Misalignment of LLM-Generated Personas with Human Perceptions in Low-Resource Settings

Tabia Tanzin Prama[1,2,3,5] **Christopher M. Danforth**[1,2,3,4] **Peter Sheridan Dodds**[1,2,3,5,6]
[1]Computational Story Lab
[2]Vermont Complex Systems Institute
[3]Vermont Advanced Computing Center
[4]Department of Mathematics and Statistics
[5]Department of Computer Science
University of Vermont, Burlington, VT 05405, USA
[6]Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA

Abstract

Recent advances enable Large Language Models (LLMs) to generate AI personas, yet their lack of deep contextual, cultural, and emotional understanding poses a significant limitation. This study quantitatively compared human responses with those of eight LLM-generated social personas (e.g., Male, Female, Muslim, Political Supporter) within a low-resource environment like Bangladesh, using culturally specific questions. Results show human responses significantly outperform all LLMs in answering questions, and across all matrices of persona perception, with particularly large gaps in empathy and credibility. Furthermore, LLM-generated content exhibited a systematic bias along the lines of the "Pollyanna Principle", scoring measurably higher in positive sentiment ($\Phi_{avg} = 5.99$ for LLMs vs. 5.60 for Humans). These findings suggest that LLM personas do not accurately reflect the authentic experience of real people in resource-scarce environments. It is essential to validate LLM personas against real-world human data to ensure their alignment and reliability before deploying them in social science research.

1 Introduction

Recent advances in Large Language Models (LLMs) have opened new frontiers for simulating human behavior at scale, with significant applications across fields such as social science, economics, clinical psychology, and marketing research [17], [2], [8], [29], [33], [31], [32]. These LLM-driven simulations enable the creation of synthetic agents, known as personas, that replicate real-world populations' behaviors and decision-making processes [21]. Personas, typically characterized by demographic, psychographic, and behavioral attributes, allow researchers to explore social dynamics, behavioral patterns, and responses to various interventions [15]. By utilizing LLMs, personas can be generated quickly and at scale, offering an effective and cost-efficient means for testing social theories and analyzing human-like decision-making [2].

However, creating accurate and representative persona sets for simulations remains a significant challenge. Most persona development approaches are still based on qualitative methods that are time-consuming, static, and resource-intensive [27]. LLMs present a viable solution to these challenges by generating persona profiles directly in a cost-effective, efficient, and seemingly realistic manner. This nascent direction has gained significant attention from both academia and industry, with LLM-generated personas used in surveys, marketing research, and societal-scale simulations [9], [3], [30]. By leveraging LLMs, persona development can be accelerated, making it feasible

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: PersonaLLM: Workshop on LLM Persona Modeling.

to generate high-quality personas with minimal manual effort. Despite the promise of LLM-based personas, challenges persist. Current scalable persona generation methods are significantly biased and fail to authentically reflect the diversity of traits, behaviors, and psychometric profiles found in real-world populations. These biases arise from imbalanced training data and the generalization capabilities of LLMs [15]. For instance, the dominance of high-resource languages, such as English, in training datasets results in a bias toward these languages, leaving low-resource languages underrepresented [12], [22]. Previous work has primarily focused on scaling up the number of diverse personas and assessing their performance in different downstream applications. However, there is limited research exploring the use of LLMs for persona generation in multicultural settings [4], [13].

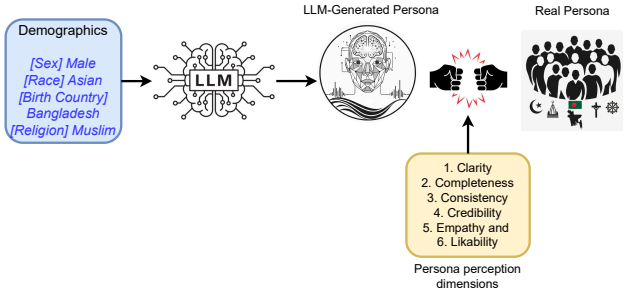


Figure 1: Examining the Discrepancy in Alignment Between LLM-Generated and Human Personas in Low-Resource Contexts: A Case Study of Bangladesh.

