Whose Personae?
Review of 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]Technical University Munich

[*]

Abstract

Synthetic personae studies have become a prominent method in AI alignment research. Whether based on inferred personae from user surveys or on LLM-generated ones (e.g., "I am a 38-year-old PhD student at MIT"), the representation and validity of these personae vary considerably across studies. This paper systematically assesses the sociodemographic attributes represented in these personae and evaluates their ecological validity. Based on a review of 69 peer-reviewed studies published between 2023 and 2025 in leading NLP and AI venues, we reveal substantial differences in diverse user representation. Most studies focus on a limited subset of demographic characteristics while excluding critical attributes like disability status, gender, and veteran identity, and only 29.5% of studies ground their persona construction in social science literature. Based on our findings, we propose a standardized framework for synthetic persona development that emphasizes representative sampling, explicit grounding in social science theory, and enhanced ecological validity. Our work provides a comprehensive assessment of current practices and offers practical guidelines based on six recommendations for improving persona-based evaluation in language model alignment research.

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, p. 1). 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 or discriminating in response to its assigned persona or user characteristics (Khan et al. 2024). Such errors could in turn pose serious risks to patient safety.

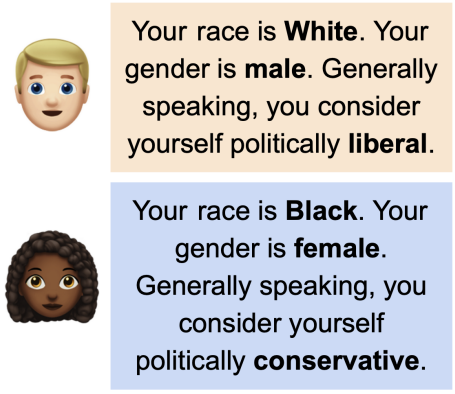

Figure 1: LLM Persona Experiment Example with quotes from Hu and Collier (2024, ACL Full Paper): *'Quantifying the Persona Effect in LLM Simulations'*.

Synthetic personae are constructed profiles using sociodemographic attributes, values, and behavioral traits. These descriptions of "imaginary people" (An et al. 2018a) range 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 integrated into 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, rather than model fine-tuning, offer versatile applications for user-based in-context personalization and the systematic auditing of model behaviors (e.g., bias evaluations). The applications of persona-based role-play extend to numerous domains, including practicing clinical patient scenarios in healthcare education (Louie et al. 2024) and developing more engaging AI companions for various user needs and usage scenarios.

Designing representative and safe personae for real-world applications requires defining both the *task* and the intended

[*]Work done as Columbia University Innovation Lab Mentor.

population of interest. Unclear task boundaries lead to overgeneralized claims and misaligned evaluations, a scenario Raji et al. (2021) refer to as the "everything and the whole wide world" benchmark problem. The same applies to personalization strategies. Only if both the *task* and *target population* are well defined and clearly specified can a truly representative synthetic persona dataset be constructed. As Talat et al. (2022) argue, a critical, yet well-known problem emerges when machine learning systems are designed to capture subjective human judgments without sufficient attention to the perspectives being represented and how those perspectives are operationalized. For LLM benchmarks, first studies evaluated the quality of LLM benchmarks (Raji et al. 2021; Reuel-Lamparth et al. 2024) and suggested more representative LLM benchmarks (Kirk et al. 2024b). While persona datasets do not serve as a performance benchmark, we need to adhere to equal transparency and representativeness standards. However, a systematic assessment of synthetic personae in LLM research remains a critical gap in the literature. In this paper, we address this shortcoming and make the following contributions:

1. **Comprehensive Literature Survey:** We evaluate 69 papers that employ synthetic personae published in top NLP and AI venues (2023-2025), highlighting sociodemographic representation patterns and identifying methodological strengths and weaknesses.
2. **Missing Ecological Validity:** We find poor ecological validity of the current state of LLM persona experiments by failing to reflect real-world demographic distributions, representative user interactions, and domain datasets.
3. **Pathways to Transparency:** We synthesize our findings from the literature survey and provide concrete guidelines for developing transparent synthetic personae, including protocols for explicit attribute reporting, social science grounding, representativeness assessment, and participatory development.

