ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Artificial Intelligence and Indigenous Language Revitalization: A Gadé Language Digitization Model

Obadiah GT

Founder/Executive Director, Institute of Gade and Classical Studies, Gaduge Popa, Nasarawa LGA, Nasarawa State & Dean, Faculty of African Languages and Development, and Department of Gadé Language, Omni University, Imo State & Managing Director/Chief Executive Officer, Microbadiah Group, Nigeria

ABSTRACT

The marginalization of low-resource indigenous languages within contemporary Artificial Intelligence (AI) and Natural Language Processing (NLP) ecosystems poses a significant threat to global linguistic diversity and cultural heritage preservation. This study presents a comprehensive AI-driven Gadé Language Digitization Model designed to support the revitalization, preservation, and technological integration of the Gadé language of North-Central Nigeria. As a tone-rich Niger-Congo language characterized by complex morphosyntactic structures, nuanced semantic expressions, and an indigenous philosophical framework rooted in Gaboism, Gadé presents substantial challenges to conventional Large Language Models (LLMs), which often fail to adequately represent tonal distinctions and culturally embedded meanings. To address these limitations, the study proposes an integrated computational framework incorporating a Tone-Aware Automatic Speech Recognition (ASR) system that maps fundamental frequency (f0) contours to native orthographic markers, alongside a Neural Machine Translation (NMT) architecture enhanced by custom subword tokenization optimized for Niger-Congo linguistic patterns. Guided by principles of community-led data sovereignty, linguistic authenticity, and ethical AI development, the model utilizes a high-quality annotated corpus developed collaboratively with native speakers, community elders, and the Institute of Gade and Classical Studies. Experimental evaluations indicate notable reductions in Word Error Rate (WER) and significant improvements in translation quality and semantic preservation when compared with baseline multilingual models. Beyond technological innovation, the proposed framework establishes a scalable digital repository, educational infrastructure, and language technology ecosystem for Gadé. The study demonstrates how AI can serve as a transformative instrument for indigenous language revitalization and provides a replicable blueprint for preserving endangered African languages in the digital age.

***Corresponding author**

GT Obadiah, Founder/Executive Director, Institute of Gade and Classical Studies, Gaduge Popa, Nasarawa LGA, Nasarawa State & Dean, Faculty of African Languages and Development, and Department of Gadé Language, Omni University, Imo State & Managing Director/Chief Executive Officer, Microbadiah Group, Nigeria

Received: August 17, 2026; **Accepted:** August 21, 2026; **Published:** August 30, 2026

Keywords: Artificial Intelligence, Gadé Language, Indigenous Language Revitalization, Natural Language Processing, Speech Recognition, Neural Machine Translation, Digital Humanities, Gaboism, Language Digitization

Introduction

Language constitutes one of humanity's most valuable intellectual and cultural assets [1]. Beyond its communicative function, language serves as a repository of collective memory, indigenous knowledge systems, social values, historical experiences, philosophical traditions, and cultural identity [2]. Every language embodies a unique worldview through which its speakers interpret reality transmit knowledge across generations, and maintain connections with their ancestral heritage. However, the contemporary era of globalization, technological transformation, urban migration, and socio-economic integration has accelerated the decline of numerous indigenous languages across the world. While dominant international languages continue to expand their influence through education, governance, media, commerce, and digital technologies, many minority and indigenous languages remain marginalized within modern communication infrastructures [3]. This imbalance has contributed to a growing linguistic crisis in which thousands of languages face varying degrees of endangerment. The consequences of language loss extend far beyond vocabulary extinction; they include the disappearance of traditional ecological knowledge, oral literature, indigenous philosophies, cultural practices, and community identities [4]. In Africa, where linguistic diversity represents one of the continent's most significant cultural resources, the challenge of language preservation has become increasingly urgent. Many indigenous languages remain under-documented, under-resourced, and digitally invisible despite being actively spoken by substantial populations [5]. As digital technologies increasingly shape access to information, education, governance, and economic participation, languages without technological representation risk further exclusion from the knowledge economy [5]. Consequently, the revitalization and digitization of indigenous languages have emerged as critical priorities for scholars, policymakers, cultural institutions, and indigenous communities seeking to preserve linguistic diversity while ensuring that future generations remain connected to their cultural heritage [6].

Within this broader context, the Gadé language of North-Central Nigeria presents both a significant challenge and a remarkable opportunity for indigenous language revitalization in the digital age. Spoken primarily among the Gadé people, the language represents a rich cultural and linguistic tradition shaped by centuries of historical development, social interaction, and intellectual creativity [6]. Embedded within Gadé are oral histories, indigenous governance systems, traditional ecological knowledge, ritual expressions, artistic performances, and philosophical concepts associated with Gaboism, a distinctive worldview that informs Gadé cultural identity [7]. Despite its cultural significance, the language faces many of the structural pressures confronting indigenous languages worldwide. Intergenerational language transmission is increasingly challenged by urbanization, formal education systems dominated by English, changing communication patterns, and the growing influence of digital media produced in major global languages. Although community-based efforts have contributed to preserving aspects of Gadé linguistic heritage, substantial gaps remain in documentation, corpus development, language standardization, educational resources, and technological integration. Existing digital infrastructures provide limited support for Gadé language users, resulting in minimal representation across websites, mobile applications, machine translation platforms, speech technologies, and artificial intelligence systems. This technological absence has significant implications because digital relevance increasingly determines linguistic visibility and long-term sustainability. Languages that cannot participate effectively within contemporary technological ecosystems often struggle to attract younger speakers who conduct much of their social, educational, and professional activities online. Therefore, the development of robust digital frameworks for Gadé is not merely a technological undertaking but a strategic cultural intervention aimed at ensuring the language's continued vitality within rapidly evolving social and technological environments [8].

Recent advances in Artificial Intelligence (AI), particularly in the fields of Natural Language Processing (NLP), machine learning, speech technology, and large language models, have transformed possibilities for language documentation, preservation, and revitalization [9-11].

AI technologies now facilitate automated transcription, speech recognition, machine translation, sentiment analysis, conversational agents, educational tutoring systems, and digital content generation across numerous languages. These innovations have demonstrated remarkable success in expanding linguistic accessibility and enhancing communication across diverse populations [12]. Nevertheless, the benefits of AI have been distributed unevenly. Most contemporary AI systems are developed using massive datasets derived from globally dominant languages such as English, Mandarin Chinese, Spanish, French, and Arabic. Consequently, low-resource indigenous languages often remain excluded from technological innovation due to insufficient digital data, limited computational resources, and inadequate research [13]. Furthermore, many existing AI architectures struggle to accurately process linguistic features commonly found in African languages, including tonal distinctions, complex morphology, contextual semantics, serial verb constructions, and culturally specific conceptual frameworks [14-17]. For tone-heavy languages such as Gadé, conventional speech recognition and machine translation systems frequently fail to preserve essential semantic distinctions encoded through pitch variation. Such limitations highlight the need for specialized computational approaches that recognize the linguistic uniqueness of indigenous languages rather than forcing them into models optimized for fundamentally different linguistic structures. Emerging research increasingly emphasizes the importance of culturally responsive AI systems capable of supporting linguistic diversity while respecting indigenous knowledge systems, community values, and data sovereignty principles. This shift provides a foundation for developing technological solutions specifically designed to address the needs and realities of indigenous language communities.

This study responds to these challenges by proposing a comprehensive Gadé Language Digitization Model that integrates artificial intelligence, community participation, linguistic documentation, and digital infrastructure development into a unified framework for language revitalization [18]. Unlike conventional language technology initiatives that focus primarily on technical outcomes, the proposed model adopts a holistic approach that recognizes language as both a cultural resource and a technological asset. Central to the framework is the development of a community-governed digital ecosystem encompassing corpus creation, orthographic standardization, speech technology development, machine translation systems, educational applications, and long-term digital preservation strategies [19]. Particular emphasis is placed on the creation of a Tone-Aware Automatic Speech Recognition system capable of accurately representing Gadé tonal structures, as well as Neural Machine Translation architectures optimized for the language's unique morphosyntactic characteristics. The model also prioritizes community-led data collection processes involving native speakers, traditional knowledge custodians, educators, and cultural institutions, thereby ensuring linguistic authenticity and ethical data governance [20] Through collaborative partnerships between researchers, technology developers, and indigenous stakeholders, the framework seeks to create sustainable pathways for language innovation while preserving cultural integrity [21]. By integrating technological advancement with community ownership, the model challenges extractive approaches to AI development and instead promotes equitable participation in the creation of digital language resources [22-24].

The significance of this research extends beyond the Gadé language community to broader discussions concerning digital inclusion, linguistic justice, cultural sustainability, and the future of indigenous languages within artificial intelligence ecosystems. As governments, universities, technology companies, and international organizations increasingly explore the social implications of AI, questions regarding linguistic representation and cultural equity have become increasingly important. The exclusion of indigenous languages from emerging technologies risks reinforcing existing inequalities and contributing to the further marginalization of already vulnerable linguistic communities [25,26]. Conversely, the intentional integration of indigenous languages into AI systems offers opportunities to democratize technological access, preserve cultural heritage, strengthen educational outcomes, and enhance community resilience. By presenting a scalable and replicable model for AI-driven language revitalization, this study contributes to ongoing efforts to bridge the gap between technological innovation and cultural preservation [27-31]. The Gadé Language Digitization Model demonstrates that indigenous languages need not remain passive observers of the digital revolution; instead, they can actively participate in shaping inclusive technological futures that reflect the full diversity of human knowledge and experience. Through this approach, the study positions the Gadé language as a pioneering example of how artificial intelligence can be harnessed ethically and effectively to safeguard endangered

linguistic heritage while fostering innovation, empowerment, and sustainable cultural development in the twenty-first century [32].

The Digital Divide in Natural Language Processing (NLP)

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) has transformed the ways in which human beings communicate, access information, and interact with digital technologies [33]. From machine translation and virtual assistants to automated speech recognition and conversational AI systems, language technologies have become deeply integrated into contemporary social, economic, and educational life [34]. However, the benefits of these technological innovations have not been distributed equitably across the world's linguistic communities. A significant digital divide persists between high-resource languages, which enjoy extensive technological support and computational representation, and low-resource indigenous languages, which remain largely absent from modern NLP infrastructures. This disparity has created a form of linguistic inequality in which access to digital opportunities is increasingly determined by the technological visibility of a language. As societies become more dependent on AI-powered communication systems, languages that lack computational resources face growing risks of marginalization, exclusion, and eventual decline [35].

The concept of the digital divide in NLP extends beyond disparities in internet access or technological infrastructure. It encompasses unequal access to digital language resources, computational tools, annotated datasets, speech corpora, machine translation systems, language models, and educational technologies. Most contemporary NLP systems are trained using enormous quantities of textual and speech data derived from a relatively small number of globally dominant languages such as English, Mandarin Chinese, Spanish, French, German, and Arabic. These languages benefit from extensive digitization efforts, large-scale research investments, and strong commercial incentives that drive continuous technological development. In contrast, thousands of indigenous and minority languages possess limited digital content, insufficient linguistic documentation, and minimal representation within AI training datasets. Consequently, modern language technologies often fail to recognize, process, or generate meaningful outputs in these languages, reinforcing existing linguistic hierarchies and limiting opportunities for digital participation among their speakers [36].

The situation is particularly pronounced across Africa, a continent recognized as one of the most linguistically diverse regions in the world. Despite hosting more than two thousand languages, only a small fraction of African languages has received significant attention within NLP research and development. Many indigenous languages remain underrepresented in digital repositories, academic databases, and commercial AI applications. This underrepresentation stems from multiple factors, including inadequate funding, limited linguistic documentation, scarcity of computational expertise, and the absence of large-scale language corpora required for training contemporary AI systems. Furthermore, many African languages possess linguistic characteristics that challenge conventional NLP architectures, including tonal distinctions, agglutinative morphology, noun-class systems, code-switching patterns, and culturally embedded semantic structures. As a result, technological solutions designed primarily for European and Asian languages often perform poorly when applied to African linguistic contexts, highlighting the need for localized and culturally responsive computational approaches [37].

The Gadé language exemplifies many of the challenges associated with this digital divide. Although the language embodies a rich cultural heritage and remains an important medium of communication among the Gadé people of North-Central Nigeria, its presence within digital ecosystems remains extremely limited. Existing AI platforms, machine translation systems, voice assistants, search engines, and educational technologies provide little or no support for Gadé language users. This absence restricts opportunities for digital literacy, online content creation, language learning, and cultural preservation. More importantly, it creates a technological environment in which younger generations increasingly engage with dominant languages while finding few incentives to use their indigenous language within digital spaces. Without deliberate intervention, such patterns may contribute to language shift, reduced intergenerational transmission, and the gradual erosion of linguistic heritage. The challenge, therefore, is not merely one of technological development but also of cultural sustainability and linguistic justice [38].

Addressing the digital divide in NLP requires a paradigm shift from language inclusion as an afterthought to language inclusion as a fundamental principle of AI development. Indigenous communities must be recognized not merely as sources of linguistic data but as active partners in the design, governance, and implementation of language technologies. For the Gadé language, this entails the creation of high-quality digital corpora, speech datasets, lexicographic resources, educational platforms, and AI models specifically tailored to its linguistic structure and cultural context [39,40]. By integrating community-led documentation, ethical data governance, and advanced computational methodologies, it becomes possible to bridge the technological gap that currently separates Gadé from mainstream digital infrastructures. Ultimately, reducing the NLP digital divide is not simply a technical objective; it is a commitment to preserving linguistic diversity, promoting digital equity, and ensuring that indigenous languages can participate meaningfully in the emerging knowledge economy of the twenty-first century.

The Ethnographic Context of the Gadé People (Obadiah, 2025; 2026)

The Gadé people, constitute a distinct and historically profound ethnolinguistic group situated within the West African Middle Belt. Primarily concentrated across the Federal Capital Territory (FCT) of Abuja, Nasarawa State, and Niger State in Nigeria, their ethnographic narrative provides an important case study in long-term migration, socio-political resilience, and independent cultural evolution within the Benue-Congo linguistic sphere. The contemporary distribution of the Gadé people is organized into distinct geopolitical clusters, traditionally classified into three mutually intelligible linguistic and regional sub-groups: the Baponoic, the Northern dialect, and the Southern sectors.

Historically bounded by complex topographical features, including the defensive hill configurations of North-Central Nigeria, the Gadé established a highly organized geographic footprint. Rather than existing as scattered agrarian settlements, historical records, heavily documented by researchers such as GT Obadiah indicate that the pre-colonial Gadé were structured into a cohesive entity known as the Gadé Federation (or the Gadé Republic). This federation was physically demarcated and protected by extensive systems of fortified city-walls (Ganuwa), reflecting an advanced understanding of military architecture and territorial sovereignty designed to withstand the turbulent slave-raiding eras. The historical memory of the Babye challenges conventional

Eurocentric historiography by utilizing a deep-time chronological framework. Oral traditions and classical texts detail an expansive migratory itinerary spanning millennia:

- **The Primordial Epoch (c. 36,000 BCE):** Gade historical scholarship positions the ultimate origin of the Babye in the Upper Nile/Cushitic region (modern-day Ethiopia/ Sudan). Driven by early regional shifts, they migrated westward to establish a primordial homeland known as Àṣam (conventionally interpreted in Gade semantics as the "Garden of Eden"). This epoch marks the foundation of their ancestral governance and early urban planning frameworks.
- **The Nok Convergence and Kano Exile Migrations:** The historical arc traces their movement through Keffi (c. 12,000 BCE) and into the provinces of the Nok Culture (c. 776 BC), where they engaged in cross-cultural technological exchanges. By 10 BC, a significant population had established deep roots in the Kano and Jigawa regions (Gadawur), where they developed an influential economy centred on agrarian production and the famous Gishirin Gade (Gadé salt industry). A defining feature of the Gadé ethnographic profile is their unique approach to authority (Morgan, 1936b). The political architecture of the Gadé Federation operates as a decentralized, democratic republic integrated with institutional monarchism.

At the apex of the administrative hierarchy sits the Gomo (the paramount ruler/Head of State). However, unlike the absolute autocracies found in neighbouring historical caliphates or emirates, the Gomo's power is strictly checked by the Bátsákpâ (the Council of Elders or Congressmen). The Bátsàkpá function as a legislative and judicial body, ensuring that executive decisions align with ancestral law and collective security.

Historically, this governance model was divided into functional portfolios—akin to modern state ministries—that independently managed defense, agrarian resource allocation, internal judicial arbitration, and external trade relations, emphasizing a societal philosophy rooted in distributed intelligence and civic consensus.

To analyze the Gadé ethnography without accounting for Gaboism is to strip the culture of its operational grammar. Gaboism represents the comprehensive, indigenous philosophical and metaphysical system of the Gadé people. It functions as a "spiritual constitution" that dictates ethical conduct, legal boundaries, and ecological management.

A critical ethnographic boundary must be drawn between general Gaboism and the esoteric institution of Gobo. Gaboism is an open, community-wide code of ethics, metaphysical belief in ancestral continuity (Gabo), and social responsibility based on generic love. Gobo, conversely, refers specifically to the highly restricted, male-dominated Secret Societies and ritual masquerades. The Gobo functions essentially as the enforcement arm of the state-a traditional judiciary that executes the legal verdicts passed by the Gomo and the Bátsàkpá [39].

Through this intricate matrix of deep historical consciousness, republican governance, and a resilient philosophical framework, the Gadé people have preserved a distinct identity (Obadiah, 2025). Digitizing their language requires an AI model that does not merely translate words, but actively encodes this profound ethnographic metadata [39].

Historical Epistemology and Language as a Cultural Archive

To build a computationally robust Natural Language Processing (NLP) framework for the Gadé language, the system must treat language not merely as a stochastic sequence of textual tokens, but as a dynamic, historical repository of a people's knowledge system. Historical epistemology—the study of how knowledge frameworks emerge, shift, and solidify over time—reveals that for marginalized and low-resource languages, the lexicon acts as a primary cultural archive. In the case of Gadé, historical trajectories, socio-political institutions, and metaphysical philosophies are directly encoded into the structural morphology and semantic mutations of the language itself.

