White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247902/

Version: Published Version

Proceedings Paper:
Bao, E., Perez, A., Parapar, J. et al. (2026) Eval4Sim: An evaluation framework for persona simulation. In: CIKM '26: Proceedings of the 35th ACM International Conference on Information and Knowledge Management. 35th ACM International Conference on Information and Knowledge Management (CIKM '26), 07-11 Nov 2026, Rome, Italy. ACM, pp. 8848-8853. ISBN: 9798400725395.

https://doi.org/10.1145/3799682.3840172

© 2026 Owner/Author. This work is licensed under a Creative Commons Attribution International 4.0 License. (https://creativecommons.org/licenses/by/4.0/)

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Eval4Sim: An Evaluation Framework for Persona Simulation

Eliseo Bao
eliseo.bao@udc.es
IRLab, CITIC, Universidade da Coruña
A Coruña, Spain

Anxo Perez
anxo.pvila@udc.es
IRLab, CITIC, Universidade da Coruña
A Coruña, Spain

Javier Parapar
javier.parapar@udc.es
IRLab, CITIC, Universidade da Coruña
A Coruña, Spain

Xi Wang
xi.wang@sheffield.ac.uk
University of Sheffield
Sheffield, United Kingdom

Abstract

Large Language Model personas, explicit profiles specifying a user's attributes, preferences, and behavioural tendencies, are increasingly used to simulate human conversations for user modelling, social reasoning, and behavioural analysis. Evaluating whether such simulations faithfully reflect human conversational behaviour is critical, yet current practice often relies on LLM-as-a-judge approaches that provide limited grounding in observable behaviour and produce opaque scalar scores. We present Eval4Sim, an evaluation framework that measures alignment between simulated and human conversations across three dimensions: *adherence*, whether persona traits are recoverable from dialogue via dense retrieval; *consistency*, whether a persona maintains a distinguishable stylistic identity via authorship verification; and *naturalness*, whether conversations exhibit human-like turn-to-turn flow via dialogue NLI. Unlike optimization-oriented metrics, each dimension takes a human corpus as a reference baseline and penalizes deviations in both directions, distinguishing insufficient persona encoding from over-optimized, unnatural behaviour. The framework is corpus-agnostic: any persona-annotated conversational dataset can serve as the reference. Evaluated over ten simulation corpora, Eval4Sim surfaces systematic trade-offs invisible to single-score methods.

CCS Concepts

• **Computing methodologies** → **Simulation evaluation**.

Keywords

Evaluation, Persona Simulation, Conversational AI, LLMs

ACM Reference Format:
Eliseo Bao, Anxo Perez, Javier Parapar, and Xi Wang. 2026. Eval4Sim: An Evaluation Framework for Persona Simulation. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 6 pages.

1 Introduction and Motivation

Persona grounding conditions a conversational agent on an explicit user profile [11, 13] specifying attributes, preferences, and behavioural tendencies. LLM personas are now widely used to simulate human conversations for user modelling, social reasoning, and behavioural analysis [5, 22], and ensuring that these simulations faithfully reflect human conversational behaviour has become a pressing evaluation challenge. A persona simulator must maintain a stable and distinguishable identity over time, expressing persona-specific cues consistently across dialogues [20], yet phenomena such as persona drift [15], inconsistency [14], and exaggerated trait expression [4] limit their realism and validity.

Existing evaluation approaches fall short in different ways. Traditional metrics such as BLEU [17], or perplexity [9] capture surface overlap rather than persona consistency [16, 20]. Specialized frameworks introduce persona-specific metrics [21], consistency labels [23], and human evaluation protocols [19], but isolate single attributes without considering trade-offs across dimensions. Recent benchmarks such as *PersoBench* [1] and *PersonaGym* [18] advance the state of the art, yet both score individual turns against gold references or LLM-as-a-judge rubrics [7, 25], which are sensitive to prompt design and can exhibit systematic scoring biases [2, 8]. Crucially, none of these approaches grounds evaluation in empirical human behaviour: they assess quality relative to a gold response or a model opinion, not relative to how humans actually converse.