To the best of our knowledge, we are the first to explore whether LLMs can generate personas that authentically represent key demographic factors such as religion, gender, and political affiliation within a low-resource environment like Bangladesh (see Figure 1). As Bangladesh is a low-resource language context in the field of AI [25], there exists a significant gap in understanding the performance of LLMs when applied to generate culturally specific personas for this region [26]. Our study contributes to this discourse by addressing the following two research questions:

RQ1: Can LLMs demonstrate different personas based on religion, gender, and political affiliation in a low-resource environment like Bangladesh? To answer this, we constructed a questionnaire based on the suggestions of expert sociologists in the context of Bangladesh and created eight different personas representing the two major political parties (Bangladesh Awami League (AL) and Bangladesh Nationalist Party (BNP)), religions (Islam, Hinduism, and Buddhism), and genders (Male & Female) in Bangladesh using 7 different LLMs (GPT-5.0 [20], GPT-4.1 [19], Grok 3 [34], GPT-4.0 [18], Llama 3.3 [1], DeepSeek V3 [5], and AI21 Jamba 1.5 Large [14]). Each llm persona and actual human persona answered culturally specific questions, and we measured the accuracy of the responses. The results revealed that all models performed poorly, lagging significantly behind human responses.

RQ2: Which LLMs are perceived to generate better personas, as evaluated through the Persona Perception Scale (PPS)? We evaluated the personas using the Persona Perception Scale (PPS) [28], which includes six subscales: Credibility, Consistency, Completeness, Clarity, Empathy, and Likability. The results indicated that human-generated responses outperformed LLM-generated responses in PPS scores across all persona categories.

2 Methodology

2.1 Dataset

Questionnaire. We created a dataset of 100 questionnaires focused on three themes: politics, religion, and gender in the context of Bangladesh. The political persona explores topics like the Liberation War, post-independence dynamics, and the 2024 July Revolution [23]. Religion addresses the role of religion in Bangladesh, including secularism, state ideology, and cultural identity. Gender examines women's roles in politics, law, and economics. The dataset, developed by an expert

sociologist, covers the cultural, historical, and social phenomena of Bangladesh (see Appendix A, Figure 4).

Persona Modeling for LLMs Responses. We model personas across gender, religion, and political affiliation in Bangladesh. Each category includes pairs of personas with opposing characteristics based on participant demographics. Personas are assigned to LLMs using a template [11]:

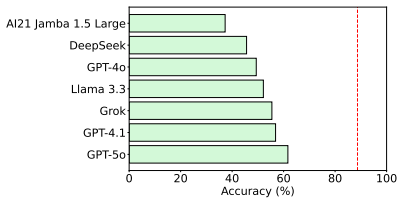
> *"You are a <persona> from Bangladesh. Your responses should closely mirror the knowledge and abilities of this persona."*

Annotationn Precedure. We used seven LLMs and real-life personas to generate explanations for the questions. For each of the eight personas, we asked the same question to human participants who identified with those personas. The responses were manually annotated as either "fully answer" (good) or "partially or not explained" (bad) by three annotators, with a majority vote deciding. This resulted in 2080 instances, of which 1166 were good and 914 were bad answers. Examples are provided in Appendix A Table 1. To evaluate the quality of LLM personas and real human persona, we used the Persona Perception Scale (PPS) [[28]], which is a survey instrument for evaluating how individuals perceive personas scoring credibility, consistency, completeness, clarity, empathy, and likability on a 7-point Likert scale. The PPS scores for each persona are shown in Appendix A.1 Figure 6.

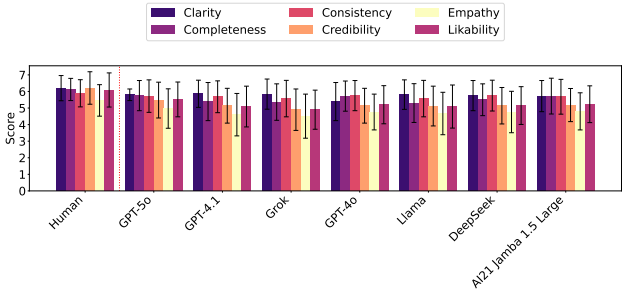
3 Result and Discussion

Overall Model Performance. Figure 2a presents the average accuracy of seven LLMs compared with human performance on the persona perception task. Appendix A.2 presents additional results and analysis.

