A Participatory Governance Framework for Culturally Adaptive LLM Personalization: Advancing Digital Humanism Through Community-Driven Safeguards*

Carine P. Mukamakuza[1,2,*,†], Ronald Kato[4,†]

[1]Carnegie Mellon University Africa
[4]Mbarara University of Science and Technology

Abstract

Large language model (LLM) personalization has enhanced recommendation systems, education and digital communication through user-tailored outputs. However, these systems increasingly amplify socio-cultural homogenization, misinformation, labor inequalities and the marginalization of local knowledge. Existing technical approaches, such as prompt engineering, bias mitigation and fine-tuning, remain insufficient as they optimize for individuals rather than communities. This paper proposes the Participatory Cultural Alignment Network (PCAN), a decentralized governance framework that embeds Digital Humanism principles into culturally adaptive personalization. PCAN integrates federated learning, community-driven value elicitation, multi-stakeholder governance and a novel Digital Humanism Score (DHS) to evaluate diversity, inclusivity and social impact. The rise of agentic AI further intensifies these challenges by enabling autonomous, multi-step decision-making with limited human oversight. As such systems become more influential, governance frameworks must ensure alignment with community values and societal well-being. Through simulated scenarios in education, democracy and labor markets, PCAN demonstrates its ability to reduce cultural bias, enhance trust and preserve local knowledge. The framework provides a practical, measurable and scalable approach to responsible LLM personalization and future AI governance.

Keywords

Digital Humanism, Large Language Models, Personalization, Federated Learning, Cultural Alignment, Community Governance, Inclusivity, Ethical AI

1. Introduction

The emergence of large language models has transformed how users interact with digital platforms [1, 2]. Modern LLMs personalize outputs by adapting recommendations, educational materials, advertising and online interactions to user preferences and behavioral data [3]. Although personalization improves efficiency and engagement, it also introduces significant risks [4]. AI systems trained on globally dominant datasets frequently prioritize Western norms, English-language assumptions and mainstream values. Consequently, local traditions, minority perspectives and culturally specific forms of knowledge may be misrepresented [5].

Digital Humanism argues that technology should serve humanity rather than merely optimize commercial efficiency [6, 7]. Within this perspective, personalization must move beyond individual preference prediction and instead respect collective values, cultural diversity and democratic participation. Existing regulatory mechanisms, such as general AI governance policies, provide broad safeguards but rarely address the cultural dimensions of personalization.

This paper introduces the Participatory Cultural Alignment Network (PCAN), a framework designed to align personalized LLM systems with local communities and human-centered values. PCAN is guided

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden
*Corresponding author.
†These authors contributed equally.
cmukamak@andrew.cmu.edu (C. P. Mukamakuza); ronaldkato19@gmail.com (R. Kato)
https://www.africa.engineering.cmu.edu/about/contact/directory/bios/mukamakuza-carine.html (C. P. Mukamakuza); https://www.linkedin.com/in/ronald-kato/ (R. Kato)
0009-0007-0812-4955 (R. Kato)

by four principles:

1. Community ownership of cultural knowledge.
2. Transparent and measurable evaluation of socio-cultural impact.
3. Inclusive governance through multi-stakeholder participation.
4. Protection of local identities within global AI ecosystems.

The framework is especially relevant in regions where global AI systems may overlook local languages, customs and social realities, including African, indigenous and multilingual contexts.

2. Related Work

Research on large language model personalization has largely concentrated on improving technical performance and response relevance [1, 8]. The most common approaches include prompt engineering, fine-tuning, reinforcement learning from human feedback (RLHF), cultural prompting, persona adaptation and Retrieval-Augmented Generation (RAG) [9]. Prompt engineering enables rapid adaptation to cultural contexts at low cost, while fine-tuning allows models to become more locally relevant by retraining on user-specific data [10]. RLHF improves alignment with user preferences and RAG strengthens outputs by incorporating external local knowledge during inference.