The Lexicalization of Deep-Time Chronology: The historical timeline of the Gadé people—spanning from the primordial epoch of Àṣam 'Garden of Life and Purposes' (c. 36,000 BCE) through migrations to Keffi (c. 12,000 BCE), Kano (c. 10 BC), and the Nok culture provinces (c. 776 BC)—is structurally embedded in the classical vocabulary of the language. When a language undergoes long-term physical displacement, historical events leave morphological footprints. For example, the toponym Àṣam in classical Gadé semantics translates to an original paradise or a foundational place of genesis. The survival of this word as a linguistic root for "origin" provides an empirical linguistic link to the group's earliest recorded migration from the Upper Nile/ Cushitic region [40].

Sociopolitical Morphosyntax and Governance Archives: The decentralized, republican governance model of the pre-colonial Gadé Federation is preserved within the institutional titles and political terminology of the language. The balance of power between the executive branch and the legislative congress is explicitly maintained through linguistic protocols:

- **The Gomo-Bátsákpâ Binary:** The term Gomo (paramount ruler) cannot be linguistically or conceptually isolated from the Bátsákpâ (Council of Elders). In traditional Gadé discourse, sentences involving executive decrees (Gomo) structurally require syntactic or contextual validation from the elder assembly (Bátsàkpá).

The Defensive Architecture of Gụvụ́: The physical presence of ancient fortified city-walls is preserved in the vocabulary of spatial orientation and civil defense within the language. Terms denoting protection, community boundaries, and civic duty frequently branch from the root word gụvụ́ archiving an era of military engineering and territorial sovereignty directly within everyday speech.

- **Metaphysical Encoding:** The Semantic Divergence of Gabo and Gobo: The most critical epistemological challenge for a machine translation or digitization model lies in the semantic precision of Gadé metaphysical vocabulary. The core philosophical framework of the culture is archived through a highly sophisticated system of near-homophones that carry drastically different societal, legal, and spiritual weights [41,42].

Standard translation algorithms trained on Indo-European language datasets often categorize non-Western spiritual terms under broad, reductive signifiers such as "religion" or "animism." However, the historical epistemology of Gadé reveals that Gabo and Gobo represent distinct structural pillars of the state: one is a philosophy of mind and ethics, while the other is an operational arm of traditional jurisprudence.

If an AI model treats the Gadé language as a simple parallel corpus to be translated into English equivalents, this rich historical archive is effectively erased. A literal translation pipeline transforms a phrase embedded with centuries of political philosophy or

migratory history into an abstracted, Westernized concept.

Therefore, a true digitization model must implement context-aware tokenization and semantic metadata tagging. By treating the language as an epistemological archive, the AI can map tokens not just to target-language words, but to the historical, political, and philosophical matrices established across the deep chronology of the Gadé people.

Statement of Problem and Research Objectives

Statement of Problem

The foundational bottleneck in contemporary Natural Language Processing (NLP) is its systemic architectural bias toward high-resource, non-tonal, morphologically linear languages (primarily Indo-European). Massively multilingual Large Language Models (LLMs) rely on massive data ingestion web-scrapes, which fundamentally excludes low-resource indigenous African languages. When applied to a highly complex language like Gadé, these standard models exhibit catastrophic failures across three critical dimensions:

- **Phonetic and Tonal Flattening:** Gadé relies on contrastive pitch variations and intricate phonetics to dictate semantic meaning, as seen in the functional distinction between m̄ gẹ dẹ ("I have said") and everyday conversational markers. Standard Automatic Speech Recognition (ASR) pipelines lack the foundational frequency (f0) tracking sensitivity required to capture these acoustic nuances, resulting in severe orthographic corruption and transcription errors.

Semantic Erasure of Epistemological Metadata: Conventional machine translation architectures force direct, literal mappings into target languages like English. This process effectively erases the deep historical, political, and philosophical frameworks archived within the Gadé lexicon. For instance, a standard model flattens the distinct sociopolitical and metaphysical boundaries between Gabo (universal ancestral philosophy and ethics) and Gobo (the judicial enforcement secret societies), incorrectly categorizing both under generic, Westernized descriptors like "religion" or "folklore."

Data Extrospection and Sovereignty Loss: Current AI deployment models treat indigenous linguistic data as an open-source, extractive commodity. This lacks the cultural safeguards necessary to protect sacred oral histories, regional dialectal variations (Baponoic, Northern dialect and Southern sectors), and historical deep-time chronologies managed by traditional institutions like the Bátsàkpá (Council of Elders). Without a specialized, community-governed computational framework, the digitization of the Gadé language risks either total algorithmic erasure or structural homogenization.

Research Objectives

To resolve these computational and linguistic limitations, this research develops a localized, ethically grounded, and phonetically sensitive AI architecture. The specific objectives of this study are as follows:

- **To Engineer a Tone-Aware Acoustic Modelling Pipeline:** Develop and fine-tune an Automatic Speech Recognition (ASR) engine capable of tracking fundamental frequency (f0) contours to accurately map and render native Gadé orthographic tone marks from high-fidelity oral recordings.
- **To Design a Contextualized Subword Tokenizer Optimized for Niger-Congo Morphology:** Replace standard byte-pair encoding (BPE) with a custom morphosyntactic tokenizer designed to handle the noun-class agreement systems, serial verb constructions, and near-homophonous semantic splits characteristic of the language.
- **To Build a Knowledge-Graph Enriched Neural Machine Translation (NMT) Model:** Construct an NMT framework that integrates cultural metadata and historical footnotes into the translation pipeline, preserving the deep-time chronological archives (e.g., Àsáṁ̄, and the governance frameworks of the (Gomo) rather than producing flat literal translations.
- **To Operationalize a Framework for Community-Led Data Sovereignty in AI:** Establish a secure, decentralized data governance model in direct partnership with local linguistic authorities and elders, creating a reproducible digital blueprint for the ethical preservation of marginalized West African languages.

Key Contributions and Significance of the Study

The significance of this study lies at the intersection of advanced computational linguistic and cultural heritage preservation. By moving away from conventional, data-hungry algorithmic paradigms, this research offers critical contributions to the fields of Natural Language Processing (NLP) for low-resource environments, algorithmic ethics, and West African historical linguistics.

Core Technical and Conceptual Contributions

- **Development of a Tone-Aware Acoustic Modelling Architecture:** This study introduces a novel approach to fine-tuning acoustic speech pipelines by mapping fundamental frequency (f0) pitch contours directly to native orthographic markers. This solves the persistent issue of phonetic and tonal flattening that occurs when low-resource, tone-dependent African languages are processed by standard Automatic Speech Recognition (ASR) engines [43].
- **Morphosyntactically Optimized Tokenization Blueprint:** Instead of relying on English-centric Byte-Pair Encoding (BPE), this paper designs a custom tokenizer tailored to the noun-class agreements, serial verb constructions, and semantic structures of the Gadé language. This structural adaptation significantly minimizes token fragmentation and improves translation accuracy.
- **Epistemological Metadata Integration:** This research introduces a Knowledge-Graph Enriched Neural Machine Translation (NMT) model [44]. By mapping complex, near-homophonous terms to a localized cultural ontology, the architecture preserves critical historical and philosophical frameworks—such as the vital distinction between Gabo (cosmological philosophy) and Gobo (judicial enforcement apparatus)—rather than forcing flat, literal translations into Western signifiers.

Practical and Societal Significance

Reversing Language Shift Among Digitally Native Youth: By deploying this digitization model into user-facing interfaces—including Text-to-Speech (TTS) systems and interactive educational technologies—this study provides actionable tools to bridge the generational gap. It equips younger generations with high-fidelity digital platforms to engage natively with their language across contemporary networks [45].

Empirical Validation of Community-Led Data Sovereignty: This study operationalizes a secure, ethical framework for

data governance that respects indigenous intellectual property. Managed in collaboration with local linguistic authorities and elders, it establishes a practical model for protecting sacred oral histories and deep-time chronological archives from extractive or unauthorized data-mining practices [46].

Scalable Blueprint for Benue-Congo Languages: Beyond its primary focus on Gadé, the algorithmic pipelines and ethical frameworks developed in this research serve as a highly portable, reproducible blueprint. It offers a clear methodology for computational linguists and indigenous communities seeking to digitize and preserve other marginalized, tone-heavy, low-resource languages across West Africa and the global South.

Literature Review and Thematic Analysis

The intersection of Artificial Intelligence (AI), Natural Language Processing (NLP), and indigenous language revitalization has emerged as one of the most significant areas of contemporary linguistic and technological research. As digital technologies increasingly shape communication, education, governance, and cultural production, the ability of a language to function within digital ecosystems has become closely linked to its long-term survival. Scholars have argued that language endangerment in the twenty-first century is no longer solely a sociolinguistic issue but also a technological challenge. Languages that remain absent from digital infrastructures risk exclusion from emerging knowledge economies, educational innovations, and global communication networks. Consequently, research efforts have increasingly focused on developing computational tools capable of supporting low-resource and endangered languages through documentation, digitization, machine translation, speech technologies, and AI-assisted learning systems [47].

Indigenous Language Endangerment and Revitalization

Language revitalization scholarship has traditionally focused on reversing language shift, strengthening intergenerational transmission, and promoting cultural continuity within indigenous communities. Early foundational studies emphasized the relationship between language loss and broader processes of cultural assimilation, colonialism, urbanization, and globalization [48]. Researchers have consistently demonstrated that language extinction often results in the disappearance of oral traditions, indigenous ecological knowledge, customary legal systems, and unique philosophical worldviews.

Contemporary revitalization frameworks increasingly advocate for multidimensional approaches that combine language documentation, educational reform, community mobilization, policy interventions, and technological innovation. Indigenous communities across the world have adopted diverse strategies including immersion schools, digital storytelling projects, community archives, online learning platforms, and mobile language applications. These initiatives recognize that language preservation must extend beyond documentation toward active language use within contemporary social environments [49].

For many African languages, however, revitalization efforts remain constrained by limited institutional support, inadequate funding, and insufficient technological resources. While significant attention has been devoted to major African languages, smaller indigenous languages often remain underrepresented in national language policies and digital development programs. This gap highlights the need for innovative approaches that leverage emerging technologies to support sustainable language revitalization [50].

Artificial Intelligence and Language Preservation

Recent advancements in Artificial Intelligence have transformed possibilities for language preservation and revitalization. AI technologies now support a wide range of linguistic applications including speech recognition, machine translation, text generation, sentiment analysis, language tutoring systems, and conversational agents. These developments have enabled researchers to process large volumes of linguistic data more efficiently while creating new opportunities for language learning and digital engagement [51].

Several studies have demonstrated the potential of AI-assisted tools for endangered language documentation. Automated transcription systems have accelerated the processing of oral recordings, while machine learning algorithms have facilitated linguistic annotation and corpus development. Similarly, AI-powered educational applications have enhanced language learning experiences by providing personalized instruction, pronunciation feedback, and interactive exercises.

Despite these advances, scholars caution that many AI systems remain heavily biased toward high-resource languages. Large Language Models (LLMs) often rely on massive datasets derived primarily from dominant global languages, resulting in limited support for minority and indigenous languages. Consequently, indigenous communities frequently remain excluded from the benefits of AI innovation. This challenge has motivated calls for the development of inclusive AI frameworks that prioritize linguistic diversity and equitable technological representation [52-54].

Low-Resource Languages and NLP Challenges

One of the most persistent themes in contemporary NLP research concerns the challenges associated with low-resource languages. NLP systems generally require substantial quantities of high-quality textual and speech data for training and evaluation. However, many indigenous languages lack the digital corpora necessary to support modern machine learning architectures [55].

Researchers identify several major obstacles affecting low-resource language development. First, limited textual documentation restricts opportunities for corpus creation and language modelling. Second, the absence of standardized orthographies complicates data collection and annotation. Third, many indigenous languages exhibit complex linguistic structures that are insufficiently represented within mainstream NLP frameworks. Features such as tonal distinctions, agglutinative morphology, serial verb constructions, and context-dependent semantics often challenge conventional computational approaches [56,57].

Tonal Languages and Computational Representation

Tone constitutes one of the most significant yet underrepresented linguistic features in NLP research. In tonal languages, variations in pitch can alter lexical meaning, grammatical function, and discourse interpretation. Failure to accurately represent tone may therefore result in substantial semantic distortions [58].

Existing speech recognition systems often prioritize segmental phonological information while overlooking tonal features. As a result, many conventional Automatic Speech Recognition (ASR) models struggle to process tone-heavy languages effectively. Scholars have increasingly emphasized the need for tone-aware computational architectures capable of integrating pitch contours into speech processing workflows.

Research on African tonal languages demonstrates that incorporating acoustic pitch information into machine learning models significantly improves transcription accuracy and semantic preservation. These findings are particularly relevant to the Gadé language, where tonal distinctions play a critical role in meaning formation. The literature therefore supports the development of specialized tone-sensitive technologies as a central component of indigenous language digitization efforts [59].

Community-Led Data Sovereignty and Ethical AI

Recent scholarship has highlighted ethical concerns surrounding indigenous data collection and AI development. Historically, many language documentation projects extracted linguistic resources from indigenous communities without ensuring meaningful community participation or long-term benefits. Such practices often resulted in the external control of indigenous knowledge systems and cultural assets.

In response, researchers have advanced the concept of Indigenous Data Sovereignty, which emphasizes community ownership, governance, and control over linguistic resources. This framework advocates collaborative approaches in which indigenous communities actively participate in data collection, annotation, storage, and technological development. Ethical AI principles further stress transparency, accountability, cultural sensitivity, and equitable benefit-sharing [60].

Community-led approaches are increasingly recognized as essential for sustainable language technology initiatives. By involving local speakers, elders, educators, and cultural institutions, researchers can ensure that technological systems reflect authentic linguistic practices while respecting cultural values and intellectual property rights [61].

Digital Humanities and Indigenous Knowledge Systems

The emergence of Digital Humanities has expanded opportunities for preserving and disseminating indigenous knowledge through digital platforms. Digital archives, virtual museums, multimedia repositories, and interactive educational resources have enabled communities to document oral traditions, historical narratives, ceremonial practices, and linguistic heritage.

Scholars argue that digitization should not merely preserve linguistic forms but should also capture the broader cultural contexts within which language functions. Indigenous knowledge systems often encode environmental wisdom, ethical principles, governance structures, and philosophical traditions that are inseparable from linguistic expression. Consequently, effective language digitization initiatives must integrate linguistic preservation with cultural documentation [62].

Research Gap and Conceptual Contribution

The Gadé language remains largely absent from contemporary NLP literature despite its linguistic significance and cultural richness. Existing research provides little guidance regarding the development of AI systems specifically designed to accommodate Gadé tonal structures, morphosyntactic complexity, and indigenous epistemological frameworks. This gap underscores the need for a comprehensive and culturally grounded digitization model [63].

The present study contributes to scholarship by proposing an integrated Gadé Language Digitization Model that combines AI innovation, linguistic documentation, community governance, educational technology, and ethical data management. By addressing both technical and cultural dimensions of language revitalization, the model advances broader efforts to create inclusive and equitable AI ecosystems for indigenous languages.

Thematic Synthesis

The literature reveals five dominant themes: (1) the urgent need to revitalize endangered indigenous languages; (2) the transformative potential of AI and NLP technologies for language preservation; (3) the structural challenges confronting low-resource and tonal languages; (4) the importance of community ownership and indigenous data sovereignty; and (5) the integration of cultural knowledge within digital preservation frameworks. Together, these themes establish the theoretical and practical foundation for the Gadé Language Digitization Model proposed in this study. They demonstrate that successful indigenous language revitalization requires not only technological innovation but also ethical governance, cultural authenticity, and sustained community participation.

Machine Learning Methodologies for Low-Resource Languages

The development of effective Natural Language Processing (NLP) systems for low-resource languages has become a major research priority within contemporary Artificial Intelligence (AI). Unlike high-resource languages such as English, Chinese, French, and Spanish, which benefit from extensive digital corpora, annotated datasets, and decades of computational research, low-resource languages often lack the linguistic resources required to train conventional machine learning models. This disparity presents significant challenges for indigenous language communities seeking to participate in emerging digital ecosystems (Devlin, 2019). Consequently, researchers have developed specialized machine learning methodologies designed to address data scarcity, linguistic complexity, and limited computational resources while improving technological support for underrepresented languages [64].

Traditional machine learning approaches relied heavily on supervised learning techniques that required large quantities of labelled data for model training. While effective for high-resource languages, these methods are often impractical for indigenous languages where annotated corpora are either unavailable or extremely limited. To overcome this challenge, researchers have increasingly adopted transfer learning methodologies, whereby models pre-trained on large multilingual datasets are adapted to low-resource languages through fine-tuning processes.

Multilingual transformer architectures such as multilingual BERT (mBERT), XLM-Roberta, and other cross-lingual language models have demonstrated promising results in transferring linguistic knowledge from resource-rich languages to resource-constrained environments. These approaches significantly reduce the amount of training data required while improving performance across a variety of NLP tasks, including text classification, machine translation, and information retrieval [65].

Another important methodology involves unsupervised and semi-supervised learning techniques. These approaches leverage unlabelled textual and speech data, which are generally more accessible than manually annotated resources. Through pattern recognition, representation learning, and contextual prediction mechanisms, machine learning models can acquire meaningful linguistic structures without extensive human supervision. Semi-supervised learning further enhances performance by combining small quantities of labelled data with larger collections of

unlabelled materials. For indigenous languages, where community-generated content may exist but remain unannotated, these methods provide practical pathways for expanding computational capabilities while minimizing annotation costs. Studies have shown that semi-supervised learning can substantially improve language modelling accuracy, especially when combined with active learning strategies that prioritize the most informative data samples for human annotation [66].

Neural Machine Translation (NMT) has emerged as one of the most transformative applications of machine learning for low-resource languages. Conventional statistical translation systems often struggle with complex grammatical structures and limited parallel corpora. In contrast, neural architectures employ deep learning mechanisms capable of capturing contextual relationships and semantic representations across languages. Recent innovations include multilingual NMT systems, zero-shot translation models, and subword tokenization techniques that enable translation between languages with minimal bilingual data. For Niger-Congo languages such as Gadé, custom subword segmentation models are particularly valuable because they can effectively handle rich morphological structures, affixation patterns, and compound word formations. By decomposing words into smaller meaningful units, these models reduce vocabulary sparsity and improve translation quality in data-constrained environments [67].