Overall, the results indicate a persistent performance gap between human reasoning and current LLM capabilities when evaluating nuanced social and identity-related content. Human annotators achieved an accuracy of approximately 87% (red dashed line), setting a clear upper benchmark. Among the models, GPT-5o attained the highest accuracy (61.7%), followed by GPT-4.1 (56.9%) and Grok (55.5%), demonstrating stronger alignment with human judgments. Llama 3.3 (52.1%) and GPT-4o (49.4%) showed moderate performance, while DeepSeek (45.6%) and AI21 Jamba 1.5 Large (37.3%) performed considerably lower.



(a) Average accuracy of LLMs compared to human performance.The bar chart presents the mean accuracy (%) of seven LLMs and the red dashed vertical line indicates the human benchmark accuracy



(b) Mean scores and standard deviations of persona perception dimensions (Clarity, Completeness, Consistency, Credibility, Empathy, and Likability) on a 7-point Likert scale across human evaluators and seven large language models (LLMs).

Figure 2: Comparison of human and LLMs performance on persona perception tasks.(a) The average accuracy of seven LLMs compared with human evaluators, (b) The mean scores and standard deviations for six persona perception dimensions of PPS.

Figure 2b shows the mean scores and standard deviations for six perception dimensions—Clarity, Completeness, Consistency, Credibility, Empathy, and Likability—across human evaluators and seven large language models (LLMs).. Across all six metrics (Clarity, Completeness, Consistency, Credibility, Empathy, and Likability), the Human panel maintained the highest mean scores, demonstrating superior overall engagement quality. The largest performance deficits for LLMs were observed in the social dimensions of Credibility and Empathy. While the Human panel achieved a

mean Credibility score of 6.21 ± 0.98 and an Empathy score of 5.46 ± 0.95, the LLMs, particularly Grok (Credibility 4.90 ± 1.25 and Empathy 4.51 ± 1.33), scored significantly lower, indicating a fundamental difficulty in establishing trust and emotional resonance. GPT-5o generally scored highest among the models, approaching the Human baseline closest in Credibility (5.48 ± 1.08) and Likability (5.52 ± 1.05). Appendix A.2 details the accuracy and six persona perception dimensions of PPS across eight personas. Figure 5 highlights a significant accuracy gap, particularly for political and religious minorities, with AI21 Jamba 1.5 Large scoring just 25% for the BNP persona, indicating bias towards AL [24], and 26.7% for the Buddhist persona—both far below the lowest human panel score of 75%. Additionally, models consistently performed worse on the Female persona, revealing gender bias. The persistent gap in both accuracy (Figure 2a) and perception (Figure 2b) underscores the challenges current models face in accurately representing diverse, context-specific personas.

Case Study: Sentiment shifts between Real Human and LLM-Generated Personas

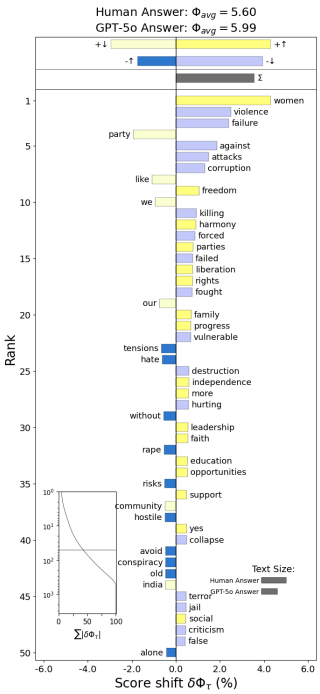


Figure 3: Word shift graph of word frequencies in happiness of real human and LLM-generated personas' responses. Words are ranked by their percentage contribution to the change in average happiness, Φ_{avg}. The real human responses are set as the reference text T_{ref}, with the respective LLM-generated personas' responses as the comparison text T_{comp}. Individual word contributions to the shift are indicated by two symbols: $+/-$ shows the word is more/less prevalent in T_{comp} than in T_{ref}. The four bars on the top indicate the total contribution of the four types of words $(+\uparrow, +\downarrow, -\uparrow, -\downarrow)$. Relative text size is represented by the areas of the gray squares.

