This timeframe captures the recent surge in persona-based LLM research that has emerged alongside advances in large language models. We employed a broad search strategy using the term "persona" in titles and abstracts to identify all studies exploring persona-based approaches.

Selection Process Two authors independently screened all identified papers using a two-stage process: initial title and abstract screening followed by full-text review. Disagreements were resolved through discussion and, when necessary, consultation with a third reviewer. During screening, we excluded studies that did not meet our computational focus, did not evaluate pretrained language models, or were not a full paper. After removal of duplicates and application of our selection criteria, our final corpus comprises 63 articles that form the foundation of our analysis.[2]

Persona Probe	Reference
"I am a woman. I have 2 kids."	Wan et al., 2023
"You are [...] from New York City."	Malik et al., 2024
"I love to go to Disney World."	Kane et al., 2023
"Speak like Muhammad Ali."	Deshpande et al., 2023
"You are a conservative person."	Shu et al., 2024
"Your race is Black."	Hu & Collier, 2024
"[A]verage in your computer skills."	Zhang et al., 2023
"Age: 73", "Openness: Extremely High."	Castricato et al., 2025

Table 1: Examples of synthetic persona descriptions from reviewed paper corpus.

Content Analysis Approach

Given the growing variety in LLM persona research, literature reviews and evaluations can help synthesize findings and identify best practices. To develop a checklist for persona-based LLM research, we used a multi-author iterative approach for codebook development and content analysis. The final version of our codebook resulted in a standardized checklist that operationalizes evaluation criteria, enabling comprehensive assessment of synthetic persona usage across our selected corpus.

In the initial phase, the first author created a preliminary codebook based on randomly selected papers from our corpus. This draft codebook contained categories addressing methodological transparency, data sources, and reproducibility considerations in synthetic persona development as informed by the ML checklists discussed earlier, as well as persona-specific features such as sociodemographic representation. We decided to include open text and qualitative assessment elements in our review, because they capture critical contextual information that a multiple choice approach might miss. For instance, the extent to which persona construction is grounded in the social science literature or assessments of the rationale for specific attribute selection requires nuanced evaluation that goes beyond binary coding. Our approach allows us to identify not only which attributes were represented but also how thoroughly researchers engaged with questions of representativeness and ecological validity.

In the second phase, the codebook was refined. This phase involved four authors of this paper, who independently coded the same subset of papers using the preliminary codebook. Following this first round of coding, we identified disagreements in the annotations between authors and revised the codebook through consensus meetings, which enabled (i) clarification of ambiguous coding categories, (ii) the addition of previously unidentified elements, and (iii) consolidation of overlapping codes.

In the third phase, we specified multiple questions on task and population of interest to better assess representativeness and specifically ecological validity. Each paper was coded by two researchers using the checklist, with disagreements resolved through discussion to maintain consistency. This iterative process resulted in our final *Checklist for Persona-based LLM Research*.

[2]Final Review Corpus: github.com/janbatzner/WhosePersonae

Typology of Personae

Our analysis reveals how researchers construct personae in LLM research in a variety of studies. Based on this analysis, we develop a typology consisting of five primary types of personae that differ in their formatting, level of explicitness, and data structure:

I am (Format: role-play) This type is based on first-person statements to explicitly define persona characteristics. These descriptions serve as direct instructions for in-context personalization, such as "I am a woman. I have 2 kids" (Wan et al. 2023). These personae often combine multiple sociodemographic attributes into one longer prompt. The first-person format simulates a user interaction with an LLM, while commonly being fully constructed. Note that this is a well-known role-playing prompting strategy (Hu and Collier 2024; Batzner et al. 2025b; Kim, Koo, and Lim 2024; Lim et al. 2023).

You are (Format: role-play) Second-person instructional statements directly assign roles to the model, such as "You are a person from New York City" (Malik, Jiang, and Chai 2024) or "You are politically conservative" (Hu and Collier 2024). This format is widely used in LLM role-playing experiments, with various applications in healthcare, education, costumer support, coaching, and AI companions (Louie et al. 2024). The second-person format is particularly prevalent in fairness and bias evaluation studies, where researchers test how models respond when explicitly instructed to adopt specific sociodemographic characteristics. This approach is often combined with explicit role-playing instructions. Hu and Collier (2024) have raised questions about the steerability differences for certain personae across different LLMs. In previous work, we highlighted potential overlaps in model responses to "I am" and "You are" persona instructions (Batzner et al. 2025a).

Preferences (Format: unstructured) This type involves simple prompts that directly state the preferences of a synthetic user persona like "I love to go to Disney World every year" (Kane and Schubert 2023). While often combined with the "I am" type of sociodemographic attributes, this type includes any format that directly prompts specific user preferences to the model.

Real Conversations (Format: chat data) Some studies are based on implicit personae that are derived from actual chat conversation data. Rather than explicitly stating sociodemographic attributes, these approaches extract persona characteristics from conversational patterns, stylistic elements, or topical preferences as exhibited in real human conversations. While providing prima facie the highest ecological validity, most works rely on modifications of the *PersonaChat* dataset (Yamashita et al. 2023). Therefore, to meaningfully evaluate the representativeness of those chat personae, the *task* and *population of interest* must be taken into account.

