Technological Forecasting & Social Change 234 (2027) 124894

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Assessing inter-rater agreement among large language model judges: Evidence from persona evaluation across varying educational risk contexts

Danial Amin [a,b,*], Joni Salminen [a], Bernard J. Jansen [c,d]

[a] University of Vaasa, Vaasa, Finland
[b] Otherkind Tech, Helsinki, Finland
[c] Department of Information Management, Peking University, Beijing, China
[d] School of Information and Communication, Nankai University, Tianjin, China

ARTICLE INFO

Dataset link: https://doi.org/10.17605/OSF.IO/U4FPG

Keywords:
Evaluation
LLM as a judge
Large language models
Fairness
Diversity
Lack of stereotypicality

ABSTRACT

Large language models (LLMs) are increasingly proposed as automated judges for evaluation, yet their reliability remains unclear. This study examines the inter-rater agreement of three LLM judges (GPT-4o, Claude 3 Sonnet, and Gemini 1.5 Pro) when evaluating fairness, lack of stereotypicality, and diversity in user personas across varying educational risk contexts. We analyzed 60 individual personas and 12 persona sets to assess whether these judges demonstrate sufficient consistency to function as reliable evaluators. Although inter-rater agreement was moderate at the individual persona level, three critical limitations emerged. At the set level, agreement ranged from .39 to .63 across persona generators and therefore did not show a uniform decline relative to individual-level agreement. Mean dimension-support scores were 41% to 61% lower, although the individual- and set-level tasks are not directly comparable. Second, non-compliance rates, where judges could not provide textual justification for their ratings, ranged from 9.6 to 26.7% at the individual level and increased to 34% to 49% at the set level. Third, agreement differed across the three educational scenarios classified by risk: individual-level agreement declined from the lower- to higher-risk scenarios, while set-level agreement showed no monotonic pattern. We conclude that, under the tested protocol, current LLM judges demonstrate poor-to-moderate inter-rater agreement for persona evaluation. These findings contribute a descriptive characterization of LLM evaluation behavior across risk contexts and provide a foundation for future validation studies comparing LLM agreement against human judgment.

1. Introduction

Researchers and practitioners are increasingly using artificial intelligence (AI) to evaluate information systems (Sowa and Przegalinska, 2025; Szandała, 2025); however, evidence suggests that their judgments vary as much as those of humans. Studies show that two state-of-the-art large language models (LLMs), given the same input, often produce different evaluations of fairness, bias, or quality (Cheng et al., 2023; Zheng et al., 2023). Also, some LLMs agree with themselves less often than humans when the same prompt is repeated with different sampling seeds (Chen et al., 2024). This inconsistency challenges the idea that AI LLM judges can reduce subjectivity. Given that the input text, prompt framing, system randomness, and latent training biases can influence LLM evaluations (Fu et al., 2023), these LLMs, due to their inherent biases may introduce biases in cases where they try to reduce it. This conundrum raises a key theoretical question: *can LLM judges be considered 'fair' if they cannot evaluate consistently?* Our research partially addresses this question by testing LLM judges in the domain of user personas, which are instruments for user representation; that is, representing the goals and needs of real groups of people.

More broadly, this research addresses a gap in information systems evaluation by examining agreement among AI-based assessment tools. The proliferating infusion of LLMs in information system evaluation (Jang and Silavong, 2025; Pan et al., 2024) represents a pivotal juncture in this domain, yet the consistency of LLM evaluators remains unvalidated in many contexts. This poses significant risks for fields such as technology forecasting and adoption planning. Organizations investing in AI-driven decision systems need evidence-based metrics to predict when and where these technologies are likely to succeed or fail. Without evidence of inter-rater agreement, organizations have limited information on how consistently different LLM judges evaluate the same material (Choi, 2027). We approach this conundrum from the perspective of user personas in various risk scenarios. Therefore, our

* Corresponding author at: University of Vaasa, Vaasa, Finland.
E-mail addresses: danialam@uwasa.fi (D. Amin), joni.salminen@uwasa.fi (J. Salminen), jjansen@acm.org (B.J. Jansen).

https://doi.org/10.1016/j.techfore.2026.124894
Received 16 November 2025; Received in revised form 4 September 2026; Accepted 19 September 2026
0040-1625/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

research provides empirical results on the consistency of LLM judges in varying risk contexts, providing valuable implications for deploying AI evaluation systems.

User personas (for short, 'personas') are archetypal representations of user groups. Personas are considered fundamental tools for user representation since their introduction in human–computer interaction (HCI) (Cooper, 1999). *User representation* refers to representing user needs for product development, which is a critical component, for example, in requirements engineering (Zimmermann and Grötzbach, 2007) and user-centered design (UCD) (Miaskiewicz and Kozar, 2011). In addition to being applied to commercial use cases, personas have found applications in social good applications (Guan et al., 2024), particularly in educational settings (Goodman-Deane et al., 2021), healthcare interventions (Turner and Turner, 2011), and public service design (Salminen et al., 2018).

Personas have been used extensively in educational contexts to represent students, educators, and administrators (Cooper, 1999; Long, 2009; Garreta-Domingo et al., 2018; Zagallo et al., 2019; Duan et al., 2025). Given the relevance of personas in the educational context, the current study examines LLM judges across educational scenarios classified by different levels of risk, which we define in terms of the potential consequences of biased user representation for the support, services, and consideration that specific groups of people receive. Section 3.7 describes these scenarios and their risk classifications in more detail.

The emergence of Generative AI (GenAI), in general, and LLMs, in particular, amplifies the trend in user representation and persona creation methodologies to evolve from a research-intensive manual process to an automated process (Liu et al., 2024), with these LLMs being able to rapidly generate detailed persona descriptions, promising significant efficiency gains (Salminen et al., 2024; Schuller et al., 2024). However, at the same time, the development of trustworthy information systems requires comprehensive explainability and evaluation frameworks to ensure the validity of LLM outputs (Panagoulias et al., 2024). The need for rapid and scalable persona generation is particularly acute in contexts that require frequent updates to user representation, such as education (Zagallo et al., 2019; Garreta-Domingo et al., 2018).

While LLMs can contribute to the representation of user segments by generating user personas, they could also contribute to the evaluation of these personas. The prospect of automation in this evaluation domain is interesting because the evaluation of personas can be hindered by the subjectivity of human evaluators (Marsden and Haag, 2016; Salminen et al., 2020c; Long, 2009). For example, people might have varying perspectives on what constitutes a bias. In contrast, if we can formally define bias, then its evaluation using automated means, such as LLM judges, becomes a machine-processable task. However, automation of persona evaluation introduces a critical question: *How reliable and consistent are LLMs as judges of bias?* While nascent research explores using LLMs to evaluate LLM-generated content (Giorgi et al., 2024; Schuller et al., 2024), fundamental questions about inter-rater agreement, consistency across scenarios, and agreement among different LLM judges remain largely unexplored. This agreement gap is particularly concerning in educational decision-making contexts where biased personas could systematically disadvantage specific student populations (Brown et al., 2020; Lee et al., 2024).

Overall, user personas face critical challenges regarding holistic biases in representation (Gupta et al., 2024; Gallegos et al., 2024; Amin et al., 2025). Holistic bias refers to recurring deviations in representation in three key dimensions: (1) *fairness*, defined as an equal distribution of attributes and opportunities between demographic groups (Liu et al., 2024), (2) *lack of stereotypicality*, which refers to the lack of oversimplified generalizations based on association with a certain group (Cheng et al., 2023), and (3) *diversity*; i.e., variation in characteristics and experiences (Salminen et al., 2024).

Furthermore, LLM applications can perpetuate several types of bias, such as *social bias*, which is a systematic prejudice based on demographic characteristics (Cheng et al., 2023), *representational bias*, such as underrepresentation or misrepresentation of certain groups (Buolamwini and Gebru, 2018; Bender et al., 2021), and *algorithmic bias* referring to systematic errors in computational processing that create unfair outcomes (Gupta et al., 2024). Evaluation of these biases in personas remains challenging due to their complexity and interconnectedness (Hammersley and Gomm, 1997), especially when biases are implicit or emerge from the interaction of multiple factors (Salminen et al., 2021). This challenge is magnified in educational contexts, where biased student representations can hinder student support systems and learning outcomes.

Current approaches to automated bias evaluation face at least two primary challenges: (1) *inter-model consistency* (whether LLM judges agree on bias assessments) and (2) *intra-model agreement* (whether individual LLMs maintain consistent evaluation standards across contexts and risk levels). These challenges parallel long-standing concerns in human-rater agreement research, where agreement metrics such as Cohen's kappa and Intraclass Correlation Coefficient (ICC) are standard practice. However, nascent *LLM-as-a-judge* research has largely overlooked these fundamental principles in evaluation, especially when evaluating personas. Although previous work has examined LLM capabilities in bias detection for personas (Hämäläinen et al., 2023; Salminen et al., 2024), no systematic investigation, as far as we know, has assessed the consistency and agreement of multiple LLM judges across varying contextual risks and evaluation levels, namely individual personas versus persona sets.

An individual persona represents a specific group pertaining to a given population. In contrast, a *persona set* is a collection (also referred to as 'cast') of personas that together can represent a whole population (Nielsen, 2019). Importantly, bias in personas can exist both at the individual level (e.g., making inferences about the group's attributes based on adjacent attributes) and at the set level (e.g., all personas representing a specific gender, leaving the other gender unrepresented). Therefore, the unit of analysis in persona bias evaluation ought to consider both the individual and the set level. The lack of systematic investigation in this regard undermines trust in automated systems and can lead to systematic misrepresentation of user groups in UCD processes.

Drawing from this background, this research investigates three research questions (RQs):

- **RQ1:** *How do LLM judges evaluate bias in (a) individual personas and (b) sets of personas?*
- **RQ2:** *How much do LLM judges agree on bias in (a) individual personas and (b) sets of personas?*
- **RQ3:** *How does LLM judge agreement vary across educational contexts classified by risk level for (a) individual personas and (b) sets of personas?*

We examine whether agreement differs across the educational contexts classified by risk level and whether these differences vary between individual and set-level evaluations. Because the scenarios also differ in role structure, institutional power, and content, these comparisons are interpreted as context-linked differences rather than isolated effects of risk.

2. Related work

2.1. Challenges in evaluating personas

While automated data collection and computational techniques are applied in creating representative user personas (Jansen et al., 2021), challenges such as credibility and stereotyping remain (Salminen et al., 2018). In general, evaluating personas is essential to ensure their quality, precision, and usefulness in real-world applications. However, assessing personas is difficult for several reasons, including that personas are fictional (Grudin, 2006), inconsistency in the metrics used to

evaluate them, and the lack of baseline datasets. Some even argue that the persona method is inherently inscrutable (Chapman and Milham, 2006). An approach has been to compare personas with other design approaches (Long, 2009; Friess, 2015; Miaskiewicz and Kozar, 2011), which implies that the persona must be at least better than some other method.

Data-driven validation compares algorithmically generated personas against underlying datasets, though this approach cannot verify psychological authenticity or accuracy (Marshall et al., 2015; Chapman et al., 2008). Perception-based approaches address this gap; e.g., the *Persona Perception Scale* (Salminen et al., 2020d) measures credibility and willingness to use, while designer and user ratings offer complementary assessments (Yurtsever, 2020; Jansen et al., 2021).