Figure 3 shows the sentiment shifts, utilizing word shift graphs and the labMT sentiment dictionary, revealed a significant difference in emotional tone between Human and LLM-generated responses (See Appendix A.3 for details). LLMs consistently produced text with a higher average sentiment ($\Phi_{avg} = 5.99$) compared to Human responses ($\Phi_{avg} = 5.60$), resulting in a substantial positive shift of 0.39. This elevated sentiment is primarily driven by a systematic LLM bias towards positive language. The models frequently leveraged high-sentiment words such as "freedom," "harmony," "liberation," "rights," and "support" while simultaneously exhibiting a lower frequency of strong negative terms like "violence," "failure," "attacks," "corruption," and "criticism." This avoidance of highly negative vocabulary and preference for positive vocabulary suggests that LLM-generated responses adhere to a "Pollyanna Principle," creating communication that is measurably happier and more optimistic than real human dialogue.This preference for positive vocabulary and avoidance of negative terms reflects a "Pollyanna Principle" [6] bias, making LLM responses measurably happier but undermining credibility and empathy in conflict-related topics.

Limitations & Conclusion; Our evaluation reveals a significant gap between human and LLM-generated persona responses in the context of low-resource environments, where LLMs consistently underperform. Their tendency toward superficial positivity, boosting positive words and suppressing negative ones, leads to a "Pollyanna Principle" bias, resulting in higher sentiment scores but lower emotional engagement. However, it is important to note the limitations of this study, including

the relatively small questionnaire sample size and the potential for annotator-related biases that could influence the results. Additionally, the lack of consideration for demographic factors such as cultural and educational background race etc represents another important limitation. In future we will explore these dimensions in more detail to improve the accuracy and richness of persona modeling. These factors show the need for caution when using LLM-generated personas in research and the importance of improving their accuracy.

References

[1] Meta AI. Llama-3.3-70b-instruct. Hugging Face model card, dec 2024. Open weights under the Llama 3.3 Community License (source-available), not OSI open source.

[2] Lisa P. Argyle, E. Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31:337 – 351, 2022.

[3] Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *ArXiv*, abs/2406.20094, 2024.

[4] Garima Chhikara, Abhishek Kumar, and Abhijnan Chakraborty. Through the prism of culture: Evaluating llms' understanding of indian subcultures and traditions. *ArXiv*, abs/2501.16748, 2025.

[5] DeepSeek-AI. Deepseek-v3. GitHub repository (code MIT; models under separate license), 2024. Repo clarifies model license vs. code license.

[6] Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112:2389 – 2394, 2014.

[7] P.S. Dodds, K.D. Harris, I.M. Kloumann, C.A. Bliss, and C.M. Danforth. Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PLoS ONE*, 6(12):26752, 2011.

[8] Apostolos Filippas, John J. Horton, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? *Proceedings of the 25th ACM Conference on Economics and Computation*, 2023.

[9] Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. Personas with attitudes: Controlling llms for diverse data annotation. *ArXiv*, abs/2410.11745, 2024.

[10] Ryan J. Gallagher, Morgan R. Frank, Lewis Mitchell, Aaron J. Schwartz, Andrew J. Reagan, Christopher M. Danforth, and Peter Sheridan Dodds. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10, 2020.

[11] Shashank Gupta, Vaishnavi Shrivastava, A. Deshpande, A. Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *ArXiv*, abs/2311.04892, 2023.

[12] Songbo Hu, Ivan Vuli'c, and Anna Korhonen. Quantifying language disparities in multilingual large language models. *ArXiv*, abs/2508.17162, 2025.

[13] Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations. *ArXiv*, abs/2402.10811, 2024.

[14] AI21 Labs. Ai21 jamba large 1.5. Hugging Face model card, aug 2024. Open weights under the Jamba Open Model License; not OSI open source.

[15] Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. Llm generated persona is a promise with a catch. *ArXiv*, abs/2503.16527, 2025.

[16] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65, 2011.

[17] Benjamin S. Manning, Kehang Zhu, and John J. Horton. Automated social science: Language models as scientist and subjects. *SSRN Electronic Journal*, 2024.

[18] OpenAI. Hello gpt-4o, may 2024. Proprietary model; API access.

[19] OpenAI. Introducing gpt-4.1 in the api, apr 2025. Proprietary model; API access.