Survey Responses (Format: tabular data) This approach constructs personae based on tabular survey data, often in csv or json format. For instance, the *OpinionQA* dataset is based on Pew Research Public Opinion Polls. Castricato et al. (2025) demonstrate this approach with structured attributes such as "Age 73, [...] Filipino, Openness: Extremely High." This type offers greater standardization and experiment control across personae but may sacrifice the ecological validity of narrative personae. One persona would therefore seek to emulate the survey choices of one respondent, which allows scalable, empirically grounded experiments.

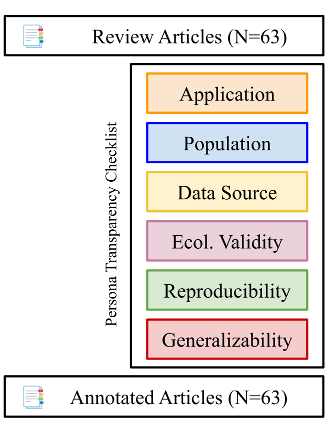

Figure 2: Persona Transparency Checklist.

Results

Checklist for Persona-based LLM Research

Based on our literature review and the iterative codebook development, our checklist for persona-based LLM research encompasses six key evaluation dimensions.

Application

Assessment Criteria: Application

- ☐ **Task Definition**: Was the measured task clearly defined?
- ☐ **Capability Categorization**: Which kind of capability was evaluated?
- ☐ **Application Domain**: What is the specific domain context?
- ☐ **Use Case Specification**: Were concrete application use cases described?

Similarly to LLM performance benchmarks, the *task* of interest needs to be clearly defined first (Raji et al. 2021). Our assessment framework examines two key dimensions. First, *task definition* and *capability classification* to evaluate whether papers explicitly state which capabilities are being evaluated. Second, *application domain* and *use case specification* to assess whether the specific deployment context and concrete implementation scenarios were described.

Task Categorization	Share	Example
Personalization	44%	Personalized RAG
Robustness	22%	Persona-consistent dialogue
Bias/Fairness	18%	Identify social biases
Domain-Specific	16%	Persona-based healthcare

Table 2: Task distribution of persona research papers.

As shown in Table 2, our analysis reveals a strong preference for broad personalization (44%, n=28), while only a subset (16%, n=10) target domain-specific applications. As Raji et al. (2021) and Kirk et al. (2024a) emphasize, without clearly defined tasks, claims about personalization or other capabilities remain fundamentally incomplete: we cannot meaningfully evaluate **what** is being personalized without specific application definitions.[†]

[†]Checklist Development Process: Papers were categorized by two researchers using open text coding followed by manual assignment to the best-fitting category. Categories are not mutually exclusive, presented percentages reflect the primary category.

Population

Assessment Criteria: Population

- ☐ **Target Population**:
  What population group was represented?
- ☐ **Sociodemographic Attributes**:
  Which demographic attributes were included?
- ☐ **Persona Type**:
  How were personae structured and presented?

After defining the specific task, research on synthetic personae must specify **who** it is personalized for. Our population assessment evaluated three critical dimensions: the identification of target populations, the selection of sociodemographic attributes, and the persona structure used to describe these personae.

As shown in Table 3, our analysis reveals a lack of population specificity. Over a third of the reviewed papers (43%, n=27) target an undifferentiated "general population," while more specific categories like occupational (8%, n=5) and healthcare populations (5%, n=3) receive much less attention.[†] This generalization mirrors the task definition problem identified earlier: without clearly specified populations, persona representativeness cannot be meaningfully addressed. General population approaches risk creating what Talat et al. (2022) describe as a fundamental disconnect between the subjective human judgments being modeled and the perspectives that are actually represented.

Target Population Category	Share	Example
General Population	43%	Global
Platform Usage	25%	Users of r/Journaling
Simulation/Fictional	11%	Movie Characters
Geographic Identity	8%	US demographic
Occupational	8%	Academics
Healthcare	5%	Diabetes Patient

Table 3: Target Population Distribution.

Our analysis further identifies the sociodemographic attributes most commonly used in synthetic personae research. Figure 4 shows gender (n=25), age (n=19), as well as race and ethnicity (n=17) appear most frequently, followed by education (n=14) and religion (n=12). These differ from attributes commonly addressed in platform content moderation guidelines,[4] such as disability status (n=5), sexual orientation (n=3), and veteran status (n=1).

[4]Content moderation criteria typically include race, ethnicity, age, religion, non-binary gender, disability, language, sexual orientation, and veteran status based on (Meta 2025). These align with sensitive personal data categories defined in EU General Data Protection Regulation (GDPR) Articles 4(13)-(15) and Article 9.

Sociodemographics	**Count**
Gender	25
Age	19
Race or Ethnicity	17
Political Views	16
Education	14
Religion	12
Non-Binary Gender	7
Economic Status	5
Language	5
Disability	5
Sexual Orientation	3
Veteran Status	1

Table 4: Explicit sociodemographic attribute mentions in reviewed papers. Half of papers (n=30) mention no sociodemographic persona attributes in their main text.

Data Source

Assessment Criteria: Data Source

- ☐ **Originality**:
  Were existing datasets reused or modified?
- ☐ **Dataset Reference**:
  Were existing datasets referenced or reused?
- ☐ **Construction Method**:
  How were the personae designed and created?

