FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 15, August, 2026, pp 215 – 220
DOI: https://doi.org/10.33003/fjs-2026-1015-5225

Ethical Safeguards in Autonomous Data Collection: Protecting Cultural Context and Privacy in Indigenous Language NLP

***Rosemary Agharese Usiobaifo, Roseline Oghogho Osaseri, Kaizer John Onyekachukwu**

Department of Computer, Faculty of Physical Sciences, University of Benin, P.O. Box 1154, Edo State, Nigeria.

*Corresponding authors' email: roseline.osaseri@uniben.edu

ABSTRACT

The increasing reliance of Natural Language Processing (NLP) systems on large-scale textual datasets has accelerated the use of automated data collection techniques, including web scraping, crawling, and autonomous extraction pipelines. Although these approaches improve scalability and support the development of Large Language Models (LLMs), they also introduce significant ethical challenges, particularly in indigenous and low-resource language contexts. Existing data collection methods often prioritize data availability and technical efficiency while overlooking privacy, cultural sensitivity, informed consent, and community-centered governance. Consequently, automated systems may collect personally identifiable information (PII) and culturally sensitive indigenous knowledge without adequate ethical safeguards. This paper proposes an ethical-by-design framework for autonomous indigenous language data collection using an LLM-based multi-agent architecture. The framework embeds ethical safeguards directly into the data collection process through context-aware filtering, hybrid PII detection, cultural sensitivity classification, and auditability mechanisms. A Design Science Research methodology was employed to develop and evaluate the framework using scenario-based validation involving privacy-sensitive and culturally sensitive textual data. The evaluation demonstrates that the framework supports transparent and accountable data collection while reducing ethical risks associated with privacy violations and the inappropriate extraction of culturally protected information. By balancing ethical safeguards with data utility, the proposed framework contributes to responsible AI, indigenous data governance, and ethical NLP, providing a foundation for transparent, culturally aware, and privacy-conscious data collection systems for indigenous and low-resource language technologies.

Received: 24 June 2026

Accepted: 02 July 2026

Published: 10 Sept. 2026

Keywords: Autonomous Data Collection; Ethical AI; Indigenous Data Governance; Large Language Models (LLMs); Multi-Agent Systems

INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) has been driven largely by the availability of massive textual datasets used to train Large Language Models (LLMs). Modern NLP systems increasingly depend on automated data collection techniques such as web scraping, crawling, and autonomous extraction pipelines to obtain large-scale language resources. While these approaches improve scalability and dataset diversity, they also introduce significant ethical concerns, particularly in the context of indigenous and low-resource languages.

Most existing data collection systems are designed primarily around technical efficiency and data availability. Publicly accessible digital content is often treated as freely reusable, with limited consideration for privacy, cultural context, or community ownership. As a result, automated pipelines may collect personally identifiable information (PII), culturally sensitive narratives, and indigenous knowledge without appropriate safeguards or contextual understanding. These concerns have contributed to growing discussions around data colonialism, where marginalized communities become sources of extractable data without meaningful participation or control in AI development processes (Couldry & Mejias, 2019; Birhane, 2020).

The problem is particularly important in indigenous language NLP. Languages such as Yoruba, Hausa, Igbo, and Nigerian Pidgin contain culturally embedded expressions, oral traditions, and community-specific meanings that may not be appropriate for unrestricted computational use. Conventional filtering systems based on simple keyword matching are often insufficient for identifying nuanced forms of sensitive or context-dependent content. Furthermore, the absence of transparent data governance mechanisms limits accountability in how language resources are collected and processed.

Recent ethical frameworks such as the CARE Principles for Indigenous Data Governance emphasize collective benefit, responsibility, and ethical stewardship of indigenous data (Carroll et al., 2020). Similarly, Responsible AI research highlights the importance of transparency, fairness, and accountability in AI systems. However, most existing approaches remain largely policy-oriented and are rarely integrated directly into autonomous NLP data collection infrastructures.