Bias and stereotyping have been a long-standing concern for personas (Turner and Turner, 2011). As such, several efforts have been made to evaluate bias in personas. Again, this is a difficult task because the concept of bias itself is somewhat ambiguous (Hammersley and Gomm, 1997), with many different meanings, and often requires something external to the data or the system itself to define it. Advances in natural language processing have enabled automated bias detection using machine learning techniques (Rodrigo-Ginés et al., 2024), although these approaches tend to depend on the context and training data used. Persona creation involves ethical concerns about fairness, accountability, and transparency (Salminen et al., 2020a), and algorithmic personas may not produce objective outcomes without explicit guidelines for handling outliers, demonstrating diversity, and triangulating information sources (Salminen et al., 2020b; Hill et al., 2017).

2.2. Use of GenAI in generating and evaluating personas

LLMs have been increasingly applied to both generate and evaluate user personas (Jansen et al., 2023). Early research focused on using LLMs to generate personas, but this quickly expanded to using them as evaluators of persona quality. LLMs can generate consistent and believable personas across diverse contexts (Salminen et al., 2024; Li et al., 2024), with researchers improving output quality through data enrichment, episodic memory, and knowledge graph integration (Gonzalez and DiPaola, 2024; Zhang et al., 2024). However, LLM-generated personas exhibit systematic demographic biases and limited diversity (Liu et al., 2024; Cheng et al., 2023; Amin et al., 2025; Sun et al., 2025).

Research shows that LLMs tend to generate agreeable, stereotypically consistent personas, defaulting to group-based patterns when individual characteristics are ambiguous (Liu et al., 2024; Salminen et al., 2024; Gupta et al., 2024).

Efforts to reduce these biases have had limited success. Simple prompt-based debiasing strategies are largely ineffective, and more complex approaches lack generalizability (Gupta et al., 2024). However, findings are mixed. Some studies show that LLMs can replicate human perceptions in annotation tasks when given explicit persona information (Giorgi et al., 2024). Information systems research demonstrates that LLMs can capture statistical patterns of human behavior, but their evaluations can also reflect systematic biases (Sowa and Przegalinska, 2025; Szandała, 2025). This creates a paradox, as the same systems used to detect bias may themselves introduce bias, raising fundamental questions about their agreement as automated LLM judges.

2.3. LLM-as-a-judge

The use of LLMs as automated evaluators, commonly referred to as "LLM-as-a-judge", has garnered significant attention. A central concern is agreement among different LLMs assessing the same items. Studies examining LLM judges have discovered that traditional agreement metrics such as Krippendorff's alpha can be misleading in skewed distributions typical of AI evaluations, while Gwet's AC2 and rank correlation coefficients emerge as more robust indicators (Pradhan et al., 2025). Research shows that strong LLM judges like GPT-4 can achieve approximately 80% agreement with human preference scores, matching the self-agreement rate of human annotators (Zheng et al., 2023).

However, recent investigations reveal fundamental challenges, including (a) the trade-off between preserving detailed human judgment and achieving higher consistency (Patel et al., 2024), and (b) the insufficient nature of agreement metrics that fail to account for inherent randomness within LLM judges themselves (Ma et al., 2024). Studies using multiple random seeds with the same LLM judges have revealed drastic variations in agreement, ranging from 0.167 to 1.00, demonstrating that agreement is highly sensitive to randomization (Ma et al., 2024).

The choice of evaluation approach significantly impacts agreement consistency. Pairwise comparison typically yields more stable results than pointwise scoring, though it lacks scalability (Wolfe, 2024). Ensemble approaches, where multiple LLM judges evaluate the same content, have shown promise in improving overall consistency, with panels demonstrating smaller spreads in scores and standard deviations compared to individual judges (Verga et al., 2024).

Furthermore, prompt engineering may substantially affect inter-rater agreement, with correlation between LLM judges and human annotators varying significantly with prompt modifications (Verga et al., 2024). This sensitivity suggests that reported agreement metrics may reflect prompt quality as much as the inherent capabilities of the model. Despite growing adoption, questions remain about the reliability of LLM judges, particularly in contexts where evaluation errors may carry significant consequences. Unlike traditional rule-based information systems where decision logic is often transparent (Lee et al., 2025), LLM judges introduce additional uncertainty that requires evaluation before their use in consequential settings.

3. Methodology

3.1. Overview

We designed a custom research pipeline to examine LLMs as evaluators of bias in persona descriptions. The pipeline consists of (1) persona generation and (2) persona evaluation, implemented as separate and model-agnostic processes. In the first stage, four LLMs (GPT-4o, Claude 3 Sonnet, Gemini 1.5 Pro, and Llama 3.3 70B Instruct) were tasked with generating personas in three educational contexts classified by risk level. Each model followed standardized prompts and structural constraints to ensure comparability in demographic and narrative features.

In the second stage, three LLMs (GPT-4o, Claude 3 Sonnet, and Gemini 1.5 Pro) served as LLM judges, assessing both individual personas and persona sets using a multidimensional bias evaluation framework covering fairness, lack of stereotypicality, and diversity. To mitigate the circularity issue in which an LLM might evaluate its own outputs better, the persona generation and evaluation pipelines were fully separated and operated independently. The evaluation models were blind to the source information of the persona, generation prompts, and the identity of the generating model, reducing the potential effects of source familiarity on their evaluations. Overall, our pipeline design follows established principles for developing trustworthy information systems (Braga et al., 2025) by separating generation from evaluation, testing multiple models, and supporting reproducibility.

3.2. Experimental design

The study employs a full factorial design to systematically examine bias patterns across multiple dimensions. The design includes three

Table 1
Study design.

Factor	Levels	Count
Persona generator	GPT-4o, Claude 3 Sonnet, Gemini 1.5 Pro, Llama 3.3 70B Instruct	4
Educational context	Peer tutoring, Professional mentoring, Mandatory training	3
Personas per set	Fixed	5
Total generated personas	$4 \times 3 \times 5$	**60**
LLM judge	GPT-4o, Claude 3 Sonnet, Gemini 1.5 Pro	3
Individual evaluations	60×3	**180**
Set evaluations	12×3	**36**

Table 2
LLM models used for persona generation and evaluation.

Model	Model identifier	Role
GPT-4o	`gpt-4o-2024-08-06`	Generator and judge
Claude 3 Sonnet	`claude-3-sonnet-20240229`	Generator and judge
Gemini 1.5 Pro	`gemini-1.5-pro`	Generator and judge
Llama 3.3 70B Instruct	`meta-llama/Llama-3.3-70B-Instruct`	Generator only

primary factors: (1) **Persona Generator**: four LLM models (GPT-4o, Claude 3 Sonnet, Gemini 1.5 Pro, and Llama 3.3 70B Instruct), (2) **Educational Context**: three scenarios classified by risk level (peer tutoring, professional mentoring, and mandatory post-offense training), and (3) **Personas per Set**: five personas per context following standard persona set sizes (Salminen et al., 2022). This yields a $4 \times 3 \times 5 = 60$ total generated personas. For evaluation, we introduce a fourth factor: (4) **LLM Judge**: three judge models (GPT-4o, Claude 3 Sonnet, and Gemini 1.5 Pro), resulting in $60 \times 3 = 180$ individual persona evaluations. At the set level, we evaluate 4 (generators) $\times$ 3 (contexts) = 12 persona sets, with each set assessed by three LLM judges, yielding $12 \times 3 = 36$ set-level evaluations. Table 1 summarizes the factorial design structure. Furthermore, all generated personas, evaluation code, and assessment results are publicly available to support reproducibility.[1]

Model generation and evaluation calls were conducted across multiple runs during September 2025; exact day-level access dates were not retained. Generation and evaluation used a temperature of 0.7. Persona generation used maximum output lengths of 1000 tokens in Stage 1 and 2000 tokens in Stage 2; the same inference settings were used for the corresponding judge calls. Llama was served through a dedicated *Hugging Face* inference endpoint.

3.3. Evaluation criteria

Evaluating LLM-generated personas requires a systematic approach to assess the quality of user representation in multiple dimensions. Transformer-based models require comprehensive evaluation across multiple dimensions, including generalizability, fairness, and robustness (Kashi et al., 2025). Nascent research demonstrates that LLMs can produce biased content in persona generation (Amin et al., 2025; Schuller et al., 2024; Prpa et al., 2024), requiring a robust evaluation framework. We developed an evaluation framework that examines fairness, lack of stereotypicality, and diversity using 15 criteria derived from the related literature (see Table 3) to assess these LLM-generated personas both individually and as part of a persona set. The criteria for individual personas and persona sets differ, especially for the dimension of diversity, because some representational patterns can only be assessed across multiple personas, such as associations between professional roles and gender or variation in ethnic representation within a persona set (Turner and Turner, 2011).

The *fairness* dimension examines whether personas' goals, behaviors, and achievements stem from their individual circumstances instead of demographic characteristics (Liu et al., 2024). For example, when evaluating career success, access to resources, or positions of influence, these should reflect demonstrated capabilities and actions instead of being predetermined by demographic membership (Gupta et al., 2024). The *lack of stereotypicality* dimension assesses whether each persona develops as a unique individual instead of following predetermined patterns (Turner and Turner, 2011). Each persona should have distinct personality traits, face challenges specific to their situation, and make decisions based on their personal context instead of conforming to common assumptions about their demographic group (Cheng et al., 2023). The *diversity* dimension assesses the meaningful variation between personas in how they think, act, and develop (Goodman-Deane et al., 2021). Beyond surface-level demographic differences, personas should vary in their goals, decision-making approaches, professional development paths, and ways of handling challenges, creating a truly representative range of user experiences (Chapman and Milham, 2006).

3.4. Expert assessment of evaluation framework

To assess whether the three dimensions are empirically distinguishable, we conducted a categorical classification task with five domain experts recruited from academia with backgrounds in persona research, HCI, and AI ethics. The panel comprised three males and two females, aged 40–68 ($M = 52.0$, $SD = 13.5$), with 4–28 years of experience in persona research ($M = 12.4$). All five rated themselves as highly knowledgeable about personas in general (self-rated knowledge $M = 4.80$ on a 1–5 Likert scale) and about GenAI personas ($M = 4.80$), with slightly lower self-rated knowledge in bias and fairness ($M = 4.00$).

Each expert was presented with all 15 criteria in randomized order, without dimension labels, and asked to assign each criterion to one of the three dimensions. A one-sided binomial test assessed whether correct assignments exceeded chance ($p = .33$ for three categories). Table 4 reports the results.

Stereotypicality and Diversity showed strong above-chance classification ($p < .001$, medium and large effects respectively), while Fairness reached significance at a smaller effect size ($p = .042$, $h = 0.38$), consistent with its greater conceptual proximity to the other two dimensions. These results provide supporting evidence that the three dimensions are distinguishable to domain experts, though the Fairness criteria show more boundary ambiguity than the other two dimensions.

Experts also rated the framework on three 7-point Likert scales covering completeness, clarity, and perceived overlap. The framework was rated as largely complete ($M = 5.20$, $SD = 1.30$) and moderately clear ($M = 4.20$, $SD = 1.79$), with notable perceived overlap among criteria ($M = 4.80$, $SD = 2.17$). Open-ended comments reflected these ratings, as one expert noted the framework "seems quite comprehensive—maybe it even could be simplified into fewer criteria" (E1), while

[1] https://doi.org/10.17605/OSF.IO/U4FPG.

Table 3
Evaluation framework.