Speech technologies represent another critical area where machine learning methodologies have demonstrated substantial promise for low-resource languages. Automatic Speech Recognition (ASR) systems traditionally require thousands of hours of annotated speech recordings to achieve acceptable performance levels. However, recent developments in self-supervised learning have significantly reduced these requirements. Models such as wav2vec and related architectures learn speech representations from large volumes of unlabelled audio data before being fine-tuned on smaller annotated datasets. This approach is particularly advantageous for indigenous languages where oral traditions remain vibrant despite limited textual documentation. By leveraging community-generated recordings, oral histories, folktales, songs, and conversational speech, self-supervised learning can facilitate the development of robust ASR systems even in highly resource-constrained contexts [68].

For tonal languages, additional methodological considerations are necessary because meaning is frequently encoded through pitch variations rather than segmental phonological structures alone. Conventional speech recognition systems often overlook tonal information, resulting in semantic ambiguities and transcription errors. To address this limitation, researchers have proposed tone-aware machine learning architectures that explicitly incorporate acoustic features such as fundamental frequency (f0), pitch contours, and tonal trajectories into model training [69]. These systems have demonstrated improved performance across several African and Asian tonal languages by preserving critical distinctions that conventional models frequently ignore. In the context of the Gadé language, tone-aware machine learning methodologies offer significant potential for enhancing speech recognition, pronunciation modelling, and semantic accuracy within AI-driven language technologies.

Recent scholarship also emphasizes the importance of community-centered machine learning frameworks for indigenous language development. Rather than relying solely on externally generated datasets, participatory approaches encourage native speakers, educators, cultural custodians, and local institutions to contribute directly to data collection, annotation, validation, and evaluation processes. Such collaborative methodologies not only improve data quality but also promote linguistic authenticity, cultural sensitivity, and indigenous data sovereignty. Community participation ensures that machine learning systems accurately reflect local linguistic practices, dialectal variations, and culturally significant expressions that might otherwise be overlooked by purely technical approaches Obadiah, et al.

Collectively, these machine learning methodologies provide a foundation for developing scalable and sustainable language technologies for low-resource languages. Through the integration of transfer learning, semi-supervised learning, neural machine translation, self-supervised speech modelling, tone-aware architectures, and community-driven data governance, researchers can address many of the challenges that have historically limited indigenous language digitization. For the Gadé language, these approaches form the computational backbone of a comprehensive AI-enabled revitalization strategy, enabling the creation of intelligent language tools that preserve linguistic integrity while facilitating broader digital inclusion in the twenty-first century [70].

Benchmarking Massively Multilingual LLMs on African Morphosyntax

The emergence of Massively Multilingual Large Language Models (LLMs) has transformed the field of Natural Language Processing (NLP) by enabling a single model to process and generate text across hundreds of languages. Models such as multilingual BERT (mBERT), XLM-RoBERTa, BLOOM, mT5, and other multilingual transformer architectures have demonstrated remarkable capabilities in cross-lingual transfer learning, machine translation, question answering, summarization, and conversational AI [71]. Their success has generated optimism regarding the possibility of extending advanced language technologies to historically underrepresented linguistic communities. However, recent studies reveal that the performance of these models varies significantly across languages and language families, particularly when evaluated against the complex grammatical structure's characteristic of many African languages.

Consequently, benchmarking multilingual LLMs on African morphosyntax has become an important area of inquiry for assessing the inclusiveness, fairness, and linguistic competence of contemporary AI systems [72].

Morphosyntax refers to the interaction between morphological structures and syntactic organization within a language. Many African languages exhibit linguistic features that differ substantially from those found in the predominantly Indo-European languages that dominate AI training datasets. These features include extensive affixation systems, noun-class structures, agreement mechanisms, serial verb constructions, tonal distinctions, reduplication patterns, and context-sensitive semantic relationships. Such characteristics present unique computational challenges because they often encode grammatical and semantic information within complex morphological forms rather than through independent lexical units. As a result, language models trained primarily on European and other high-resource languages frequently struggle to accurately represent and process African morphosyntactic patterns [73].

Research evaluating multilingual LLMs on African languages consistently highlights disparities in model performance. While

multilingual architectures demonstrate impressive capabilities in languages with abundant digital resources, their effectiveness declines considerably when applied to low-resource African languages. One contributing factor is data imbalance within training corpora. Most multilingual models are trained on vast quantities of text from dominant global languages, whereas African languages constitute only a small fraction of available training data. Consequently, model representations tend to favour grammatical structures, lexical patterns, and syntactic conventions associated with high-resource languages. This imbalance often results in reduced accuracy for tasks involving morphological analysis, part-of-speech tagging, dependency parsing, and semantic interpretation in African linguistic contexts Several benchmarking initiatives have sought to quantify these limitations through systematic evaluation frameworks [74]. These benchmarks assess model performance across diverse NLP tasks while accounting for language-specific grammatical features. Findings indicate that multilingual LLMs frequently exhibit difficulties in handling agglutinative and polysynthetic structures common in many African languages. In particular, models often fail to accurately segment morphologically complex words, identify grammatical relationships encoded through affixes, or preserve semantic information embedded within morphological markers. Such deficiencies can lead to translation errors, incorrect syntactic interpretations, and reduced performance in downstream language applications. The challenges become even more pronounced when evaluating languages with limited digital representation, where sparse training data further constrain model effectiveness [75].

An additional concern involves the treatment of serial verb constructions, a prominent feature in numerous Niger-Congo languages. Serial verb constructions allow multiple verbs to occur within a single clause without explicit conjunctions, creating semantic relationships that may be difficult for conventional language models to interpret. Benchmarking studies reveal that multilingual LLMs frequently misrepresent these constructions during translation and text generation tasks, resulting in semantic distortions and loss of contextual meaning. Similar challenges arise with noun-class systems, where grammatical agreement patterns depend on complex categorical distinctions that may not align with structures commonly encountered in high-resource languages. These findings underscore the importance of developing evaluation frameworks specifically designed to measure linguistic competence within African morphosyntactic environments [76].

For tonal languages, benchmarking becomes even more complex because morphosyntactic meaning is often intertwined with tonal variation. Most text-based multilingual LLMs operate on orthographic representations that omit or inadequately encode tonal information. Consequently, models may generate grammatically plausible outputs while failing to preserve crucial semantic distinctions conveyed through tone. This limitation has significant implications for languages such as Gadé, where tonal patterns contribute directly to lexical meaning, grammatical interpretation, and discourse coherence. Researchers increasingly advocate the integration of tone-sensitive evaluation metrics capable of measuring a model's ability to preserve both morphosyntactic and tonal information simultaneously [77].

Recent advances in benchmarking methodologies emphasize culturally and linguistically grounded evaluation strategies. Rather than relying exclusively on standard NLP metrics developed for high-resource languages, scholars have proposed task-specific benchmarks that reflect the structural realities of African languages. These include assessments of morphological segmentation, noun-class agreement, tonal preservation, proverb interpretation, culturally embedded semantic expressions, and indigenous discourse patterns. Such approaches recognize that language competence extends beyond lexical prediction and requires sensitivity to the linguistic and cultural contexts within which meaning is constructed. Community participation in benchmark development further enhances evaluation validity by ensuring that performance metrics reflect authentic language use rather than externally imposed standards [78].

Speech Synthesis and Tonal Modelling in Benue-Congo Languages

Speech synthesis and tonal modelling have emerged as critical areas of research in the development of language technologies for African languages, particularly those belonging to the Benue-Congo language family. As one of the largest linguistic branches within the Niger-Congo phylum, Benue-Congo encompasses hundreds of languages spoken across West, Central, and Southern Africa. These languages exhibit remarkable linguistic diversity while sharing several structural characteristics, including extensive tonal systems, complex morphology, noun-class features, and context-dependent semantic patterns. Among these features, tone remains one of the most challenging aspects for computational modelling because variations in pitch frequently carry lexical, grammatical, and pragmatic meaning. Consequently, effective speech synthesis systems for Benue-Congo languages must move beyond conventional text-to-speech approaches and incorporate sophisticated tonal modelling mechanisms capable of preserving linguistic accuracy and naturalness.

Speech synthesis, commonly referred to as Text-to-Speech (TTS) technology, involves the automatic conversion of written text into intelligible and natural-sounding speech. Modern TTS systems have achieved impressive performance in major world languages through the application of deep learning techniques, neural acoustic modelling, and end-to-end speech generation architectures. However, these advances have not been evenly distributed across the global linguistic landscape. Most state-of-the-art speech synthesis systems are optimized for high-resource languages with extensive digital corpora, leaving many African languages technologically underserved. For Benue-Congo languages, the scarcity of annotated speech datasets, standardized orthographies, and computational resources has significantly constrained the development of high-quality speech technologies. As a result, speakers of these languages often lack access to voice-enabled applications, digital assistants, accessibility tools, and educational technologies available in more widely represented languages [79].

Research on tonal modelling has increasingly focused on developing computational methods capable of preserving pitch information throughout the speech generation process. Early approaches relied on rule-based systems that manually encoded tonal patterns according to predefined linguistic descriptions. While effective for limited applications, these systems struggled to accommodate contextual variation and natural speech dynamics. More recent studies have adopted statistical and neural methodologies that learn tonal patterns directly from speech data. Deep learning architectures such as recurrent neural networks, transformer-based models, and neural vocoders have demonstrated significant potential in modelling complex tonal relationships while generating more natural and contextually appropriate speech outputs. These systems can capture subtle interactions between tone, syntax, morphology, and discourse context, thereby

improving both intelligibility and linguistic authenticity [80].

The importance of tonal modelling becomes particularly evident when examining languages within the Benue-Congo family. Many of these languages utilize multiple level and contour tones that interact with grammatical structures in highly systematic ways. Tonal alternations may signal tense, aspect, plurality, emphasis, focus, or other grammatical distinctions. Furthermore, tone often operates within broader phonological processes involving tone spreading, tone assimilation, downdrift, and tonal terracing. These phenomena introduce additional layers of complexity that challenge conventional computational frameworks. Effective speech synthesis systems must therefore account not only for isolated tonal categories but also for dynamic tonal interactions occurring across words, phrases, and discourse units. Failure to model these relationships accurately can result in speech outputs that sound unnatural or misrepresent intended meanings [81].

Digital Sovereignty, Data Colonization, and Indigenous Agency

The rapid expansion of Artificial Intelligence (AI), Big Data, and digital technologies has generated unprecedented opportunities for the documentation, preservation, and revitalization of indigenous languages. However, these technological developments have also raised critical concerns regarding ownership, control, access, and governance of indigenous linguistic and cultural data. As indigenous communities increasingly participate in digital ecosystems, questions surrounding digital sovereignty, data colonization, and indigenous agency have become central to contemporary debates in language technology, digital humanities, and ethical AI. Scholars argue that while digitization can empower marginalized communities by enhancing visibility and access to knowledge, it can also reproduce historical patterns of extraction and exclusion if indigenous peoples are denied meaningful control over their digital resources. Consequently, the development of language technologies for indigenous communities must be guided not only by technical considerations but also by principles of equity, self-determination, and cultural autonomy [82].

Digital sovereignty refers to the right of communities, nations, or cultural groups to exercise authority over their digital assets, technological infrastructures, and information systems. Within indigenous contexts, digital sovereignty extends beyond questions of data storage and cybersecurity to encompass ownership of linguistic resources, cultural knowledge, oral traditions, and intellectual heritage represented within digital environments. It affirms the principle that indigenous communities should retain the authority to determine how their knowledge is collected, stored, shared, interpreted, and utilized by external actors. This perspective challenges conventional models of technological development in which data is often treated as an open resource available for unrestricted extraction and commercial exploitation. For indigenous language communities, digital sovereignty represents a mechanism for protecting cultural integrity while ensuring that technological innovation aligns with community priorities and values [83].

The principle of Indigenous Data Sovereignty (IDS) has emerged as a leading response to concerns surrounding data colonization and digital exclusion. Indigenous Data Sovereignty advocates for indigenous governance over data relating to indigenous peoples, territories, cultures, and languages. It emphasizes that indigenous communities possess inherent rights to determine how their data is collected, managed, interpreted, and shared. Several international frameworks have reinforced these principles by recognizing indigenous rights to cultural preservation, self-determination, and control over traditional knowledge systems. Within AI research, Indigenous Data Sovereignty encourages the development of ethical protocols that prioritize transparency, informed consent, community benefit, accountability, and equitable access to technological outcomes [84].

For the Gadé language, the principles of digital sovereignty and indigenous agency are particularly relevant to the proposed digitization model. As efforts intensify to develop AI-powered tools for speech recognition, machine translation, corpus development, and language learning, it is essential that Gadé communities retain meaningful control over linguistic resources and cultural content. The digitization process should be governed through collaborative partnerships involving native speakers, traditional authorities, educators, researchers, and cultural institutions such as the Institute of Gade and Classical Studies. Such partnerships can ensure that data collection methodologies respect cultural sensitivities, intellectual property rights, and community expectations. Moreover, locally governed digital repositories can provide secure mechanisms for preserving linguistic materials while supporting educational, research, and cultural objectives.

Linguistic Profiling and Computational Challenges of Gadé

The successful development of Artificial Intelligence (AI)-driven language technologies depends fundamentally on a comprehensive understanding of the linguistic characteristics of the target language. Linguistic profiling provides the descriptive foundation necessary for designing computational systems capable of accurately representing, processing, and generating language data. For indigenous and low-resource languages, this process is particularly important because existing computational frameworks are often developed around linguistic assumptions derived from high-resource languages, many of which differ substantially in phonology, morphology, syntax, semantics, and discourse structure. Consequently, language technologies designed without consideration of language-specific features frequently exhibit poor performance, semantic inaccuracies, and cultural misrepresentation. In the case of Gadé, a language spoken predominantly in North-Central Nigeria, linguistic profiling is essential for identifying the structural characteristics that influence computational processing and for developing AI systems capable of supporting effective language digitization and revitalization [85-87].

Phonological Characteristics of Gadé

One of the most significant features of the Gadé language is its tonal nature. Tone functions as a phonemic feature, meaning that variations in pitch can alter the meaning of words, grammatical relationships, and communicative intent. Unlike non-tonal languages where pitch primarily serves prosodic functions such as emphasis or emotional expression, tonal distinctions in Gadé contribute directly to semantic interpretation. A single lexical form may convey multiple meanings depending on the tonal pattern applied. This characteristic creates substantial challenges for computational systems because conventional speech recognition and text-processing technologies often fail to capture or represent tonal information accurately (GWF, et al). For example, Yé 'Search, find' and Yè 'Capacitated, Intentional'–Ụ nì yè "S/he is capacitated, intended".

In addition to tonal complexity, Gadé possesses a rich phonological inventory comprising vowels, consonants, and suprasegmental features that interact in linguistically significant ways. Certain phonological distinctions may not correspond directly to sounds

represented in English or other widely digitized languages. As a result, speech technologies trained on dominant language datasets often struggle to recognize or synthesize Gadé speech with acceptable levels of accuracy. Effective computational modelling therefore requires the development of acoustic representations that explicitly account for tonal contours, pitch variations, pronunciation patterns, and phonological processes unique to the language.

Morphological Structure and Word Formation

Morphology represents another critical dimension of Gadé linguistic structure. Like many Benue-Congo languages, Gadé exhibits productive word formation processes that enable speakers to express complex meanings through morphological modifications. Words may incorporate prefixes, suffixes, reduplication patterns, and other morphological elements that contribute grammatical and semantic information. These structures frequently encode distinctions relating to tense, aspect, plurality, possession, emphasis, and relational meaning.

From a computational perspective, rich morphology presents several challenges. Traditional NLP systems often rely on word-based processing approaches that assume relatively stable lexical boundaries. In morphologically complex languages, however, a single word may contain information equivalent to an entire phrase or clause in a less morphologically rich language. This increases vocabulary size and contributes to data sparsity problems, particularly in low-resource environments. Machine learning systems trained on limited datasets may therefore struggle to recognize morphological variations and generate linguistically appropriate outputs. Addressing these challenges requires computational models capable of analyzing words at subword and morphemic levels rather than relying exclusively on whole-word representations.

Syntactic Features and Sentence Structure

The syntactic organization of Gadé constitutes another important consideration for computational language development. While sharing certain characteristics with other Niger-Congo languages, Gadé possesses distinctive sentence structures, grammatical relationships, and discourse patterns that influence meaning construction. Word order, argument structure, modification strategies, and clause-linking mechanisms contribute to the overall grammatical architecture of the language.

Particularly significant are constructions that may differ from the syntactic conventions commonly represented in multilingual language models. For example, contextual interpretation often depends upon grammatical relationships that cannot be adequately captured through direct word-to-word translation. Such features create difficulties for machine translation systems that rely heavily on statistical correspondences derived from high-resource language pairs. Furthermore, syntactic ambiguity may arise when computational models attempt to apply grammatical assumptions derived from unrelated linguistic traditions. Consequently, language-specific syntactic analysis is necessary to ensure accurate parsing, translation, and text generation.

Semantic Complexity and Indigenous Knowledge Representation

The semantic system of Gadé reflects the cultural experiences, historical development, and worldview of its speakers. Many lexical items and expressions derive their meaning from indigenous knowledge systems, social practices, environmental interactions, and philosophical concepts embedded within Gadé society. Cultural expressions associated with Gaboism, oral traditions, customary institutions, kinship structures, and ecological knowledge frequently contain layers of meaning that extend beyond literal translation.

This semantic richness presents important challenges for computational systems. Conventional machine translation models often prioritize lexical equivalence while overlooking contextual, cultural, and conceptual dimensions of meaning. As a result, culturally significant expressions may be mistranslated, oversimplified, or rendered incomprehensible. The problem is particularly pronounced when AI systems are trained primarily on datasets derived from unrelated cultural contexts. Effective language technologies for Gadé must therefore incorporate mechanisms capable of preserving semantic nuance, contextual interpretation, and culturally embedded knowledge structures.