This paper proposes an ethical-by-design framework for autonomous data collection in indigenous language NLP using an LLM-based multi-agent architecture. The framework embeds ethical safeguards directly into system operations through privacy-preserving filtering, cultural sensitivity classification, and auditability mechanisms. Rather than applying ethical review retrospectively, the proposed approach enables context-aware decision-making during data discovery, extraction, and validation.

The main contributions of this paper are:

i. The development of an ethical-by-design framework for autonomous indigenous language data collection.

ii. The integration of privacy protection and cultural sensitivity mechanisms into LLM-based agent systems.
iii. The introduction of transparent audit and accountability mechanisms for traceable data processing decisions.

The remainder of this paper is organized as follows. Section 2 reviews related work on data colonialism, indigenous data governance, and ethical NLP. Section 3 presents the proposed framework and methodology. Section 4 discusses implementation and evaluation, while Section 5 concludes the paper and outlines future research directions.

Related Work

The growing dependence of NLP systems on large-scale textual data has intensified concerns regarding the ethical implications of automated data collection. Existing research highlights that many contemporary AI systems rely on large-scale web scraping practices that prioritize scalability while providing limited consideration for privacy, consent, or cultural context. Couldry and Mejias (2019) describe this phenomenon as *data colonialism*, where digital activity is transformed into extractable economic resources. In AI development, this often results in marginalized communities contributing linguistic and cultural data without corresponding control or benefit.

These concerns are particularly significant in African and indigenous language contexts. Birhane (2020) argues that indiscriminate data extraction practices risk reproducing historical forms of colonial inequality through algorithmic systems. Low-resource language communities are frequently underrepresented in NLP research while simultaneously becoming targets for large-scale data harvesting. This creates tensions between the need for language resources and the ethical responsibilities associated with data collection.

In response to these concerns, Indigenous Data Sovereignty (IDS) frameworks emphasize the rights of communities to govern the collection, ownership, and use of their cultural and linguistic data (Kukutai & Taylor, 2016). The CARE Principles for Indigenous Data Governance further promote collective benefit, authority to control, responsibility, and ethics in data management practices (Carroll et al., 2020). Unlike traditional FAIR data principles, which focus primarily on accessibility and interoperability, CARE emphasizes ethical stewardship and community-centered governance. However, despite their relevance, these principles are rarely operationalized within autonomous NLP data collection systems.

Research in Responsible AI has also introduced frameworks aimed at improving transparency, accountability, fairness, and privacy in AI systems. Floridi et al. (2018) proposed AI4People as a governance-oriented ethical framework, while Gebru et al. (2021) introduced "Datasheets for Datasets" to improve transparency in dataset creation and documentation. Although these approaches improve awareness of ethical risks, they remain largely documentation-oriented and do not directly address real-time ethical filtering during autonomous data collection.

Another major challenge in NLP data collection involves privacy and cultural sensitivity. User-generated online content frequently contains personally identifiable information (PII), contextual identity markers, and culturally embedded narratives that may not be suitable for unrestricted computational use. Conventional filtering systems based on rule-based matching are often inadequate for detecting nuanced or context-dependent forms of sensitive content, particularly in multilingual environments characterized by code-switching and informal language use.

Despite increasing recognition of these issues, current NLP data pipelines continue to exhibit three major limitations. First, ethical safeguards are commonly applied retrospectively rather than integrated directly into system architecture. Second, existing systems lack context-aware mechanisms capable of distinguishing between public and culturally sensitive indigenous language content. Third, most automated data collection pipelines provide limited transparency regarding how information is selected, filtered, or excluded.

This paper addresses these limitations by proposing an ethical-by-design framework that integrates privacy protection, cultural sensitivity classification, and transparent auditability directly into an LLM-based autonomous data collection architecture for indigenous language NLP.