We address this gap with Eval4Sim[1], a corpus-agnostic framework for evaluating persona-grounded conversation simulation against a human reference corpus (Figure 1). The framework jointly evaluates three dimensions: *adherence*, whether persona traits are recoverable from dialogue; *consistency*, whether speakers maintain a stable stylistic identity; and *naturalness*, whether conversations follow human-like interaction patterns. Rather than rewarding maximal scores in isolation, Eval4Sim measures alignment to human behaviour and penalizes deviations in both directions. This is important because persona simulation involves trade-offs: making persona traits too recoverable may lead to unnatural self-disclosure, while optimizing fluent dialogue may weaken stylistic identity. Eval4Sim provides reproducible evaluations over ten simulation corpora, enabling researchers to diagnose where simulators diverge from human conversational patterns and to plug in new reference corpora or simulated datasets.

2 The Eval4Sim Framework

Eval4Sim operates on two inputs: (i) a *reference corpus* of human-to-human persona-grounded conversations that establishes the behavioural baseline, and (ii) one or more *simulation corpora* to

[1] Code and documentation: https://github.com/IRLab-UDC/eval4sim.

This work is licensed under a Creative Commons Attribution 4.0 International License.
CIKM '26, Rome, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11

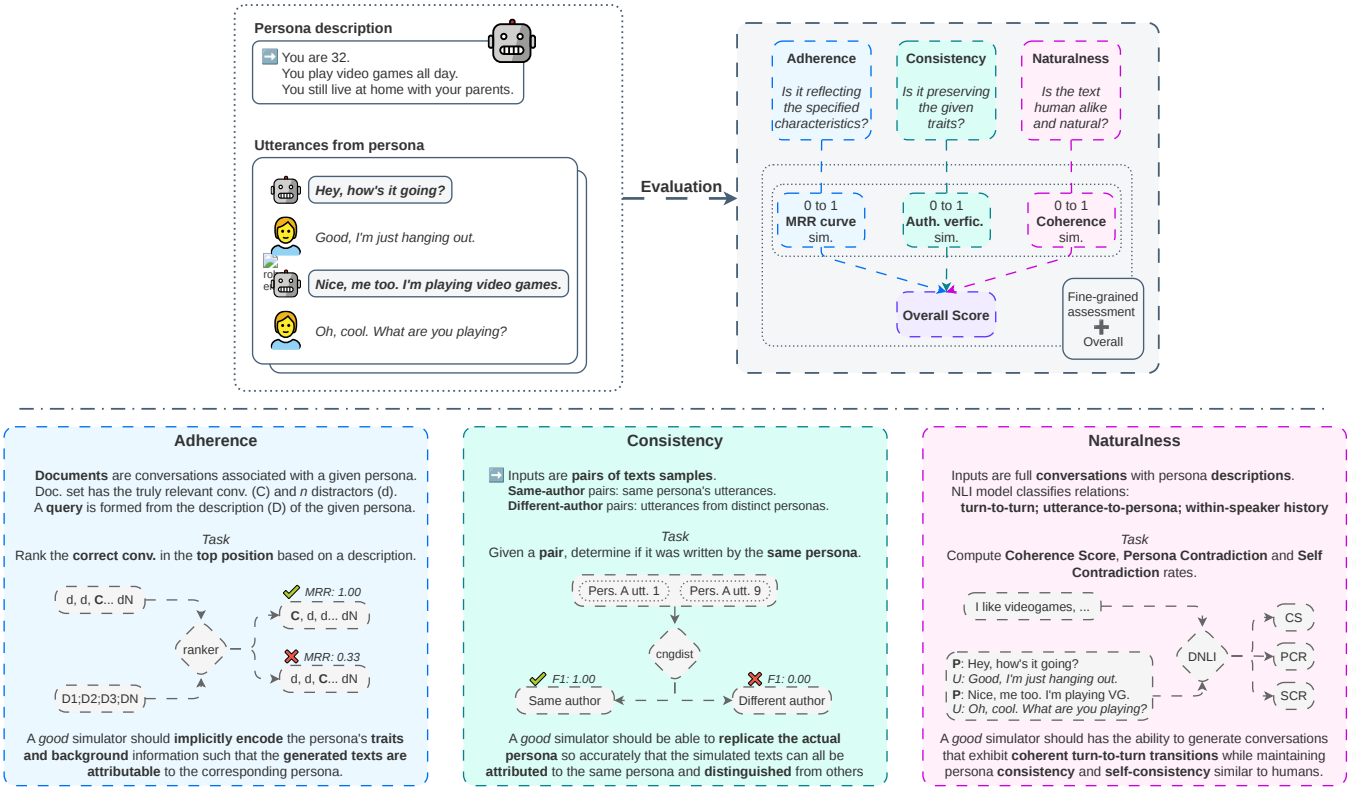


Figure 1: General overview of the Eval4Sim evaluation framework for persona simulation.