RELATED WORKS

Persona as a Method 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 relied on surveys, interviews, and ethnographic studies but suffered from 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). However, researchers often did not assess whether these computationally-created personae accurately capture the underlying user population (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 4), where one-third of the studies target undifferentiated "general populations". The lack of representativeness assessment has therefore been a long-standing 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 intended user group, as well as how the model's performance could vary depending on user characteristics. For instance, they described facial recognition models that register different error rates depending on the color of the skin. citetgebru2021datasheets's "Datasheets for Datasets" framework established a template for thorough dataset documentation, ranging from motivation to composition, preprocessing, uses, distribution, and maintenance. They refer to datasheets for hardware components and advocate for more equal transparency in ML research. A range of other ML checklists have also been proposed, like REFORMS for ML-based science (Kapoor et al. 2024) and BetterBench (Reuel-Lamparth et al. 2024) for LLM performance benchmarks, and recommendations for ML dataset curation (Orr and Crawford 2024; Zhao et al. 2024). The REFORMS checklist in particular comprises of 32 questions across eight steps of conducting and reporting a Machine Learning project. REFORMS was developed through a consensus process involving domain experts from various fields to ensure broad applicability. Finally, Reuel-Lamparth et al. (2024)'s assessment of AI benchmarks revealed substantial quality differences among common benchmarking practices, showing that small changes in documentation and transparency standards can significantly improve benchmark quality and usability.

Our Synthetic Personae Dataset Transparency Checklist builds upon the above practices, while addressing the unique challenges of LLM persona datasets. Like previous checklist frameworks, our checklist emphasizes methodological transparency and reproducibility. However, we specifically focus on dimensions critical to persona-based evaluation: application domain clarity, population representation, data source integrity, user interactions, and ecological validity considerations. 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

Paper Dataset

Our study employs a systematic literature review approach to analyze the landscape of synthetic personae studies in LLM research. We followed a structured three-stage process to identify and screen relevant papers for analysis. This process is illustrated in Figure 2.

Search For our initial search, we collected papers published between 2023 and 2025, since we focus exclusively on recent advances in LLM alignment. Our initial search on Google Scholar yielded a corpus of 8,150 papers. We then filtered for peer-reviewed articles published by April 2025 containing the concept "*persona*" in the title or abstract, reducing our corpus to 329 papers. Notably, a vast majority of papers are still preprints underlining the contemporary relevance of persona-based LLM research.

Screening In the screening phase, each remaining article was reviewed by two authors to determine eligibility according to our selection criteria. First, we focused on papers with computational experiments, excluding purely qualitative or theoretical works. Second, all papers must evaluate at least one pretrained language model. Third, we focused on full paper publications only, excluding extended abstracts, workshop papers, and work-in-progress. Those should be published in the proceedings of top-tier AI and NLP conference venues: ICML, NeurIPS, ICLR, CHI, AAAI, FAccT, AIES, and the *ACL Anthology. While these are all indexed on Google Scholar, two authors manually checked the proceedings of these venues to ensure all relevant articles are included in our corpus. We selected these venues for their impact on the community and their influence in shaping research directions in conversational AI. During this process, we also identified and removed duplicate articles published in multiple venues.

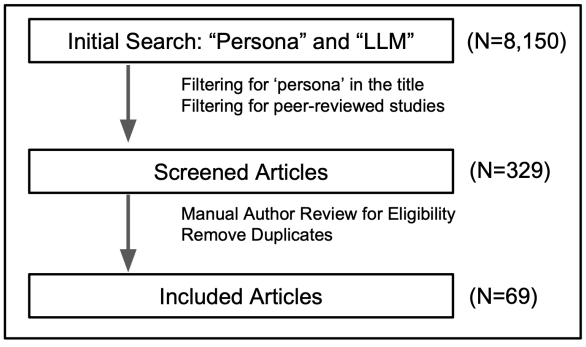

Figure 2: Paper Inclusion Process. Every filtered paper in the second stage was reviewed by the first-author and one annotator.