MATERIALS AND METHODS

Proposed Ethical Framework

This paper proposes an ethical-by-design framework for autonomous indigenous language data collection using a Large Language Model (LLM)-based multi-agent architecture. Unlike conventional NLP pipelines that apply ethical review after data collection, the proposed framework integrates ethical safeguards directly into system operations. This enables context-aware decision-making during data discovery, extraction, validation, and storage.

The framework consists of four interconnected components: ethical governance, agent intelligence, processing and validation, and accountability and auditability. Together, these components ensure that ethical constraints are continuously enforced throughout the data lifecycle.

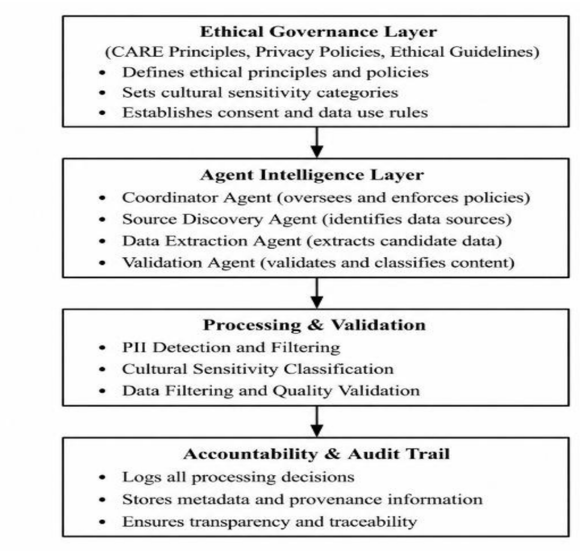

Figure 1: Ethical-by-Design Framework Architecture

The framework is designed around three primary objectives:

i. protection of personally identifiable information (PII).
ii. identification of culturally sensitive content.
iii. transparent and auditable data processing.

Ethical Governance Layer

The Ethical Governance Layer functions as the policy core of the framework. It defines the ethical rules and constraints that guide all autonomous system operations. The layer incorporates principles derived from Responsible AI research and Indigenous Data Sovereignty frameworks, particularly the CARE Principles for Indigenous Data Governance (Carroll et al., 2020).

To support context-aware filtering, the framework adopts a three-level cultural sensitivity classification model.

Table 1: Cultural Sensitivity Categories

Category	Description	Example
Public	Information intended for open dissemination	News articles, educational materials
Sensitive	Community-linked or context-dependent content	Informal narratives, community discussions
Restricted	Private or culturally protected knowledge	Sacred rituals, restricted traditional practices

This classification model provides structured guidance for autonomous decision-making and reduces dependence on unrestricted data extraction practices.

Agent Intelligence Layer

The Agent Intelligence Layer consists of LLM-based autonomous agents responsible for source discovery, language identification, data extraction, content validation, and metadata generation.

Unlike traditional scraping systems, these agents operate under explicit ethical constraints embedded within prompts and decision rules. Prompt engineering techniques are used to instruct agents to avoid collecting explicit personal identifiers, identify culturally sensitive information, and apply context-aware reasoning during filtering operations.

A centralized Coordinator Agent supervises all agent activities and enforces system-wide ethical policies. The Coordinator Agent validates outputs, resolves classification conflicts, and ensures consistency in ethical decision-making.

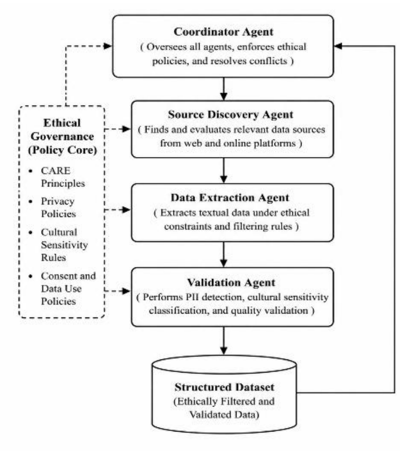

Figure 2: Ethical Multi-Agent Architecture

This layered agent structure enables adaptive and context-sensitive processing while maintaining centralized ethical oversight.