The data source assessment examines how researchers generated the personae used in their studies. Here, we focused on dataset originality, reference sources, and construction methods. Our analysis shows reliance on existing resources, with 33% (n=21) of reviewed studies using unmodified datasets like *PersonaChat* (Zhu et al. 2023; Lee, Oh, and Lee 2023; Kim, Koo, and Lim 2024) and an additional 16% (n=10) implementing only minor modifications to existing persona collections like *SyntheticPersonaChat* (Chen et al. 2025).

Ecological Validity

The ecological validity assessment examines whether synthetic personae and experimental designs reflect real-world human populations and usage scenarios. Our assessment framework distinguishes between empirical grounding, which examines whether personae are based on verifiable demographic data and social science research; and ecological validity, which assesses whether interaction settings reflect real-world deployment contexts. Our analysis reveals gaps across all dimensions: 65% (n=41) of papers did not explicitly discuss the representativeness of their personae in the main text of their papers. Similarly, 60% (n=38) of studies employed fully constructed interaction settings unlikely to reflect how users would naturally interact with LLMs in practice. A common example is when researchers directly inject demographic traits from survey responses as descriptions into the model like "Suppose there is a person who is politically liberal and opposes increased military expansion" (Liu, Diab, and Fried 2024). While such approaches allow researchers to observe how the model behaves under the prompted persona, such personae are rarely introduced by real-world users in this format. These findings highlight opportunities to strengthen the ecological validity of research relying on synthetic personae, potentially improving the applicability of findings to diverse real-world contexts.

Assessment Criteria: Ecological Validity

- ☐ **Representativeness**: Reflects distribution of relevant user demographic?
- ☐ **Empirical Grounding**: Empirical evidence like social science or user data?
- ☐ **Interaction Ecology**: Experiment reflective of human-AI interactions?

Reproducibility

Assessment Criteria: Reproducibility

- ☐ **Code Repository**: Is the experiment code publicly shared?
- ☐ **Dataset Availability**: Complete persona dataset provided?
- ☐ **Documentation Completeness**: Documentation sufficient to reproduce experiment?

Our reproducibility assessment evaluates whether synthetic personae datasets can be independently built upon by other researchers. This evaluation became necessary due to gaps in documentation practices we encountered across our corpus. While 78% (n=50) of the reviewed papers included any supplementary material link, predominantly to GitHub code repositories (70%, n=44), the remaining papers provided no link to their persona datasets. Among papers that included dataset links, we observed various limitations. For instance, repositories included only exemplary probes rather than complete datasets, provided incomplete generation scripts, or included limited documentation. This lack of transparency hinders evaluation and meta-analysis efforts (Gebru et al. 2021; Reuel et al. 2024) and poses critical challenges for assessing representativeness. These findings originally prompted our decision to conduct an expert-annotated paper review rather than attempt to aggregate or compare the actual personae datasets directly.

Generalizability

We split the last section into baselines and transparency. Our baselines assessment evaluates whether researchers benchmark their experiments against existing methods and across different demographic groups. Notably, papers commonly did not compare model performance across different social groups or against existing persona datasets or established performance baselines, limiting their ability to demonstrate methodological improvements or evaluating bias.

Assessment Criteria: Baselines

- ☐ **Dataset Comparison**: Compared against established persona datasets?
- ☐ **Social Group Analysis**: Evaluated differences across social groups?
- ☐ **Performance Baselines**: Compared personae to performance baselines?

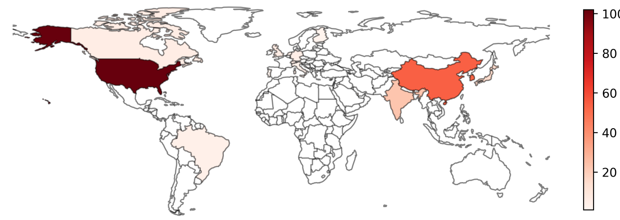

Figure 3: Global Author Location Distribution: The major author university affiliations in our corpus are the USA (102, 34%), China (54, 18%), South Korea (52, 17%), India (23, 8%), Singapore (17, 6%), and Japan (15, 5%), among others.

Assessment Criteria: Transparency

- ☐ **Funding Transparency**: Are the funding sources clearly disclosed?
- ☐ **Ethics Discussion**: Ethical considerations of persona design included?
- ☐ **Geographic Distribution**: Regional knowledge of author team?
- ☐ **Positionality Statement**: Do authors acknowledge their positionality?
- ☐ **Limitations Acknowledgment**: Discussed limitations of their personae explicitly?

Lastly, we examine researchers' transparency practices regarding funding, ethics, and limitations in their persona-based studies. While the importance of positionality statements varies depending on application domain (e.g., more critical for culturally-sensitive applications), the analysis found that none of the 63 reviewed papers included an explicit positionality statement.[5] Although most papers included limitations sections discussing persona constraints, none contained explicit acknowledgments of how author backgrounds might influence design decisions. Our review corpus shows a notable geographic concentration, with 34%

[5]Similar transparency limitations have been observed in popular LLM benchmark studies (Kraft, Simon, and Schimmler 2025).