Content Analysis Approach

Given the current fragmented state of LLM persona research, systematic and evaluations are critical. 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 systematic assessment of synthetic persona usage across our selected corpus (Section: Checklist for Persona-based LLM Research).

Persona Probe	Reference
"I am a woman. I have 2 kids."	Wan et al., 2023
"You are a person from New York City."	Malik et al., 2024
"I love to go to Disney World every year."	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. Your gender is female."	Hu & Collier, 2024
"You are above average in your computer skills."	Zhang et al., 2023
"Age 73", "Filipino", 'Openness: Extremely High."	Castricato et al., 2025

Table 1: Probes of Personae: One-Line Examples from Persona Papers in our review corpus.

In the initial phase, the first author created a preliminary codebook based on 25 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 above (p. 2), as well as persona-specific features such as sociodemographic representation. We decided to include open text and qualitative assessment elements in our checklist, because they capture critical contextual information that multiple choice might miss. For instance, the extent to which persona construction is grounded in social science literature grounding or the rationale for specific attribute selection requires nuanced evaluation beyond binary coding. This approach allowed us to identify not only which attributes were represented but also how thoroughly researchers engaged with questions of representativeness and validity.

In the second phase, the codebook was refined. This phase involved four annotators (all are 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 – a collaborative discussion that led to (i) clarification of ambiguous coding categories, (ii) the addition of previously unidentified elements, and (iii) consolidation of overlapping codes.

In the third phase, a second round of testing was conducted with all four annotators coding an additional subset of papers. We specified multiple questions on task and population of interest to better assess representativeness and specifically *ecological validity*[1]. Each paper was ultimately coded by two authors using the finalized checklist, with disagreements resolved with the first author to maintain consistency throughout the corpus. This iterative process resulted in our final *Checklist for Persona-based LLM Research*.

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

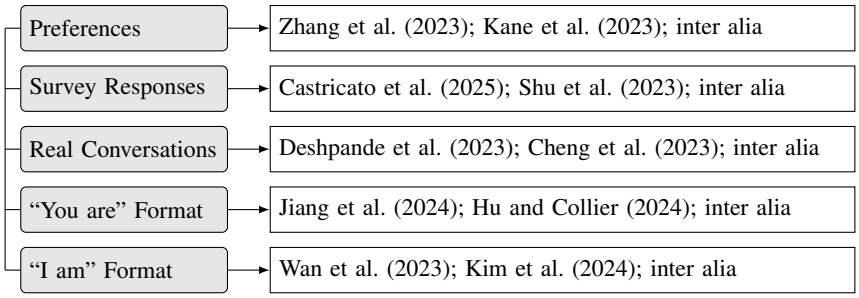

Figure 3: A typology of synthetic personae formats with representative papers.

RESULTS

Typology of Personae

Our analysis revealed 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.

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. Moreover, recent scholarship highlighted potential overlaps in model responses to "I am" and "You are" persona instructions (Batzner et al. 2024).

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 (biographic) 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. Therefore, to 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 typology 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.

Checklist for Persona-based LLM Research

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

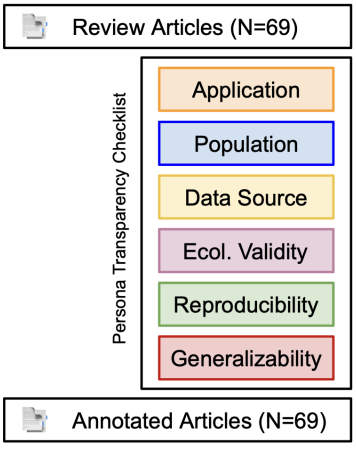

Figure 4: Personae Dataset Transparency Checklist.