evaluate. Both contain multi-turn dialogues where each speaker is associated with a persona description. The framework is corpus-agnostic: any dataset with speaker-level persona annotations can serve as the reference. An expected input sample is shown below:

```
(Alex 🚹) User 1 persona traits:
    I like to dance at the club.
    I run a dog obedience school.
(Emma 🚺) User 2 persona traits:
    I love to meet new people.
    I have a turtle named Timothy.

Alex (🚹) : Hi, I'm Alex. What's your name?
Emma (🚺) : Hi, I'm Emma. Nice to meet you.
Alex (🚹) : What are you interested in?
Emma (🚺): I enjoy meeting people and playing frisbee.
```

Each dimension produces a human similarity score in $[0, 1]$, where 1.0 indicates perfect alignment with the reference. The aggregate E4S score is the unweighted mean of the three, treating the dimensions as diagnostics of equal standing; practitioners with domain-specific priorities may reweight accordingly.

1. Adherence via Dense Retrieval. We define *adherence* as the ability of a simulator to implicitly expresses a persona's traits such that the generated conversation remains attributable to that persona. Formally, let $C = \{c_1, \ldots, c_n\}$ be a set of conversations and $\mathcal{P}$ the set of personas, where each $p \in \mathcal{P}$ is associated with a set of relevant conversations $C_p \subseteq C$. Given p as a query, the retrieval objective is to rank conversations in C_p above those in $C \setminus C_p$. We use ColBERT [12] to retrieve conversations from persona descriptions, representing each conversation by the utterances of the evaluated speaker, since the query persona refers to that speaker. We evaluate MRR [6] over increasingly difficult pools, each containing one relevant conversation and k distractors, yielding an MRR degradation curve per corpus.

Importantly, adherence should not be maximized blindly: a simulator may over-expose persona traits (e.g., "As a teacher, I..."), making retrieval artificially easy but dialogue unnatural. We therefore compare each degradation curve to the human reference and penalizes deviations in either direction. Curves below the reference indicate under-encoding of persona cues, while curves above it indicate over-explicit persona signaling. Standard area-based scores are unsuitable because they reward systematically higher retrieval performance. Instead, we compute a scale-sensitive similarity between the simulation curve and the human reference curve:

$$\mathrm{Adh}_{\mathrm{sim}} = \frac{\sum_i w_i \cdot a_i \cdot b_i}{\max\left(\sum_i w_i \cdot a_i^2,\ \sum_i w_i \cdot b_i^2\right)} \tag{1}$$

Here, a_i and b_i are the MRR values at pool size p_i for the reference and simulation, respectively, and δ_i are span weights accounting for unequal spacing between pool sizes. The score is 1 when both curves match exactly and decreases when the simulated curve falls either below or above the human reference.

2. Consistency via Authorship Verification. Whereas adherence measures whether persona *content* is recoverable from the conversation, consistency measures whether each speaker maintains a stable *style* distinguishable from other speakers. We define *consistency* as the ability of a simulator to maintain a stable and distinguishable speaker identity across conversations: utterances from the same persona should exhibit coherent stylistic regularities while remaining separable from other personas. We implement this as authorship verification: given a pair of text samples (t_1, t_2), the task is to predict whether both were produced by the same persona or by different personas.