Processing and Validation Mechanisms

The framework operationalizes ethical safeguards during data transformation and filtering through a hybrid validation approach.

Privacy Protection

PII filtering combines rule-based pattern matching with LLM-based contextual reasoning.

Rule-based mechanisms detect structured identifiers such as phone numbers, email addresses, and numerical identifiers.

However, because personal information may also appear implicitly within narrative text, contextual reasoning is used to identify indirect identity references.

For example:

Input:

"Mr. Adewale from Ibadan shared his phone number during the meeting."

Filtered Output:

"[NAME] from [LOCATION] shared his phone number during the meeting."

This hybrid approach improves both filtering efficiency and contextual accuracy.

Cultural Sensitivity Filtering

The framework evaluates textual content using the predefined cultural sensitivity categories. Based on classification outcomes, public content is retained, sensitive content is flagged for review, and restricted content is excluded.

Table 2: Ethical Decision Rules

Condition	System Action
PII detected	Filter or reject
Sensitive cultural content	Flag for review
Restricted cultural content	Reject
Public content without PII	Accept

These rules enable consistent and traceable ethical decision-making within the autonomous pipeline.

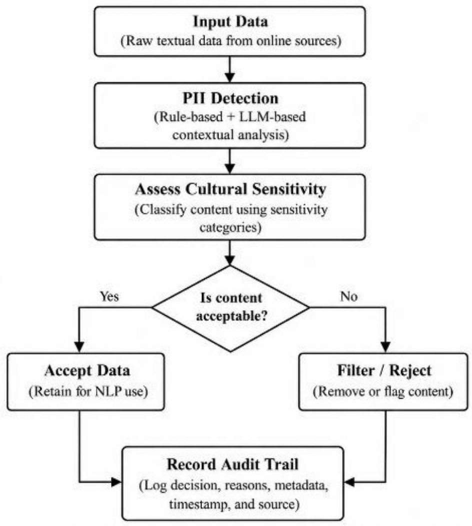

Figure 3: Ethical Processing and Validation Pipeline

Accountability and Auditability

To improve transparency and reproducibility, the framework incorporates an audit trail mechanism that records all major processing decisions. Each processed data item is associated with metadata including source information, language identification, sensitivity classification, filtering decisions, and processing history.

The audit mechanism enables researchers to trace why particular content was retained, filtered, or rejected. This addresses one of the major limitations of existing large-scale autonomous data collection systems, where filtering decisions are often opaque and difficult to reproduce.

By integrating explainability and accountability directly into system architecture, the proposed framework supports more transparent and governance-aware NLP data collection practices for indigenous and low-resource languages.

RESULTS AND DISCUSSION

The proposed framework was evaluated using a Design Science Research validation approach focused on demonstrating the feasibility of integrating ethical safeguards into autonomous NLP data collection systems. Rather than conducting large-scale empirical benchmarking, the evaluation emphasized scenario-based analysis involving privacy-sensitive and culturally sensitive textual data representative of indigenous language environments.

The implementation combined rule-based filtering mechanisms, LLM-driven contextual reasoning, cultural sensitivity classification, and audit trail generation.

Evaluation focused on four key criteria: privacy protection, cultural sensitivity awareness, transparency, and consistency of ethical decision-making.

Table 3: Evaluation Criteria

Criterion	Evaluation Focus
Privacy Protection	Detection and filtering of PII
Cultural Awareness	Identification of sensitive cultural content
Transparency	Availability of audit information
Consistency	Stability of ethical decision rules

To evaluate privacy protection, the framework processed textual inputs containing both explicit and implicit personally identifiable information (PII). Rule-based filtering successfully detected structured identifiers such as phone numbers and email addresses, while LLM-based contextual analysis improved detection of indirect identity references embedded within narrative text. The hybrid approach reduced the limitations commonly associated with purely rule-based systems, particularly in multilingual and informal communication contexts.