Application

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 *task classification* to evaluate whether papers explicitly stated which capabilities were being evaluated and which use case is investigated. Second, *application domain* and *use case specification* to assess whether the specific deployment context and concrete implementation scenarios were described.

Assessment Criteria: Application

Task Definition: Did the paper clearly define what it measured?
Task Classification: Which capability was evaluated (e.g., personalized responses, bias mitigation, factual consistency)?
Application Domain: What is the specific domain context in which personae was applied (e.g., customer support, healthcare, education)?
Use Case Specification: What was the concrete usage scenarios described?

Table 2: Assessment Criteria on Personae Application.

As shown in Table 3, our analysis reveals a strong preference for general capability evaluation rather than domain-specific applications. Most papers focus on general capabilities, such as robustness and consistency (28.5%) or personalization (22.6%), while only 12.9% 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.

Task Categorization	Share	Example
Robustness	28.5%	Persona-consistent dialogue
Personalization	22.6%	Personalized RAG
Bias/Fairness	21.0%	Identify social biases
General Purpose	17.7%	Rewriting tweets
Domain-Specific	12.9%	Persona-based healthcare

Table 3: Task of Persona papers as categorized by authors.

Population

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 format used to describe these personae (Table 4).

As shown in Table 5, our analysis reveals a lack of population specificity. Over a third of the reviewed papers (36.5%) target an undifferentiated "general population," while more specific categories like occupational (7.7%) and healthcare populations (5.8%) receive minimal attention. This generalization mirrors the task definition problem identified earlier: without clearly specified populations, claims about persona representativeness are not supported. 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.

Assessment Criteria: Population

Target Population: What population group the personae were intended to represent (e.g., general population, platform users, geographic region)?
Sociodemographic Attributes: Which demographic characteristics were included in personae (e.g., gender, age, race/ethnicity, education, occupation)?
Persona Description Format: How were personae structured and presented (e.g., "I am a woman", Wan et al. (2023))?

Table 4: Assessment Criteria on Personae Population of interest.

Target Population Category	Share	Example
General Population	36.5%	Global
Geographic Identity	21.2%	US demographic
Platform Usage	17.3%	Users of r/Journaling
Simulation/Fictional	11.5%	Movie Characters
Occupational	7.7%	Academics
Healthcare	5.8%	Diabetes Patient

Table 5: Target Population Categorization.

Our analysis further identifies the sociodemographic attributes most commonly used in synthetic personae research. Figure 5 shows race and ethnicity as well as political views (0.29) appear most frequently, followed by age (0.27) and education (0.22). These differ notably from attributes typically addressed in platform content moderation guidelines (Meta 2025), such as disability status (0.12), sexual orientation (0.06), and non-binary gender identities (0.02). Across all platform content moderation-relevant attributes, we calculate a mean probability of only Mean $F(\text{Attribute})_{\text{platform}} = 0.145$ across all studies.[2]

Data Source

The data source assessment examines how researchers generated the personae used in their studies. Here, we focused on dataset originality, reference sources, and construction

[2]Content moderation criteria examined include race/ethnicity, age, religion, gender (including non-binary identities), 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.