(102 authors) affiliated with US institutions and 18% (54 authors) with Chinese institutions. Notably, 40% of the papers we reviewed have at least one US-based co-author, compared to 19% for China.

Pathways Toward Enhanced Transparency

Based on our review of synthetic personae in LLM research, we propose the following six recommendations to enhance the transparency, quality, and representativeness of synthetic persona experiments:

(1) Application: Define task of interest clearly

Researchers must clearly define specific tasks for which personae are designed instead of making overly global claims (Table 2). Stating the "intended use" (Mitchell et al. 2019) and the "motivation for dataset creation" (Gebru et al. 2021) as recommended in ML-based research should equally apply to persona experiments in LLM research. The domain of interest needs to be defined to select use case-specific performance metrics instead of generic measures, e.g., healthcare applications need different evaluation criteria than applications in educational or customer service domains. Therefore, synthetic personae should be created to meet the specific domain and context requirements, such as clinical accuracy for healthcare or pedagogical appropriateness for educational tools.

(2) Population: Specify Demographic Target Group

Researchers should explicitly define which demographic target group their personae represent instead of relying on generic or generalized descriptions (Table 3). Based on the task, domain, and use case defined earlier, the representativeness of synthetic personae depends on the population of interest. In ML-based research, an insufficient definition of the target group has been identified as a common limitation. Information on the distribution of subpopulations by sociodemographic aspects and a reflection on representativeness of these groups are required (Kapoor et al. 2024). When constructing persona datasets, the relevant subset of sociodemographic aspects is dependent on its application. Our analysis highlights that to identify the target population, e.g., user communities on the social media platform Reddit (Pal, Das, and Srihari 2024), researchers must carefully select relevant persona attributes in that particular context.

(3) Data Source: Empirically Ground the Data

After the task and the target user population are defined, the synthetic persona dataset can be created. While the lack of transparency in dataset creation is an open challenge in ML research (Kapoor et al. 2024; Gebru et al. 2021; Reuel et al. 2024), persona datasets are a particularly sensitive domain. As the majority of studies in our review were motivated by personalization, transparency on the data sources is essential to evaluate representativeness. We recommend documenting the persona construction process, including which datasets were used, modified, or created to construct the synthetic personae. The methods and sampling approach should be

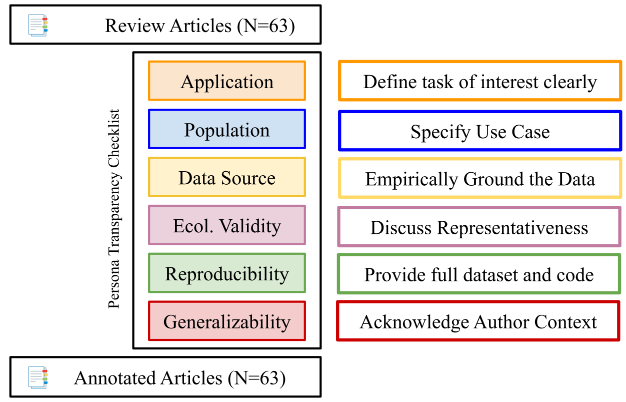

Figure 4: Pathways Toward Enhanced Transparency: Recommendations for Synthetic Persona Construction.

stated clearly, along with a disclosure of LLM-generated elements. Moreover, we recommend to base persona attributes on real demographic data, census information, or user statistics whenever possible, with appropriate references.

(4) Ecological Validity: Discuss Representativeness

Empirically grounded user data does not guarantee ecological validity. Whether an experiment can generalize to real-world user interactions (Schmuckler 2001) cannot be determined solely from user demographics or platform statistics. Therefore, researchers should evaluate population representativeness and ecological validity as distinct considerations. Real user interactions with LLMs may differ substantially from experimental settings, even when demographic characteristics are accurately represented. While achieving ecological validity in large-scale LLM experiments presents challenges, researchers should explicitly discuss the interaction context and provide evidence for how their experimental design relates to real-world usage patterns.

(5) Reproducibility: Provide Full Dataset and Code

Computational reproducibility, including code availability, dataset access, documentation, and reproduction scripts (Kapoor et al. 2024; Mitchell et al. 2019; Reuel et al. 2024), remains an ongoing challenge in ML-based research. Our review found that many persona datasets were built upon similar underlying sources, highlighting opportunities to strengthen documentation practices. Researchers could enhance reproducibility by providing comprehensive documentation in public repositories, including persona generation code, final datasets, and statistical distributions of demographic attributes. When using LLM-generated personae, we recommend releasing complete datasets rather than selected examples or prompts alone, which would facilitate meta-analyses and replication studies.

(6) Generalizability: Acknowledge Author Context

While ethics statements have been increasingly integrated into ML research, we recommend enhanced transparency through researcher positionality statements and funding disclosures. Such statements should discuss potential impacts on generalizability, addressing the absence of positionality acknowledgments in our corpus despite their importance in research involving human representation.

Limitations

First, our literature corpus has several constraints. We focused exclusively on leading AI conferences (2023-2025) and identified relevant contributions through keyword searches for "persona." While this approach helped us identify key studies, it likely excluded relevant work published in other venues, timeframes, and studies using alternative terminology, particularly from product development, marketing, or social science research. Additionally, excluding non-peer-reviewed preprints and workshop papers means we may not have captured the most recent scholarship.