Oral Tradition and Speech-Centred Knowledge Systems

A distinguishing characteristic of the Gadé language community is the central role of oral tradition in cultural transmission. Historical narratives, genealogies, folktales, proverbs, songs, ritual discourse, and philosophical teachings have traditionally been preserved and transmitted through spoken communication. These oral forms serve as repositories of communal knowledge and cultural identity.

For computational linguistics, the predominance of oral traditions presents both opportunities and challenges. On one hand, oral resources provide valuable data for speech recognition, speech synthesis, and corpus development. On the other hand, oral materials often exist in formats that require extensive recording, transcription, annotation, and digitization before they can be incorporated into machine learning systems. Variations in pronunciation, speaking styles, and performance contexts further complicate data processing. The creation of high-quality speech corpora therefore represents a foundational requirement for AI-enabled Gadé language technologies.

Data Scarcity and Corpus Limitations

Perhaps the most significant computational challenge affecting Gadé is the scarcity of digital linguistic resources. Modern AI systems depend heavily on large quantities of structured data for training, validation, and evaluation. However, many indigenous languages, including Gadé, lack the extensive digital corpora available for high-resource languages. Existing textual materials may be fragmented across educational documents, religious texts, cultural publications, personal archives, and community records [88-97].

The limited availability of digitized data constrains the development of language models, machine translation systems, speech technologies, and educational applications. Furthermore, the absence of standardized annotation protocols complicates corpus construction and quality assurance processes. Overcoming these limitations requires coordinated efforts involving data collection, digitization, linguistic annotation, and community participation. The development of a comprehensive Gadé digital corpus is therefore a strategic priority within the broader digitization framework proposed in this study.

Computational Implications for AI Development

The linguistic characteristics discussed above collectively shape the computational requirements for Gadé language technologies. Tonal complexity necessitates tone-aware speech recognition

and synthesis systems. Rich morphology requires subword tokenization and morphology-sensitive language models. Distinctive syntactic structures demand customized parsing and translation frameworks. Semantic complexity calls for culturally informed NLP methodologies capable of preserving indigenous meanings and knowledge systems. Meanwhile, the predominance of oral traditions emphasizes the importance of speech-centered computational resources.

These considerations suggest that conventional AI architectures cannot simply be transferred to Gadé without substantial adaptation. Instead, effective digitization requires the development of specialized computational frameworks tailored to the language's structural and cultural characteristics. Such frameworks must integrate linguistic expertise, community knowledge, and advanced machine learning methodologies in order to achieve meaningful and sustainable outcomes.

The linguistic profile of Gadé reveals a language of considerable structural richness, cultural significance, and computational complexity. Its tonal phonology, productive morphology, distinctive syntax, culturally embedded semantics, and strong oral traditions present both challenges and opportunities for AI-driven language revitalization. While these characteristics complicate the application of conventional NLP methodologies, they also underscore the importance of developing specialized computational approaches capable of preserving linguistic authenticity and cultural integrity. Understanding the linguistic foundations of Gadé is therefore essential for the design of effective speech technologies, machine translation systems, digital corpora, and educational applications. The insights generated through this linguistic profiling exercise provide the conceptual and technical basis for the Gadé Language Digitization Model proposed in subsequent sections of this study.

Phonological Architecture and Contrastive Tonal Topology

The phonological architecture of the Gadé language constitutes a foundational component of its linguistic identity and a primary determinant of its computational behavior within Natural Language Processing (NLP) systems. As with many Benue-Congo languages, Gadé exhibits a structurally complex sound system in which segmental and suprasegmental features operate in close interdependence. However, what distinguishes Gadé most significantly is its contrastive tonal topology, wherein pitch variation functions not as a secondary prosodic feature but as a primary lexical and grammatical encoding mechanism. This tonal system introduces a multilayered phonological structure that requires specialized modelling approaches for accurate representation in computational systems. Understanding this architecture is essential for designing speech recognition, speech synthesis, and machine translation systems capable of preserving both semantic integrity and phonetic fidelity.

At the segmental level, Gadé phonology is organized around a system of consonants and vowels that interact with syllable structures to form meaningful lexical units. While the language shares typological similarities with other Niger-Congo languages, including relatively simple syllable templates and a preference for open syllables in certain phonotactic environments, its distinctive identity emerges through the interaction of these segmental units with tonal patterns. Consonantal articulation and vowel quality provide the structural scaffolding upon which tonal contrasts are superimposed. This dual-layered structure—segmental and tonal—creates a phonological system in which meaning is distributed across multiple acoustic dimensions, thereby increasing both expressive richness and computational complexity.

The tonal topology of Gadé can be conceptualized as a system of pitch-based contrasts that operate at both lexical and grammatical levels. In such a system, tonal variation is not random but systematically organized, with specific pitch contours associated with distinct semantic outcomes. Minimal pairs differentiated solely by tone demonstrate the phonemic status of pitch in the language, where identical segmental sequences yield entirely different meanings depending on tonal configuration. This characteristic places Gadé within the class of tone languages in which pitch functions as an integral component of lexical identity. Furthermore, tonal patterns may extend beyond individual lexical items to influence phrase-level and discourse-level interpretation, indicating a hierarchical organization of tonal behavior that spans multiple linguistic domains.

From a computational perspective, contrastive tonal topology introduces significant challenges for acoustic modelling and speech processing. Conventional Automatic Speech Recognition (ASR) systems often rely on spectral features that prioritize segmental phonetic information while treating pitch as a secondary or optional parameter. In contrast, Gadé requires a modelling framework in which fundamental frequency (f0) trajectories are treated as primary discriminative features. The failure to incorporate tonal contrasts into acoustic representations can lead to severe semantic ambiguity, as identical phonetic transcriptions may correspond to multiple meanings depending on tonal realization. Consequently, tone-aware modelling is not an enhancement but a necessity for accurate linguistic representation in Gadé language technologies.

The dynamic nature of tonal realization further complicates computational modelling. Tonal contours in Gadé may undergo contextual modifications due to phonological processes such as tone sandhi, downdrift, assimilation, and register shifts. These processes involve systematic alterations of tonal patterns depending on surrounding phonological or syntactic environments.

Another critical dimension of Gadé phonological architecture lies in the interaction between tone and morphological structure [98]. In many instances, tonal distinctions serve as markers of grammatical categories such as tense, aspect, mood, plurality, or focus. This morphotonemic relationship implies that tone is not only a lexical feature but also a grammatical signalling mechanism. Consequently, any computational system designed for Gadé must account for the dual functional role of tone as both a lexical and morphological feature. Failure to do so may result in incomplete or inaccurate language modelling, particularly in tasks involving translation or syntactic parsing.

Lexical Disambiguation: Semantic Distinctions in Gabo and Gobo

Lexical disambiguation in the Gadé language presents important insights into how meaning is structured, differentiated, and contextually interpreted within indigenous semantic systems. A key illustration of this phenomenon is the contrast between the lexical items Gabo and Gobo, which, despite their phonological similarity, carry distinct semantic and cultural meanings depending on tonal configuration, contextual usage, and discourse environment. This distinction highlights the broader linguistic principle that meaning in Gadé is not solely encoded through segmental phonemes but is significantly shaped by tonal variation and pragmatic context.

At a structural level, Gabo and Gobo exemplify minimal or near-minimal lexical pairs were subtle differences in vowel quality, tone, or phonetic realization result in divergent semantic interpretations.

In tonal languages such as Gadé, such lexical contrasts are not incidental but systematic, forming part of a broader lexical network where tone serves as a primary disambiguating feature. As a result, accurate interpretation of meaning requires sensitivity to both acoustic properties and contextual cues. Without tonal or contextual awareness, computational systems risk conflating semantically distinct expressions, leading to errors in translation, speech recognition, and semantic modelling.

From a semantic perspective, these lexical distinctions are deeply embedded within cultural and epistemological frameworks associated with Gaboism, where language functions as a carrier of indigenous worldview and social meaning. Words like Gabo and Gobo may encode different conceptual domains, social references, or metaphorical associations depending on usage. This reinforces the idea that lexical meaning in Gadé is not fixed but dynamically constructed through interaction between speaker intention, tonal realization, and cultural context. Such complexity challenges traditional NLP models that rely primarily on surface-level lexical matching without accounting for deeper semantic variability [99].

Morphosyntactic Complexity: Noun-Class Systems and Serial Verbs

The morphosyntactic structure of the Gadé language reflects a high degree of grammatical complexity characteristic of many languages within the Benue-Congo branch of the Niger-Congo family. Two of the most significant features contributing to this complexity are the noun-class system and serial verb constructions. These grammatical mechanisms play a central role in sentence formation, agreement patterns, and semantic interpretation, while simultaneously presenting considerable challenges for computational modelling in Natural Language Processing (NLP) systems. Understanding these structures is essential for designing AI-driven language technologies that can accurately represent Gadé syntax and preserve its linguistic integrity in digital environments.

The noun-class system in Gadé organizes nouns into semantically and morphologically defined categories that influence agreement across related grammatical elements such as verbs, adjectives, pronouns, and determiners (Sterk, 1994). Unlike simple gender systems found in many Indo-European languages, noun-class systems in Niger-Congo languages are typically more extensive and semantically nuanced, often encoding distinctions related to animacy, shape, size, collectivity, or abstract conceptual categories [100]. In Gadé, noun classes are not merely classificatory but function as integral components of grammatical structure, ensuring cohesion across syntactic units through agreement markers. This system requires that modifiers and related grammatical elements conform to the class of the noun they reference, thereby creating a network of morphosyntactic dependencies that enhance meaning precision but increase structural complexity.

From a computational perspective, noun-class systems pose significant challenges for NLP systems that are primarily designed around languages with limited agreement morphology. Standard tokenization and parsing algorithms often fail to capture the hierarchical relationships embedded in noun-class agreements, leading to errors in syntactic analysis and machine translation outputs. Accurate modelling of Gadé noun classes therefore requires specialized annotation schemes and dependency parsing frameworks that can represent agreement relations across sentence structures. Additionally, machine learning models must be trained to recognize morphological markers that signal noun-class membership, even when such markers are phonetically subtle or context-dependent.

Serial verb constructions represent another defining feature of Gadé morphosyntax, contributing to the language's expressive richness and syntactic flexibility [101]. In serial verb constructions, multiple verbs are combined within a single clause without overt markers of subordination or coordination. These verbs typically share a common subject and collectively express a sequence of actions, cause-effect relationships, manner specifications, or directional meanings. Rather than being treated as independent clauses, serial verb constructions function as integrated semantic units that convey complex event structures in a compact grammatical form. This feature allows speakers to encode detailed actions and relationships efficiently, reflecting a high degree of linguistic economy and cognitive sophistication.

The computational modelling of serial verb constructions presents substantial difficulties for conventional NLP systems. Standard syntactic parsers, particularly those trained on Indo-European language datasets, often interpret serial verbs as separate clauses or incorrectly insert artificial conjunctions, thereby distorting the intended meaning. This misinterpretation arises because many NLP architectures assume explicit syntactic markers for clause boundaries, which are absent in serial verb structures. As a result, machine translation systems may produce fragmented or semantically inaccurate outputs when processing Gadé sentences containing serial verbs. Addressing this issue requires the development of language-specific parsing models that can recognize serial verb sequences as unified predicate structures rather than independent syntactic units.

The interaction between noun-class systems and serial verb constructions further increases the morphosyntactic complexity of Gadé. Noun-class agreement patterns may extend across multiple verbs within serial constructions, creating layered dependencies that challenge linear parsing approaches. Additionally, tonal variation may interact with both noun-class marking and verb serialization, adding another layer of phonological influence on syntactic interpretation. This multi-dimensional interaction between morphology, syntax, and phonology highlights the need for integrated computational models capable of capturing cross-linguistic dependencies across different linguistic levels.

Dialectal Variability Across Baponu, Northern, and Southern Sectors

Dialectal variation constitutes an important dimension of the internal structure of the Gadé language, reflecting its geographic spread, historical development, and sociocultural differentiation. The language is not monolithic; rather, it exists as a continuum of mutually intelligible but phonetically, lexically, and syntactically differentiated varieties across major regional groupings, commonly identified as the Baponu, Northern, and Southern sectors. These dialectal divisions are shaped by patterns of settlement, migration, intergroup contact, and localized cultural practices, all of which contribute to subtle but meaningful linguistic variation. Understanding this dialectal diversity is essential for computational modelling, as it directly influences data consistency, corpus design, and the performance of AI-driven language technologies [62].

The Baponu dialectal cluster is often regarded as a central or reference variety due to its relatively widespread usage and perceived linguistic stability in comparison to other regional forms [62]. It typically serves as a communicative bridge among speakers from different sectors, facilitating inter-dialectal intelligibility. Nevertheless, even within the Baponu variety,

internal variation exists in pronunciation, lexical preference, and tonal realization. Such variation reflects localized identity markers and sociolinguistic stratification, demonstrating that dialects function not only as linguistic systems but also as expressions of cultural belonging. For computational systems, the Baponu variety often provides a useful baseline for corpus development, although reliance on a single dialect risk underrepresenting the full linguistic diversity of the language [62].

The Northern sector exhibits distinctive phonological and lexical features that set it apart from other dialects. These may include variations in tonal contour realization, vowel quality shifts, and the presence of region-specific lexical items that encode environmental, cultural, or social realities unique to northern communities. In some cases, syntactic preferences may also differ subtly, particularly in sentence structuring and the use of discourse markers. These differences, while not necessarily impeding mutual intelligibility, introduce challenges for language standardization and computational consistency. For machine learning systems, such variability increases the complexity of training datasets and requires careful balancing to avoid bias toward more dominant dialectal forms [62].

The Southern sector demonstrates its own set of linguistic characteristics, often influenced by distinct patterns of language contact, cultural practices, and historical development. Lexical borrowing, phonological adaptation, and stylistic variation may be more pronounced in this variety, depending on the extent of interaction with neighbouring linguistic communities.

Additionally, tonal patterns in the Southern dialect may exhibit subtle shifts in pitch range or contour distribution, which can affect meaning interpretation in tone-sensitive computational systems. These variations underscore the importance of capturing fine-grained phonetic detail in speech corpora used for AI training, particularly in tasks involving Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) synthesis.

From a computational linguistics perspective, dialectal variability presents significant challenges for corpus design, AI and computing model training, and evaluation. Machine learning systems typically perform best when trained on large, homogeneous datasets; however, the presence of multiple dialects introduces data heterogeneity that can reduce model accuracy if not properly managed. In the case of Gadé, failure to account for dialectal variation may result in biased language models that overfit to a single dialect while performing poorly on others. This issue is particularly critical in low-resource settings, where data scarcity already limits the robustness of AI systems. Consequently, dialect-aware modelling approaches are essential for ensuring equitable and accurate language representation.

Addressing dialectal variability requires the implementation of stratified corpus development strategies that capture linguistic data from all major sectors, including Baponu, Northern, and Southern varieties. Such corpora should be annotated with metadata indicating dialectal origin, phonological features, and contextual usage patterns. This enables machine learning models to learn both shared linguistic structures and dialect-specific variations. Additionally, multi-dialectal modelling approaches, such as shared encoder architectures with dialect-specific adapters, can improve system adaptability while preserving linguistic diversity. These techniques allow AI systems to generalize across dialects without erasing their distinct identities.

Dialectal variability also has important implications for language standardization efforts. While standardization can facilitate literacy development, educational material production, and computational consistency, it must be approached cautiously to avoid privileging one dialect over others. In the Gadé context, inclusive standardization frameworks should aim to preserve dialectal diversity while establishing functional norms for communication and digital representation. Community involvement is essential in this process to ensure that decisions regarding orthography, vocabulary selection, and grammatical conventions reflect collective linguistic preferences rather than external imposition [69].

The dialectal variability across the Baponu, Northern, and Southern sectors of Gadé reflects the dynamic and adaptive nature of the language. While these variations enrich linguistic diversity and cultural expression, they also introduce significant computational challenges for AI-driven language technologies. Effective digital representation of Gadé requires dialect-sensitive approaches that integrate comprehensive data collection, advanced machine learning techniques, and community-led linguistic governance. By acknowledging and incorporating dialectal diversity, language technologies can better support inclusive revitalization efforts and ensure that all speech communities within the Gadé linguistic continuum are accurately represented in the digital age [70].

Discussion and Methodology: The Gade Digitization Framework

The Gade Digitization Framework represents an integrated methodological response to the linguistic, computational, and cultural challenges identified in previous sections of this study. It is designed as a hybrid system that combines Artificial Intelligence (AI), community-based linguistic documentation, and Indigenous Data Sovereignty principles to support sustainable revitalization of the Gade language. Unlike conventional NLP pipelines that prioritize data abundance and statistical generalization, the Gade framework is grounded in low-resource language realities, emphasizing data quality, tonal precision, morphosyntactic integrity, and cultural authenticity. This section presents the conceptual design, methodological architecture, and operational workflow of the proposed system.

Conceptual Foundation of the Framework

The Gadé Digitization Framework is built on four interdependent theoretical pillars: (1) computational linguistics for low-resource languages, (2) tone-aware speech processing, (3) community-driven data governance, and (4) culturally grounded AI design. These pillars ensure that technological development does not occur in isolation from linguistic reality or cultural context. The framework rejects extractive data practices and instead promotes participatory co-creation, where native speakers, linguists, and technologists jointly shape the structure and content of the digital language ecosystem.

At its core, the framework recognizes that Gadé is not merely a dataset to be processed but a living linguistic system embedded within cultural, philosophical, and historical contexts. Therefore, the digitization process must preserve not only lexical and syntactic structures but also tonal meaning, oral traditions, and culturally embedded semantic knowledge systems such as Gaboism.

Methodological Architecture

The methodology is structured into five interconnected layers:

Data Acquisition Layer

This layer focuses on systematic collection of linguistic data from diverse sources, including:

- Oral narratives (folktales, proverbs, historical accounts) otherwise known as pipirige
- Everyday conversational speech
- Ritual and ceremonial discourse (Gápyā, and tụvā)
- Educational materials and community records (Ibé yà Kakaranta nị iré nẹ).

Data acquisition is conducted through community participation, ensuring ethical consent, representativeness, and dialectal inclusivity across Baponu, Northern, and Southern varieties. Audio recordings are prioritized to capture tonal and prosodic features that are often absent in written corpora.