Although these approaches enhance personalization, they present important limitations [11, 12]. Prompt engineering often produces temporary and inconsistent results, while fine-tuning requires centralized datasets that may not adequately represent local communities [13]. RLHF can unintentionally reinforce majority opinions and dominant social norms, thereby excluding minority perspectives [14]. RAG injects relevant knowledge but does not guarantee cultural fairness and meaningful participation by local communities [15]. Existing governance policies provide broad legal and ethical safeguards, yet they rarely include mechanisms for direct community involvement in defining acceptable outputs [16, 17].

The literature also highlights broader societal risks created by highly personalized systems [18]. Political recommendation algorithms may deepen polarization by creating echo chambers that repeatedly expose users to similar viewpoints [19]. In labor markets, personalization may encourage skill-biased automation and widen inequalities by favoring dominant economies over local employment realities [19]. In education, many systems prioritize Western curricula and overlook local history, indigenous knowledge and multilingual learning practices [20]. And as thus, indigenous and minority communities are frequently underrepresented in AI-generated content and recommendations.

Despite growing concern about these issues, current research has not yet produced a unified framework that combines cultural adaptation, social impact assessment and decentralized governance [21]. Most existing studies address only one aspect of the problem, such as cultural relevance and fairness, without integrating them into a single personalization pipeline [22, 23]. This gap motivates the development of the Participatory Cultural Alignment Network, which introduces community data contribution, Digital Humanism Score evaluation, multi-stakeholder governance and adaptive personalization within one coherent framework.

3. Methodology

3.1. The Participatory Cultural Alignment Network Workflow

The workflow of the Participatory Cultural Alignment Network is illustrated in Figure 1. The framework integrates four main stages: community data contribution, Digital Humanism Score evaluation, multi-stakeholder governance and adaptive personalization. Community knowledge is collected through federated learning and anonymized local data sources, then evaluated according to diversity, inclusivity, value alignment and social impact. A governance board subsequently reviews these outcomes and establishes policies before the system produces culturally adaptive recommendations and content.

This study adopts a system-oriented approach to operationalize the Participatory Cultural Alignment Network (PCAN) as a practical and reproducible framework for culturally adaptive LLM personalization as illustrated in Figure 1. Rather than presenting PCAN as a purely conceptual pipeline, the methodology formalizes it as a modular, multi-layered architecture that integrates machine learning, community governance and continuous evaluation. The framework is designed to balance global model efficiency with local cultural specificity by combining shared foundational models with decentralized adaptation mechanisms.

The methodological design follows a pipeline structure in which data flows through four interconnected stages: model generation, contextual adaptation, governance filtering and socio-cultural evaluation. Each stage is explicitly defined with clear inputs, outputs and update mechanisms to enhance reproducibility. In addition, the system supports iterative feedback loops, enabling continuous refinement based on community input and measured outcomes. This approach allows PCAN to move beyond static personalization toward a dynamic system that evolves in response to changing cultural contexts and societal needs.

3.2. System Architecture and Design Assumptions

To address limitations identified in earlier versions, PCAN is implemented as a modular LLM-based personalization system composed of four configurable layers: the Base Model Layer, Adaptation Layer, Governance Layer and Evaluation Layer. This layered design separates core language capabilities from cultural adaptation and policy enforcement, thereby enabling flexibility, scalability and clear responsibility boundaries across system components.