Second, despite employing a two-author screening process with iterative discussions, our analysis relies on qualitative coding. While this research design enabled the iterative design of the persona transparency checklist, the results inevitably shaped by the authors' perspectives and understanding.

Conclusion

Synthetic personae studies have become a prominent method in AI alignment research. Whether based on user surveys or LLM-generated ones, the diversity representation and validity of these personae vary considerably across studies. Synthetic persona datasets provide a valuable resource for aligning, personalizing, and evaluating language models. We conducted a literature review of 63 persona studies from leading AI venues, informed by existing ML research checklists, and derive six recommendations for creating representative and transparent synthetic persona datasets in LLM research.

Our analysis identifies opportunities to strengthen persona representativeness in existing research designs: 43% (n=27) of studies target undifferentiated "general populations," while 35% (n=22) explicitly discuss representativeness. Addressing these areas could enhance the ecological validity of persona-based evaluations and improve the generalizability to real-world deployment scenarios. By synthesizing established ML documentation frameworks with our literature review findings, we developed a persona-specific transparency checklist that emphasizes the application, population, data source, ecological validity, reproducibility, and generalizability. As LLMs gain greater importance in high-stakes domains, evaluating persona datasets for representativeness and ecological validity becomes increasingly important.

Ethics and Adverse Impacts Statement

This study examines published research papers using publicly available information and does not involve human subjects or personal data collection. While our work aims to improve the representativeness and ethical use of synthetic personas, we acknowledge that highlighting demographic attributes risks reinforcing categorizations of human identity that may oversimplify intersectional experiences. We acknowledge the dual-use potential of user persona datasets, which could be exploited for malicious purposes such as targeted manipulation or discriminatory profiling, emphasizing the importance of ethical guidelines and access controls.

Acknowledgements

This research was supported by the Federal Ministry of Education and Research of Germany (BMBF) under grant 16DII131 "Weizenbaum Institut für die vernetzte Gesellschaft" and the German Research Foundation (DFG), "Schwerpunktprogramm: Resilienz in Vernetzten Welten" (SPP 2378, Projekt ReNO, 2023-2027). We acknowledge Columbia University's Institute for Social and Economic Research and Policy, Quantitative Methods in the Social Sciences, and Columbia Data Science Institute. This work benefited from feedback received at the ACM Conference on Fairness, Accountability, and Transparency (FAccT) Doctoral Consortium. We thank Antonia Döring, Carlo Uhl, Jonathan Reti, Merle Uhl, Elena Krumova, and Monserrat Lopéz Pérez for their valuable input and feedback.

References

Agrawal, H.; Mishra, A.; Gupta, M.; and Mausam. 2023. Multimodal Persona Based Generation of Comic Dialogs. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14150–14164. Toronto, Canada: Association for Computational Linguistics.

Ahmad, Z.; Mishra, K.; Ekbal, A.; and Bhattacharyya, P. 2023. RPTCS: A Reinforced Persona-aware Topic-guiding Conversational System. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3482–3494. Dubrovnik, Croatia: Association for Computational Linguistics.

An, J.; Kwak, H.; and Jansen, B. J. 2017. Personas for content creators via decomposed aggregate audience statistics. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 632–635.

An, J.; Kwak, H.; Jung, S.; Salminen, J.; Admad, M.; and Jansen, B. 2018. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Trans. Web*, 12(4).

Batzner, J.; Stocker, V.; Schmid, S.; and Kasneci, G. 2025a. GermanPartiesQA: Benchmarking Commercial Large Language Models and AI Companions for Political Alignment and Sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Batzner, J.; Stocker, V.; Schmid, S.; and Kasneci, G. 2025b. Sycophancy Claims about Language Models: The Missing Human-in-the-Loop. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.

Benary, M.; Wang, X. D.; Schmidt, M.; Soll, D.; Hilfenhaus, G.; Nassir, M.; Sigler, C.; Knödler, M.; Keller, U.; Beule, D.; et al. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11): e2343689–e2343689.

Castricato, L.; Lile, N.; Rafailov, R.; Fränken, J.-P.; and Finn, C. 2025. PERSONA: A Reproducible Testbed for Pluralistic Alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, 11348–11368.

Chen, R.; Wang, J.; Yu, L.-C.; and Zhang, X. 2023. Learning to memorize entailment and discourse relations for persona-consistent dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11, 12653–12661.

Chen, Y.; Wei, W.; Fan, S.; Xu, K.; and Chen, D. 2025. CoMIF: Modeling of Complex Multiple Interaction Factors for Conversation Generation. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 7355–7366. Abu Dhabi, UAE: Association for Computational Linguistics.

Cheng, J.; Sabour, S.; Sun, H.; Chen, Z.; and Huang, M. 2023. PAL: Persona-Augmented Emotional Support Conversation Generation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 535–554. Toronto, Canada: Association for Computational Linguistics.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1504–1532. Toronto, Canada: Association for Computational Linguistics.

Choi, H. K.; and Li, Y. 2024. PICLe: eliciting diverse behaviors from large language models with persona in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, 8722–8739.