Data Processing and Annotation Layer

Collected data undergoes multi-level annotation, including: Phonetic transcription with tone marking, Morphological segmentation, Syntactic parsing, Semantic tagging, Dialect classification metadata.

Special emphasis is placed on tone annotation using fundamental frequency (f0) tracking, ensuring that tonal distinctions are preserved in digital form. Annotation is performed through a hybrid model combining automated tools and human linguistic validation by native speakers and trained annotators.

Corpus Construction Layer

The annotated data is organized into a structured digital corpus comprising: Parallel text-speech datasets, Monolingual Gadé corpora, Bilingual alignment datasets (Gadé–English and other regional languages), Dialect-specific sub-corpora.

This corpus serves as the foundational resource for training machine learning models. It is continuously expanded through iterative community contributions and periodic validation cycles.

AI Model Development Layer

This layer integrates multiple machine learning components tailored for Gadé: Tone-aware Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), Text-to-Speech (TTS) systems, Context-aware language modelling.

Model architectures incorporate transformer-based systems enhanced with tone embeddings, subword tokenization, and morphosyntactic constraints. Self-supervised learning techniques are used to mitigate data scarcity, while transfer learning enables adaptation from multilingual models [75].

Deployment and Application Layer

The final layer focuses on real-world application of the developed technologies through: Mobile language learning applications, Digital dictionaries and lexicons, Speech-enabled educational tools, Conversational AI assistants [76]. Cultural preservation platforms. These applications are designed for both offline and online use to accommodate varying levels of digital access within Gadé-speaking communities.

Community-Centred Methodology

A defining feature of the Gadé Digitization Framework is its emphasis on community participation. Unlike traditional NLP systems that rely primarily on external datasets, this model positions native speakers as co-researchers and data custodians. Elders, educators, and cultural practitioners play a central role in validating linguistic data, ensuring cultural accuracy, and guiding ethical decisions related to language representation.

This participatory approach enhances data authenticity while reinforcing Indigenous Data Sovereignty. It also ensures that technological outputs align with community expectations, linguistic norms, and cultural sensitivities. The framework therefore functions not only as a computational system but also as a social and cultural collaboration platform [80].

Addressing Key Computational Challenges

- The Gadé Digitization Framework directly responds to previously identified challenges:
- Tonal complexity is addressed through tone-aware ASR and acoustic modelling.
- Morphological richness is handled using subword tokenization and morphological analyzers.
- Dialectal variation is managed through multi-dialect corpus segmentation and adaptive modelling.
- Data scarcity is mitigated via transfer learning, self-supervised learning, and community-driven data expansion.
- Semantic depth is preserved through culturally informed annotation and contextual embedding techniques.

Ethical and Governance Structure

The framework incorporates a governance model based on transparency, accountability, and community ownership. Data usage policies are determined collaboratively with Gadé stakeholders, ensuring that linguistic resources are not exploited without consent or benefit-sharing. Intellectual property rights over linguistic data are jointly held by the community and institutional partners such as the Institute of Gade and Classical Studies (2026), Gade-English Dictionary Contributors.

Thematic Discussion of Gade Words Formation and Datasets

The Gadé Digitization Framework demonstrates that effective AI-driven language revitalization requires more than computational innovation; it demands an integrated approach that combines technology, linguistics, ethics, and community engagement. By aligning machine learning methodologies with indigenous knowledge systems, the framework addresses both technical limitations and socio-cultural imperatives [80]. The methodology presented in this section establishes a comprehensive pathway for the digitization and revitalization of the Gadé language. By integrating AI technologies with community governance and linguistic expertise, the framework offers a scalable and sustainable model for indigenous language preservation. It provides a foundation for developing tone-aware, morphologically sensitive, and culturally aligned language technologies that can support long-term linguistic survival in the digital age.

See Appendix C: Glossary of Cultural Metadata and philosophical Terms.

Adakpụ Ancestral Guidance – The Unending Voice of Ancestors (Obadiah, 2026)

From time immemorial to the contemporary era, every Gade community has acknowledged the existence of ancestral spirits, deities, or the Supreme Being who watches over and protects members of families and communities, irrespective of their social status. Although there are numerous deities and spiritual beings, they are ultimately connected to a single Supreme Being known as Sigapa, the Owner of Heaven and Earth.

What is Sigapa?

Like many other words in the Gade language, Sigapa may be understood through its lexical and philosophical formation. Gade adjectives and related expressions are often formed from root or base words through the use of prefixes, suffixes, and other affixational processes. The word Sigapa is interpreted on the existential principles of form and substance.

The element "Sị" in Sigapa functions as a conjunction or relational particle meaning "and" or "with." It suggests a relationship between two entities or states of existence. In this sense, Sị represents a phrase of existential connection.

This understanding may be appreciated alongside Plato's Theory of Forms, in which every physical object is regarded as an imperfect reflection of an eternal and perfect reality. Every beautiful object reflects the ideal Form of Beauty (Yériyé/Sèré); every just law reflects the ideal Form of Justice (Kẹ́wá); and every manifestation of darkness or evil reflects the concepts of Darkness (Gùbí) or Evil (Bìbìñsá). According to Plato, genuine knowledge is attained through the understanding of these eternal Forms rather than their imperfect physical manifestations.

Within Gade philosophy of forms, Sigapa represents the supreme knowledge and eternal Form from which every other form and substance derives its existence. The following illustrations demonstrate this philosophical relationship:

- Sị gada — together with the earth, south, or downward direction, where gada figuratively represents the earthly realm rather than the literal meaning of "going downward" (Sí gada). The tonal distinction should therefore be carefully observed.
- Sị gapa — together with heaven, north, or the upward direction, where gapa figuratively represents heaven or the sky (Nkyi).

Rather than treating Sị gada and Sị gapa as separate expressions referring to God or a deity, Gade philosophical thought recognises their conceptual unity in the single expression Sigapa (Sị + gapa), representing the Supreme Being who transcends both heaven and earth.

According to Gade creation mythology, the ancestors were entrusted with authority to inhabit, cultivate, and govern the earthly realm (gada), while gapa remained the dwelling place of Sigapa, the Supreme Being. Within Gade philosophical tradition, collectively known as Adakpụ, the ancestors serve as custodians of culture, traditions, Gaboism, moral values, social order, and the continuity of communal existence.

The origin of spirits and deities within Gade cosmology has been preserved through multiple channels, including oral tradition, written texts, folklore, cultural practices, religious rituals, symbols of worship, and various forms of spiritual embodiment. These ancestral beings may be invoked through prayers, praise, veneration, and sacrificial rites performed for specific purposes and at prescribed times according to established cultural traditions.

The term Adakpụ literally means "Ancestral Father" or "Ancient Father," and may also be understood as "Great-grandfather." Etymologically, the word is derived from two Gade lexical roots: Adā, meaning father, and kpụ, meaning ancient or old.

The origin of this cosmological concept can only be appreciated through an understanding of the foundational history of the Gadé people. The earliest ancestors—collectively known as Bàpọ́nè, meaning those who came before us and now rest in the grave—were the originators of the Gade language and its philosophical vocabulary. Their legacy continues through language, customs, and collective memory.

A more comprehensive discussion of these ideas may be found in Adakpụ Mythology, Gaboic Theory of Evolution, Religious Theories of Creation, and the Big Bang Theory, where the origin of life is examined in relation to the Garden of Life and Purposes (Asam).

The Journey of the Soul and the Origin of Adakpụ

According to Gadé cosmology, the departure of the soul from the physical body does not mark the end of existence. Rather, it signifies the beginning of a spiritual journey through successive stages that culminate in the afterlife. This transition begins at burial and continues as the physical body gradually returns to the earth through natural decomposition.

Within Gadé traditional knowledge, it is generally believed that the soul requires three (3) years to complete its transition into the spiritual realm and reunite with the ancestors. At the completion of this three-year period, the surviving members of the deceased's family perform rites of remembrance at the grave. These rites include prayers, praise, sacrifices, and supplications, through which the departed is honoured and invoked to continue protecting the family and community. Depending on the circumstances, these ceremonies may also include solemn declarations against injustice or evil believed to exist among the living.

Upon completing this transitional period, the departed soul is regarded as having fully entered the ancestral realm and may subsequently participate in the process of reincarnation. According to Gadé belief, there may be a spiritual contest regarding lineage, whereby a soul may reincarnate through either the paternal or maternal line. This reincarnated manifestation is known as Gukwu, representing the return of an ancestor in a new earthly existence.

The veneration of Gukwu forms an important aspect of Gade ancestral philosophy. It is generally understood that every Gadé person participates in this continuous cycle of existence, thereby ensuring that the legacy, virtues, responsibilities, and identity of the ancestors are perpetuated across generations.

The Place of Origin and the Qualifications for Becoming an Ancestor (Adakpụ)

Consistent with the philosophical conception of forms, the origin of Adakpụ is rooted in the universal knowledge of existence itself. Within Gadé cosmology, the dwelling place of the ancestors is known as Àsám̄ (Asam), commonly translated as the Garden of Life and Purposes.

Asam is regarded as the spiritual abode of departed souls, ancestral beings, and other spiritual entities. It is the realm in which purification and sanctification occur after death and where souls receive the consequences of their earthly lives. In Gadé mythology, Asam is also remembered as the primordial dwelling place of humanity. However, following the breakdown of divine order, Sigapa reserved Asam as the abode of souls in the afterlife.

According to this understanding, every soul ultimately returns to Asam after death, where it receives its reward according to the life it lived on earth. Those who lived righteous lives continue in goodness, while those who practised evil continue to bear the

consequences of their conduct. This philosophical conception bears similarities to the broader ideas of heaven and hell, although it remains uniquely expressed within the framework of Gadé cosmology. Souls that suffer punishment are nevertheless not excluded from the possibility of reincarnation.

Within Gade philosophy, Adakpụ cannot exist independently of the living community. The relationship between the living and the ancestors is reciprocal: the living honours the ancestors through remembrance and ritual, while the ancestors continue to guide, protect, and influence the living. Consequently, every member of the Gadé community is, by virtue of birth and participation in communal life, regarded as possessing the potential to become an ancestor after death.

The process of reincarnation through Gukwu represents the continuation of this relationship. Gukwu is understood as the renewed manifestation of an ancestral soul, preserving the continuity of lineage, identity, and purpose across successive generations.

This philosophy is reflected in the songs, chants, and ritual performances associated with the Gunyi Festival and the Inā Festival. During these ceremonies, the departure of the Adakpụ masquerade, accompanied by male and female ancestral escorts, symbolically represents the ancestors' return to Asam, the Garden of Life and Purposes. These performances serve not merely as cultural expressions but as living repositories of ancestral knowledge, transmitting Gadé cosmological beliefs faithfully from one generation to the next.

The Unending Voice of Ancestors

Within Gade philosophy, every Iyé—the soul, heart, and mind—bears the enduring voice of the ancestors. These voices are not merely echoes of the past; they constitute the living embodiment of inherited wisdom, moral consciousness, and collective memory transmitted from generation to generation. Consequently, the voices of the ancestors continue to shape the character, identity, and destiny of the Gadé people.

The philosophical message of Adakpụ is founded upon the principles of love, peace, justice, patience, perseverance, communal responsibility, continuity of purpose, development, and harmonious coexistence. These principles encourage unity within families and communities, foster interregional relationships, and promote peaceful cooperation among nations. The ancestral voice therefore calls humanity to preserve social order while pursuing collective progress and the common good.

The voices of the ancestors unequivocally reject terrorism, unlawful killings, oppression, injustice, and every form of evil that threatens human dignity and communal stability. According to Gadé moral philosophy, evil inevitably attracts consequences, and individuals or communities that perpetuate injustice ultimately become subject to moral and spiritual accountability.

Central to this philosophy is the Kẹ́wá System, the indigenous legal and judicial tradition of the Gadé people. The Kẹ́wá System embodies the ancestral commitment to justice, equity, truth, and social harmony. It provides the ethical foundation upon which communal order is maintained and disputes are resolved. In many respects, the codification of modern legal systems reflects similar aspirations for justice, although expressed within different historical and institutional contexts.

The enduring voice of Adakpụ therefore transcends ritual observance alone. It speaks through the customs, institutions, laws, moral values, language, and collective memory of the people. It reminds every generation that culture is sustained not merely by preserving traditions but by living according to the ethical principles inherited from the ancestors.

Every culture and civilization possess a body of knowledge and philosophical traditions that contribute to humanity's understanding of justice, morality, governance, and existence. Such indigenous knowledge systems deserve careful study and should not be dismissed merely because they originated outside the framework of modern Western scholarship. When examined critically and responsibly, many of these traditions reveal principles that are consistent with natural justice, social order, and human development.

Within the Gadé knowledge system, the voices and works of those who have departed remain significant because their wisdom continues to guide the living. They are honoured as Adakpụ, whose enduring legacy is preserved through memory, culture, and communal practice. In this sense, the contributions of ancestors are not fundamentally different from those of national heroes, statesmen, scholars, and freedom fighters whose ideas continue to shape society long after their deaths.

Accordingly, the philosophical voice of Adakpụ promotes national unity, justice, impartial leadership, merit-based appointments, peaceful coexistence, and responsible governance. It rejects tribalism, nepotism, favouritism, discrimination, pogroms, and every practice that undermines the dignity of human beings or weakens the foundations of society. The philosophy ultimately calls upon both leaders and citizens to pursue justice and the common good as enduring responsibilities inherited from preceding generations.

The African Knowledge Systems, African Philosophy, and Indigenous Intellectual Traditions be accorded greater recognition within national educational curricula into AI Language presentation, preservation, documentation and transformation. Such inclusion would contribute significantly to preserving Africa's rich philosophical heritage while encouraging comparative studies between indigenous knowledge systems and other philosophical traditions across the world.

System Architecture Overview

The proposed Gadé Language Digitization Model is conceived as a layered, modular, and interoperable artificial intelligence architecture for transforming heterogeneous Gadé linguistic resources into validated computational knowledge and subsequently exposing that knowledge through language technologies and community-facing applications. The architecture is designed around the principle that indigenous-language AI cannot be reduced to a single neural model. Rather, reliable language intelligence requires an ecosystem in which data acquisition, corpus governance, linguistic annotation, lexical representation, speech processing, natural language processing, machine learning, application services, and human validation operate as interconnected but distinguishable components [65].

The system architecture therefore separates the linguistic resource layer from the computational intelligence layer and the application layer. This separation allows each component to evolve independently while maintaining well-defined interfaces

between components. It also makes it possible to identify which parts of the proposed system constitute established technologies and which parts represent Gadé-specific adaptations, datasets, annotation schemes, integration mechanisms, or empirical research contributions [52].

At the highest level, the architecture follows a data-to-intelligence-to-application pathway. Raw or previously documented Gadé materials are first ingested into governed repositories. They are then transformed into structured textual, speech, tonal, phonological, and lexical resources. These resources feed computational models such as tone-aware automatic speech recognition, language models, NLP components, and lexical intelligence services. An AI Language Intelligence Layer subsequently integrates outputs from these components and exposes them through search, translation, dictionary, speech, educational, and research interfaces.

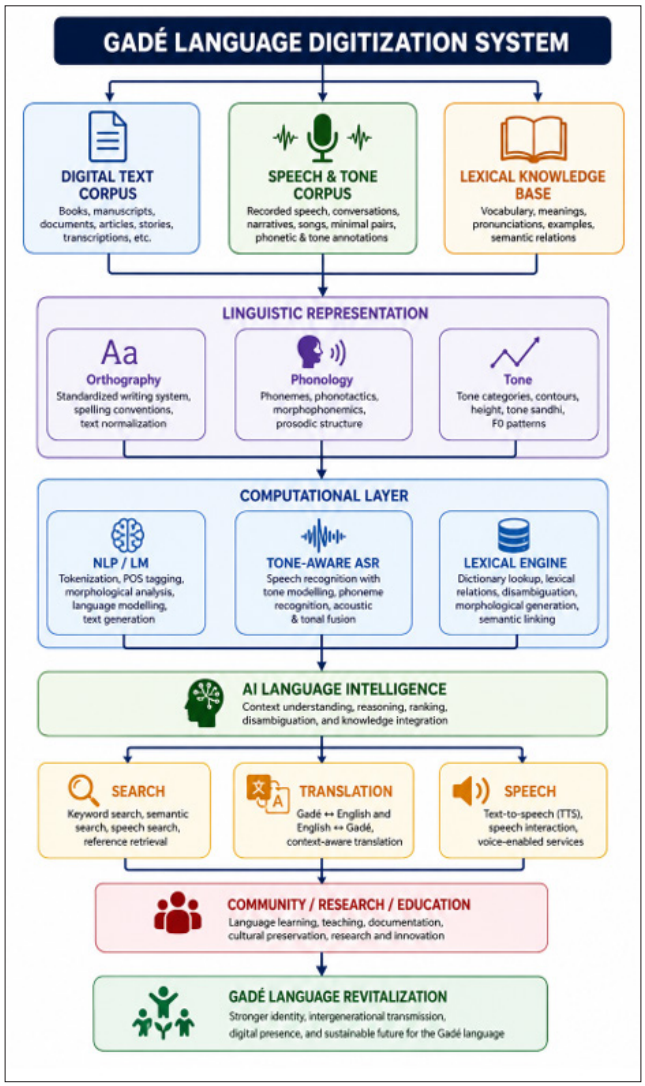


Figure 1: Gade Language Digitalization Model. Originally compiled by the Author, GT Obadiah on July 29, 2026 at 10:23pm (Nigerian Local Time)

Architectural Design Principles

The system architecture is governed by several design principles intended to ensure that the resulting platform is scientifically defensible, linguistically meaningful, technically extensible, and appropriate for an indigenous-language context.

Modularity: Each major subsystem should have a clearly defined responsibility and interface. The speech-processing subsystem, for example, should be independently replaceable without requiring the reconstruction of the entire corpus or application layer.

Interoperability: The system should support standardized machine-readable formats and APIs so that datasets and models can be reused by multiple applications and research environments [87].

Human-in-the-loop validation: AI-generated transcripts, lexical matches, translations, and annotations should be treated as computational suggestions until validated under the applicable linguistic and governance rules.