We use the character 4-gram TF–IDF baseline from PAN 2023 [3], which represents texts through TF–IDF weighted character n-grams, computes cosine similarities between pairs, and calibrates them into verification scores. This yields a *separability* signal: if a simulator realizes personas consistently, same-persona pairs should be easier to distinguish from different-persona pairs. However, as with adherence, the goal is not to maximize separability. Excessive stylistic uniformity may make personas artificially easy to identify while producing repetitive, less human-like dialogue. We aggregate the five official PAN metrics (F1, AUC, Brier, c@1, $F^u_{0.5}$) [2] into a single Consistency score. Human similarity penalizes deviation in either direction:

$$\text{Con}_{\text{sim}} = 1 - \frac{|\text{Consistency}_{\text{sim}} - \text{Consistency}_{\text{ref}}|}{\text{Consistency}_{\text{ref}}} \tag{2}$$

3. Naturalness via NLI Distributions. We define *naturalness* as the ability of a simulator to generate conversations with human-like turn-to-turn dynamics while remaining compatible with the speaker's persona and internally self-consistent. We implement this dimension through dialogue-focused Natural Language Inference (NLI). Given a conversation with utterances $\{u_t\}_{t=1}^{T}$ and speaker-specific persona descriptions $\mathcal{P}_s = \{p_{s,1}, \ldots, p_{s,n}\}$, we apply a DeBERTa-based NLI model fine-tuned on conversational data [23][3] to two types of pairs: (i) consecutive turn pairs (u_{t-1}, u_t), which capture local dialogue flow, and (ii) persona–utterance pairs $(p_{s,i}, u_t)$, which capture compatibility between a speaker's utterances and their assigned persona. From these predictions, we derive three diagnostic metrics:

- **Coherence Score (CS):** the mean turn-to-turn coherence score over consecutive utterance pairs, assigning 1.0 to entailment, 0.5 to neutral, and 0.0 to contradiction.
- **Persona Contradiction Rate (PCR):** the fraction of persona–utterance pairs predicted as contradiction with confidence at least 0.7.[4]
- **Self-Contradiction Rate (SCR):** the fraction of within-speaker utterance pairs, computed within a 5-turn history window, predicted as contradiction with confidence at least 0.7.

To obtain a compact Naturalness score, we aggregate CS, (1 − PCR), and (1 − SCR). We estimate the weights of each component from the human reference corpus: for each component, we compute its standard deviation across individual PersonaChat conversations and normalize the three standard deviations to sum to one. The intuition is that components with greater variability in natural dialogue provide more discriminative information, whereas near-constant components should contribute less. Since std(1 − PCR) = std(PCR) and std(1 − SCR) = std(SCR), the same weights apply to the non-contradiction terms. This yields weights of 0.565 for CS, 0.126 for (1 − PCR), and 0.309 for (1 − SCR):

$$\text{Naturalness} = 0.565 \times \text{CS} + 0.126 \times (1 - \text{PCR}) + 0.309 \times (1 - \text{SCR}). \tag{3}$$

Finally, as in the other dimensions, we convert this raw Naturalness score into a human-similarity score by comparing it with the reference corpus:

$$\text{Nat}_{\text{sim}} = 1 - \frac{|\text{Naturalness}_{\text{sim}} - \text{Naturalness}_{\text{ref}}|}{\text{Naturalness}_{\text{ref}}} \tag{4}$$

3 Experiments and Results

Experimental Setup. We instantiate Eval4Sim using Persona Chat [24] as the reference corpus, a widely used collection of real persona-grounded human-to-human dialogues. We evaluate ten simulation corpora spanning two origins. The first two are published datasets: Synthetic-Persona-Chat (SPC) [10], which reuses PersonaChat persona descriptions with synthetic dialogues, and SPC-New [10], which uses fully synthetic personas and conversations. Both were produced with a Generator–Critic framework, where a generator creates candidate persona-grounded dialogues and a critic filters or selects them according to quality and persona-faithfulness criteria. The remaining eight corpora were generated in the experiments of this work by prompting open-source LLMs on the same PersonaChat personas: Qwen3 (1.7B, 4B, 14B, 30B) and Gemma 3 (1B, 4B, 12B, 27B). For each LLM, five independent conversation sets were generated under different random seeds. All reported scores are averaged at the corpus level across these five runs. Table 2 summarises corpus statistics, while Table 1 reports the scores for all dimensions and the Eval4Sim overall score.