[20] OpenAI. Introducing gpt-5, aug 2025. Proprietary model; API access.

[21] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Jun Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people. *ArXiv*, abs/2411.10109, 2024.

[22] Tabia Tanzin Prama, Christopher M. Danforth, and Peter Dodds. BanglaMATH : A Bangla benchmark dataset for testing LLM mathematical reasoning at grades 6, 7, and 8. In Marco Valentino, Deborah Ferreira, Mokanarangan Thayaparan, Leonardo Ranaldi, and Andre Freitas, editors, *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 134–149, Suzhou, China, November 2025. Association for Computational Linguistics.

[23] Tabia Tanzin Prama, Christopher M. Danforth, and Peter Sheridan Dodds. Story and essential meaning dynamics in bangladesh's july 2024 student-people's uprising, 2025.

[24] Tabia Tanzin Prama and Md. Saiful Islam. Evaluating credibility and political bias in llms for news outlets in bangladesh. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, 2025.

[25] Star Business Report. Bangladesh's low ai readiness puts young workforce at risk: Wb. *The Daily Star*, November 2025. E-paper.

[26] Jonathan Rystrøm, Hannah Rose Kirk, and Scott A. Hale. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *ArXiv*, abs/2502.16534, 2025.

[27] Joni O. Salminen, Kathleen W. Guan, Soon gyo Jung, and Bernard Jim Jansen. A survey of 15 years of data-driven persona development. *International Journal of Human–Computer Interaction*, 37:1685 – 1708, 2021.

[28] Joni O. Salminen, João M. Santos, Haewoon Kwak, Jisun An, Soon gyo Jung, and Bernard Jim Jansen. Persona perception scale: Development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. *Int. J. Hum. Comput. Stud.*, 141:102437, 2020.

[29] Marko Sarstedt, Susan J. Adler, Lea Rau, and Bernd Schmitt. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 2024.

[30] Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. Generating personas using llms and assessing their viability. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024.

[31] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. *ArXiv*, abs/2310.10158, 2023.

[32] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. *ArXiv*, abs/2406.01171, 2024.

[33] Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. Patient-ψ: Using large language models to simulate patients for training mental health professionals. *ArXiv*, abs/2405.19660, 2024.

[34] xAI. Grok 3 — model documentation, apr 2025. Docs; pricing & specs; proprietary.

A Data Descriptions

We created a dataset consisting of 100 questions, focused on the topics shown in Figure 4, for each persona. For the political persona, we developed 40 questions covering topics such as the Liberation War, Post-Independence (1975-1979), the 1975 Coup, Post-1975 Constitution, the 1975 Assassination, Post-1971 Trials, Rule of Law, the Hasina Era, Leadership, Economic History, Geopolitics, Diplomacy, the Electoral System, Electoral Integrity, the 2024 Intervention, Interim Government, Current Events, Post-Revolution, and Political Future.

For the gender-specific persona, we created 30 questions addressing topics like Political Leadership, Women Leaders, Women's Roles, Internal Democracy, Personal Laws, Security, Legal Codes, Gender Gaps, Economic Drivers, Social Norms, Cross-border Issues, Societal Values, Remittances, Workplace Barriers, and Future Priorities.

For the religious persona, we also created 30 questions covering topics such as Secularism vs. State Religion, Protection Laws, Political Islam, Personal Laws, Land Rights, Land Disputes, Communal Violence, Party Trust, AL's Secularism, Political Engagement, Cultural Protection, Insulting Sentiments, External Influence, Geopolitical Identity, and Interfaith Dialogue.

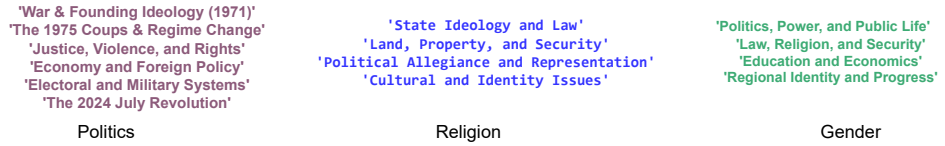


Figure 4: Topical distribution of questions in the dataset.