The framework was also evaluated using culturally sensitive scenarios involving indigenous narratives and community-linked discussions. Public informational content such as news articles was retained without modification, while culturally sensitive narratives were flagged for review. Restricted content associated with sacred or protected cultural practices was excluded from the processing pipeline. These outcomes demonstrate the framework's ability to distinguish between different levels of contextual sensitivity during autonomous data collection.

Transparency and accountability were assessed through the audit trail mechanism integrated into the framework. For each processed data item, the system recorded source information, language identification, sensitivity classification, filtering actions, and processing timestamps.

This audit mechanism improved traceability by enabling researchers to inspect how and why data processing decisions were made. Such transparency is particularly important in autonomous AI systems where filtering processes are often opaque.

The evaluation further demonstrated that LLM-based agents can support ethically guided automation when operating under explicit governance constraints. By embedding ethical rules into prompt design and centralized oversight mechanisms, the framework extends agentic AI systems beyond task automation toward governance-aware data processing.

Despite these strengths, several limitations remain. The framework depends on LLM-based interpretation for contextual reasoning, meaning classification outcomes may vary depending on prompt quality and model behavior. In addition, cultural sensitivity is inherently context-dependent and may differ across communities and linguistic environments. The absence of direct community participation in defining sensitivity boundaries also limits the framework's ability to fully capture indigenous knowledge governance requirements.

Nevertheless, the results demonstrate that ethical safeguards can be operationalized within autonomous NLP infrastructures without entirely sacrificing scalability or data utility. The framework therefore provides a practical foundation for developing transparent, privacy-conscious, and culturally aware data collection systems for indigenous and low-resource language technologies.

CONCLUSION

This paper examined the ethical challenges associated with autonomous data collection for indigenous language Natural Language Processing (NLP). While automated web scraping and LLM-driven extraction systems improve the scalability of NLP data acquisition, they also introduce risks related to privacy, cultural sensitivity, and unequal data governance practices.

To address these challenges, this study proposed an ethical-by-design framework for autonomous indigenous language data collection using an LLM-based multi-agent architecture. The framework integrates privacy-preserving filtering, cultural sensitivity classification, and auditability mechanisms directly into system operations, enabling ethical safeguards to function as part of the data collection process rather than as retrospective controls.

The evaluation demonstrated that the framework can support context-aware filtering, transparent decision-making, and privacy protection while maintaining practical usability for NLP applications. The study further highlights the potential of governance-aware AI agents for responsible data collection in indigenous and low-resource language environments.

Future research should focus on large-scale empirical validation, community-informed governance mechanisms, and adaptation of the framework across diverse multilingual settings.

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623. https://doi.org/10.1145/3442188.3445922

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed, 17*(2), 389-409. https://script-ed.org/article/algorithmic-colonization-of-africa/

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. *Data Science Journal, 19*(1), 43. https://doi.org/10.5334/dsj-2020-043

Couldry, N., & Mejias, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92. https://doi.org/10.1145/3458723

Hutchinson, B. (2024). *Modeling the sacred: Considerations for NLP on religious and culturally sensitive text*. Findings of the Association for Computational Linguistics: NAACL 2024. https://aclanthology.org/2024.findings-naacl.65/

Kukutai, T., & Taylor, J. (2016). *Indigenous data sovereignty: Toward an agenda*. ANU Press. https://doi.org/10.22459/CAEPR38.11.2016

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2). https://doi.org/10.1177/2053951716679679

Ricaurte, P. (2022). Ethics for the majority world: AI and the question of violence at scale. *Media, Culture & Society*, *44*(4), 582-599. https://doi.org/10.1177/01634437221099612

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI Now Report 2018*. AI Now Institute. https://ainowinstitute.org/reports.html