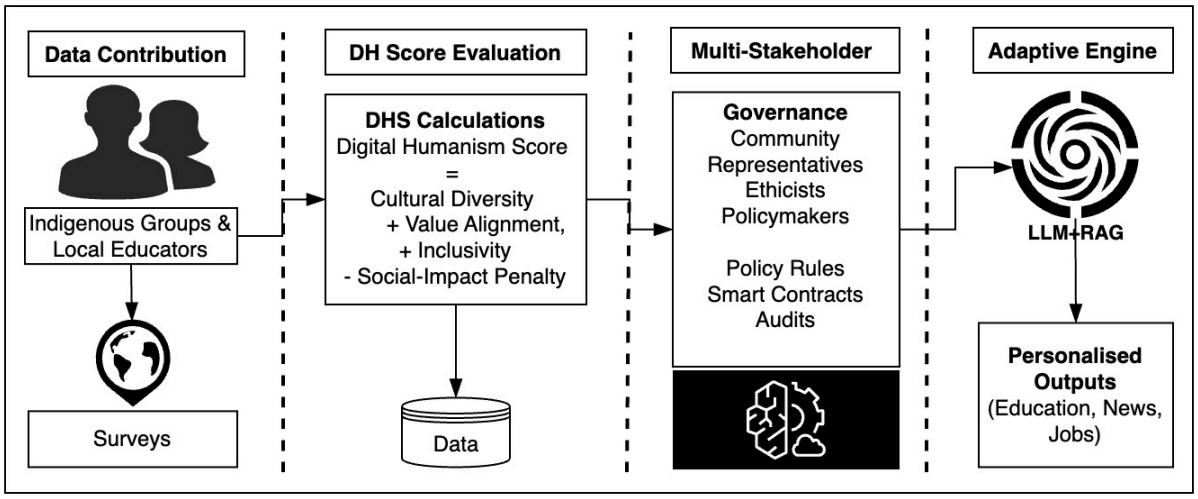

Figure 1: The Participatory Cultural Alignment Network Workflow

3.2.1. Base Model Layer

This layer consists of a shared pretrained LLM (GPT-style transformer) that provides general linguistic and reasoning capabilities. It is consistent across all users, ensuring efficiency and a common foundation for personalization.

3.2.2. Adaptation Layer (Task + Community Specific)

This layer enables personalization through task-specific prompts, Retrieval-Augmented Generation (RAG) with community data and lightweight tuning methods (LoRA/PEFT). It allows dynamic cultural adaptation without retraining the base model.

3.2.3. Governance Layer

This layer enforces alignment with community values by defining output constraints, applying policies and enforcing DHS-based thresholds. It enables multi-stakeholder control through configurable rules.

3.2.4. Evaluation Layer (DHS Engine)

The DHS engine evaluates outputs before deployment, measuring diversity, alignment, inclusivity and harm. Outputs below threshold are rejected or refined, ensuring responsible and culturally appropriate responses.

3.3. Quantitative Simulations

The framework is evaluated across three domains: education, democracy and labor. Education assesses cultural adaptation in learning content, democracy evaluates reduction of polarization and diversity of viewpoints and labor examines fairness in reskilling recommendations.

Table 1
Evaluation Metrics Used in the Simulations

Domain	Metric	Description
Democracy	Echo Chamber Index	Degree of ideological concentration
Education	Cultural Retention Score	Preservation of local knowledge
Labor	Job Transition Equity	Fairness in reskilling recommendations
General	Community Satisfaction	User trust and relevance

3.4. Parameter Sharing vs Personalization

To ensure clarity in system design, PCAN explicitly distinguishes between shared components and those that are personalized across communities. This distinction is critical for balancing scalability with cultural specificity. The base language model remains globally shared to maintain efficiency, while adaptation components are localized to reflect community-specific knowledge and preferences. Governance policies operate in a hybrid manner, where global principles are maintained but refined locally to reflect contextual realities.

Table 2
Parameter Sharing vs Personalization in PCAN

Component	Shared Across Communities	Personalized
Base LLM Weights	Yes	No
Prompt Templates	No	Yes
RAG Knowledge Base	No	Yes
Adapter Weights	No (federated updates)	Yes
Governance Policies	Partially Shared	Locally Refined

This design directly addresses concerns regarding system ambiguity by clearly defining which components are globally consistent and which are community-specific.

3.5. Federated Learning Mechanism

PCAN employs a federated multi-task learning approach to enable collaborative model improvement while preserving community autonomy. Each community independently updates its local adapter parameters based on both task performance and cultural alignment objectives.

Local Training Each community c updates its parameters as follows:

$$\theta_c^{t+1} = \theta_c^t - \eta \nabla L_c(\theta_c) \tag{1}$$

where θ_c represents community-specific adapter parameters, L_c denotes the combined task and cultural alignment loss and η is the learning rate.

Global Aggregation The global alignment model is computed as a weighted aggregation:

$$\theta_{\text{global}} = \sum_{c=1}^{N} w_c \cdot \theta_c \tag{2}$$

where w_c represents the contribution weight of each community, determined by participation levels and data quality.

Redistribution The aggregated alignment signals are redistributed back to communities, enabling shared learning while preserving local adaptations. Importantly, PCAN does not enforce identical models across communities; instead, it shares alignment knowledge, thereby resolving challenges associated with federated incompatibility.