Adherence. Figure 2 shows MRR degradation as the distractor pool grows. PersonaChat drops from near-perfect retrieval at small pools to roughly 0.3 at 1000 distractors. Several LLM-generated corpora closely track this human pattern, suggesting that their persona cues are recoverable without becoming unnaturally explicit. The Generator–Critic datasets deviate more clearly: SPC remains consistently below the human curve, indicating weaker persona attribution, while SPC-New shows a sharp early drop before partially converging. Overall, direct generation with modern LLMs yields more human-aligned persona encoding than the older Generator–Critic corpora. Gemma 3 12B is closest to the human baseline (−1.70%), followed by Qwen3 30B (−2.60%). SPC and SPC-New are least aligned (−18.90% and −16.30%).

Consistency. The datasets closest to PersonaChat under the stylometry-based verifier are Qwen3 14B (−2.40%), SPC (−4.50%), and SPC-New (−4.70%). This contrasts with adherence, where larger modern LLMs perform best, showing that model capacity alone does not guarantee human-like stylistic stability. Notably, Qwen3 1.7B attains a higher raw consistency score than PersonaChat but lower similarity (0.919), illustrating the bidirectional nature of the

[2] We refer the reader to the official PAN overview [3] for full metric definitions.
[3] `zayn1111/deberta-v3-dnli`; 95.36% accuracy on dialogue NLI.
[4] We use a confidence threshold to avoid counting uncertain NLI predictions as contradictions, since explicit persona contradictions are rare in well-formed human dialogues.

Table 1: Overall E4S evaluation. Similarity scores measure alignment to PERSONACHAT (1.000). E4S is the unweighted mean of the three similarity scores. Bold: 1st, underline: 2nd, *italic*: 3rd per column.

	Adherence			Consistency			Naturalness			E4S		
Dataset	Score	Off.	Pos.	Score	Off.	Pos.	Score	Off.	Pos.	Score	Off.	Pos.
PersonaChat	*1.000*	—	—	*1.000*	—	—	*1.000*	—	—	*1.000*	—	—
Qwen3 30B	0.974	-2.60%	2	0.917	-8.30%	6	**0.962**	-3.80%	1	**0.951**	-4.90%	1
Gemma 3 12B	**0.983**	-1.70%	1	0.877	-12.30%	8	*0.959*	-4.10%	3	0.940	-6.00%	2
Gemma 3 4B	0.948	-5.20%	4	0.901	-9.90%	7	0.962	-3.80%	2	*0.937*	-6.30%	3
Qwen3 4B	0.912	-8.80%	7	*0.946*	-5.40%	4	0.951	-4.90%	5	0.936	-6.40%	4
Qwen3 14B	0.919	-8.10%	5	**0.976**	-2.40%	1	0.892	-10.80%	9	0.929	-7.10%	5
Gemma 3 27B	*0.970*	-3.00%	3	0.874	-12.60%	9	0.919	-8.10%	7	0.921	-7.90%	6
Gemma 3 1B	0.918	-8.20%	6	0.867	-13.30%	10	0.953	-4.70%	4	0.913	-8.70%	7
Qwen3 1.7B	0.840	-16.00%	8	0.919	-8.10%	5	0.932	-6.80%	6	0.897	-10.30%	8
SPC	0.811	-18.90%	10	0.955	-4.50%	2	0.904	-9.60%	8	0.890	-11.00%	9
SPC-New	0.837	-16.30%	9	*0.953*	-4.70%	3	0.874	-12.60%	10	0.888	-11.20%	10

Table 2: Dataset statistics for PERSONACHAT and all simulations. Desc.: avg. items and words per description. Convs.: avg. messages and words per conversation.

Dataset	Samples	Desc.		Convs.	
		$\overline{\lvert items \rvert}$	$\overline{len.}$	$\overline{\lvert mess. \rvert}$	$\overline{len.}$
PERSONACHAT [24]	1,000	4.50	33.87	15.60	11.71
SPC [10]	968	4.50	28.11	27.55	11.05
SPC-NEW [10]	11,001	5.00	56.77	17.44	13.32
QWEN3 1.7B	645.4	4.50	28.13	29.83	12.29
QWEN3 4B	954.0	4.50	28.13	35.22	20.05
QWEN3 14B	946.6	4.50	28.11	31.27	22.34
QWEN3 30B	967.0	4.50	28.11	32.34	19.35
GEMMA 3 1B	743.2	4.50	28.02	25.47	20.71
GEMMA 3 4B	872.8	4.50	27.34	31.72	16.63
GEMMA 3 12B	964.0	4.50	28.11	32.37	18.94
GEMMA 3 27B	942.6	4.50	28.07	31.92	22.18