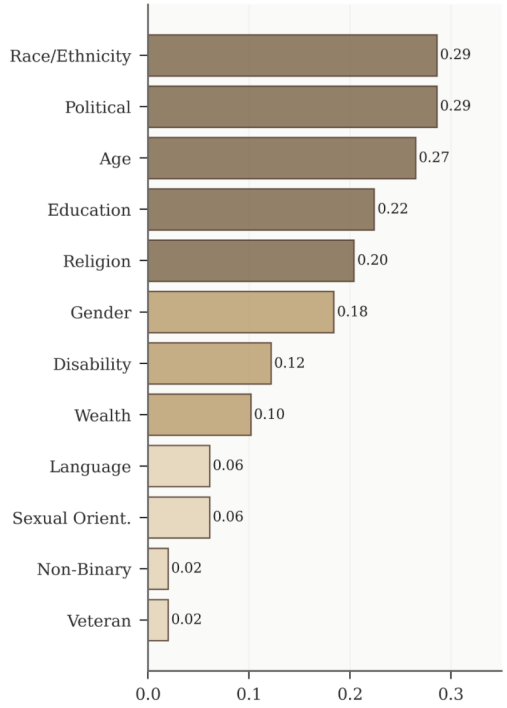

Figure 5: Proportion of reviewed papers explicitly including sociodemographic aspects to be part of their Persona Datasets.

Assessment Criteria: Data Source
Originality: Did the paper provide a method only or create a new persona dataset (e.g., method only using existing dataset, modified existing dataset, original persona dataset)?
Dataset Reference: Were existing datasets referenced or reused (e.g., Persona-Chat, OpinionQA)?
Construction Method: How were the synthetic personae designed and created (e.g., LLM-generated, algorithmic, human-written, web-scraped)?

Table 6: Assessment Criteria on Personae Data Source.

methods (Table 6). Our analysis shows reliance on existing resources, with 22.5% 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 29.6% implementing only minor modifications to existing persona collections like *SyntheticPersonaChat* (Chen et al. 2025). When examining construction methods, we find that a majority of studies (55.6%) used LLMs to generate their personae.

Ecological Validity

The ecological validity assessment examines whether synthetic personae and experimental designs reflect real-world human populations and usage scenarios. Our assessment approach distinguishes between empirical grounding, whether personae are based on verifiable demographic data or social science theory, and ecological validity, whether the interaction settings mirror authentic deployment contexts (Table 7).

Assessment Criteria: Ecological Validity
Representativeness: How did the paper address or discuss whether the personae strategy reflects real-world demographic distributions (e.g., not discussed, survey, social science theory-based)?
Theoretical Grounding: Was the persona construction explicitly grounded in empirical evidence such as social science literature or real user data?
Interaction Ecology: How realistically did the experimental setting reflect actual human-AI interactions in deployment scenarios?

Table 7: Assessment Criteria on Personae Ecological Validity.

Our analysis reveals gaps in both dimensions: 60.8% of papers did not explicitly discuss the representativeness of their personae in the main text of their papers. Similarly, 56.8% 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 (e.g. "Suppose there is a person who is politically liberal and opposes increased military expansion"; (Liu, Diab, and Fried 2024)). While such approaches allows researchers to observe how the model behaves under the prompted persona, personae are rarely invoked by real-world users in this manner. These findings suggest that much current research using synthetic personae as a method lacks the ecological validity necessary to draw meaningful conclusions about model performance in real-world contexts with diverse users.

Reproducibility

Our reproducibility assessment evaluates whether synthetic personae datasets can be independently built upon by other researchers (Table 8). This evaluation became necessary due to the poor documentation practices we encountered across our corpus. While 76% of the reviewed papers included links to code repositories, the remaining papers provided no link to their persona datasets. Of those that did include links, we found repositories that were frequently empty, incomplete, or poorly maintained. Multiple papers referenced datasets that, for instance, offered only limited example probes rather than complete datasets or provided generation scripts without adequate documentation. This lack of transparency hinders evaluation and meta analysis efforts (Gebru et al. 2021; Reuel-Lamparth et al. 2024). These findings are what originally prompted our decision to conduct an expert-annotated paper review rather than attempt to aggregate or compare the actual personae datasets directly.

Our review revealed that code repositories on GitHub associated with persona studies are frequently empty, incomplete, or highly heterogeneous in their data formats and

documentation. This lack of standardization impedes reproducibility efforts and comparative computational analyses across studies. Furthermore, representativeness cannot be universally defined and quantified but must be assessed individually for each context and study, depending on the *task* and *population of interest* (Raji et al. 2021). Simply aggregating diverse persona datasets without accounting for their intended applications risks suggesting misleading evaluations.