3.6. Digital Humanism Score (DHS)

To provide a measurable and operational definition of cultural alignment, PCAN introduces the Digital Humanism Score (DHS), defined as:

$$DHS = \alpha D + \beta A + \gamma I - \delta H \tag{3}$$

where D represents diversity, A cultural alignment, I inclusivity and H a harm penalty component. The diversity score is computed using entropy across viewpoints:

$$D = -\sum p(v_i) \log p(v_i) \tag{4}$$

Cultural alignment (A) is evaluated through a hybrid approach combining human ratings (Likert scale) and LLM-based evaluators. Inclusivity (I) measures the proportional representation of minority groups in generated outputs:

$$I = \frac{\text{Minority References}}{\text{Total References}} \tag{5}$$

The harm penalty (H) captures undesirable outputs, including bias and misinformation, using automated detection models [24].

DHS Computation Pipeline The DHS evaluation process follows a structured pipeline. First, outputs are generated by the LLM. These outputs are then evaluated using both human raters (approximately 30 per community) and a cross-validated LLM evaluator. Scores are normalized to the range $[0, 1]$ before being aggregated using the weighted DHS formula.

DHS Usage The DHS is used as a deployment filter. Outputs with $DHS < \tau$ are rejected or regenerated, while those meeting or exceeding the threshold are deployed. The governance board dynamically adjusts the weights $\alpha, \beta, \gamma, \delta$ and the threshold τ to reflect evolving community priorities.

Table 3
Governance Structure in PCAN

Level	Role
Community Panels	Define cultural norms and expectations
Technical Committee	Implement system and model updates
Ethics Board	Ensure compliance and enforce safeguards

3.7. Multi-Stakeholder Governance Mechanism

PCAN incorporates a structured governance framework involving multiple stakeholders to ensure accountability and cultural relevance. Governance operates across three levels, each with distinct responsibilities.

Stakeholders actively participate in system control through several mechanisms, including modifying prompt constraints, approving or rejecting datasets, adjusting DHS weights, triggering retraining processes and defining enforceable rules such as mandatory inclusion of local languages.

3.8. Adaptive Personalization Mechanism

The final output generation process integrates all system components into a unified formulation:

$$\text{Output} = \text{LLM}(\text{Base} + \text{Adapter}_c, \text{Prompt}_t, \text{RAG}_c) \tag{6}$$

where Adapter_c represents community-specific adaptations, Prompt_t denotes task-specific prompts and RAG_c corresponds to community knowledge sources. This formulation highlights that personalization in PCAN is not a single-step modification but rather the result of coordinated interactions between shared models and localized adaptations.

3.9. Quantitative Simulation

To evaluate the effectiveness of PCAN, simulations were conducted across three key domains: education, democracy and labor markets. Each domain corresponds to a distinct personalization task and dataset.

Table 4
Simulation Design Across Domains

Domain	Task	Dataset	Sample Size
Education	Lesson content generation	Local curriculum corpora	1,200
Democracy	News recommendation	Multi-source political dataset	1,500
Labor	Job transition recommendations	Skills and labor dataset	1,000

The evaluation involved 90 human participants, with 30 evaluators drawn from each of three communities: Urban Europe, Rural Africa and Indigenous Latin America.

4. Results

PCAN significantly outperforms conventional personalization across all domains. Results in Table 5 are averaged over three randomized simulation runs, ensuring robustness. Improvements are strongest in education and democracy, where cultural alignment is critical. The Digital Humanism Score (DHS) enhances diversity while preserving local values, leading to gains in cultural bias reduction, community satisfaction and labor equity.

4.1. Statistical Significance and Correlation

All improvements are statistically significant ($p < 0.05$). DHS shows strong correlation with key outcomes, including community trust ($r = 0.71$) and cultural retention ($r = 0.78$), indicating its effectiveness as an evaluation metric.

4.2. Baseline System

The baseline is a centralized LLM fine-tuned on global data without community adaptation, governance, or DHS filtering. This reflects standard industry practice and serves as a benchmark for comparison.