Cunha, R.; Castro Ferreira, T.; Pagano, A.; and Alves, F. 2024. A Persona-Based Corpus in the Diabetes Self-Care Domain - Applying a Human-Centered Approach to a Low-Resource Context. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1353–1369. Torino, Italia: ELRA and ICCL.

Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1236–1270.

Do, X. L.; Kawaguchi, K.; Kan, M.-Y.; and Chen, N. 2025. Aligning Large Language Models with Human Opinions through Persona Selection and Value–Belief–Norm Reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2526–2547.

Gao, J.; Lian, Y.; Zhou, Z.; Fu, Y.; and Wang, B. 2023a. LiveChat: A Large-Scale Personalized Dialogue Dataset Automatically Constructed from Live Streaming. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15387–15405. Toronto, Canada: Association for Computational Linguistics.

Gao, S.; Borges, B.; Oh, S.; Bayazit, D.; Kanno, S.; Wakaki, H.; Mitsufuji, Y.; and Bosselut, A. 2023b. PeaCoK: Persona Commonsense Knowledge for Consistent and Engaging Narratives. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6569–6591.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Ghandeharioun, A.; Yuan, A.; Guerard, M.; Reif, E.; Lepori, M. A.; and Dixon, L. 2024. Who's asking? User personas and the mechanics of latent misalignment. arXiv:2406.12094.

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Ha, J.; Jeon, H.; Han, D.; Seo, J.; and Oh, C. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. arXiv:2402.15265.

Hao, J.; and Kong, F. 2025. Enhancing Emotional Support Conversations: A Framework for Dynamic Knowledge Filtering and Persona Extraction. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 3193–3202. Abu Dhabi, UAE: Association for Computational Linguistics.

Hong, M.; Zhang, C.; Chen, C.; Lian, R.; and Jiang, D. 2024. Dialogue Language Model with Large-Scale Persona Data Engineering. *arXiv preprint arXiv:2412.09034*.

Hu, T.; and Collier, N. 2024. Quantifying the Persona Effect in LLM Simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10289–10307.

Hu, Z.; Chan, H. P.; Li, J.; and Yin, Y. 2025. Debate-to-Write: A Persona-Driven Multi-Agent Framework for Diverse Argument Generation. arXiv:2406.19643.

Huang, Q.; Zhang, Y.; Ko, T.; Liu, X.; Wu, B.; Wang, W.; and Tang, H. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11, 12916–12923.

Hwang, E.; Shwartz, V.; Gutfreund, D.; and Thost, V. 2024. A Graph per Persona: Reasoning about Subjective Natural

Language Descriptions. In *Findings of the Association for Computational Linguistics ACL 2024*, 1928–1942.

Inaba, M. 2024. PersonaCLR: Evaluation Model for Persona Characteristics via Contrastive Learning of Linguistic Style Representation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 674–685.

Jandaghi, P.; Sheng, X.; Bai, X.; Pujara, J.; and Sidahmed, H. 2024. Faithful Persona-based Conversational Dataset Generation with Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, 15245–15270.

Jiang, H.; Zhang, X.; Cao, X.; Breazeal, C.; Roy, D.; and Kabbara, J. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 3605–3627. Mexico City, Mexico: Association for Computational Linguistics.

Jung, S.-G.; An, J.; Kwak, H.; Ahmad, M.; Nielsen, L.; and Jansen, B. J. 2017. Persona generation from aggregated social media data. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, 1748–1755.

Kane, B.; and Schubert, L. 2023. We Are What We Repeatedly Do: Inducing and Deploying Habitual Schemas in Persona-Based Responses. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10998–11016.

Kapoor, S.; Cantrell, E. M.; Peng, K.; Pham, T. H.; Bail, C. A.; Gundersen, O. E.; Hofman, J. M.; Hullman, J.; Lones, M. A.; Malik, M. M.; Nanayakkara, P.; Poldrack, R. A.; Raji, I. D.; Roberts, M.; Salganik, M. J.; Serra-Garcia, M.; Stewart, B. M.; Vandewiele, G.; and Narayanan, A. 2024. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18): eadk3452.

Khan, A. A.; Alam, S.; Wang, X.; Khan, A. F.; Neog, D. R.; and Anwar, A. 2024. Mitigating Sycophancy in Large Language Models via Direct Preference Optimization. In *2024 IEEE International Conference on Big Data (BigData)*, 1664–1671. IEEE.

Kim, D.; Ahn, Y.; Kim, W.; Lee, C.; Lee, K.; Lee, K.-H.; Kim, J.; Shin, D.; and Lee, Y. 2023a. Persona Expansion with Commonsense Knowledge for Diverse and Consistent Response Generation. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1139–1149. Dubrovnik, Croatia: Association for Computational Linguistics.

Kim, D.; Ahn, Y.; Lee, C.; Kim, W.; Lee, K.-H.; Shin, D.; and Lee, Y. 2023b. Concept-based Persona Expansion for Improving Diversity of Persona-Grounded Dialogue. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3471–3481. Dubrovnik, Croatia: Association for Computational Linguistics.

Kim, H.; Ong, K.; Kim, S.; Lee, D.; and Yeo, J. 2024a. Commonsense-augmented Memory Construction and Management in Long-term Conversations via Context-aware Persona Refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 104–123.