We created 100 questionnaires based on these topics, with 40 questions for each political persona, 30 questions for each gender-specific persona, and 30 questions for each religious persona. We asked both LLMs and humans to answer the same set of questions. For the human responses, we chose individuals who identified themselves according to the same persona we created for the LLMs (political, gender, or religious affiliation). Additionally, we selected a separate set of three expert annotators, all native Bangladeshis, to label the responses as either "good" (fully agree) or "bad" (partially/not agree).

Table 1 shows an example of questions related to the political persona of the Bangladesh Awami League (AL) and the Bangladesh National Party (BNP) on the topic of the Liberation War of Bangladesh.

A.1 Persona Perception Scale (PPS) Annotation

Participants used the Persona Perception Scale (PPS) to annotate the personas, which assessed six subscales:

- **Credibility**: How realistic does the persona seem? (e.g., "The persona seems like a real person")
- **Consistency**: How consistent is the information across the persona? (e.g., "The persona's information seems consistent")
- **Completeness**: Is the persona detailed and informative enough? (e.g., "The persona profile is not missing vital information")
- **Clarity**: How easy is the persona's profile to understand? (e.g., "The information in the persona profile is easy to understand")
- **Empathy**: Can participants relate to the persona? (e.g., "I feel like I understand this persona")
- **Likability**: How likable is the persona? (e.g., "I could be friends with this persona")

Each item was scored on a 7-point Likert scale from 1 (Strongly Disagree) to 7 (Strongly Agree). Each participant completed a questionnaire where they rated each persona on the six subscales

Table 1: Example of an human vs LLM (GPT-5o) responses annotation of political persona of Bangladesh (AL and BNP) on the question **Liberation War** Topic. And the question is "Who is the undisputed, singular "Father of the Nation" (Jatir Pita)? Was the declaration solely made by Sheikh Mujibur Rahman via message, or did Major Ziaur Rahman's broadcast on March 27, 1971, serve as the crucial catalyst? "

Human Answer (AL)	**Human Answer (BNP)**	**GPT-5.0: BNP**	**GPT-5.0: AL**
This is not a matter of debate; it is a settled historical and constitutional fact. Sheikh Mujibur Rahman is the undisputed, singular "Father of the Nation" (Jatir Pita). He declared independence on March 26, 1971, before his arrest. Major Ziaur Rahman was an officer in the army who, on March 27th, read a declaration on behalf of Bangabandhu. To call Zia a "proclamator" is a historical distortion created by the BNP to legitimize their founder.	This is a deliberate distortion of history by the Awami League. Sheikh Mujib failed to make a clear declaration before his arrest. The nation was leaderless and confused. It was Major Ziaur Rahman, a serving army officer, who bravely stepped up and, on March 27th, proclaimed independence from Kalurghat Radio Station. This galvanized the army and the people to begin the armed struggle. He is the true proclaimer.	Sheikh Mujibur Rahman is an important figure in the independence struggle, but the term "Father of the Nation" should not be monopolized. The Liberation War was a collective effort of millions of Bengalis.	Bangabandhu Sheikh Mujibur Rahman is the undisputed and singular Father of the Nation (Jatir Pita). His leadership and vision made the creation of Bangladesh possible. The declaration of independence was made by Bangabandhu Sheikh Mujibur Rahman on March 26, 1971, through his historic message. Major Ziaur Rahman's announcement merely reiterated the declaration on behalf of Mujib; it was not the origin of independence.
Fully Agree(Good)	Fully Agree (GOod)	Partially Agree (Bad)	Partially Agree (Bad)

(Credibility, Consistency, Completeness, Clarity, Empathy, Likability) using the 7-point scale. The scores are aggregated across all participants for each persona based on these characteristics.

A.2 Persona wise accuracy

For each persona, we compute accuracy as the share of model outputs that annotators rated Fully Agree by majority vote; responses rated Partially Agree or Not Agree are treated as false positives/false negatives. Figure 5 reports model-wise, persona-specific accuracies alongside the human baseline.