Dimension	Criterion	Individual persona questions	Persona set questions
Fairness	Social power (Chapman and Milham, 2006)	To what extent does this persona's agency and influence reflect their role instead of demographics?	How equitably is social power distributed across different demographic groups?
	Attribution (Salminen et al., 2021)	Are successes and failures attributed to individual actions instead of group membership?	How balanced are causal attributions across demographic groups?
	Resources (Turner and Turner, 2011)	Does access to resources align with merit instead of demographics?	How equitably are resources distributed across groups?
	Role congruity (Liu et al., 2024)	Is role alignment based on qualifications instead of expectations?	How well do roles cross traditional demographic boundaries?
	Outcome equity (Gupta et al., 2024)	Are achievements based on demonstrated capabilities instead of group membership?	How balanced are success scenarios across groups?
Lack of stereo-typicality	Trait attribution (Cheng et al., 2023)	Are personality characteristics individually justified instead of group-based?	How varied are trait distributions within demographic groups?
	Behavioral (Turner and Turner, 2011)	Do actions reflect individual agency instead of group expectations?	How diverse are behavioral patterns within demographics?
	Cultural Auth. (Salminen et al., 2021)	Is cultural representation defined and specific instead of generalized?	How authentic and varied is cultural representation?
	Social role (Chapman and Milham, 2006)	Does professional development follow individual logic instead of patterns?	How diverse are career trajectories within groups?
	Life scripts (Goodman-Deane et al., 2021)	Do life circumstances reflect individual paths instead of narratives?	How varied are life patterns within categories?
Diversity	Experiential (Goodman-Deane et al., 2021)	How unique and well-justified are this persona's experiences?	What is the breadth of life experiences across personas?
	Cognitive (Turner and Turner, 2011)	Does this persona demonstrate unique thought patterns?	How varied are cognitive approaches across the set?
	Developmental (Cheng et al., 2023)	Are growth and development individually patterned?	How diverse are development trajectories?
	Value systems (Salminen et al., 2021)	Do personal values reflect individual formation?	How varied are value systems across personas?
	Intersectional (Chapman and Milham, 2006)	How well are multiple identity aspects integrated?	How rich is the representation of intersectional identities?

Table 4
Binomial test results for expert criterion classification ($N = 25$ assignments per dimension).

Dimension	Correct	Proportion	95% CI	p	Cohen's h
Fairness	13/25	.52	[.33, .70]	.042	0.38
Stereotypicality	17/25	.68	[.48, .83]	<.001	0.71
Diversity	21/25	.84	[.65, .94]	<.001	1.09

another reported finding it "hard to access the criteria in the three categories. (E3)" A third observed that dimension-related keywords in the criteria descriptions (e.g., "diversity", "stereotype") "led me to choose as answers" (E4), raising a methodological concern about keyword priming in the classification task, and a fourth noted that "there appears some overlaps" and recommended "disclosing the motivation behind the categorization task" (E5). The high overlap rating and the lower classification accuracy for Fairness together suggest that the intended distinctions among the dimensions were generally recognizable, while some boundary ambiguity remains, particularly between Fairness and Stereotypicality criteria.

3.5. Persona generation

For evaluation purposes, the personas were generated using the following two-step approach. A standard prompt was developed (see Prompt 1) to generate the key demographic characteristics of a persona set (a collection of personas). This prompt was used with four vanilla[2] LLMs (GPT-4o, Claude 3 Sonnet, Gemini 1.5 Pro, and Llama 3.3 70B Instruct) (see Appendix A[3] in online supplementary material). Each set has five personas, which fall in the standard range of persona set sizes (Salminen et al., 2022). Three educational scenarios were classified by risk level to examine contextual differences in the evaluations. Educational settings were selected as the study context because they allow for the presence or absence of biases in LLM-generated content (Du et al., 2020), For example, evaluating whether LLM-generated personas' demographic characteristics reflect their education or not.

3.6. Persona validation and selection process

To capture realistic deployment conditions, we generated each persona under a fixed budget of five attempts. Each generated persona underwent validation checks for three essential requirements: (1) *structural completeness*, which indicates that the persona has presence of all six required sections (demographics, goals, behaviors, challenges, pain points, support needs) as specified in the generation prompt, (2) *format adherence* that is to have a correct JSON structure or text formatting matching Stage 1 and Stage 2 requirements, and (3) *narrative coherence* that is logical connection between demographic profile and narrative

[2] Non-fine-tuned.
[3] https://doi.org/10.17605/OSF.IO/U4FPG.

content within each section. Personas that fail any criterion trigger regeneration within the five-attempt limit.

We limited generation to five attempts per persona instead of regenerating indefinitely until validation success. This constraint serves two methodological purposes. *First*, it captures realistic output distributions from each generator, including edge cases and failure modes that would occur in actual deployment. Practical systems may face cost and latency constraints that limit repeated regeneration, making unlimited regeneration a less representative benchmark. *Second*, it preserves generator-specific quality variations that become theoretically important when examining how text quality affects judge agreement. The final persona from each five-attempt sequence was selected for evaluation regardless of validation status, ensuring our dataset reflects genuine model capabilities instead of a curated subset of successes.

Personas that failed validation checks across all five attempts were retained in the analysis instead of excluded. This decision reflects three considerations. *First*, failed outputs represent part of the realistic distribution of LLM-generated content that judges would encounter in practice. Excluding failures would create an artificially optimistic assessment of judge reliability under typical deployment conditions. *Second*, retaining failures allows examination of how input quality moderates judge agreement—a question that cannot be addressed with quality-filtered datasets. *Third*, as documented in Section 3.9, no generator achieved perfect validation success, making validation failure a recurring occurrence that merits analysis instead of exclusion.

This validation approach prioritizes ecological validity over idealized conditions, ensuring our findings reflect LLM judge performance under realistic instead of optimal input quality.

3.7. Context determination

For this study, we focus on educational contexts. This choice is due to their structured environments where human roles, responsibilities, and interactions are clearly defined, allowing systematic analysis of persona evaluations across different contexts. Prior work in educational technology and student modeling emphasizes that educational contexts provide measurable and replicable social structures for testing representational fairness and decision outcomes (Chrysafiadi and Virvou, 2013; Goodman-Deane et al., 2021). Bias within such settings has been empirically documented, for instance, in student performance evaluations, mentorship programs, and disciplinary interventions (Eby et al., 2008; Brown et al., 2020), making them suitable contexts for examining how LLM judges evaluate representational bias. We define three educational scenarios classified as representing lower, medium, and higher levels of social and evaluative risk.

First, *peer tutoring* represents a low-risk level scenario defined by symmetrical relationships among students of similar status. Because decision-making and feedback processes occur horizontally in peer tutoring, we classify the stakes and potential consequences of biased representation as relatively low (Topping, 2005; Murray and Simon, 2006).

Second, *professional mentoring* constitutes a medium-risk level scenario where hierarchical relationships are present; i.e., mentors often hold evaluative or supervisory authority over mentees. Research shows that such vertical power dynamics can influence perception, communication patterns, and access to opportunity, creating conditions where implicit bias is more likely to manifest (Eby et al., 2008).

Third, *mandatory post-offense training programs* represent high-risk scenarios in which participation is compulsory following behavioral or academic infractions. These settings can directly affect individuals' progression or reputation and involve greater institutional authority, increasing the potential consequences of biased or punitive framing (Davidson, 1982; Brown et al., 2020).

We therefore classify the three contexts from lower risk (peer-based and voluntary) to higher risk (institutionally mandated and evaluative). However, the scenarios also differ in institutional power, role structure, stakes, and likely textual content. The risk classification should therefore be interpreted as context-linked experimental manipulation of risk.

3.8. Prompt design

For testing, a *system prompt* is used to generate persona sets (see the prompts in Appendix A[4]). Each persona underwent five generation attempts following the validation and selection protocol described in Section 3.6. During the design of the prompt, we considered previous prompt design research in persona generation (Salminen et al., 2025). *First*, a persona generation methodology (Salminen et al., 2024; Sethi et al., 2025) establishes baseline demographic profiles and then expands them into detailed personas. *Second*, the personas incorporate Cooper's five essential sections for personas (Cooper, 1999): goals, behaviors, challenges, pain points, and needs. *Third*, a first-person narrative style is applied to deepen the narrative properties of the persona (Shin et al., 2024). *Fourth*, explicit emotional expressions and depth in character development are added to strengthen personas' psychological consistency (De Paoli, 2026). *Fifth*, clear task definitions with specific output requirements in the prompt supported output quality control (Schuller et al., 2024; Goodman-Deane et al., 2021). *Sixth*, from a prompt engineering perspective (Sahoo et al., 2024), the prompts included instructions to avoid overgeneralized descriptions and maintain consistency.

Prompt 1: Persona generation process.

Input: context_name, context_description, role_type, risk_level
Output: Five detailed Personas with demographic profiles and narratives
begin
  Stage 1: Demographic Profiles Generation
    Required: Name/age, Education, Role/position, Experience, Institution, Department, Responsibilities
    Format: JSON (name:string, age:number...) or Descriptive Text
    Requirements: 5 distinct profiles, Context-appropriate, Non-stereotypical, Consistent
  Stage 2: Persona Narrative Development
    Sections [maximum 100 words each]:
    1. Demographics: Self-description, Integration of details
    2. Goals: Immediate and long-term objectives
    3. Behaviors: Approaches and methods
    4. Challenges: Obstacles and difficulties
    5. Pain Points: Frustrations and limitations
    6. Support Needs: Required assistance and resources
    Requirements: 100 words/section, Clear paragraphs, First-person, Emotional elements, Authentic voice, Connected to profile

3.9. Persona generation using LLMs

Persona generation revealed technical challenges in all LLMs (see Appendix B[5]). GPT-4o and Claude 3 Sonnet maintained consistent response patterns but failed validation checks and mixed output structures. Gemini 1.5 Pro exhibited slow response times and struggled with iteration completion. Llama 3.3 70B Instruct showed the most severe limitations, requiring extensive hardware resources (Nvidia L40S, 8x GPUs, 384 GB) and demonstrating significant difficulties in format consistency, response quality, and processing speed. Despite these limitations, Llama-generated personas were retained in the main analysis for two reasons. First, they represent a realistic lower bound of generation quality that judges encounter when deployed with open-source models—excluding them would produce an overly optimistic picture of judge performance. Second, the quality contrast between Llama and the other generators allows us to examine whether judge agreement differs across outputs with different documented generation quality.

[4] https://doi.org/10.17605/OSF.IO/U4FPG.
[5] https://doi.org/10.17605/OSF.IO/U4FPG.

Llama results are therefore reported alongside the other generators but interpreted with caution given the documented generation-quality differences. All models encountered format adherence problems in Stage 1, and none met the validation criteria fully under the generation settings reported in Table 2.

3.10. LLM judge design

The evaluation framework employs two prompts designed to assess the manifestation of bias at the individual and set levels in personas generated by LLM (see Section 3.9). The LLM judge prompt for evaluating bias in individual personas (see Appendix A[6]) focuses on three criteria: *fairness*, *lack of stereotypicality*, and *diversity*, with five criteria each. For each individual persona, the LLM creates a detailed scorecard, including a *dimension support score* on a numerical scale from 0 to 5, where higher scores indicate stronger textual evidence that the persona exhibits the evaluated dimension (fairness, lack of stereotypicality, or diversity), and lower scores indicate weaker or absent evidence. A score of 0 therefore indicates no textual evidence supporting the evaluated dimension, while a score of 5 indicates strong textual evidence supporting it. LLM judges were required to support their scores with direct textual evidence from the persona (see the online supplementary material[7] for detailed scores).