Kim, J.; Koo, S.; and Lim, H.-S. 2024. PANDA: Persona Attributes Navigation for Detecting and Alleviating Overuse Problem in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12005–12026.

Kim, M.; Kim, M.; Kim, H.; Kwak, B.-w.; Kang, S.; Yu, Y.; Yeo, J.; and Lee, D. 2024b. Pearl: A Review-driven Persona-Knowledge Grounded Conversational Recommendation Dataset. In *Findings of the Association for Computational Linguistics ACL 2024*, 1105–1120.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4): 383–392.

Kirk, H. R.; Whitefield, A.; Rottger, P.; Bean, A. M.; Margatina, K.; Mosquera-Gomez, R.; Ciro, J.; Bartolo, M.; Williams, A.; He, H.; et al. 2024b. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37: 105236–105344.

Kraft, A.; Simon, J.; and Schimmler, S. 2025. Social Bias in Popular Question-Answering Benchmarks. arXiv:2505.15553.

Kumar, S.; Gupta, R.; Akhtar, M. S.; and Chakraborty, T. 2024. Adding SPICE to Life: Speaker Profiling in Multiparty Conversations. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 415–425. Torino, Italia: ELRA and ICCL.

Lee, J.; Oh, M.; and Lee, D. 2023. P5: Plug-and-Play Persona Prompting for Personalized Response Selection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 16571–16582.

Lee, Y.-J.; Lee, D.; Youn, J.; Oh, K.-J.; Ko, B.; Hyeon, J.; and Choi, H.-J. 2024. Stark: Social Long-Term Multi-Modal Conversation with Persona Commonsense Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12137–12162.

Li, J.; Peris, C.; Mehrabi, N.; Goyal, P.; Chang, K.-W.; Galstyan, A.; Zemel, R.; and Gupta, R. 2024. The steerability of large language models toward data-driven personas. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7290–7305. Mexico City, Mexico: Association for Computational Linguistics.

Li, Y.; Hu, Y.; Sun, Y.; Xing, L.; Guo, P.; Xie, Y.; and Peng, W. 2023. Learning to know myself: A coarse-to-fine persona-aware training framework for personalized dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11, 13157–13165.

Lim, J.; Kang, M.; Kim, J.; Kim, J.; Hur, Y.; and Lim, H.-S. 2023. Beyond candidates: adaptive dialogue agent utilizing persona and knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7950–7963.

Liu, A.; Diab, M.; and Fried, D. 2024. Evaluating Large Language Model Biases in Persona-Steered Generation. In *Findings of the Association for Computational Linguistics ACL 2024*, 9832–9850.

Liu, P.; Huang, Z.; Zhang, X.; Wang, L.; de Melo, G.; Lin, X.; Pang, L.; and He, L. 2023. A disentangled-attention based framework with persona-aware prompt learning for dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11, 13255–13263.

Louie, R.; Nandi, A.; Fang, W.; Chang, C.; Brunskill, E.; and Yang, D. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10570–10603. Miami, Florida, USA: Association for Computational Linguistics.

Mahajan, K.; and Shaikh, S. 2024. Persona-aware Multi-party Conversation Response Generation. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 12712–12723. Torino, Italia: ELRA and ICCL.

Malik, M.; Jiang, J.; and Chai, K. M. 2024. An Empirical Analysis of the Writing Styles of Persona-Assigned LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19369–19388.

Meta. 2025. Community Standards Enforcement Report. Accessed: 20 May 2025.

Miaskiewicz, T.; and Kozar, K. A. 2011. Personas and user-centered design: How can personas benefit product design processes? *Design studies*, 32(5): 417–430.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Mondal, I.; S, S.; Natarajan, A.; Garimella, A.; Bandyopadhyay, S.; and Boyd-Graber, J. 2024. Presentations by the Humans and For the Humans: Harnessing LLMs for Generating Persona-Aware Slides from Documents. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2664–2684. St. Julian's, Malta: Association for Computational Linguistics.

Mullick, A.; Bose, S.; Saha, R.; Bhowmick, A.; Goyal, P.; Ganguly, N.; Dey, P.; and Kokku, R. 2024. On The Persona-based Summarization of Domain-Specific Documents. In *Findings of the Association for Computational Linguistics ACL 2024*, 14291–14307.

Occhipinti, D.; Tekiroglu, S. S.; and Guerini, M. 2024. PRODIGy: a PROfile-based DIalogue Generation dataset. arXiv:2311.05195.

Orr, W.; and Crawford, K. 2024. Building Better Datasets: Seven Recommendations for Responsible Design from Dataset Creators. *Journal of Data-centric Machine Learning Research*.

Pal, S.; Das, S.; and Srihari, R. K. 2024. Beyond Discrete Personas: Personality Modeling Through Journal Intensive Conversations. arXiv:2412.11250.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18.

Peng, L.; and Shang, J. 2024. Quantifying and Optimizing Global Faithfulness in Persona-driven Role-playing. arXiv:2405.07726.

Pillai, R. G.; Fokkens, A.; and van Atteveldt, W. 2025. Engagement-driven Persona Prompting for Rewriting News Tweets. In *Proceedings of the 31st International Conference on Computational Linguistics*, 8612–8622.

Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the everything in the whole wide world benchmark. *Advances in Neural Information Processing Systems*.

Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. J. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. *Advances in Neural Information Processing Systems*, 37: 21763–21813.

Salewski, L.; Alaniz, S.; Rio-Torto, I.; Schulz, E.; and Akata, Z. 2023. In-context impersonation reveals large language models' strengths and biases. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 72044–72057.

Salminen, J.; Guan, K.; Jung, S.-g.; Chowdhury, S. A.; and Jansen, B. J. 2020a. A literature review of quantitative persona creation. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14.

Salminen, J.; Jansen, B. J.; An, J.; Kwak, H.; and Jung, S.-G. 2018. Are personas done?: Evaluating the usefulness of personas in the age of online analytics. *Persona Studies*, 4(2): 47–65.

Salminen, J.; Liu, C.; Pian, W.; Chi, J.; Häyhänen, E.; and Jansen, B. J. 2024. Deus ex machina and personas from large language models: investigating the composition of AI-generated persona descriptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20.

Salminen, J.; Santos, J. M.; Kwak, H.; An, J.; Jung, S.-g.; and Jansen, B. J. 2020b. Persona perception scale: development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. *International Journal of Human-Computer Studies*, 141: 102437.

Schmuckler, M. A. 2001. What is ecological validity? A dimensional analysis. *Infancy*, 2(4): 419–436.

Sengupta, A.; Akhtar, M. S.; and Chakraborty, T. 2024. Persona-aware Generative Model for Code-mixed Language. *Transactions on Machine Learning Research*.

Shea, R.; and Yu, Z. 2023. Building Persona Consistent Dialogue Agents with Offline Reinforcement Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1778–1795.

Shu, B.; Zhang, L.; Choi, M.; Dunagan, L.; Logeswaran, L.; Lee, M.; Card, D.; and Jurgens, D. 2024. You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5263–5281. Mexico City, Mexico: Association for Computational Linguistics.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Sun, C.; Yang, K.; Gangi Reddy, R.; Fung, Y.; Chan, H. P.; Small, K.; Zhai, C.; and Ji, H. 2025. Persona-DB: Efficient Large Language Model Personalization for Response Prediction with Collaborative Data Refinement. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 281–296. Abu Dhabi, UAE: Association for Computational Linguistics.

Takayama, J.; Ohagi, M.; Mizumoto, T.; and Yoshikawa, K. 2025. Persona-Consistent Dialogue Generation via Pseudo Preference Tuning. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 5507–5514. Abu Dhabi, UAE: Association for Computational Linguistics.

Talat, Z.; Blix, H.; Valvoda, J.; Ganesh, M. I.; Cotterell, R.; and Williams, A. 2022. On the Machine Learning of Ethical Judgments from Natural Language. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 769–779. Seattle, United States: Association for Computational Linguistics.

Tanprasert, T.; Fels, S. S.; Sinnamon, L.; and Yoon, D. 2024. Debate chatbots to facilitate critical thinking on youtube: Social identity and conversational style make a difference. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–24.

Wan, Y.; Zhao, J.; Chadha, A.; Peng, N.; and Chang, K.-W. 2023. Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9677–9705.

Wang, Z.; Mao, S.; Wu, W.; Ge, T.; Wei, F.; and Ji, H. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 257–279. Mexico City, Mexico: Association for Computational Linguistics.

Weidinger, L.; Barnhart, J.; Brennan, J.; Butterfield, C.; Young, S.; Hawkins, W.; Hendricks, L. A.; Comanescu, R.; Chang, O.; Rodriguez, M.; et al. 2024. Holistic safety and responsibility evaluations of advanced ai models. *arXiv preprint arXiv:2404.14068*.

Wu, S.; Fung, M.; Qian, C.; Kim, J.; Hakkani-Tur, D.; and Ji, H. 2024. Aligning LLMs with Individual Preferences via Interaction. arXiv:2410.03642.

Yamashita, S.; Inoue, K.; Guo, A.; Mochizuki, S.; Kawahara, T.; and Higashinaka, R. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, 852–861.

Yeo, S.; Lim, G.; Gao, J.; Zhang, W.; and Perrault, S. T. 2024. Help Me Reflect: Leveraging Self-Reflection Interface Nudges to Enhance Deliberativeness on Online Deliberation Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–32. ACM.

Zhang, X.; Brown, H.-F.; and Shankar, A. 2016. Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5350–5359.

Zhao, D.; Andrews, J.; Papakyriakopoulos, O.; and Xiang, A. 2024. Position: Measure Dataset Diversity, Don't Just Claim It. In *International Conference on Machine Learning*, 60644–60673. PMLR.

Zhou, J.; Pang, L.; Shen, H.; and Cheng, X. 2023. SimOAP: Improve Coherence and Consistency in Persona-based Dialogue Generation via Over-sampling and Post-evaluation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9945–9959. Toronto, Canada: Association for Computational Linguistics.

Zhou, W.; Li, Q.; and Li, C. 2023. Learning to Predict Persona Information for Dialogue Personalization without Explicit Persona Description. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 2979–2991. Toronto, Canada: Association for Computational Linguistics.

Zhu, L.; Li, W.; Mao, R.; Pandelea, V.; and Cambria, E. 2023. PAED: Zero-shot persona attribute extraction in dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9771–9787.