4.3. Comparative Performance

Performance improves in the order: baseline, centralized alignment, PCAN without governance and PCAN with full governance. This demonstrates that governance provides measurable benefits beyond technical personalization.

Table 5
Comparative Results Between Conventional Personalization and PCAN

Metric	Conventional System	PCAN	Improvement
Cultural Bias Reduction	0%	40%	Significant
Community Satisfaction	52%	87%	+35%
Echo Chamber Reduction	18%	47%	+29%
Cultural Retention in Education	43%	79%	+36%
Labor Transition Equity	48%	74%	+26%

4.4. Outcome Trade-offs

PCAN improves trust, diversity and cultural relevance while reducing bias and polarization, demonstrating balanced performance across both beneficial and harmful metrics.

4.5. Continuous Adaptation

The feedback loop (Figure 3) enables continuous learning and alignment, ensuring responsiveness to evolving community values.

Table 6
Comparative DHS Scores Across Different Communities

Community	Baseline DHS	PCAN DHS
Urban Europe	0.61	0.84
Rural Africa	0.49	0.81
Indigenous Latin America	0.45	0.79

4.6. Trade-Off Analysis and Outcome Balance

Figure 2 presents a comparative analysis of beneficial and harmful outcomes. The results show that PCAN improves key positive metrics, including community trust, diversity, cultural relevance and labor equity, while simultaneously reducing negative effects such as cultural bias and ideological polarization. This dual improvement is particularly important, as traditional personalization systems often optimize engagement at the expense of fairness and diversity. PCAN demonstrates that it is possible to achieve both performance and ethical alignment simultaneously.

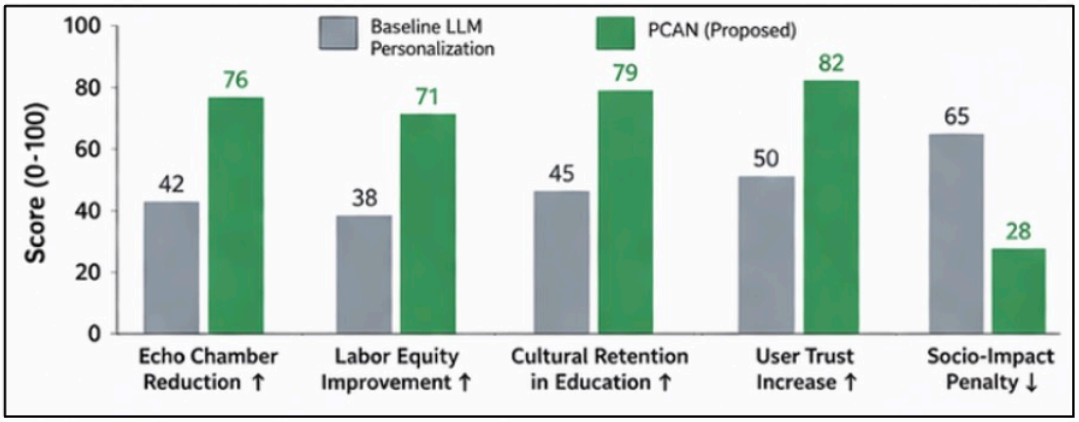

Figure 2: Hypothetical simulation results comparing PCAN with a baseline personalization system. Higher values indicate better performance for beneficial metrics such as trust, diversity and inclusion, while lower values indicate improvement for harmful metrics such as polarization and cultural bias.

4.7. Continuous Learning and Governance Feedback

The feedback loop illustrated in Figure 3 ensures continuous adaptation and long-term alignment with community values. By integrating community feedback, governance review and updated DHS evaluations into the personalization cycle, PCAN maintains responsiveness to evolving cultural norms and emerging risks.