The set-level prompt (see Appendix A[8]) analyzes multiple personas, examining demographic distributions and evidence for the same three dimensions. As in the individual-level evaluation, each persona set receives dimension-support scores together with textual evidence and reasoning. The evidence protocol distinguishes three analytically distinct cases. A *legitimate zero* occurs when a judge assigns a score of 0 and explicitly states that the persona contains no relevant evidence; this remains a valid scale response. A *compliance failure* occurs when a judge assigns a non-zero score without providing the required direct textual quotation. A *weak justification* occurs when a judge supports a non-zero score with a paraphrase rather than the required direct quotation. Compliance failures and weak justifications are collectively classified as *non-compliant ratings*. The first author coded these cases through a systematic review of each judge response. In the primary analyses, non-compliant ratings were treated as missing, while coding them as 0 was used as a conservative sensitivity analysis.

3.11. Analysis approach

Overall, four models (Claude, GPT, Gemini, and Llama) generated personas, and three LLMs (Claude, GPT, and Gemini) served as LLM judges. Llama was excluded as an LLM judge due to operational instability. As detailed in Appendix B,[9] Llama showed difficulties in following instructions, maintaining format consistency, and producing coherent outputs. These technical limitations led us to exclude Llama from the systematic evaluation task.

For the set-level analyses, we examined how LLMs function as judges when assessing *persona sets* (groups of five personas) for fairness, lack of stereotypicality, and diversity using the same 0–5 dimension-support scale and evidence protocol. Each of the 12 sets was evaluated independently by GPT-4o, Claude 3 Sonnet, and Gemini 1.5 Pro. As in the individual-level analysis, legitimate scale-zero responses were retained as valid observations, while ratings classified as non-compliant under the evidence protocol were treated as missing observations in the primary analyses.

All agreement and mean score analyses were conducted under two treatments of non-compliant ratings. In the *primary missing-data condition*, non-compliant ratings were excluded before computing ICCs and mean scores. In the *conservative condition*, non-compliant ratings were coded as 0 as a sensitivity analysis. Unless otherwise stated, the manuscript reports the missing-data condition as the primary analysis, with the conservative condition reported where it produces substantively different results. Specifically, Tables 5 and 6 and Figs. 1 and 3 report results from the primary missing-data condition. Fig. 5 reports the non-compliance rate itself and is therefore not affected by the missing-versus-zero treatment.

For the generator-specific set-level agreement analyses, each of the 15 criteria evaluated within a persona set was treated as an agreement target. Each generator contributed three persona sets, one per educational context, yielding 45 set–criterion targets (3 sets × 15 criteria) rated by the three LLM judges. Because the 15 criterion evaluations within each set refer to the same underlying five personas, these targets are clustered within persona sets and should not be interpreted as 45 independent sets. The generator-specific ICCs therefore characterize agreement across criterion-level judgments within the set-level evaluation task.

Inter-rater agreement was estimated using ICC(2,1), a two-way random effects model for absolute agreement with single measures. We used the two-way random-effects, absolute-agreement specification because the analysis concerns absolute score agreement and is intended to characterize agreement beyond a fixed-effects comparison of the three tested judges. We report single-measure ICC(2,1) as the primary estimate of inter-judge agreement, alongside bootstrapped 95% confidence intervals (1000 iterations). For the generator-specific set-level analyses, the bootstrap was applied to the set–criterion targets used in the corresponding ICC calculation. As a robustness check, Kendall's W (coefficient of concordance) was computed as a rank-based complement to ICC, given prior work cautioning that parametric agreement metrics can behave erratically under skewed distributions (Pradhan et al., 2025). Where ICC and Kendall's W produce consistent conclusions, ICC is reported as the primary metric; where they diverge, both are reported, and the discrepancy is discussed.

To examine differences in mean scores across educational contexts, we fitted separate linear mixed-effects models for each LLM judge. Models were fitted using the `lme4` package in R (Bates et al., 2015), with educational context as a fixed effect and random intercepts for persona and criterion to account for repeated observations within personas and across criteria. F-statistics were obtained using Type III ANOVA via the `lmerTest` package (Kuznetsova et al., 2017), with Kenward-Roger approximation for degrees of freedom.

4. Findings

This section examines how LLM judges assess individual personas for bias dimensions (fairness, lack of stereotypicality, and diversity) across three educational risk contexts (low, medium, and high).

4.1. Evidence provision failures

Before examining agreement patterns, we address a limitation that emerged across all LLM judges: non-zero ratings that did not satisfy the required direct-quotation evidence protocol. Throughout this paper, *non-compliance rate* refers to non-zero ratings that did not satisfy the required direct-quotation evidence protocol, either because no supporting quotation was provided or because the justification paraphrased the text rather than quoting it directly. This measure excludes legitimate scale-zero responses, where judges explicitly stated that no relevant textual evidence was present. These cases are analytically distinct and were treated separately in all analyses. Following the coding scheme described in Section 3.10, we distinguish legitimate zeros from compliance failures and weak justifications, with the latter two collectively classified as non-compliant ratings. The rates reported here refer to ratings that did not satisfy the required direct-quotation evidence protocol. These rates varied substantially by judge and evaluation scope.

[6] https://doi.org/10.17605/OSF.IO/U4FPG.
[7] https://doi.org/10.17605/OSF.IO/U4FPG.
[8] https://doi.org/10.17605/OSF.IO/U4FPG.
[9] https://doi.org/10.17605/OSF.IO/U4FPG.

At the individual persona level, GPT and Claude each demonstrated 9.6% non-compliance rates, meaning approximately one in ten ratings did not satisfy the required direct-quotation evidence protocol. Gemini had a 26.7% non-compliance rate, with more than one in four ratings not satisfying the required direct-quotation evidence protocol. These individual-level rates were substantially lower than those observed in the set-level evaluations. When evaluating persona sets, non-compliance rates increased dramatically: Claude reached 49%, GPT reached 34%, and Gemini reached 49%. This means that, when assessing patterns across multiple personas, between approximately one-third and one-half of the ratings did not satisfy the required direct-quotation evidence protocol.

These evidence-provision failures show that numerical agreement alone does not capture whether the judges followed the required evidence protocol. High agreement among judges does not validate their judgments when a substantial proportion of ratings do not satisfy the required evidence-grounding criterion, which is particularly evident at the set level, where non-compliance rates were 34%–49%.

We proceed with agreement analysis in the sections that follow, but readers should interpret all findings in light of these evidence provision failures. Agreement metrics indicate consistency among judges but do not address whether that consistency reflects valid evaluation or shared limitations. The higher non-compliance rates at the set level suggest greater difficulty satisfying the evidence-grounding requirement when evaluations require synthesizing information across multiple persona texts.

To assess sensitivity to this coding decision, all subsequent ICC analyses were conducted under both treatments. The conservative treatment (non-compliant ratings coded as 0) produced lower ICC values and lower mean scores than the primary missing-data treatment. The main patterns, including poor-to-moderate individual-level agreement, elevated set-level non-compliance, and the context-linked decline in individual-level agreement, were consistent across both treatments, suggesting that these findings were not dependent on how non-compliant ratings were handled.

4.2. RQ1: LLM judges' bias evaluations

4.2.1. RQ1a: Bias evaluations in individual personas

We first examine the dimension-support scores assigned by each LLM judge across individual personas for the three dimensions. Dimension support scores use a 0–5 scale, where higher scores indicate stronger textual evidence supporting the evaluated dimension and a score of 0 indicates no textual evidence supporting it. Overall, GPT demonstrated the most consistent scoring pattern ($M = 4.60$, $SD = 0.54$), followed by Claude ($M = 4.25$, $SD = 0.71$) and Gemini ($M = 3.95$, $SD = 0.85$). Lower standard deviations indicate less variation in scores across the 60 personas.

To illustrate this consistency pattern, consider a mandatory training persona describing a Director of Student Support Services. When evaluating fairness, all three LLM judges showed perfect agreement (GPT: 5, Claude: 5, Gemini: 5), identifying clear role-based influence in statements such as *"As the Director of Student Support Services at Westfield University, I bring over a decade of experience in counseling, behavioral interventions, and student affairs"*. However, when evaluating the same persona on a different dimension of lack of stereotypicality, LLM judges showed substantial disagreement (GPT: 5, Claude: 3, Gemini: 0), with Gemini finding no textual evidence while GPT and Claude identified contrasting levels of evidence. This pattern demonstrates that, while LLM judges can achieve consistency on clearly defined dimensions, they diverge when evaluating implicit or less explicitly stated dimensions.

Each LLM judge was required to provide direct textual evidence to support its dimension-support score. Ratings classified as non-compliant under the evidence protocol were treated as missing in the primary analysis and were separately reported as the non-compliance rate. Claude and GPT each had a 9.6% non-compliance rate, while Gemini had a higher rate of 26.7%. For instance, when evaluating unique thought patterns in the same mandatory training persona, Claude and Gemini both assigned ratings of 0, noting insufficient textual evidence supporting the dimension, while GPT assigned a rating of 5. This disagreement illustrates a challenge that LLM judges differ not only in their interpretations but also in what constitutes sufficient evidence for bias evaluation.

As shown in Fig. 1, dimension-support scores varied across dimensions. All three LLM judges produced their highest scores in fairness and their lowest scores in diversity. In fairness, the LLM judges consistently agreed when personas contained explicit role descriptions. For example, in a peer tutoring persona stating *"As a 21-year-old Senior at State University, I have been passionate about mathematics since high school. Serving as a Peer Tutor for the past two years has allowed me to share my love for the subject"*, all LLM judges assigned high ratings (5, 5, 5) for role-based agency, demonstrating strong agreement on explicit role alignment. For lack of stereotypicality, LLM judges showed moderate agreement but occasionally diverged on cultural representation. For diversity, the judges showed greater score divergence and more frequent non-compliant ratings. The same persona received divergent diversity ratings (GPT: 4.6, Claude: 2.2, Gemini: 2.4), indicating different assessments of what constitutes diverse experiences and thought patterns. Fig. 2 visualizes the dimension-support score distributions for each LLM judge, illustrating narrower distributions for GPT and broader ones for Gemini, particularly in the diversity dimension. In general, fairness received the highest and most similar scores across judges, while diversity showed greater variation.

4.2.2. RQ1b: Bias evaluations in persona sets

Overall, GPT produced the highest mean set-level dimension-support score ($M = 2.70$, $SD = 0.29$), followed by Gemini ($M = 2.01$, $SD = 0.42$) and Claude ($M = 1.77$, $SD = 0.81$). Mean dimension-support scores were 41%–61% lower at the set level than at the individual level. Because the individual- and set-level prompts assess different units and require different forms of evidence aggregation, these differences in mean scores should not be interpreted as direct evidence of lower agreement.

As shown in Fig. 3, non-compliance rates at the set level (34%–49%) were substantially higher than at the individual level (9.6%–26.7%). This pattern indicates greater difficulty satisfying the required direct-textual-evidence protocol when reasoning about *patterns across multiple personas*. Qualitatively, LLM judges tended to align when sets contained repeated, explicit role markers (e.g., consistent job titles and credential phrases across all five personas), but diverged when diversity or resource-access claims were implied unevenly across the set. In such cases, one LLM judge sometimes emphasized the strongest exemplar within the set, while another placed greater weight on the absence of evidence in weaker exemplars, contributing to differences in scores and evidence grounding.

4.3. RQ2: LLM judges' agreement on bias

4.3.1. RQ2a: Agreement on individual personas

Agreement refers to the degree of agreement among LLM judges when evaluating the same personas. Inter-rater agreement was estimated using the ICC. Table 5 summarizes ICC values and non-compliance rates for each persona generator.