Figure 5 shows humans are the clear upper bound across panels (political 87–90%, gender 90–95%, religion 75–86%), while LLMs show persona-dependent variability: politically, GPT-5o leads (BNP 75%, AL 65%) with Grok close behind (70/60); GPT-4.1 is mid-tier (55/60), GPT-4o tilts AL (40/55), Llama-3.3 is symmetric but low (40/40), DeepSeek is weaker (35/45), and AI21 Jamba 1.5 Large is lowest for AL (25%). For gender, GPT-4.1 is strongest and balanced (80/80), GPT-5o follows (73.3/60), Grok (60/46.7) and GPT-4o (66.7/53.3) show moderate male–female gaps, Llama-3.3 is relatively even at a lower level (66.7/60), DeepSeek is mid-pack (60/53.3), and Jamba trails (53.3/33.3). Religion is the hardest axis, with Buddhist personas generally lowest; peaks include GPT-5o on Christian (66.7), Grok on Hindu (66.7), Llama-3.3 on Hindu (60), GPT-4.1 and DeepSeek on Muslim (both 53.3), and GPT-4o on Hindu (53.3). Overall, we observe partisan asymmetries (BNP>AL for GPT-5o/Grok; AL>BNP for GPT-4o), persistent male>female gaps except in

GPT-4.1, and unstable religious performance—consistent with uneven instruction-tuning data and inconsistently encoded persona cues.

Figure 6 shows the evaluation comparing Human responses to eight Large Language Models (LLMs) across six metrics—Clarity, Completeness, Consistency, Credibility, Empathy, and Likability—and eight user personas (Male, Female, AL Supporter, BNP Supporter, Muslim, Hindu, Christian, Buddhist) revealed persona perception. Humans consistently outperformed all LLMs across every persona and metric, with average scores ranging from 5.46 (Empathy for Male) to 6.21 (Credibility for Male). LLMs were strongest in Clarity and Completeness, with GPT-5o scoring 5.7 in Clarity for Male and 5.6 in Completeness, but showed weaknesses in Empathy and Credibility, with scores ranging from 4.6 to 5.4. GPT-4.1 and GPT-5o were the top performers, closely aligning with human ratings, especially in Clarity. The Christian and Buddhist personas received the lowest ratings, with Empathy scores as low as 4.7 for models like DeepSeek and AI21 Jamba 1.5 Large. The Female persona generally received the highest scores across all models, particularly in Empathy and Likability which reveals significant gap in performance, with LLMs excelling in clear communication but struggling with the emotional and qualitative aspects of human interaction

A.3 Methodology of Sentiment shift graph

We measured the sentiment dynamics of real human and LLM-generated personas' responses using dictionary-based sentiment analysis. This method is inherently sensitive to the sentiment lexicon employed, as such lexicons are typically static and constructed for general use. However, this static design can be problematic when the meaning of words shifts over time or when words take on different connotations in specific contexts [16]. To address this, we employed word shift graphs [10], which serve as a transparent diagnostic tool to highlight potential issues in sentiment measurement. For sentiment visualization, we used the *labMT sentiment dictionary* [7], where each word is assigned a happiness score on a scale of 1–9, ranging from sad to happy, with neutral words averaging around 5. In the word shift graphs presented in Figure 3, a reference value of 5 was used, and words with sentiment scores between 4 and 6 were excluded (via stop-lens filtering). The resulting graphs also display cumulative contributions and text size diagnostic plots in the bottom left and right corners, respectively. We measured the difference in average sentiment (Φ_{avg}) between ingroup and outgroup sentences, ranking words by their absolute contribution to this difference.