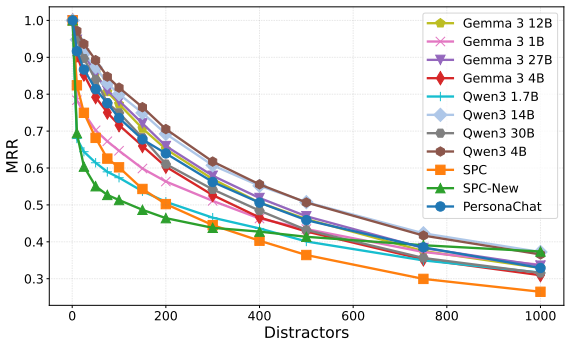


Figure 2: MRR degradation across candidate pool sizes for PERSONACHAT and ten simulated datasets.

metric: the goal is not maximal verifier performance, but matching the human level of stylistic separability.

Naturalness. QWEN3 30B and GEMMA 3 4B lead in naturalness (both −3.80%), with the top five systems within 1.1% of the human baseline. The clearest trade-off is QWEN3 14B: it ranks first in consistency but ninth in naturalness (−10.80%), suggesting that highly distinguishable stylistic identity may come at the cost of human-like dialogue flow. The Generator–Critic datasets rank near the bottom: SPC-NEW is least aligned (−12.60%), consistent with overly regular conversational flow relative to the human baseline. Since the Naturalness weights are estimated from variability in PERSONACHAT, these results should be interpreted as alignment to the baseline rather than absolute conversational quality.

Overall: E4S. Table 1 confirms that no simulator matches human behaviour uniformly across dimensions. QWEN3 30B ranks first overall (0.951), combining strong adherence (rank 2) with the best naturalness alignment (rank 1), at the cost of weaker consistency (rank 6). The Generator–Critic datasets show the opposite profile: competitive consistency but the weakest adherence and poor naturalness, resulting in the lowest overall scores. QWEN3 14B achieves the best consistency yet ranks ninth on naturalness, while GEMMA 3 4B outperforms GEMMA 3 27B, confirming that scale alone does not determine alignment. These trade-offs reinforce the need for multi-dimensional evaluation: optimizing one dimension can move a simulator away from the human baseline on others.

4 Resource Description and Conclusions

EVAL4SIM is publicly released at https://github.com/IRLab-UDC/eval4sim. As a corpus-agnostic resource, users can replace PERSONACHAT with any persona-annotated conversational corpus and evaluate new simulators against the corresponding human reference. Experiments over ten simulation corpora illustrate its diagnostic value. First, LLM-generated corpora achieve stronger adherence alignment than prior Generator–Critic datasets, while consistency follows a different ordering: larger models do not necessarily yield better stylistic alignment. Second, naturalness results show that several simulators deviate from the human reference through overly regular conversational flow, meaning high apparent coherence does not always imply human-like dialogue. Third, no simulator matches human behaviour uniformly across dimensions: QWEN3 30B achieves the best overall balance, but still trades off consistency against adherence and naturalness. These findings underscore the need for multi-dimensional evaluation, since optimizing a single aspect of persona simulation can move a system away from the human baseline on others.

Acknowledgments

The first author acknowledges the support of the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia (grant ED481A-2024-079). All authors affiliated with IRLab and CITIC acknowledge funding from the Ministry of Science, Innovation and Universities of the Government of Spain (projects PID2022-137061OB-C21, PID2025-167749OB-C22), as well as from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia (grant GRC ED431C 2025/49). CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

GenAI Usage Disclosure

During the preparation of this manuscript, generative AI tools were employed solely for light editing purposes, including proofreading, grammar correction, vocabulary improvement, and overall language polishing. All substantive ideas, analyses, experiments, and written content were created by the co-authors without direct text generation from any AI model.