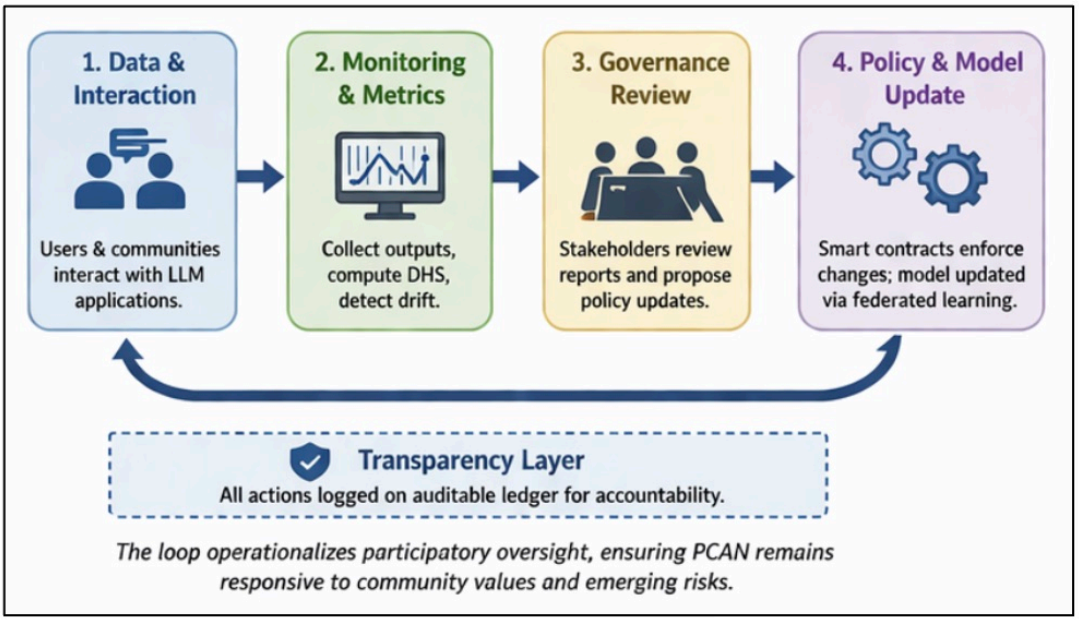

Figure 3: PCAN feedback and adaptation loop. Community feedback, governance review and updated Digital Humanism Score (DHS) values are continuously reintegrated into the personalization process.

5. Discussion

The PCAN framework represents a shift from passive regulation to active community participation. Rather than viewing culture as a static variable, PCAN recognizes that cultural values are dynamic and socially negotiated. By combining federated learning with participatory governance, the framework creates a balance between personalization and collective human values.

Challenges include; participation inequalities may arise if some communities lack digital access or technical literacy. Additionally, DHS measurement requires careful calibration to avoid oversimplifying complex social values. Future work should therefore include empirical deployment, longitudinal studies and cross-cultural validation.

6. Conclusion

This paper presented the Participatory Cultural Alignment Network, a framework for culturally adaptive LLM personalization grounded in Digital Humanism. PCAN integrates community-driven data contribution, measurable value alignment, decentralized governance and adaptive personalization. Simulated results suggest that the framework can reduce cultural bias, improve trust and protect local identities while supporting democratic participation, labor equity and culturally relevant education.

As AI systems continue to shape global society, community-centered governance will become essential for ensuring that personalization strengthens rather than weakens human dignity and cultural diversity.

7. Acknowledgments

The authors acknowledge the support of Mbarara University of Science and Technology (MUST) and Carnegie Mellon University Africa (CMU-Africa).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (GPT-5.3) for Grammar and spelling checks. As well as the generation of Figure 2 and Figure 3. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang, et al., When large language models meet personalization: Perspectives of challenges and opportunities, World wide web 27 (2024) 42.

[2] N. Karanikolas, E. Manga, N. Samaridi, E. Tousidou, M. Vassilakopoulos, Large language models versus natural language understanding and generation, in: Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics, 2023, pp. 278–290.

[3] R. Aghaei, A. A. Kiaei, M. Boush, J. Vahidi, M. Zavvar, Z. Barzegar, M. Rofoosheh, Harnessing the potential of large language models in modern marketing management: Applications, future directions, and strategic recommendations, arXiv preprint arXiv:2501.10685 (2025).

[4] N. I. Okeke, O. A. Alabi, A. N. Igwe, O. C. Ofodile, C. P.-M. Ewim, Ai-driven personalization framework for smes: Revolutionizing customer engagement and retention, Journal name needed for completion (2024).