Lastly, our analysis revealed transparency gaps in current research practice. Notably, 24% of the examined papers provided no link to their persona datasets whatsoever. Among those that did include dataset links, we observed various limitations: some offered only exemplary probes rather than complete datasets, others provided incomplete attribute lists, and few included comprehensive documentation of their persona development methodology. This lack of transparency poses critical challenges for the assessment of representativeness claims.

Assessment Criteria: Reproducibility
Code Repository: Was code for persona generation or experiments publicly shared (e.g., GitHub, Huggingface)?
Dataset Availability: Were the complete persona dataset directly provided, referenced, or were only generation scripts shared?
Documentation Completeness: Did documentation sufficiently explain how to reproduce the persona dataset and experiment results?

Table 8: Assessment Criteria on Personae Reproducibility.

Generalizability

We split the last section into baselines and author positionality. Our baselines assessment evaluates whether or how researchers benchmark their persona experiments against existing methods and across different demographic groups (Table 9). Notably, 74.3% of papers did not compare model performance across different social groups, making it impossible to detect potential disparities in how LLMs respond to diverse demographic personae. Similarly, many studies lacked comparisons with existing persona datasets or established performance baselines (Cheng, Durmus, and Jurafsky 2023; Dev, Rashidi, and Garg 2023; Cunha et al. 2024), limiting their ability to demonstrate methodological improvements or isolating instruction-following capabilities from bias.

Lastly, the assessment examines how researchers acknowledge their own positioning and limitations in persona design (Table 10). 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 69 reviewed papers included an explicit positionality statement. Although most papers included limitations sections discussing persona constraints, none contained explicit acknowledgments of how author backgrounds might influence design decisions. Additionally,

Assessment Criteria: Baselines
Dataset Comparison: Whether authors compare their approach with other persona datasets or methods
Social Group Analysis: Whether performance differences across different social groups are examined (e.g., comparing model response quality for different demographic personas)
Performance Baselines: Whether general performance baselines for persona adoption are established (e.g., instruction-following success rates)

Table 9: Assessment Criteria on Personae Baselines.

62.1% of authors are affiliated with US American institutions, which may raise questions with regard to the global representation in persona development.

Assessment Criteria: Positionality
Positionality Statement: Whether authors acknowledge their own social positioning and how it might influence persona design
Funding Transparency: Whether funding sources (academic, industry, government) are clearly disclosed
Geographic Distribution: Regional representation in research teams developing synthetic personae
Ethics Discussion: Whether papers include explicit ethical considerations of persona design choices
Limitations Acknowledgment: Whether papers explicitly discuss limitations of their persona approach

Table 10: Assessment Criteria on Author Positionality.

Figure 6: Global Author Location Distribution: In our corpus the top author residences are the US (108, 33.8%), China (66, 20.7%), South Korea (52, 16.3%), India (23, 7.2%), and Singapore (17, 5.3%).

PATHWAYS TOWARD ENHANCED TRANSPARENCY

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

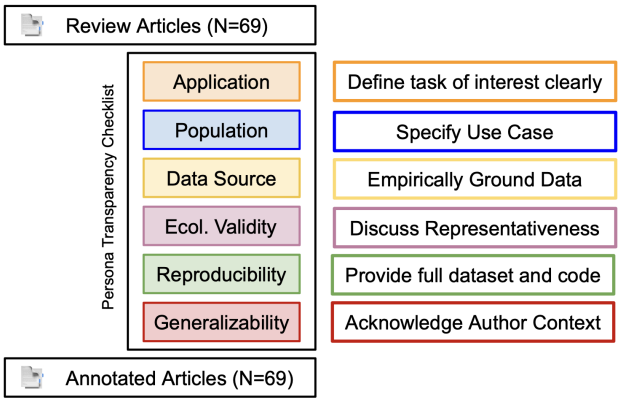

Figure 7: Pathways to Transparency: Recommendations for Synthetic Persona Construction.