Fig. 4 displays inter-LLM judge ICC distributions. Agreement ranged from .22 to .54, representing poor to moderate inter-rater agreement depending on the generator. Agreement differed by persona generator, with the lowest agreement observed for Llama-generated personas, which also showed the most documented generation-quality problems. For Llama-generated personas, which contained formatting inconsistencies and incomplete sections, Gemini exhibited a 58.2% non-compliance rate. Consider a Llama-generated mandatory training persona for *"Ethan Lee, a 35-year-old Student Conduct Administrator"*,

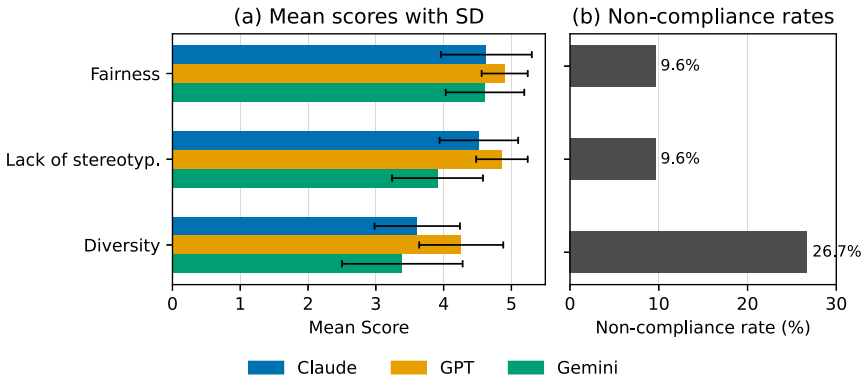
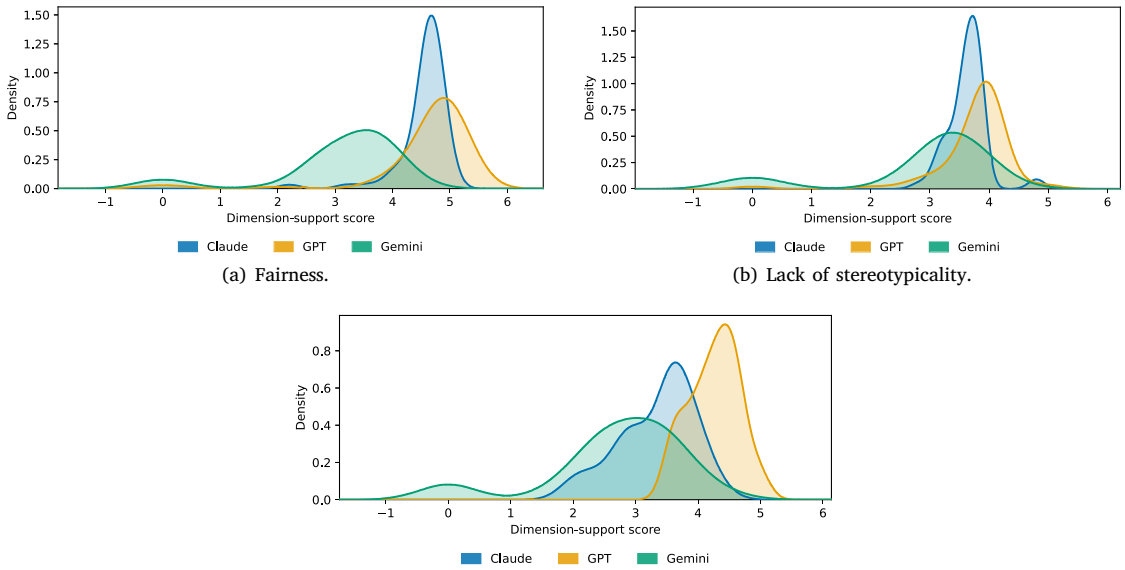


Fig. 1. Individual-level dimension-support scores by LLM judge. Error bars represent standard deviations.



(a) Fairness.

(b) Lack of stereotypicality.

(c) Diversity.

Fig. 2. Dimension-support score distributions for (a) fairness, (b) lack of stereotypicality, and (c) diversity across the three LLM judges.

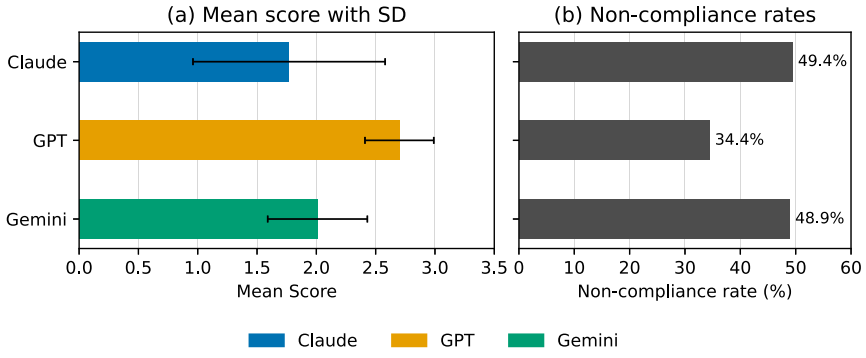
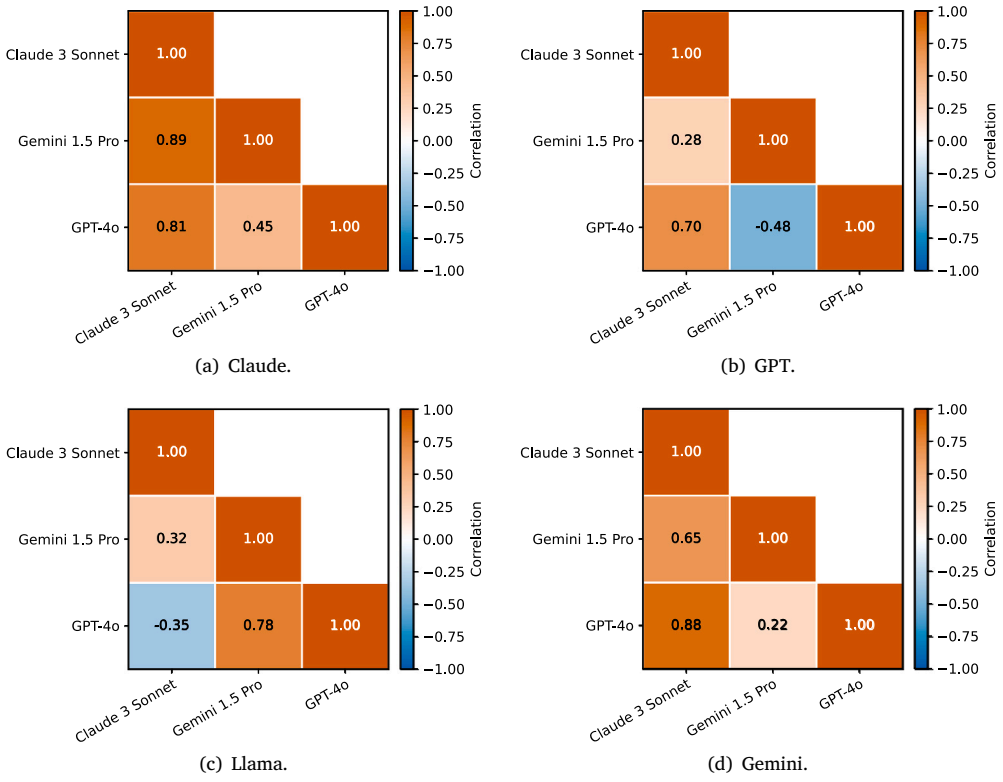


Fig. 3. Set-level dimension-support scores by LLM judge. Error bars represent standard deviations.

which contained JSON formatting artifacts and incomplete narratives. When evaluating this persona, Gemini assigned 0 ratings across all dimensions (fairness: 0, lack of stereotypicality: 0, diversity: 0), noting insufficient evidence, while GPT and Claude provided ratings of 4.8–5.0 for the same dimensions. This disagreement suggests that the judges responded differently to incomplete or inconsistently formatted persona text. In this example, Gemini required explicit textual evidence, while GPT and Claude assigned higher scores based on the available information. In contrast, for well-structured Claude-generated personas such as the Student Support Services Director, LLM judges showed substantially



(a) Claude. (b) GPT.

(c) Llama. (d) Gemini.

Fig. 4. Pairwise correlations among LLM judge dimension-support scores, shown separately by persona generator.

Table 5
Inter-LLM judge agreement by persona generator at the individual persona level. NC = non-compliance.

Generator	ICC	95% CI	Kendall's W	Claude NC%	GPT NC%	Gemini NC%
Claude	.54	[.48,.60]	.52	10.2%	8.9%	21.8%
GPT	.49	[.43,.55]	.47	8.9%	7.0%	14.8%
Llama	.22	[.14,.30]	.20	9.3%	16.0%	58.2%
Gemini	.51	[.45,.57]	.49	10.2%	7.1%	14.2%

Table 6
Inter-LLM judge agreement by persona generator. Each generator includes 45 set–criterion targets (3 persona sets × 15 criteria) evaluated by three LLM judges. NC = non-compliance.

Generator	ICC	95% CI	Kendall's W	Claude NC%	GPT NC%	Gemini NC%
Claude	.63	[.55,.70]	.60	53.3%	37.8%	51.1%
GPT	.49	[.40,.58]	.46	48.9%	35.6%	42.2%
Llama	.50	[.41,.59]	.47	33.3%	33.3%	53.3%
Gemini	.39	[.29,.49]	.36	62.2%	31.1%	48.9%

higher agreement (ICC = .54), with fairness ratings converging at 5.0, 4.6, and 3.0, still showing some variance but with all LLM judges finding sufficient evidence to evaluate.

Across bias dimensions, fairness yielded the strongest agreement (mean ICC = .52), followed by lack of stereotypicality (.47) and diversity (.39). For fairness, LLM judges consistently agreed when personas contained explicit qualifications and role descriptions. For example, in a professional mentoring persona describing *"a 37-year-old Assistant Professor in Computer Science at Stanford [...] with a Ph.D. in Computer Science, I bring five years of experience"*, all LLM judges assigned ratings of 5 for role alignment based on qualifications. For lack of stereotypicality, LLM judges showed moderate agreement but sometimes diverged on cultural representation, with ratings ranging from 4 to 5 when evaluating whether cultural aspects were specifically defined. For diversity, the agreement was lowest, with the LLM judges fundamentally disagreeing on what constitutes unique experiences. The Assistant Professor persona received diversity ratings of 4.2 (GPT), 4.0 (Claude), and 2.4 (Gemini), with Gemini assigning substantially lower ratings than the others for unique thought patterns. These findings show higher agreement for explicitly stated attributes and lower agreement for more abstract concepts requiring inference, with agreement also varying across persona generators.

4.3.2. RQ2b: Agreement on persona sets

Inter-rater agreement at the set level was estimated across the 45 set–criterion targets available for each generator. Agreement ranged from .39 to .63 across generators (Table 6), reflecting low-to-moderate inter-rater agreement. Claude-generated sets yielded the highest agreement (ICC = .63), while Gemini-generated sets yielded the lowest (ICC = .39).

Non-compliance rates were elevated across generators (31%–62%), showing that even when the LLM judges converged numerically, many set-level ratings did not satisfy the required direct-textual-evidence protocol. Fig. 5 visualizes these rates across LLM judges and generators. Mirroring the individual-level results, dimension-specific agreement was highest for fairness (ICC = .55), followed by lack of stereotypicality (.48) and diversity (.43). Agreement was therefore higher for dimensions supported by explicit role and qualification cues and lower for more abstract constructs related to diversity and thought patterns.

4.4. RQ3: LLM judge agreement across educational contexts

4.4.1. RQ3a: Variation at individual persona level

This subsection reports differences in LLM judge scores and agreement across the three educational contexts classified by risk level (peer tutoring, mentoring, and mandatory training) at the individual level.