Data provenance: Every important linguistic resource should retain information about origin, transformation, annotation, validation, and release version [88].

Scalability: The architecture should support gradual growth from an initial low-resource prototype to larger corpora, additional models, more applications, and potentially distributed infrastructure [89].

Security and access control: The system should support authentication, authorization, role-based access control, audit logging, encryption, and differentiated access to resources.

Reproducibility: Corpus versions, preprocessing pipelines, model configurations, evaluation datasets, and experimental parameters should be identifiable and reproducible where data-access restrictions permit.

Community alignment: Technical architecture should support community-defined linguistic and cultural governance rather than assuming that every resource is unrestricted public data.

Layered Architecture

The proposed architecture can be represented as a sequence of seven functional layers. These layers are conceptual rather than necessarily corresponding to seven physical servers. In an implementation, multiple layers may reside within the same service or may be distributed across cloud, on-premises, or hybrid infrastructure [90].

Table 1: Layered Architecture of the Proposed Model as compiled by the Authour, GT Obadiah on July 29, 2026 at 11:32pm (Nigerian Local Time)

Layer	Primary Responsibility	Representative Components
Data Acquisition Layer	Collects and ingests Gadé linguistic resources.	Audio, text, video-derived transcripts, lexical contributions, metadata.
Corpus and Governance Layer	Stores, versions, validates, and governs language resources.	Text corpus, speech corpus, lexical database, provenance, access control.
Linguistic Representation Layer	Represents language structure in machine-readable form.	Orthography, phonemes, tone, morphology, syntax, semantics.

Computational Modelling Layer	Learns patterns from validated language resources.	ASR, language models, NLP, embeddings, classification and ranking.
AI Language Intelligence Layer	Integrates computational outputs and applies linguistic context.	Candidate ranking, contextual disambiguation, retrieval, reasoning.
Application and Interface Layer	Exposes language capabilities to users.	Search, dictionary, translation, speech, educational interfaces.
Monitoring and Feedback Layer	Captures errors, corrections, usage, and model performance.	Evaluation, human feedback, audit logs, model monitoring.

Data Acquisition Layer

The Data Acquisition Layer is the entry point through which Gadé linguistic material enters the digital ecosystem. Its purpose is not merely to upload files but to establish a controlled relationship between a source item, its contributor, its linguistic context, its consent status, and its intended use [91].

The layer may receive: digitized books, manuscripts, articles, and educational materials; typed Gadé texts and contemporary digital documents; speech recordings from individual speakers; controlled pronunciation recordings; conversational and narrative recordings where appropriately consented; lexical entries and dictionary contributions; phonological and tonal annotations; translations and example sentences; expert corrections of existing corpus records; metadata describing speakers, sources, contexts, and collection conditions.

Audio ingestion should perform basic integrity and quality checks before the recording becomes eligible for annotation. Text ingestion should similarly identify encoding problems, unsupported characters, duplicate records, and potentially inconsistent orthographic representations. Importantly, raw source data should be preserved separately from normalized or machine-generated derivatives.

Corpus and Governance Layer

The Corpus and Governance Layer constitutes the authoritative resource-management environment of the architecture. It contains the Digital Text Corpus, Gadé Speech and Tone Corpus, Lexical Database, metadata, provenance records, annotation layers, dataset versions, and access policies.

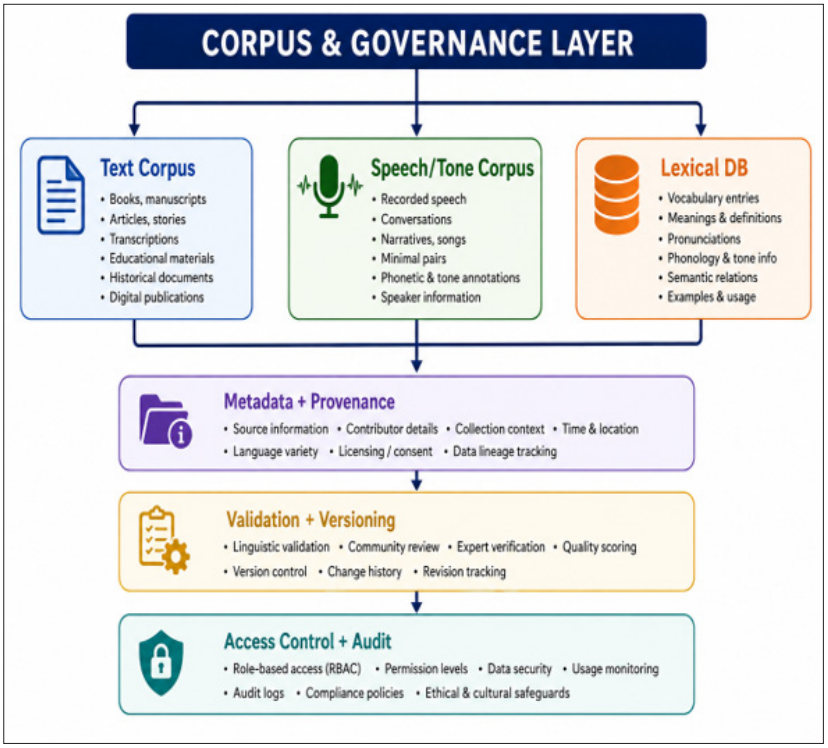
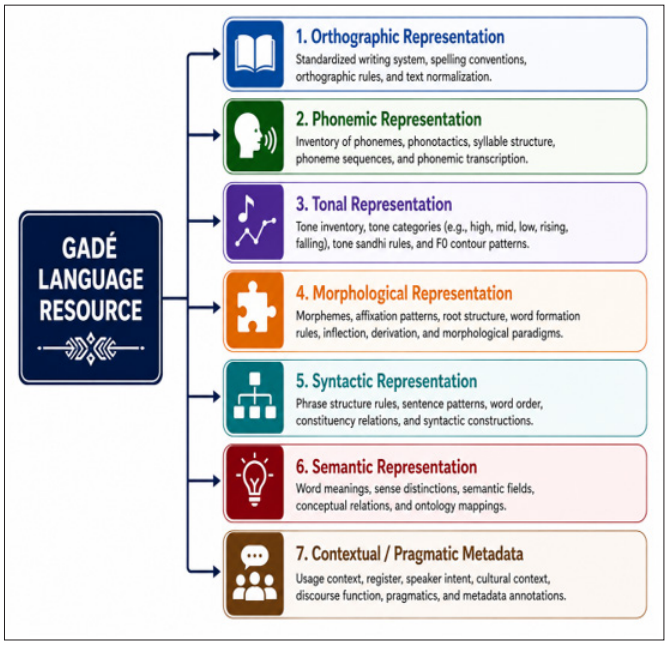


Figure 2: Corpus and Governance Layer. Originally compiled by the Author, GT Obadiah on July 21, 2026 at 4:23pm (Nigerian Local Time)



The layer should maintain a clear distinction between authoritative annotations and machine-generated suggestions. For example, a tone label approved by a qualified reviewer should be stored as a validated annotation, whereas a tone prediction generated by a neural model should be stored as a model output until human review confirms it.

This separation is particularly important for scientific reproducibility. If an ASR experiment produces an improved result after corpus correction, researchers should be able to determine whether the improvement resulted from model architecture, additional training data, revised annotations, preprocessing changes, or another factor.

Linguistic Representation Layer

The Linguistic Representation Layer transforms corpus resources into explicit representations of Gadé language structure. This layer is essential because machine-learning models require structured inputs, while linguistic phenomena frequently contain information that cannot be adequately represented by raw text alone [92].

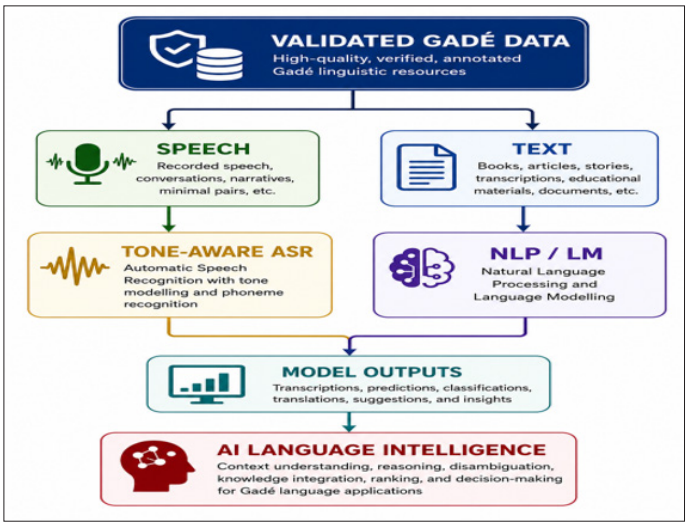


Figure 3: Linguistic Representation Layer. Originally compiled by the Author, GT Obadiah on July 21, 2026 at 6:29pm (Nigerian Local Time)

Orthographic representation establishes the written form. Phonemic representation captures relevant sound structure. Tonal representation captures linguistically meaningful pitch distinctions were established through Gadé linguistic analysis. Morphological, syntactic, and semantic representations can subsequently support increasingly sophisticated NLP applications [93].

The architecture should permit multiple annotation layers to coexist. This is preferable to forcing all linguistic information into a single representation because different research questions require different levels of analysis [94].

Computational Modelling Layer

The Computational Modelling Layer contains the machine-learning and statistical components that learn from the validated resources. Its major components include Tone-Aware Automatic Speech Recognition, NLP models, language models, lexical ranking, and other task-specific models.

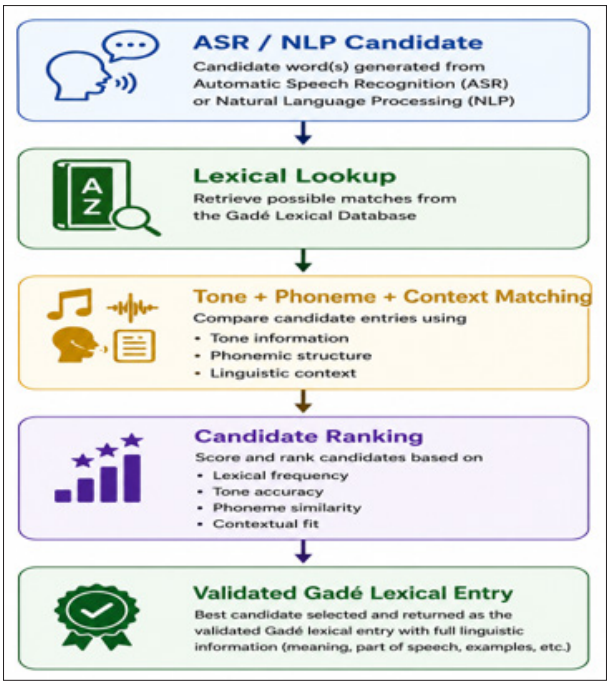


Figure 4: Computational Modelling Layer. Originally compiled by the Author, GT Obadiah on July 21, 2026 at 8:23pm (Nigerian Local Time)

The architecture intentionally avoids making a single model responsible for every language task. A specialized ASR model can optimize speech recognition, while a language model can optimize contextual sequence prediction and a lexical engine can provide authoritative dictionary information. The intelligence layer then combines these outputs (Page, 2021).

Tone-Aware ASR as a Specialized Subsystem

The Tone-Aware ASR subsystem is designed to address a central challenge in computational processing of tonal language. It receives acoustic speech, extracts acoustic and F0-related features, encodes temporal information, and produces phonemic, tonal, and orthographic hypotheses.

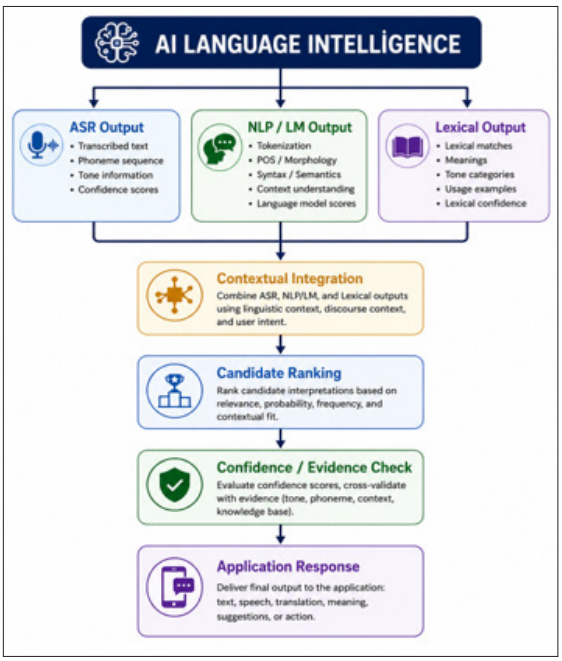


Figure 5: Tone-Aware ASR as a Specialized Subsystem. Originally compiled by the Author, GT Obadiah on July 21, 2026 at 9:23pm (Nigerian Local Time)

The architectural advantage is that tonal information becomes an explicit computational variable. The exact model configuration should be empirically determined through experimentation rather than assumed in advance. Depending on corpus size and computational constraints, the implementation may use transfer learning, self-supervised speech representations, a Conformer, Transformer-based encoders, or other suitable architectures [95].

NLP and Language Modelling Subsystem

The NLP and Language Modelling subsystem operates primarily on the Digital Text Corpus and lexical resources. It can provide tokenization, morphological analysis, part-of-speech information, named-entity recognition, language modelling, semantic representation, retrieval, classification, question answering, and translation where sufficient data exist [95]. For a low-resource language, the architecture should prioritize data-efficient learning. Transfer learning and multilingual pretraining may provide useful initialization, but Gadé-specific adaptation must be evaluated carefully because a multilingual model can carry assumptions from higher-resource languages that are inappropriate for Gadé.

P (w_t | w_1, ..., w_{t-1})

The language model can provide contextual probabilities that help resolve ambiguous ASR hypotheses, rank lexical candidates, and improve downstream language processing.

Lexical Intelligence Subsystem

The Lexical Intelligence subsystem provides authoritative lexical grounding for the AI architecture. It connects lexical forms to meanings, pronunciations, phonemic forms, tonal information, grammatical categories, examples, variants, and provenance [96].

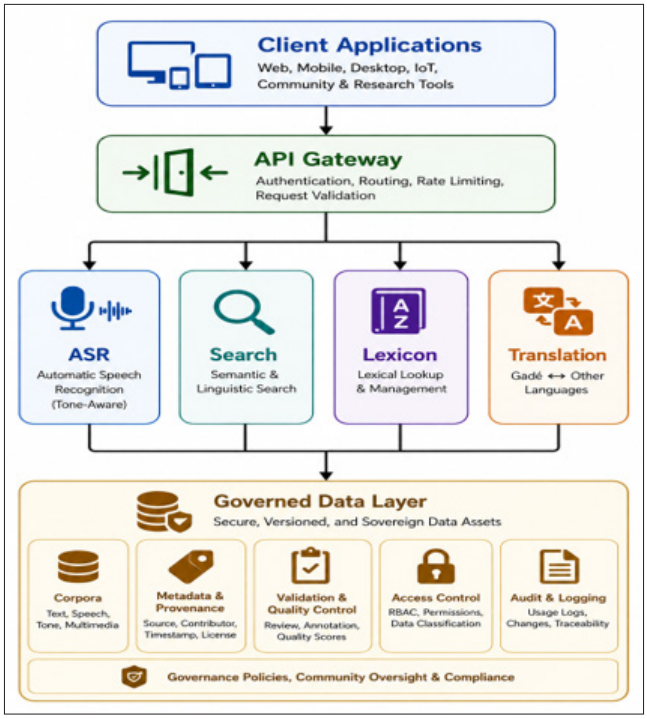


Figure 6: Lexical Intelligence Subsystem. Originally compiled by the Author, GT Obadiah on July 21, 2026 at 11:23pm (Nigerian Local Time)

This subsystem can reduce the risk of unconstrained generative systems producing plausible-looking but linguistically unsupported Gadé words. Where a lexical candidate does not exist in the validated database, the system should be capable of marking the output as uncertain rather than presenting it as established language knowledge [87].

AI Language Intelligence Layer

The AI Language Intelligence Layer functions as the integration and decision-support layer of the proposed system. It does not replace the underlying corpus, ASR, NLP, or lexical systems. Instead, it coordinates their outputs according to context, confidence, linguistic evidence, and application requirements.

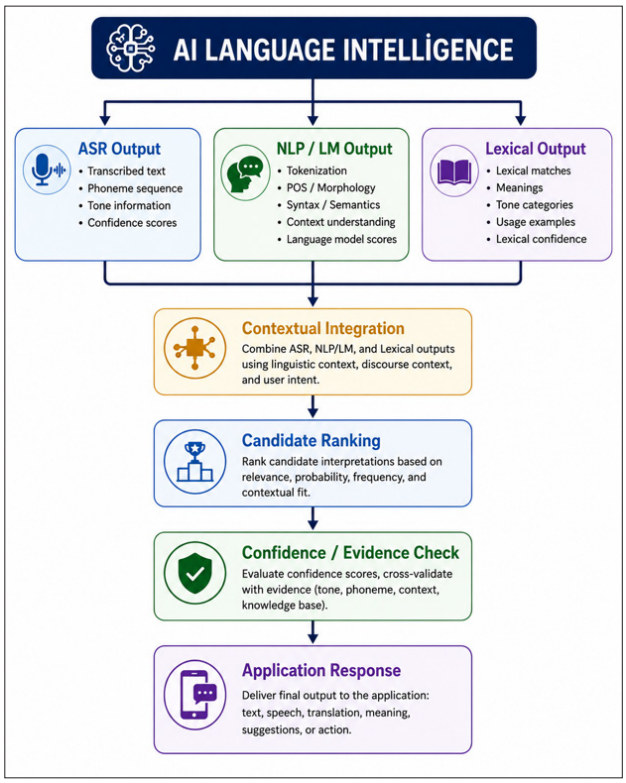


Figure 7: AI Language Intelligence Layer. Originally compiled by the Author, GT Obadiah at August 9, 2026 at 3:13am (Nigerian Local Time)

For example, if the ASR produces several candidates that are acoustically similar, the intelligence layer can compare the candidates against the lexical database, tonal annotations, phonemic structure, and language-model context. The result is a grounded ranking rather than a decision based solely on acoustic probability [67].