(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 3). 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. First, 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 domain and context-specific 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 4). 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 (Gebru et al. 2021) and a justification for the claimed representativeness of these groups (Kapoor et al. 2024) are required. When constructing persona datasets, the relevant subset of sociodemographic aspects is dependent on its application. Our analysis highlights that to identify target population, e.g., user communities on the social media platform Reddit (Pal, Das, and Srihari 2024), researchers must carefully select persona attributes tailored to that particular context.

(3) Data Source: Empirically Ground Data

After the task and the 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-Lamparth et al. 2024), persona datasets are a particularly sensitive domain. As the 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 transform the data into structured personae. The methods and sampling approach should be stated clearly, along with a disclosure of synthetic elements. Therefore, we recommend text-based disclosure instead of a dichotomous yes/no disclosure. 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. The representativeness of a real user interaction or how the experiment could generalize to a real-world user interaction (Schmuckler 2001), is not captured by raw user statistics or platform log data. Therefore, researchers should evaluate the population and ecological validity separately. Real user interactions with LLMs for the particular use case of interest might deviate from the selected experiment setting. While ecological validity can conflict with large-scale LLM experiments, researchers should move towards explicitly discussing the interaction ecology and potential empirical evidence for real-world user behavior.

(5) Reproducibility: Provide Full Dataset and Code

Computational reproducibility, particularly poor code availability, dataset access, documentation, and reproduction scripts (Kapoor et al. 2024; Mitchell et al. 2019; Reuel-Lamparth et al. 2024), present yet another challenge in ML-based research. Our analysis revealed that 24% of synthetic personae papers provided no dataset links, while the ones provided often showed poor documentation practices. Our review showed that the majority of persona datasets was built upon the same datasets, underscoring the need for better documentation practices. The code for persona generation, the final dataset, the statistical distributions of demographic attributes, should all be well documented in a public code repository. While the majority of *ACL Anthology studies in our subset released their code, no study published with ACM CHI did. Notably, when only LLM-generated personae are used, we advocate for releasing the full dataset, not just selected examples or the prompts, to allow meta-analyses and replicability studies.

(6) Generalizability: Acknowledge Author Context

While ethical considerations in ML research have been prominently highlighted (Gebru et al. 2021; Kapoor et al. 2024; Mitchell et al. 2019), we expand on that by suggesting researcher positionality statements, an acknowledgment of how researchers' backgrounds may influence persona design decisions. This addresses a near-universal absence of positionality statements in our corpus, despite their importance in research involving human representation. Drawing

on Figure 6, the author distribution reflects a strong focus on US and Chinese user populations. Authors should refrain from claiming a general global population (Table 4), but acknowledge the scope and limitations of their study.

LIMITATIONS

First, the corpus of literature we reviewed is limited as we focus solely on top-tier AI conferences (2023-2025). On the other hand, we identified relevant contributions based on a keyword search ("persona"). This approach helped us identifying key studies, yet likely excluded relevant work published in other venues, time frames, and those using alternative terminology particularly from product development or marketing. Moreover, the exclusion of non-peer-reviewed preprints and workshop papers implies that we might not have captured the most recent, emerging scholarship. Second, our study did not examine end-user perceptions of synthetic personae or how different persona construction methods affect user trust and acceptance. Future work could interview particular user populations of interest and include their perception of representation. Lastly, including user interactions would allow further analysis of ecological validity.