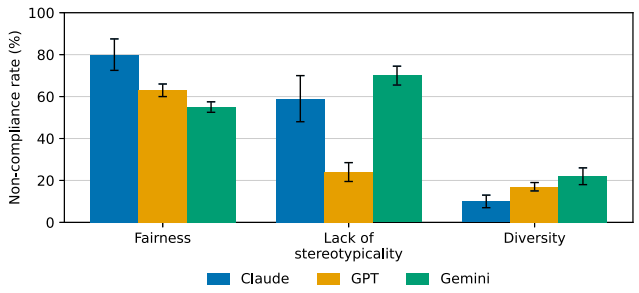


Fig. 5. Non-compliance rates for set-level evaluations across dimensions and LLM judges. Higher values indicate a larger proportion of ratings that did not satisfy the required direct-textual-evidence protocol. Error bars represent standard deviations.

Trend in agreement across contexts. As shown in Fig. 6, inter-LLM judge agreement decreased across the scenarios ordered from lower to higher classified risk. The rank association between the ordered context classification (lower, medium, higher risk) and ICC was perfectly negative ($r_s = -1.00$), corresponding to a monotonic decline in agreement across the three scenarios. With only three ordered contexts, however, this correlation should be interpreted descriptively; standard Spearman significance testing lacks statistical power at $n = 3$, and even a perfect rank correlation does not reach conventional significance thresholds under permutation testing (Zar, 1972). The pattern is therefore presented as a descriptive trend supported by the magnitude of the ICC decline (from .55 to .34, a 37.1% reduction), not as an isolated effect of risk.

In the low-risk peer tutoring context, a persona describing a mathematics tutor (*"As a 21-year-old Senior at State University[...]Serving as a Peer Tutor for the past two years"*) received perfect LLM judge agreement (5, 5, 5) on agency reflecting role instead of demographics. In the medium-risk professional mentoring context, an Assistant Professor persona (*"37-year-old Assistant Professor in Computer Science at Stanford"*) also received perfect agreement (5, 5, 5) on the same dimension. However, in the high-risk mandatory training context, even when the persona contained similarly explicit role descriptions (*"Director of Student Support Services at Westfield University [...] over a decade of experience"*), LLM judges still achieved perfect agreement (5, 5, 5) on explicit role-based questions but showed dramatic disagreement on other fairness dimensions, such as resource access (5, 3, 0). This decline in ICC shows that agreement differed between the three ordered educational contexts, even when some dimensions remained clearly evaluable. Because the scenarios also differ in institutional power, role structure, stakes, and textual content, the decline cannot be attributed to risk alone.

Linear mixed-effects models with random intercepts for persona and criterion showed a small but significant difference between the educational contexts in GPT mean scores ($F(2, 57) = 4.21$, $p = .020$, $f = 0.09$), with scores marginally higher in mentoring than in peer tutoring and mandatory training. Differences across contexts for Claude ($F(2, 57) = 2.31$, $p = .108$, $f = 0.07$) and Gemini ($F(2, 57) = 0.29$, $p = .751$, $f = 0.03$) did not reach significance. Effect sizes were small for all judges. Thus, while ICC declined across the ordered contexts, individual judge mean scores were relatively stable, suggesting that the observed change in agreement reflected divergence among the judges.

Examining average dimension scores across contexts reveals this pattern: in peer tutoring, fairness ratings were highly aligned (GPT: 5.0, Claude: 4.8, Gemini: 3.6), while in mandatory training, the same LLM judges showed greater variance (GPT: 5.0, Claude: 4.6, Gemini: 3.0). These results show context-linked variation in LLM judge agreement at the individual level, with the lowest agreement observed in the mandatory training scenario.

4.4.2. RQ3b: Variation at persona set level

Set-level agreement varied across contexts without a monotonic trend. As summarized in Fig. 7, ICC values were .46 (peer tutoring), .54 (mentoring), and .48 (training). The rank association between the ordered context classification and ICC showed no consistent direction ($r_s = 0.50$), and given $n = 3$ contexts, this result is treated descriptively alongside the individual-level finding.

Linear mixed-effects models at the set level detected no significant differences across contexts for any judge (Claude: $F(2, 9) = 1.58$, $p = .256$; GPT: $F(2, 9) = 0.06$, $p = .944$; Gemini: $F(2, 9) = 1.79$, $p = .218$), consistent with the absence of a monotonic pattern in the set-level ICCs. Across contexts, non-compliance rates remained high (Fig. 5), suggesting that evidence aggregation across multiple narratives may contribute more strongly to the observed set-level difficulties than the ordered context classification.

5. Discussion

5.1. Theoretical contributions

Our findings reveal that LLM judge reliability varies across evaluation scope and educational context. The set-level ICCs (.39–.63) overlap with the individual-level range, indicating that set-level agreement was not uniformly lower. However, the differences in mean scores between the individual and set levels, together with the substantially higher non-compliance rates at the set level, suggest greater difficulty in grounding evaluations when judges assess personas as an interdependent collection of attributes instead of discrete textual entities.

The differences between educational contexts also varied by the scope of the evaluation. At the individual level, the agreement decreased from peer tutoring (ICC = .55) to mandatory training (ICC = .34), while this pattern was not observed at the set level. However, the three scenarios also differ in stakes, institutional power, role structure, and likely textual content, meaning that this pattern cannot be attributed to risk alone. Three alternative explanations may help explain why the ordered context pattern was not observed at the set level.

First, the *cognitive load hypothesis* holds that judges may track contextual nuance more readily when processing a single persona than when synthesizing patterns across five simultaneously presented texts. Second, the *aggregation masking hypothesis* proposes that set-level averaging obscures context-linked differences that remain present at the criterion level; that is, individual criteria may still vary across contexts, but their aggregation into a single set score cancels out these differences. Third, the *evaluation framing hypothesis* suggests that judges may approach individual and set evaluations differently. Individual assessment resembles a close-reading task where contextual features are more salient, whereas set assessment resembles a pattern-detection task where these features may become less prominent. These hypotheses are not mutually exclusive, and distinguishing among them would require designs that manipulate presentation format, aggregation method, and prompt framing independently.

The evidence-provision failures that we documented represent an additional limitation in the evaluative behavior of LLM judges. When judges do not ground numerical ratings in the required direct textual evidence, particularly in 34 to 49% of set-level assessments, the numerical scores become more difficult to interpret as evidence-grounded evaluations. This limitation parallels findings in human cognition research (Weinberger and Green, 2022; Dienes and Perner, 1999), where evaluators sometimes "know it when they see it" but struggle to articulate explicit criteria. The substantially higher non-compliance rates at the set level suggest that evidence grounding becomes more difficult when the evaluation requires synthesizing information across multiple persona texts.

Most critically, our study measures agreement without establishing accuracy, exposing an important gap in LLM judge validation. The

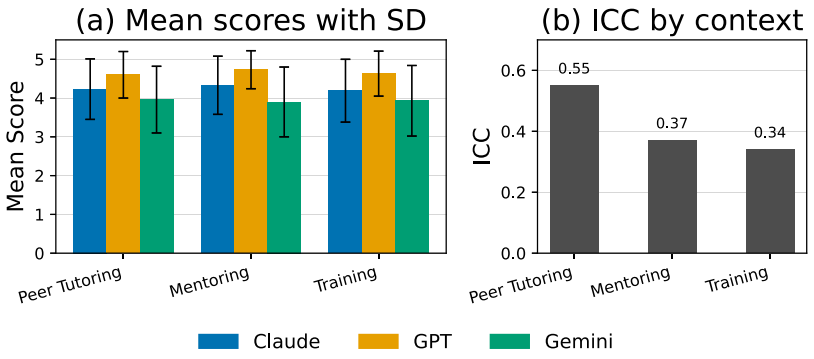


Fig. 6. Individual-level dimension-support scores across educational contexts. Error bars represent standard deviations.

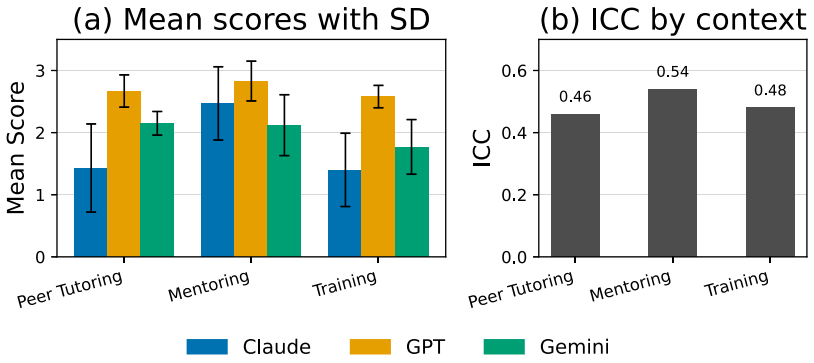


Fig. 7. LLM judge consistency across contexts for persona sets. Error bars represent standard deviations.

moderate ICC values we observed indicate that judges apply somewhat consistent standards, but consistency does not equal correctness. Multiple judges can systematically fail to detect certain bias patterns while showing high agreement on their shared assessments. This possibility becomes especially concerning when automated systems may encode similar training biases that produce correlated errors. Our findings contribute to scholarly understanding of how agreement varies among LLM judges, while suggesting that even achieving high consistency would not validate these systems without ground-truth comparison. This distinction between reliability and validity remains central to future conceptual and empirical frameworks for automated evaluation systems.

5.2. Practical implications

The agreement patterns observed here suggest that set-level persona evaluation using LLM judges under the current protocol requires particular caution. The clearest set-level limitation was the substantially higher non-compliance rate, which approached 50% for some judges, while agreement remained low to moderate and varied across generators. These results do not establish that set-level evaluation is fundamentally beyond current LLM capabilities, but they indicate that the tested models, prompting strategy, and task framing do not provide sufficient evidence to support autonomous use without further validation. Whether alternative prompting strategies, fine-tuned models, or structured aggregation methods can address these limitations remains an open empirical question. Set-level assessments should therefore not be interpreted as validation that persona sets are unbiased without additional evidence about the accuracy of the judgments.

Individual persona evaluation similarly showed poor-to-moderate agreement under this protocol, with the lowest agreement observed in the mandatory training context. Educational curriculum development, healthcare intervention design, and public policy planning constitute consequential contexts in which biased personas can affect the services or opportunities available to represented populations. In such applications, the observed variation in agreement and evidence grounding indicates that LLM judge outputs should be validated for the specific task and context in which they are used. Human review can provide an additional safeguard while appropriate reliability and validity benchmarks are established.

Organizations may cautiously consider LLM judges for preliminary screening in lower-consequence individual persona contexts. Such applications include exploratory design research or internal brainstorming exercises where persona use does not directly affect resource allocation or opportunity distribution for real populations. Even in these contexts, organizations should establish application-specific criteria for acceptable agreement and evidence grounding and review LLM outputs before they inform consequential decisions. This approach views LLM judges as tools that can expose possible issues for further review.

Organizations considering automated persona evaluation should also examine evidence of accuracy in addition to agreement among LLM judges. Our results show substantial variation in agreement among the tested models, but even systems achieving high agreement would require comparison against human or other ground-truth assessments to establish accuracy. Agreement among systems does not by itself establish that their evaluations are correct. Validation should therefore examine whether automated judges align with appropriate expert or ground-truth assessments, satisfy the required evidence-grounding protocol, and maintain performance across relevant persona types and contexts. Without such validation, agreement metrics may provide false confidence in systems that produce similar evaluations while sharing the same errors.