Application and Interface Layer

The Application and Interface Layer converts computational language capabilities into services usable by researchers, learners, educators, community members, and other authorized users. The architecture should expose these functions through responsive web interfaces, mobile applications, APIs, and potentially embedded educational or institutional systems [68].

Core interfaces include: Gadé dictionary and lexical search; full-text corpus search; speech-to-text interaction; Gadé-to-English and English-to-Gadé translation where validated resources permit; pronunciation and speech-learning interfaces; educational language-learning tools; research and corpus exploration interfaces; developer APIs for approved third-party applications.

The interface layer should display confidence or provenance information where appropriate. For example, an authoritative dictionary entry can be distinguished visually and semantically from an automatically generated translation or an unverified model suggestion [95].

API and Service-Oriented Integration

An API-oriented architecture allows the language resources and models to be reused without duplicating the underlying datasets. The Lexical API, Corpus Search API, ASR API, Translation API, and Model Inference API can provide controlled access to approved services.

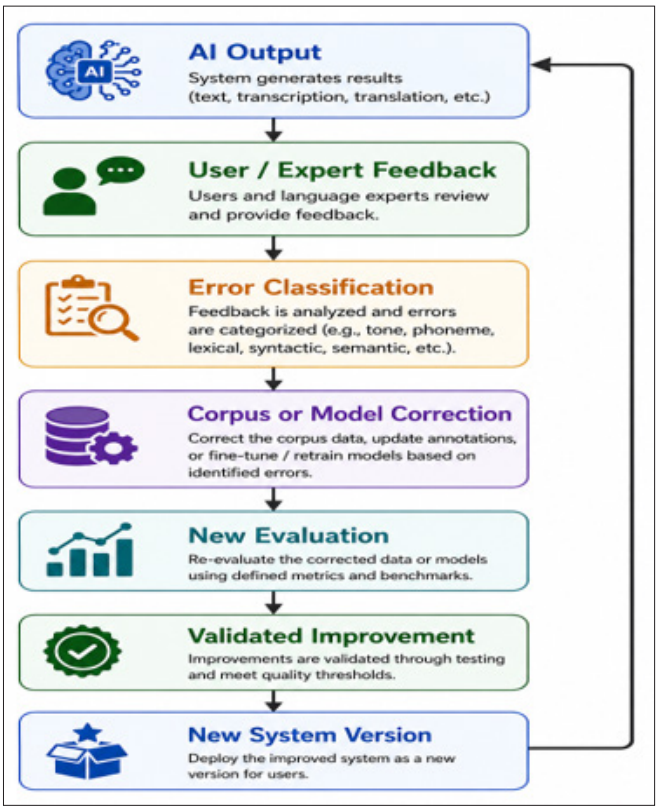


Figure 8: API and Service-Oriented Integration. Originally compiled by the Author, GT Obadiah at August 9, 2026

An API gateway can enforce authentication, authorization, rate limiting, logging, request validation, and routing. This is particularly important if the system is later opened to external developers or integrated into educational, governmental, cultural, or research platforms [96,97].

Security Architecture

Security is an architectural requirement rather than an optional implementation feature. The platform may contain contributor information, private recordings, culturally sensitive materials, research datasets, and privileged administrative functions.

The security architecture should therefore include: identity and authentication management; role-based access control (RBAC); least-privilege authorization; resource-level access policies; encryption of data in transit and at rest; audit logging; secure API authentication; input validation and output sanitization; backup and disaster-recovery mechanisms; dataset and model version integrity controls.

Access control should operate at more than the application level. Where appropriate, individual corpus resources should carry access classifications so that an API cannot accidentally expose a restricted recording simply because a user is authorized to access the general platform [97].

Monitoring, Evaluation and Feedback Layer

The Monitoring and Feedback Layer closes the architectural loop. It captures model errors, user corrections, annotation disagreements, search failures, translation feedback, and evaluation metrics. This information can be used to identify gaps in the corpus and guide subsequent annotation and model-improvement cycles.

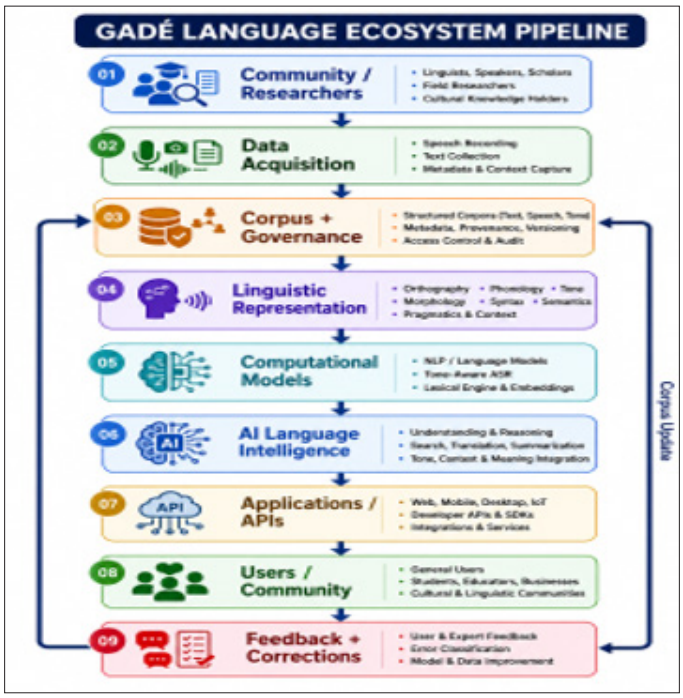


Figure 9: Monitoring, Evaluation and Feedback Layer. Originally compiled by the Author, GT Obadiah at August 9, 2026

This feedback mechanism should be governed so that arbitrary user feedback does not automatically alter authoritative language resources. Corrections should pass through the appropriate validation workflow before being incorporated into the governed corpus [98].

Data Flow Across the Architecture

The system can be understood as a set of bidirectional flows rather than a strictly linear pipeline. Data flow from community resources into computational models, while validated corrections and new contributions flow back into the corpus.

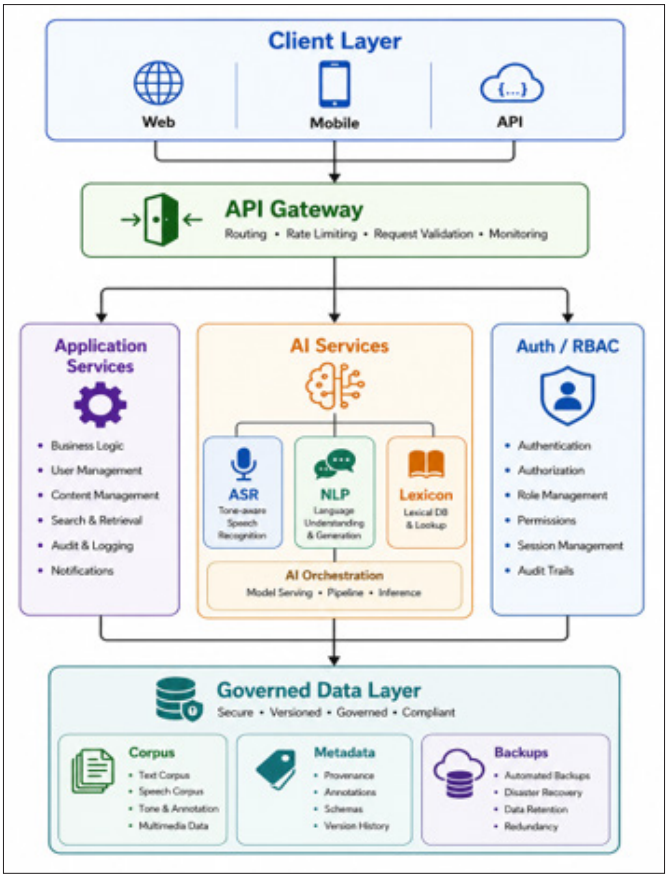


Figure 10: Data Flow Across the Architecture. Originally compiled by the Author, GT Obadiah at August 9, 2026

This cyclic structure is important because language revitalization is not a one-time digitization event. New vocabulary, new recordings, corrections, educational materials, and linguistic research can continuously enrich the system [98].

Deployment Architecture

The proposed model is compatible with cloud, on-premises, or hybrid deployment. A hybrid architecture may be particularly suitable where some linguistic resources require restricted storage while public services can operate through scalable cloud infrastructure.

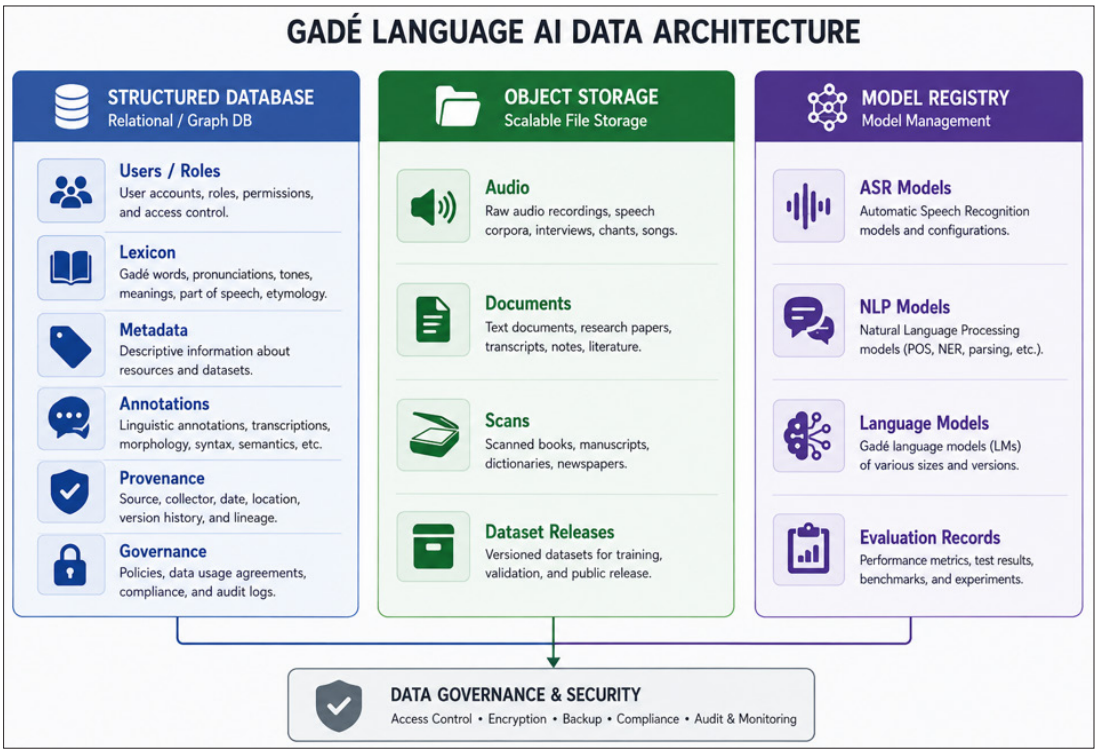


Figure 11: Deployment Architecture. Originally compiled by the Author, GT Obadiah at August 9, 2026

Model inference services can be separated from the primary data stores to support independent scaling. For example, a speech-recognition service may require GPU resources while lexical search may operate efficiently on conventional compute infrastructure. Separating these workloads can reduce operating costs and improve scalability.

Storage Architecture

The system should distinguish among structured data, unstructured media, and derived machine-learning artifacts. A relational or document-oriented database can manage lexical entries, metadata, users, permissions, and provenance. Object storage can maintain audio, video, scanned documents, and corpus files. Model registries can manage trained model versions and associated metadata [99].

The separation supports maintainability and enables large audio and document collections to grow without forcing all resources into a single database [93].

Model and Dataset Versioning

The architecture should treat datasets and models as versioned scientific artifacts. A model release should reference the corpus version used for training, preprocessing configuration, hyperparameters, training date, evaluation results, and known limitations.

Model Version = f (Corpus Version, Preprocessing, Architecture, Parameters, Training Configuration)

This relationship makes it possible to reproduce or investigate changes in performance. If a new corpus release introduces additional speakers or revised tone annotations, the effect of that change can be evaluated separately from changes to the neural architecture [93].

Evaluation Architecture

Evaluation should occur at both component and system levels. Component evaluation determines whether individual subsystems function correctly, while end-to-end evaluation determines whether the integrated architecture produces useful and linguistically valid results.

Table 2: Evaluation Architecture

Component	Primary Metrics	Human/Expert Evaluation
ASR	WER, CER, PER	Transcript and pronunciation review
Tone	Tone accuracy, F1, confusion matrix	Linguistic/tone expert review
Lexicon	Precision, recall, retrieval accuracy	Lexical validity
Translation	BLEU/other automatic metrics where appropriate	Bilingual human assessment
Search	Precision, recall, ranking metrics	Relevance assessment
NLP	Task-specific F1/accuracy	Linguistic validity
End-to-end system	Task success and error rates	Community/user evaluation

The evaluation architecture should explicitly recognize that automatic metrics do not fully capture indigenous-language validity. A model can obtain a numerically acceptable score while producing an orthographically inappropriate, culturally misleading, or linguistically unsupported result. Expert and community evaluation should therefore complement automatic evaluation [94].

Fault Isolation and Error Propagation

A modular architecture makes it possible to identify where an error originated. For example, an incorrect translated word may originate from an ASR error, a lexical mismatch, an incorrect language-model prediction, or the translation model itself [80].

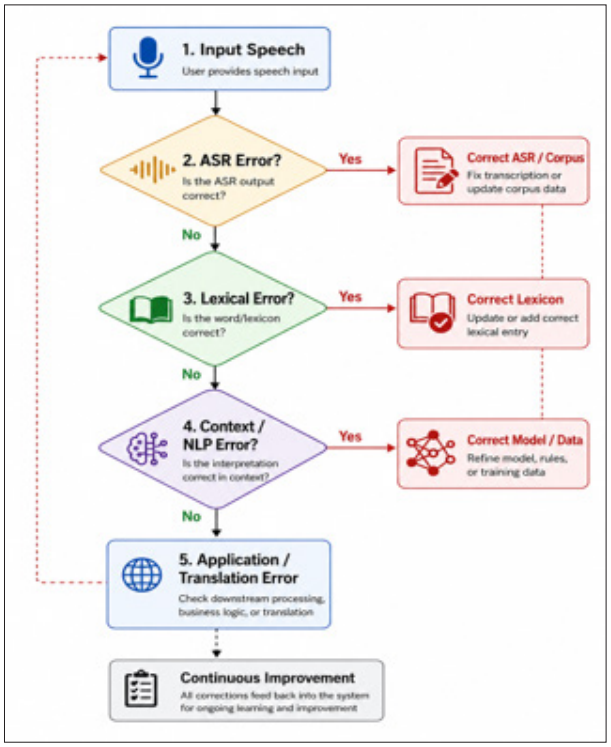


Figure 12: Fault Isolation and Error Propagation. Originally compiled by the Author, GT Obadiah at August 9, 2026

This diagnostic capability is particularly valuable for a low-resource language because improving the system may depend more on correcting a small number of high-impact linguistic resources than on increasing model size [29].

Scalability and Extensibility

The architecture is intended to accommodate gradual expansion. Initial development may focus on a small validated corpus and a limited dictionary, followed by progressively larger speech datasets, more extensive lexical resources, additional NLP tasks, and new user interfaces [69].

Potential future extensions include: Gadé text-to-speech; multimodal language models; mobile offline ASR; interactive language-learning systems; educational assessment tools; semantic search; question-answering systems; digital archival systems; cross-dialect modelling where linguistically justified; Gadé conversational agents; research APIs and developer ecosystems.

The modular design ensures that these future capabilities can be added without fundamentally restructuring the core corpus and governance infrastructure.

Distinguishing Established Technology from the Gadé-Specific Contribution

For scientific clarity, the architecture should distinguish between established computational technologies and the proposed Gadé-specific research contribution. Transformers, Conformers, ASR, NLP, databases, APIs, access-control systems, and cloud infrastructure are established technologies. Their use in the system does not by itself constitute novelty [70].

The potential research contribution instead arises from how these technologies are adapted, integrated, evaluated, and grounded in Gadé-specific linguistic resources. Relevant contributions include the construction of a validated Gadé speech and tone corpus, the annotation methodology, tone-aware ASR design, integration of tonal and phonemic information with lexical context, community-led data governance, and empirical evaluation of the resulting architecture.

Table 3: Gadé-Specific Contribution

Established Technology	Potential Gadé-Specific Contribution
Transformer / Conformer architectures	Their adaptation and evaluation for Gadé speech and tonal structure.
Automatic Speech Recognition	Tone-aware Gadé ASR corpus, annotation and decoding strategy.
Language modelling	Gadé-specific corpus adaptation and contextual modelling.
Digital dictionary systems	Governed Gadé lexical database with phoneme/tone/provenance layers.
Cloud/database/API infrastructure	Architecture supporting community-governed Gadé language resources.
Human-in-the-loop AI	Community and linguistic validation workflow for Gadé computational resources.

Relationship to Community-Led Data Ingestion and Sovereign Corpus Annotation

The system architecture described in this subsection provides the technical structure within which the community-led data ingestion and sovereign corpus annotation process operates. The latter supplies governed linguistic resources to the architecture, while the architecture supplies the mechanisms required to store, annotate, version, secure, process, and expose those resources [71].