References

[1] Saleh Afzoon, Zahra Jamali, Usman Naseem, and Amin Beheshti. 2026. PersoBench: Benchmarking Personalized Response Generation in Large Language Models. *IEEE Transactions on Artificial Intelligence* (2026), 1–10. doi:10.1109/TAI.2026.3689823

[2] Negar Arabzadeh and Charles L.A. Clarke. 2025. A Human-AI Comparative Analysis of Prompt Sensitivity in LLM-Based Relevance Judgment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 2784–2788. doi:10.1145/3726302.3730159

[3] Janek Bevendorff, Mara Chinea-Ríos, Marc Franco-Salvador, Annina Heini, Erik Körner, Krzysztof Kredens, Maximilian Mayerl, Piotr Pęzik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2023. Overview of PAN 2023: Authorship Verification, Multi-author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023*. Springer, Dublin, Ireland, 459–481. doi:10.1007/978-3-031-28241-6_60

[4] Pranav Bhandari, Nicolas Fay, Michael J Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025. Can LLM Agents Maintain a Persona in Discourse?. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 29213–29229. doi:10.18653/v1/2025.emnlp-main.1487

[5] Young Min Cho, Sunny Rai, Lyle Ungar, João Sedoc, and Sharath Guntuku. 2023. An Integrative Survey on Mental Health Conversational Agents to Bridge Computer Science and Medical Perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11346–11369. doi:10.18653/v1/2023.emnlp-main.698

[6] Nick Craswell. 2009. *Mean Reciprocal Rank*. Springer US, Boston, MA, 1703–1703. doi:10.1007/978-0-387-39940-9_488

[7] Michael Desmond, Zahra Ashktorab, Werner Geyer, Elizabeth M. Daly, Martín Santillán Cooper, Qian Pan, Rahul Nair, Nico Wagner, and Tejaswini Pedapati. 2025. EvalAssist: LLM-as-a-judge simplified. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 3453, 3 pages. doi:10.1609/aaai.v39i28.35351

[8] Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)* (Padua, Italy) *(ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 218–229. doi:10.1145/3731120.3744588

[9] Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. 2025. What is Wrong with Perplexity for Long-context Language Modeling?. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=fL4qWkSmtM

[10] Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful Persona-based Conversational Dataset Generation with Large Language Models. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, Elnaz Nouri, Abhinav Rastogi, Georgios Spithourakis, Bing Liu, Yun-Nung Chen, Yu Li, Alon Albalak, Hiromi Wakaki, and Alexandros Papangelis (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 114–139. https://aclanthology.org/2024.nlp4convai-1.8/

[11] Carl Jung. 2016. *Psychological Types* (1 ed.). Routledge, London. doi:10.4324/9781315512334

[12] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, Virtual Event, China, 39–48. doi:10.1145/3397271.3401075

[13] Mark R. Leary and Ashley Batts Allen. 2011. Personality and Persona: Personality Processes in Self-Presentation. *Journal of Personality* 79, 6 (2011), 1191–1218. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.2010.00704.x doi:10.1111/j.1467-6494.2010.00704.x

[14] Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM Generated Persona is a Promise with a Catch. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*. https://openreview.net/forum?id=qh9eGtMG4H

[15] Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and Controlling Persona Drift in Language Model Dialogs. *arXiv preprint arXiv:2402.10962* (2024).

[16] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2122–2132. doi:10.18653/v1/D16-1230

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. doi:10.3115/1073083.1073135

[18] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating Persona Agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 6999–7022. doi:10.18653/v1/2025.findings-emnlp.368

[19] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1702–1723. doi:10.18653/v1/N19-1170

[20] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 16612–16631. doi:10.18653/v1/2024.findings-emnlp.969

[21] Shiki Uehara, Shota Takahashi, Tatsuya Yamazaki, and Koichiro Yoshino. 2021. Fundamental Exploration of Evaluation Metrics for Persona Characteristics of Text Utterances. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Singapore and Online, 189–194. doi:10.18653/v1/2021.sigdial-1.19

[22] Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025. User Behavior Simulation with Large Language Model-based Agents. *ACM Trans. Inf. Syst.* 43, 2, Article 55 (Jan. 2025), 37 pages. doi:10.1145/3708985

[23] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3731–3741. doi:10.18653/v1/P19-1363

[24] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, 2204–2213. doi:10.18653/v1/P18-1205

[25] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2020, 29 pages.