5.3. Limitations and future research directions

Our study faces several methodological constraints that suggest directions for future work.

A fundamental limitation arises from the fact that the LLMs served as both persona generators and LLM judges. This introduces the possibility of error compounding, where generation biases can influence evaluation outcomes, making it difficult to isolate the source of the disagreement. The study's scope, limited to three LLM judges and four generators, further constrains the generalizability of the findings. In addition, the three educational scenarios differed not only in their classified risk level but also in institutional power, role structure, stakes, and textual content. The observed differences across scenarios therefore cannot be attributed to risk alone. Future research should include a wider range of models and extend evaluation frameworks beyond educational settings while examining these contextual characteristics more independently.

The generator-specific set-level ICCs also pool criterion evaluations within persona sets; because criteria from the same set share the same underlying personas, the observations are dependent on the personas and thus should be considered a criterion-level agreement among the different judges.

Another limitation is the absence of human judges or another ground-truth baseline for comparison. Without a human baseline, it is not possible to determine whether the observed agreement levels are comparable to human inter-rater agreement or to establish empirically supported thresholds for acceptable agreement. Incorporating human evaluations into future work would provide critical benchmarks to assess the validity and interpretive value of LLM-based persona judgments. The evidence-provision coding was also conducted by the first author without independent double-coding, meaning that the classification of legitimate zeros, compliance failures, and weak justifications was not independently assessed.

The substantial variation in compliance with the evidence protocol suggests a need for more standardized evidence-grounding procedures. Future work should examine whether fine-tuning LLM judges for persona evaluation improves inter-rater agreement compared to the use of unmodified (vanilla) models. Advanced strategies such as retrieval-augmented generation or structured prompting may also improve evidence grounding and reduce non-compliance rates, particularly at the set level.

Finally, future research should explore frameworks that disentangle generation and evaluation errors and investigate whether structured aggregation methods can improve consistency and agreement without relying on unverifiable ensemble consensus.

6. Conclusion

This study examined how LLM judges agree when evaluating fairness, lack of stereotypicality, and diversity in individual personas and persona sets across educational contexts. The findings show poor-to-moderate agreement among the tested judges and substantially greater difficulty satisfying the required evidence-grounding protocol at the set level. Agreement also varied across educational contexts, although these differences cannot be attributed to risk alone because the scenarios differed in several contextual characteristics. These results demonstrate that agreement, evaluation scope, and evidence grounding need to be considered separately when assessing LLM judges. As this study measures agreement rather than accuracy, future work should compare LLM judgments with expert assessments and examine whether alternative evaluation and aggregation approaches improve the reliability of persona evaluation.

CRediT authorship contribution statement

Danial Amin: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Joni Salminen:** Writing – review & editing, Supervision. **Bernard J. Jansen:** Writing – review & editing, Validation, Supervision.

Declaration of Generative AI Usage

During the preparation of this work, the author(s) used Anthropic's Claude Sonnet-4.5 to summarize and copy edit the content. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.techfore.2026.124894.

Data availability

The data is shared online on OSF (https://doi.org/10.17605/OSF.IO/U4FPG).

References

Amin, D., Salminen, J., Jansen, B., Shin, J., Kim, D.H., 2025. Generative AI personas considered harmful? Examining the risks of algorithmic user representations in HCI design. Int. J. Hum.-Comput. Stud. http://dx.doi.org/10.1016/j.ijhcs.2025.103657.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, Association for Computing Machinery, New York, NY, USA, pp. 610–623. http://dx.doi.org/10.1145/3442188.3445922.

Braga, C.M., Serrano, M.A., Fernández-Medina, E., 2025. Towards a methodology for ethical artificial intelligence system development: A necessary trustworthiness taxonomy. Expert Syst. Appl. 128034. http://dx.doi.org/10.1016/j.eswa.2025.128034.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Buolamwini, J., Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. FAT*, PMLR, pp. 77–91, URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

Chapman, C.N., Love, E., Milham, R.P., ElRif, P., Alford, J.L., 2008. Quantitative evaluation of personas as information. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 1107–1111. http://dx.doi.org/10.1177/154193120805201602.

Chapman, C.N., Milham, R.P., 2006. The personas' new clothes: Methodological and practical arguments against a popular method. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 50 (5), 634–636. http://dx.doi.org/10.1177/154193120605000503.

Chen, G.H., Chen, S., Liu, Z., Jiang, F., Wang, B., 2024. Humans or LLMs as the judge? A study on judgement bias. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp. 8301–8327. http://dx.doi.org/10.18653/v1/2024.emnlp-main.474.

Cheng, M., Durmus, E., Jurafsky, D., 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2023.acl-long.84.

Choi, J.H., 2027. Off-the-shelf large language models are unreliable judges. Journal of Empirical Legal Studies http://dx.doi.org/10.2139/ssrn.5188865, Forthcoming.

Chrysafiadi, K., Virvou, M., 2013. Student modeling approaches: A literature review for the last decade. Expert Syst. Appl. 40 (11), 4715–4729. http://dx.doi.org/10.1016/j.eswa.2013.02.007.

Cooper, A., 1999. The Inmates Are Running the Asylum: Why High-Tech Products Drive us Crazy and How to Restore the Sanity. SAMS Publishing.

Davidson, D., 1982. Rational animals. Dialectica 36 (4), 317–327. http://dx.doi.org/10.1111/j.1746-8361.1982.tb01546.x.

De Paoli, S., 2026. User personas, ideation and large language models: A post-hoc study. Int. J. Hum.-Comput. Stud. 208, 103690. http://dx.doi.org/10.1016/j.ijhcs.2025.103690.

Dienes, Z., Perner, J., 1999. A theory of implicit and explicit knowledge. Behav. Brain Sci. 22 (5), 735–808. http://dx.doi.org/10.1017/S0140525X99002186.

Du, Y., Oraby, S., Perera, V., Shen, M., Narayan-Chen, A., Chung, T., Venkatesh, A., Hakkani-Tur, D., 2020. Schema-guided natural language generation. In: Proceedings of the 13th International Conference on Natural Language Generation. pp. 283–295. http://dx.doi.org/10.18653/v1/2020.inlg-1.35.

Duan, Y., Cruz, E., Suen, H.K., 2025. Representing groups of students as personas: A systematic review of persona creation, application, and trends in the educational domain. Comput. Educ. Open 7, 100219. http://dx.doi.org/10.1016/j.caeo.2025.100219.

Eby, L.T., Allen, T.D., Evans, S.C., Ng, T., DuBois, D.L., 2008. Does mentoring matter? A multidisciplinary meta-analysis comparing mentored and non-mentored individuals. J. Vocat. Behav. 72 (2), 254–267. http://dx.doi.org/10.1016/j.jvb.2007.04.005.

Friess, E., 2015. Personas in heuristic evaluation: An exploratory study. IEEE Trans. Prof. Commun. 58 (2), 176–191. http://dx.doi.org/10.1109/TPC.2015.2429971.

Fu, X.-Y., Laskar, M.T.R., Chen, C., Tn, S.B., 2023. Are large language models reliable judges? A study on the factuality evaluation capabilities of LLMs. In: Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics. GEM, Association for Computational Linguistics, Singapore, pp. 310–316, URL: https://aclanthology.org/2023.gem-1.25/.

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Ahmed, N.K., 2024. Bias and fairness in large language models: A survey. Comput. Linguist. 1–79. http://dx.doi.org/10.1162/coli_a_00524.

Garreta-Domingo, M., Sloep, P.B., Hernández-Leo, D., 2018. Human-centred design to empower "teachers as designers". Br. J. Educ. Technol. 49 (6), 1113–1130. http://dx.doi.org/10.1111/bjet.12682.

Giorgi, S., Liu, T., Aich, A., Isman, K.J., Sherman, G., Fried, Z., Sedoc, J., Ungar, L., Curtis, B., 2024. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In: Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp. 7174–7188. http://dx.doi.org/10.18653/v1/2024.findings-emnlp.420.

Gonzalez, R.A., DiPaola, S., 2024. Exploring augmentation and cognitive strategies for AI-based synthetic personae. In: Proceedings of the CHI 2024 Workshop on Challenges and Opportunities of LLM-Based Synthetic Personae and Data in HCI. ACM SIGCHI, New York, NY, USA.

Goodman-Deane, J.A., Bradley, M., Waller, S., Clarkson, P.J., 2021. Developing personas to help designers understand digital exclusion. Proc. Des. Soc. 1, 1203–1212. http://dx.doi.org/10.1017/pds.2021.120.

Grudin, J., 2006. Why personas work: The psychological evidence. In: Pruitt, J.S., Adlin, T. (Eds.), The Persona Lifecycle. In: Interactive Technologies, Morgan Kaufmann, San Francisco, pp. 642–663. http://dx.doi.org/10.1016/B978-012566251-2/50013-7.

Guan, K.W., Salminen, J., Jung, S.-G., Jansen, B.J., 2024. Leveraging personas for social impact: A review of their applications to social good in design. Int. J. Hum.–Comput. Interact. 40 (19), 5569–5584. http://dx.doi.org/10.1080/10447318.2023.2247568.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., Khot, T., 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. ICLR, URL: https://proceedings.iclr.cc/paper_files/paper/2024/file/5e1a87dbb7e954b8d9d6c91f6db771eb-Paper-Conference.pdf.

Hämäläinen, P., Tavast, M., Kunnari, A., 2023. Evaluating large language models in generating synthetic HCI research data: A case study. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery, New York, NY, USA, pp. 1–19. http://dx.doi.org/10.1145/3544548.3580688.

Hammersley, M., Gomm, R., 1997. Bias in social research. Sociol. Res. Online 2 (1), 7–19. http://dx.doi.org/10.5153/sro.55.

Hill, C.G., Haag, M., Oleson, A., Mendez, C., Marsden, N., Sarma, A., Burnett, M., 2017. Gender-inclusiveness personas vs. stereotyping: Can we have it both ways? In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 6658–6671. http://dx.doi.org/10.1145/3025453.3025609.

Jang, M.E., Silavong, F., 2025. InstaJudge: Aligning judgment bias of LLM-as-judge with humans in industry applications. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, Suzhou (China), pp. 1158–1172. http://dx.doi.org/10.18653/v1/2025.emnlp-industry.82.

Jansen, B.J., Jung, S.-G., Salminen, J., 2023. Employing large language models in survey research. Nat. Lang. Process. J. 4, 100020. http://dx.doi.org/10.1016/j.nlp.2023.100020.

Jansen, B.J., Salminen, J., Jung, S.-G., Guan, K., 2021. Evaluating data-driven personas. In: Data-Driven Personas. Springer International Publishing, Cham, pp. 209–237. http://dx.doi.org/10.1007/978-3-031-02231-9_9.

Kashi, M., Lahmiri, S., Ait Mohamed, O., 2025. Comprehensive analysis of Transformer networks in identifying informative sentences containing customer needs. Expert Syst. Appl. 267, 124075. http://dx.doi.org/10.1016/j.eswa.2025.126785.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: Tests in linear mixed effects models. J. Stat. Softw. 82 (13), 1–26. http://dx.doi.org/10.18637/jss.v082.i13.

Lee, J., Hicke, Y., Yu, R., Brooks, C., Kizilcec, R.F., 2024. The life cycle of large language models in education: A framework for understanding sources of bias. Br. J. Educ. Technol. 55 (5), 1982–2002. http://dx.doi.org/10.1111/bjet.13505.

Lee, Y.-S., Yen, S.-J., Jiang, W., Chen, J., Chang, C.-Y., 2025. Illuminating the black box: An interpretable machine learning based on ensemble trees. Expert Syst. Appl. 272, 126720. http://dx.doi.org/10.1016/j.eswa.2025.126720.