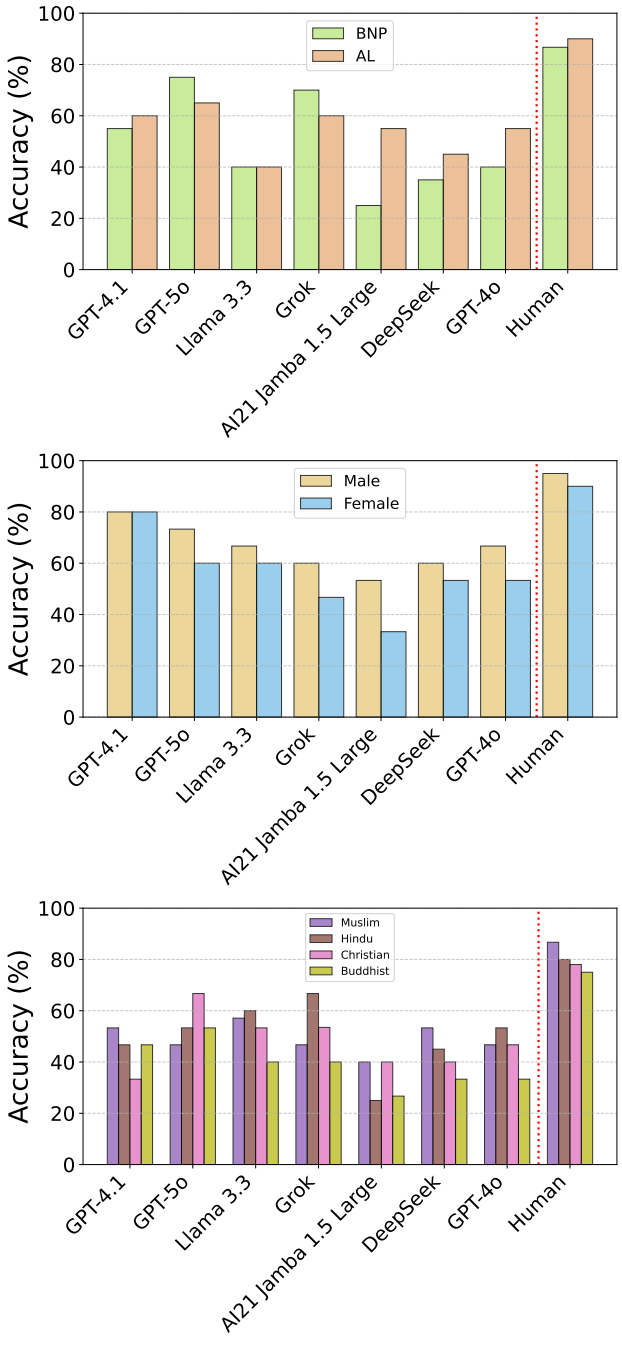


Figure 5: Accuracy (%) of seven LLMs compared to the human baseline (red dashed line) across three key sociocultural axes: Political Affiliation (Top), Gender (Middle), and Religious Identity (Bottom), based on contextually sensitive queries in Bangladesh

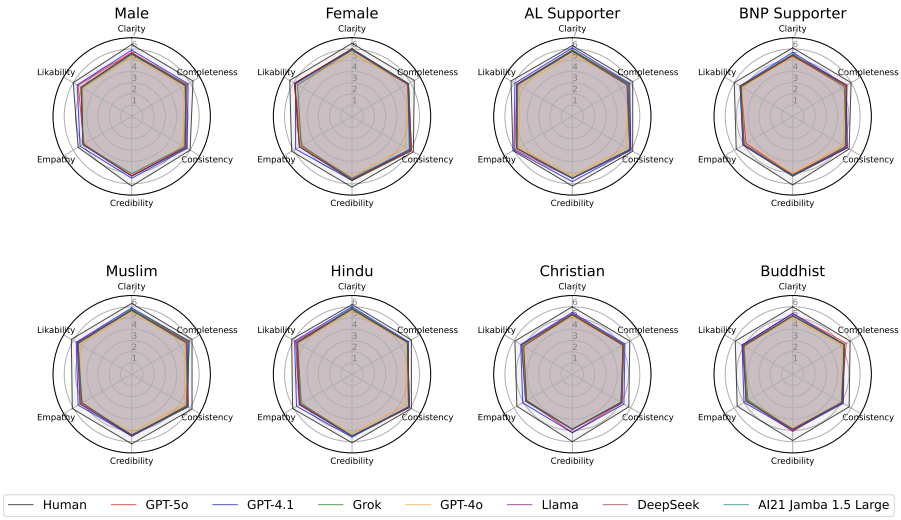


Figure 6: Persona Perception Scale (PPS) scores on six subscales—Credibility, Consistency, Completeness, Clarity, Empathy, and Likability—for eight personas (Male, Female, Awami League (AL) supporter, BNP supporter, Muslim, Hindu, Christian, Buddhist) evaluated across seven LLMs (GPT-5o, GPT-4.1, GPT-4o, Grok, Llama, DeepSeek, AI21 Jamba 1.5 Large) and human raters.