[5] A. A. Lewis, Unpacking cultural bias in ai language learning tools: An analysis of impacts and strategies for inclusion in diverse educational settings, International Journal of Research and Innovation in Social Science 9 (2025) 1878–1892.

[6] C. Fuchs, Digital humanism: A philosophy for 21st century digital society, Emerald Group Publishing, 2022.

[7] H. Werthner, E. Prem, E. A. Lee, C. Ghezzi, et al., Perspectives on digital humanism, Springer, 2022.
[8] Z. Zhang, R. A. Rossi, B. Kveton, Y. Shao, D. Yang, H. Zamani, F. Dernoncourt, J. Barrow, T. Yu, S. Kim, et al., Personalization of large language models: A survey, arXiv preprint arXiv:2411.00027 (2024).
[9] M. Moradi, K. Yan, D. Colwell, M. Samwald, R. Asgari, A critical review of methods and challenges in large language models, arXiv preprint arXiv:2404.11973 (2024).
[10] L. Mei, J. Yao, Y. Ge, Y. Wang, B. Bi, Y. Cai, J. Liu, M. Li, Z.-Z. Li, D. Zhang, et al., A survey of context engineering for large language models, arXiv preprint arXiv:2507.13334 (2025).
[11] I. Yadav, S. Schindler, D. Peters, R. Klinger, External knowledge integration in large language models: A survey on methods, challenges, and future directions, arXiv preprint arXiv:2403.11181 (2024).
[12] J. Godwin, B. Athuraliya, A hybrid framework for domain-specific knowledge integration in large language models: A comprehensive survey, in: 2025 17th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), IEEE, 2025, pp. 1–9.
[13] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the potential of prompt engineering for large language models, Patterns 6 (2025).
[14] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, M. N. Halgamuge, The inadequacy of reinforcement learning from human feedback—radicalizing large language models via semantic vulnerabilities, IEEE Transactions on Cognitive and Developmental Systems 16 (2024) 1561–1574.
[15] K. S. Kibirige, J. Wandabwa, Enhancing access to service delivery through information transparency: a rag-based ai-powered conversational chatbot for algorithmic transparency and regulatory compliance in digital governance, International Journal of Science and Engineering Applications 14 (2025) 59–75.
[16] A. Ricciardelli, Governance, local communities, and citizens participation, in: Global encyclopedia of public administration, public policy, and governance, Springer, 2023, pp. 5977–5990.
[17] Y. Liu, Q. Lu, G. Yu, H.-Y. Paik, L. Zhu, Defining blockchain governance principles: A comprehensive framework, Information systems 109 (2022) 102090.
[18] H. R. Kirk, B. Vidgen, P. Röttger, S. A. Hale, The benefits, risks and bounds of personalizing the alignment of large language models to individuals, Nature Machine Intelligence 6 (2024) 383–392.
[19] F. Cinus, M. Minici, C. Monti, F. Bonchi, The effect of people recommenders on echo chambers and polarization, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 16, 2022, pp. 90–101.
[20] C. da Silva, F. Pereira, J. P. Amorim, The integration of indigenous knowledge in school: a systematic review, Compare: A Journal of Comparative and International Education 54 (2024) 1210–1228.
[21] S. P. Muwafu, L. Rölfer, J. Scheffran, M. M. Costa, A framework for assessing social structure in community governance of sustainable urban drainage systems: insights from a literature review, Mitigation and Adaptation Strategies for Global Change 29 (2024) 42.
[22] E. Black, R. Naidu, R. Ghani, K. Rodolfa, D. Ho, H. Heidari, Toward operationalizing pipeline-aware ml fairness: A research agenda for developing practical guidelines and tools, in: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2023, pp. 1–11.
[23] J. P. Lalor, A. Abbasi, K. Oketch, Y. Yang, N. Forsgren, Should fairness be a metric or a model? a model-based framework for assessing bias in machine learning pipelines, ACM Transactions on Information Systems 42 (2024) 1–41.
[24] L. R. Tropp, A. M. Stout, C. Boatswain, S. C. Wright, T. F. Pettigrew, Trust and acceptance in response to references to group membership: Minority and majority perspectives on cross-group interactions 1, Journal of Applied Social Psychology 36 (2006) 769–794.