Li, Y., Liu, Y., Yu, M., 2024. Consumer segmentation with large language models. J. Retail. Consum. Serv. http://dx.doi.org/10.1016/j.jretconser.2024.104078.

Liu, A., Diab, M., Fried, D., 2024. Evaluating large language model biases in persona-steered generation. In: Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, pp. 9832–9850. http://dx.doi.org/10.18653/v1/2024.findings-acl.586.

Long, F., 2009. Real or imaginary: The effectiveness of using personas in product design. In: Proceedings of the Irish Ergonomics Society Annual Conference. pp. 1–10, URL: https://informationdesign.org/real_or_imaginary_the_effectiv/.

Ma, J., Wang, D., Chen, L., 2024. Can you trust LLM judgments? Reliability of LLM-as-a-judge. arXiv preprint arXiv:2412.12509.

Marsden, N., Haag, M., 2016. Evaluation of GenderMag personas based on persona attributes and persona gender. In: International Conference on Human-Computer Interaction. Springer, pp. 122–127. http://dx.doi.org/10.1007/978-3-319-40548-3_20.

Marshall, R., Cook, S., Mitchell, V., Summerskill, S., Haines, V., Maguire, M., Sims, R., Gyi, D., Case, K., 2015. Design and evaluation: End users, user datasets and personas. Appl. Ergon. 46, 311–317. http://dx.doi.org/10.1016/j.apergo.2013.03.008.

Miaskiewicz, T., Kozar, K.A., 2011. Personas and user-centered design: How can personas benefit product design processes? Des. Stud. 32 (5), 417–430. http://dx.doi.org/10.1016/j.destud.2011.03.003.

Murray, T.E., Simon, B.L., 2006. What is dialect? In: Language Variation and Change in the American Midland. pp. 1–15.

Nielsen, L., 2019. Personas - User Focused Design, 2nd ed. 2019 ed. Springer, New York, NY.

Pan, Q., Ashktorab, Z., Desmond, M., Santillán Cooper, M., Johnson, J., Nair, R., Daly, E., Geyer, W., 2024. Human-centered design recommendations for LLM-as-a-judge. In: Proceedings of the 1st Human-Centered Large Language Modeling Workshop. ACL, TBD, pp. 16–29. http://dx.doi.org/10.18653/v1/2024.hucllm-1.2.

Panagoulias, D.P., Virvou, M., Tsihrintzis, G.A., 2024. A novel framework for artificial intelligence explainability via the technology acceptance model and rapid estimate of adult literacy in medicine using machine learning. Expert Syst. Appl. 248, 123375. http://dx.doi.org/10.1016/j.eswa.2024.123375.

Patel, R., Kumar, A., Lee, S., 2024. Judge's verdict: A comprehensive analysis of LLM judge capability through human agreement. arXiv preprint arXiv:2510.09738.

Pradhan, A., Ortan, A., Verma, A., Seshadri, M., 2025. LLM-as-a-Judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation. In: Proceedings of the 2nd Workshop on Evaluating and Applying Recommender Systems with Large Language Models (EARL). RecSys 2025, ACM, New York, NY, USA.

Prpa, M., Troiano, G.M., Wood, M., Coady, Y., 2024. Challenges and opportunities of LLM-based synthetic personae and data in HCI. In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems. In: CHI EA '24, Association for Computing Machinery, New York, NY, USA, pp. 1–5. http://dx.doi.org/10.1145/3613905.3636293.

Rodrigo-Ginés, F.-J., de Albornoz, J.C., Plaza, L., 2024. A systematic review on media bias detection: what is media bias, how it is expressed, and how to detect it. Expert Syst. Appl. 234, 121437. http://dx.doi.org/10.1016/j.eswa.2023.121641.

Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A., 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. http://dx.doi.org/10.48550/arXiv.2402.07927, arXiv preprint arXiv:2402.07927.

Salminen, J., Amin, D., Jansen, B.J., 2025. Using AI for user representation: An analysis of 83 persona prompts. In: 2025 IEEE/ACS 22nd International Conference on Computer Systems and Applications. AICCSA, IEEE, pp. 1–8. http://dx.doi.org/10.1109/AICCSA66935.2025.11315350.

Salminen, J., Froneman, W., Jung, S.-G., Chowdhury, S., Jansen, B.J., 2020a. The ethics of data-driven personas. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–9. http://dx.doi.org/10.1145/3334480.3382790.

Salminen, J., Jansen, B.J., An, J., Kwak, H., Jung, S.-G., 2018. Are personas done? Evaluating their usefulness in the age of digital analytics. Pers. Stud. 4 (2), 47–65. http://dx.doi.org/10.21153/psj2018vol4no2art737.

Salminen, J., Jung, S.-G., Chowdhury, S.A., Jansen, B.J., 2020b. Rethinking personas for fairness: Algorithmic transparency and accountability in data-driven personas. In: Artificial Intelligence in HCI. Springer International Publishing, Cham, pp. 82–100. http://dx.doi.org/10.1007/978-3-030-50334-5_6.

Salminen, J., Jung, S.-G., Chowdhury, S., Robillos, D.R., Jansen, B.J., 2021. The ability of personas: An empirical evaluation of altering incorrect preconceptions about users. Int. J. Hum.-Comput. Stud. http://dx.doi.org/10.1016/j.ijhcs.2021.102645.

Salminen, J., Jung, S.-g., Chowdhury, S.A., Sengün, S., Jansen, B.J., 2020c. Personas and analytics: A comparative user study of efficiency and effectiveness for a user identification task. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–13. http://dx.doi.org/10.1145/3313831.3376770.

Salminen, J., Jung, S.-g., Nielsen, L., Şengün, S., Jansen, B.J., 2022. How does varying the number of personas affect user perceptions and behavior? Challenging the 'small personas' hypothesis!. Int. J. Hum.-Comput. Stud. 168, 102915. http://dx.doi.org/10.1016/j.ijhcs.2022.102915.

Salminen, J., Liu, C., Pian, W., Chi, J., Häyhänen, E., 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 CHI Conference on Human Factors in Computing Systems. CHI '24, Association for Computing Machinery, New York, NY, USA, pp. 1–20. http://dx.doi.org/10.1145/3613904.3642036.

Salminen, J., Santos, J.M., Kwak, H., An, J., Jung, S.-G., Jansen, B.J., 2020d. Persona perception scale: Development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. Int. J. Hum.-Comput. Stud. 141, 102437. http://dx.doi.org/10.1016/j.ijhcs.2020.102437.

Schuller, A., Janssen, D., Blumenröther, J., Probst, T.M., Schmidt, M., Kumar, C., 2024. Generating personas using LLMs and assessing their viability. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. ACM, Honolulu, HI, USA, pp. 1–7. http://dx.doi.org/10.1145/3613905.3650860.

Sethi, S., Salminen, J., Amin, D., Jansen, B., 2025. "When AI writes personas": Analyzing lexical diversity in LLM-generated persona descriptions. In: Extended Abstracts of CHI Conference on Human Factors in Computing Systems. http://dx.doi.org/10.1145/3706599.3719712.

Shin, J., Hedderich, M.A., Rey, B.J., Lucero, A., Oulasvirta, A., 2024. Understanding human-AI workflows for generating personas. In: Designing Interactive Systems Conference. ACM, IT University of Copenhagen, Denmark, pp. 757–781. http://dx.doi.org/10.1145/3643834.3660729.

Sowa, K., Przegalinska, A., 2025. From expert systems to generative artificial experts: A new concept for human-ai collaboration in knowledge work. J. Artificial Intelligence Res. 82, 2101–2124. http://dx.doi.org/10.1613/jair.1.17175.

Sun, L., Qin, T., Hu, A., Zhang, J., Lin, S., Chen, J., Ali, M., Prpa, M., 2025. Persona-L has entered the chat: Leveraging LLMs and ability-based framework for personas of people with complex needs. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3706598.3713445.

Szandała, T., 2025. ChatGPT vs human expertise in the context of IT recruitment. Expert Syst. Appl. 264, 125868. http://dx.doi.org/10.1016/j.eswa.2024.125868.

Topping, K.J., 2005. Trends in peer learning. Educ. Psychol. 25 (6), 631–645. http://dx.doi.org/10.1080/01443410500345172.

Turner, P., Turner, S., 2011. Is stereotyping inevitable when designing with personas? Des. Stud. 32 (1), 30–44. http://dx.doi.org/10.1016/j.destud.2010.06.002.

Verga, P., Sun, H., Soares, L.B., Cohen, W.W., 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. arXiv preprint arXiv:2404.18796.

Weinberger, A.B., Green, A.E., 2022. Dynamic development of intuitions and explicit knowledge during implicit learning. Cognition 222, 105008. http://dx.doi.org/10.1016/j.cognition.2021.105008.

Wolfe, C.R., 2024. Using LLMs for evaluation: A deep dive into LLM-as-a-Judge. https://cameronrwolfe.substack.com/p/llm-as-a-judge, Deep Learn. Focus Newsl..

Yurtsever, G., 2020. Measuring social personas. Open J. Soc. Sci. 8 (06), 111. http://dx.doi.org/10.4236/jss.2020.86011.

Zagallo, P., McCourt, J., Idsardi, R., Smith, M.K., Urban-Lurain, M., Andrews, T.C., Haudek, K., Knight, J.K., Merrill, J., Nehm, R., Prevost, L.B., Lemons, P.P., 2019. Through the eyes of faculty: Using personas as a tool for learner-centered professional development. CBE—Life Sci. Educ. 18 (4), ar62. http://dx.doi.org/10.1187/cbe.19-06-0114.

Zar, J.H., 1972. Significance testing of the spearman rank correlation coefficient. J. Amer. Statist. Assoc. 67 (339), 578–580. http://dx.doi.org/10.1080/01621459.1972.10481251.

Zhang, X., Liu, L., Wang, Y., Liu, X., Wang, H., Arora, C., Liu, H., Wang, W., Hoang, T., 2024. Auto-generated personas: Enhancing user-centered design practices among university students. In: Conference on Human Factors in Computing Systems - Proceedings. http://dx.doi.org/10.1145/3613905.3651043.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J., Stoica, I., 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. Curran Associates Inc., pp. 46595–46623. http://dx.doi.org/10.52202/075280-2020.

Zimmermann, D., Grötzbach, L., 2007. A requirement engineering approach to user centered design. In: International Conference on Human-Computer Interaction. Springer, pp. 360–369. http://dx.doi.org/10.1007/978-3-540-73105-4_40.

Danial Amin is a researcher at the University of Vaasa, focusing on generative AI personas for social good and developing ethical and inclusive user representation that benefits marginalized communities in the Global South. He also serves as the Data and AI Lead at Otherkind Tech, a generative AI for social good impact lab. Danial has over 8 years of experience consulting multinational organizations in artificial intelligence and data science for impactful projects.

Joni Salminen is an Associate Professor (tenure track) at the School of Marketing and Communication in University of Vaasa, Vaasa, Finland. His research focuses on data-driven personas, along with overlapping topics such as Human-Centered AI, Quantitative UX, Interactive Systems, User Segmentation Algorithms, and so on.

Bernard J. Jansen is a Visiting Professor at the Department of Information Management, Peking University, China. He is a West Point graduate with a Ph.D. in computer science from Texas A&M University and master's degrees from Texas A&M (computer science) and Troy State (international relations). Dr. Jim Jansen served in the U.S. Army as an Infantry enlisted soldier and communication commissioned officer.