CONCLUSION

Synthetic personae studies have become a prominent method in AI alignment research. Whether based on inferred personae from user surveys or on LLM-generated artificial 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 enhanced our review of 69 persona studies from leading AI venues with a review of existing ML research checklists. Our paper applies our findings to ML checklist practices, deriving six recommendations for creating representative and transparent synthetic persona datasets in LLM research. Our analysis reveals substantial gaps in existing research on persona representativeness: 36.5% of studies target undifferentiated "general populations," only 29.5% ground persona construction in social science literature, and 60.8% do not discuss representativeness. These shortcomings limit the ecological validity of persona-based evaluations and raise questions about the relevance and generalizability of the findings for real-world usage and deployment scenarios. By synthesizing established ML documentation frameworks and the findings of our literature review, we develop a persona-specific transparency checklist, which emphasizes task specification, population definition, empirical grounding, ecological validity assessment, reproducibility, and transparent limitations documentation. As LLMs gain increasing importance in high-stakes domains, evaluation persona datasets for their representativeness and ecological validity for that particular use case is crucial. Therefore, we invite future work to submit their persona datasheets along with their code to our living *"Whose Personae?"* repository which we maintain and update [anonymized, will be released after review].

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 are fully aware that rich persona libraries could be hijacked to engineer targeted disinformation. Therefore, further research on embedding provenance and watermarks for such libraries is crucial. This should be complemented by periodic third-party audits that trace and flag suspicious downstream use.

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 thank Columbia University's Institute for Social and Economic Research and Policy (ISERP), the Columbia Data Science Institute, and Quantitative Methods in the Social Sciences.

We thank Jonathan Reti, Carlo Uhl, Merle Uhl, Elena Krumova, Monserrat Lopez Perez.

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. 2018a. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Trans. Web*, 12(4).

An, J.; Kwak, H.; Jung, S.-g.; Salminen, J.; Admad, M.; and Jansen, B. 2018b. Imaginary people representing real numbers: Generating personas from online social media data. *ACM Transactions on the Web (TWEB)*, 12(4): 1–26.

Batzner, J.; Stocker, V.; Schmid, S.; and Kasneci, G. 2024. Germanpartiesqa: Benchmarking commercial large language models for political bias and sycophancy. *arXiv preprint arXiv:2407.18008*.

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, 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. arXiv:2305.18189.

Cho, W. I.; Lee, Y. K.; Bae, S.; Kim, J.; Park, S.; Kim, M.; Hahn, S.; and Kim, N. S. 2023. When crowd meets persona: Creating a large-scale open-domain persona dialogue corpus. *arXiv preprint arXiv:2304.00350*.

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.

Chu, E.; Vijayaraghavan, P.; and Roy, D. 2018. Learning Personas from Dialogue with Attentive Memory Networks. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2638–2646. Brussels, Belgium: Association for Computational Linguistics.

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.

Dev, J.; Rashidi, B.; and Garg, V. 2023. Models of applied privacy (MAP): A persona based approach to threat modeling. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15.

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, 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.

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, 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, 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 Multiparty 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.

Pang, R. Y.; Schroeder, H.; Smith, K. S.; Barocas, S.; Xiao, Z.; Tseng, E.; and Bragg, D. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–20.

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-Lamparth, 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. ???? Persona-Aware Generative Model for Code-Mixed Language. *Available at SSRN 4602608*.

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. 2024a. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In *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.

Wang, Z.; Mao, S.; Wu, W.; Ge, T.; Wei, F.; and Ji, H. 2024b. 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.

Zhang, X. A. 2022. What constitutes great IDEA? An examination of corporate diversity communication on facebook and external and internal stakeholder reactions. *Public Relations Review*, 48(5): 102254.

Zhang, Z.; Wen, J.; Guan, J.; and Huang, M. 2022. Persona-Guided Planning for Controlling the Protagonist's Persona in Story Generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3346–3361.

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. 2023a. SimOAP: Improve Coherence and Consistency in Persona-based Dialogue Generation via Over-sampling and Post-evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9945–9959.

Zhou, J.; Pang, L.; Shen, H.; and Cheng, X. 2023b. 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, H.; Wang, H.; and Carroll, J. M. 2019. Creating Persona Skeletons from Imbalanced Datasets-A Case Study using US Older Adults' Health Data. In *Proceedings of the 2019 on designing interactive systems conference*, 61–70.

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.