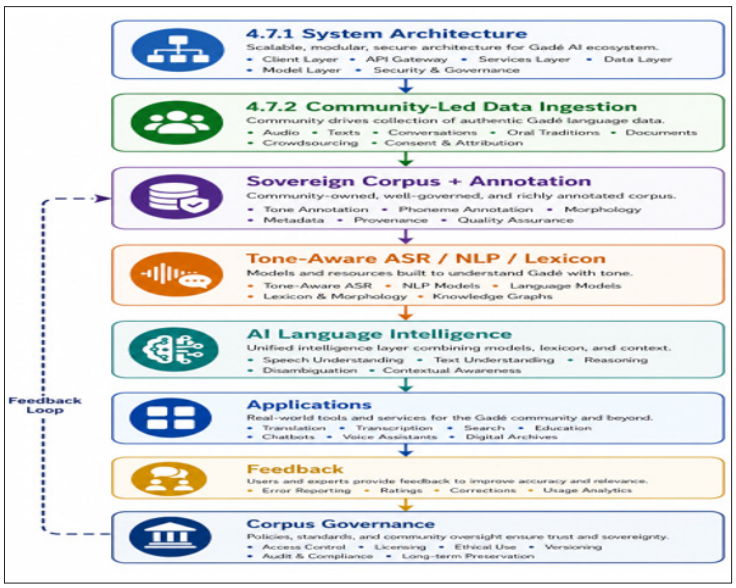


Figure 13: Fault Isolation and Error Propagation. Originally compiled by the Author, GT Obadiah at August 9, 2026

This relationship is fundamental to the proposed model. The system architecture cannot be separated from the governance architecture because the technical system determines how data are represented and accessed, while governance determines which representations are authoritative and under what conditions they may be used [71].

Architectural Significance

The significance of the proposed system architecture lies in its treatment of Gadé language digitization as an integrated computational ecosystem. Rather than isolating speech recognition, dictionary development, text digitization, or translation as independent projects, the architecture creates shared infrastructure through which these resources can reinforce one another [72].

A validated lexical database can improve ASR candidate ranking. Improved ASR can increase the availability of transcribed speech, subject to human validation. The growing text corpus can improve language modelling. Better language modelling can improve ASR contextual decoding and search. Community corrections can improve corpus quality, and improved corpus quality can support subsequent model versions. The architecture therefore creates a virtuous cycle of documentation, computation, validation, and application [73].

Language Resources → Models → Applications → Feedback → Improved Language Resources

This architecture provides the technical foundation for the wider objective of artificial intelligence and indigenous language revitalization. Its central proposition is that AI capability should emerge from a governed and linguistically grounded resource ecosystem rather than from model size alone [74].

The proposed Gadé Language Digitization System Architecture is a layered, modular, interoperable framework linking data acquisition, corpus governance, linguistic representation, computational modelling, AI language intelligence, applications, and feedback. Its principal components include the Digital Text Corpus, Gadé Speech and Tone Corpus, Lexical Database, Tone-Aware ASR, NLP and Language Modelling, AI Language Intelligence Layer, and Search, Translation, and Speech Interfaces [75].

The architecture deliberately separates source data from derived annotations and model outputs, supports provenance and versioning, accommodates human and community validation, and provides security and access-control mechanisms for governed linguistic resources. It is designed to scale from an initial research prototype to a broader language technology ecosystem while preserving the distinction between established computational technologies and the Gadé-specific research contributions that require empirical validation [76].

Within the broader research model, the architecture provides the technical framework through which community-led data ingestion and sovereign corpus annotation become operational. Together, these components establish a pathway from Gadé linguistic heritage to structured digital resources, computational intelligence, practical language applications, and ultimately improved digital accessibility and revitalization of the Gadé language.

Community-Led Data Ingestion and Sovereign Corpus Annotation
Community-Led Data Ingestion and Sovereign Corpus Annotation constitute a foundational component of the proposed Gadé Language Digitization Model because the quality, legitimacy, provenance, and long-term usefulness of an indigenous-language artificial intelligence system depend fundamentally on how its linguistic data are collected, described, validated, governed, and reused. In a low-resource language context, the corpus is not simply a technical input to a machine-learning model. It is a computational representation of linguistic knowledge, including vocabulary, pronunciation, tonal distinctions, oral traditions, personal names, place names, cultural expressions, narratives, and forms of everyday communication. Consequently, the process through which these resources enter the computational ecosystem must place the Gadé-speaking community and qualified linguistic experts at the centre of data governance and annotation.

The proposed approach treats community participation as an operational component of the data pipeline rather than as a symbolic consultation exercise. Community members may contribute recordings, texts, lexical items, translations, explanations of meaning, pronunciation variants, cultural context, and corrections to machine-generated annotations. Linguists and trained annotators provide methodological consistency and quality assurance, while technical personnel implement secure ingestion, storage, version control, annotation interfaces, and model integration. The resulting corpus is therefore conceived as a governed knowledge infrastructure in which technical accessibility is balanced against linguistic authority, consent, cultural sensitivity, provenance, and community-defined access conditions.

The term 'sovereign corpus' is used in this study to describe a corpus whose governance, provenance, access conditions, permitted uses, and representation of community knowledge are explicitly defined rather than being determined solely by the technical platform on which the data are stored. It does not imply that all data must be publicly accessible or that ownership is automatically transferred to a particular institution. Instead, sovereignty is operationalized through documented decision-making authority, consent, provenance, access control, usage restrictions, correction mechanisms, and community participation in determining how Gadé linguistic resources are represented and reused [55].

Rationale for a Community-Led Corpus

Conventional machine-learning pipelines frequently begin with the assumption that data are available, homogeneous, and sufficiently labelled. Such assumptions are problematic for indigenous languages. A low-resource language may have limited digitized text, relatively small speech datasets, inconsistent orthographic conventions, incomplete dictionaries, sparse phonemic annotation, and substantial gaps in computational resources. More importantly, a technically large dataset can still be linguistically weak if its annotations are generated without adequate knowledge of the language, if dialectal variation is erased, or if culturally significant meanings are reduced to inappropriate translations [25].

Community-led ingestion addresses these limitations by treating speakers and knowledgeable language users as contributors to the construction of the corpus. This approach recognizes that linguistic competence is distributed across the community. A speaker may know pronunciation patterns that are absent from written sources; an elder may provide culturally embedded meanings for an expression; a teacher may identify an orthographic problem; a researcher may resolve a phonological ambiguity; and a younger speaker may contribute contemporary vocabulary and digital communication patterns. Bringing these forms of knowledge into a common, governed corpus can substantially increase the representational coverage of the resulting AI system [49].

The community-led approach also reduces the risk of constructing a corpus that is technically sophisticated but linguistically detached from actual language use. Instead of beginning with a predetermined computational vocabulary and asking speakers to fit their language into it, the proposed methodology begins with documented language practices and then develops computational representations around them [50]

Principles of Data Sovereignty

The proposed sovereign corpus framework is based on several interdependent principles. These principles are methodological safeguards rather than claims that a single universal governance model applies to every indigenous community.

Community participation: Gadé speakers, language experts, educators, cultural knowledge holders, researchers, and other authorized contributors should have defined opportunities to participate in data collection, annotation, validation, correction, and governance.

Provenance: Every corpus item should retain sufficient information to establish where it came from, who contributed it, when it was collected, how it was transformed, and which validation stage it has passed.

Purpose Limitation: Data should be collected and processed for explicitly documented purposes. A recording contributed for linguistic documentation should not automatically be treated as unrestricted training data for unrelated applications.

Access Control: Different resources may require different access levels, ranging from public resources to restricted research materials or culturally sensitive content.

Correctability: Contributors and authorized reviewers should have a documented mechanism for reporting errors, requesting corrections, contesting annotations, and where appropriate requesting restriction or removal.

Transparency: The corpus should make its annotation rules, metadata conventions, quality standards, and governance procedures sufficiently transparent to authorized users.

Intergenerational Continuity: The corpus should be designed as a long-term language resource rather than a short-term dataset for a single AI experiment [51].

Community-Led Data Ingestion Architecture

Community-led ingestion should be implemented as a controlled sequence of collection, consent, preprocessing, annotation, validation, governance review, and release. The pipeline should preserve the original contribution while allowing derived computational representations to be generated without destroying the source record.

The ingestion architecture deliberately separates raw source data from processed and annotated derivatives. The original audio, text, or other contribution should remain preserved as a source object with immutable provenance metadata, while normalized transcripts, segmentation, phonemic labels, tone labels, translations, and machine-generated suggestions are stored as derivative layers. This separation makes it possible to correct an annotation without losing the original evidence [52].

Types of Community-Contributed Data

The corpus should accommodate multiple forms of linguistic evidence, including:

- spoken-word recordings and controlled pronunciation samples;
- natural conversations and narratives, subject to consent and access conditions;
- traditional stories, oral histories, and culturally significant expressions;
- lexical items, definitions, example sentences, and usage notes;
- phonemic and tonal judgements supplied by trained speakers or linguists;
- orthographic corrections and alternative spellings;
- translations and explanations of culturally embedded meanings;
- personal names, place names, traditional terminology, and domain-specific vocabulary;
- educational materials and contemporary written Gadé texts;
- corrections to AI-generated transcripts, translations, or dictionary suggestions.

The system should not assume that all data types have identical sensitivity or identical rights of reuse. A public vocabulary item and a culturally restricted narrative should therefore be represented within the same technical infrastructure but governed according to different access and usage rules [55].

Consent and Contributor Registration

Before a contribution enters the governed corpus, the system should associate it with a consent and metadata record appropriate to the collection context. Consent should be understandable to contributors and should explain, in appropriate language, what is being collected, the intended uses, who may access it, whether it may be used for AI training, whether derivative resources may be produced, and what mechanisms exist for correction or withdrawal where applicable.

Contributor registration should avoid unnecessarily exposing personal information. A unique contributor identifier can be used internally, while public or model-facing datasets can use pseudonymized identifiers. Sensitive personal information should not be embedded in training records unless there is a documented and legitimate reason to retain it.

For group conversations, narratives involving third parties, or community events, the ingestion process should include procedures for dealing with multiple participants and potentially identifiable individuals. The technical objective of building a high-quality corpus should not override privacy and consent requirements [56].

Sovereign Annotation Schema

A sovereign corpus requires a rich annotation schema because the meaning of a linguistic item cannot always be captured by a single orthographic transcription. The proposed annotation framework should therefore support multiple layers that can be independently reviewed.

Table 4: Sovereign Annotation Schema

Annotation Layer	Primary Function	Validation Requirement
Orthographic	Represents the standardized or source spelling of the Gadé form.	Orthographic reviewer or language expert.
Phonemic	Represents phoneme sequence and relevant phonological structure.	Trained annotator and/or linguist.
Tonal	Records linguistically relevant tone categories or contours.	Tone-competent speaker/linguist.
Morphological	Identifies roots, affixes, reduplication, and other morphological structure where applicable.	Linguistic review.
Syntactic	Records grammatical relations and sentence-level structure where required.	Qualified linguistic review.

Semantic	Captures meaning, sense distinctions, and contextual interpretation.	Native-speaker/contextual validation.
Translation	Provides equivalent or explanatory translations into another language.	Bilingual/community review.
Pragmatic/Cultural	Records context, usage constraints, register, or culturally specific interpretation.	Authorized cultural/language reviewers.
Provenance	Records source, contributor, date, transformation, and validation history.	Metadata/curation review.

The annotation architecture should permit multiple valid analyses where linguistic evidence is uncertain or variation exists. Forcing a single annotation in the presence of genuine variation can introduce artificial certainty into the corpus. Therefore, uncertain fields should be capable of being marked as provisional, disputed, unverified, or requiring further research.

Tone and Phonological Annotation

Tone-aware annotation is particularly important for the proposed Gadé ASR pipeline. However, tone labels must be established from Gadé linguistic evidence rather than assumed from the annotation conventions of another language. Where tone is linguistically contrastive, annotators should identify the relevant tonal category or contour and, where possible, associate the annotation with a time-aligned acoustic region.

F0 measurements can provide acoustic evidence for tonal analysis, but F0 should not automatically be equated with linguistic tone. Pitch may vary because of speaker characteristics, intonation, emotional state, prosody, recording conditions, or other factors. Consequently, the annotation protocol should combine acoustic evidence with linguistic judgement.

For the AI system, the distinction between acoustic pitch and phonological tone is important. The former is a measurable signal property, whereas the latter is a linguistic representation. A robust corpus should therefore preserve both levels rather than collapsing them into a single label.

Annotation Workflow and Double Review

To improve reliability, the proposed corpus should use a staged annotation workflow. Initial annotation can be performed by trained community annotators or linguists, followed by independent review. Disagreements should be retained as part of the quality-control process rather than silently overwritten.

For a subset of the corpus, inter-annotator agreement should be calculated. Agreement can be measured using suitable statistics for categorical or sequence-based annotations. The purpose is not merely to obtain a numerical score but to identify annotation categories that are consistently misunderstood and therefore require better guidelines or additional linguistic investigation.

Machine-Assisted Annotation and Human Oversight

Artificial intelligence can reduce the cost of corpus development by generating preliminary transcripts, segmentation, phoneme candidates, lexical matches, translations, or tone predictions. However, machine-generated annotations should be explicitly marked as provisional until they have passed the appropriate human validation stage.

This creates a controlled human-in-the-loop cycle. As the corpus improves, the ASR and NLP systems may generate increasingly useful suggestions, while human reviewers correct the errors. Those validated corrections can subsequently be used for further model adaptation. The process therefore becomes iterative rather than a one-time data collection exercise.

A critical safeguard is that model predictions should never be allowed to silently overwrite authoritative annotations. The system should maintain a distinction between source evidence, human-approved annotations, machine suggestions, and derived model outputs.

Provenance and Version Control

Every corpus object should possess a persistent identifier and a provenance record. The provenance record should document the origin of the item, its transformations, annotation events, reviewers, validation status, and release history. Version control is particularly important because linguistic resources evolve as additional speakers contribute evidence and as orthographic or analytical decisions are refined.

Corpus Item = Source + Metadata + Annotation Layers + Provenance + Governance State

A versioned corpus allows researchers to reproduce experiments against a known dataset version. It also enables the project to determine which model was trained on which annotation release. This is essential for scientific reproducibility and for tracing errors discovered after model deployment.

Access Control and Resource Classification

The sovereign corpus should implement differentiated access rather than a binary public/private model. A possible classification is:

Table 5: Access Control and Resource Classification

Level	Description	Typical Use
Public	Resources approved for unrestricted public access.	Dictionary entries, validated educational materials, selected texts.
Registered	Resources accessible to approved users after account registration.	Research corpus, detailed linguistic annotations.
Restricted	Resources requiring explicit authorization.	Sensitive recordings, contributor-linked material.
Protected	Culturally sensitive material subject to community-defined restrictions.	Materials with specific cultural or custodial conditions.

These levels are illustrative and should be defined through consultation with the relevant Gadé institutions, language authorities, researchers, and contributors. The key design principle is that the computational platform must be capable of enforcing the governance decisions established for each resource.

Data Quality Assurance

Corpus quality should be measured at several levels. Technical

quality includes audio signal quality, file integrity, segmentation accuracy, transcription completeness, and metadata consistency. Linguistic quality includes orthographic accuracy, phonemic accuracy, tonal accuracy, lexical validity, translation adequacy, and contextual appropriateness. Governance quality includes consent completeness, provenance integrity, access classification, and review status.

- audio quality screening and clipping/noise detection;
- duplicate and near-duplicate detection;
- orthographic consistency checks;
- lexical validity checks against the governed dictionary;
- phoneme and tone annotation review;
- metadata completeness checks;
- speaker and recording-condition balance;
- human review of machine-generated annotations;
- periodic corpus audits and correction cycles.

A quality score may be assigned to individual records, but the score should complement rather than replace expert review. A high technical score does not guarantee linguistic correctness, and a linguistically valuable recording should not be discarded solely because it was captured under imperfect field conditions if its limitations are properly documented.

Handling Linguistic Variation

A major risk in indigenous-language corpus construction is the inadvertent standardization of variation. Speakers may differ in pronunciation, lexical choice, grammar, or tone. Such variation can represent dialectal diversity, generational differences, register, or ordinary language change. The corpus should therefore preserve relevant variation rather than treating every deviation from a preferred form as an error.

Metadata should allow the system to record relevant linguistic variety where known. A lexical entry can consequently contain a standardized form together with validated variants, pronunciation alternatives, contextual restrictions, and regional or speaker-group information where ethically and methodologically appropriate.

This approach also benefits machine learning. A model trained only on a narrow variety may perform poorly when presented with another legitimate variety. Preserving variation can therefore improve both linguistic documentation and computational robustness [35].

Community Governance Structure

Community-led ingestion requires institutional arrangements capable of making decisions about data. A proposed governance structure may include a language and cultural advisory group, linguistic experts, technical administrators, community contributors, and designated data stewards.

The purpose of such a structure is to ensure that technical decisions affecting the corpus can be connected to linguistic and community decisions. The exact institutional arrangement should be established by the participating Gadé stakeholders rather than imposed by the software architecture [66].

Community Incentives and Contributor Recognition

Sustainable corpus development requires mechanisms for recognizing the contributions of speakers, annotators, translators, teachers, researchers, and cultural knowledge holders. Recognition may include contributor attribution where appropriate, compensation for professional annotation or collection work, acknowledgements in research outputs, contributor certificates, community access to derived tools, and opportunities for training in digital language documentation.

Contributor recognition also has scientific value because it strengthens provenance. Where appropriate and consented to, corpus records can retain information about the role of the contributor, such as speaker, translator, annotator, reviewer, or linguistic consultant. This makes it possible to distinguish original language evidence from subsequent computational interpretation [33].

Security and Integrity of the Sovereign Corpus

Because the corpus may contain valuable cultural and linguistic resources, technical safeguards should be incorporated into the platform. These may include encryption in transit and at rest, role-based access control, authentication, audit logging, secure backups, versioned storage, integrity checks, and controlled export mechanisms.

The security model should also protect against unauthorized modification of validated linguistic resources. A technically sophisticated AI platform can produce unreliable results if its underlying corpus is silently altered or contaminated. Therefore, approved corpus releases should be cryptographically or otherwise integrity-protected where appropriate, and all material changes should generate an auditable record [26].

Relationship to the Tone-Aware ASR Pipeline

Community-led ingestion directly supports the Tone-Aware ASR component described in the wider Gadé Language Digitization Model. ASR performance depends not only on the neural architecture but also on the quality and representativeness of the training labels. In a tonal language, incorrect tone labels can teach the model incorrect associations between acoustic pitch patterns and lexical representations.

This creates a continuous improvement loop in which community validation becomes part of the model development lifecycle. Importantly, the feedback loop should not be interpreted as allowing the model to determine linguistic truth. Rather, the model proposes outputs and the human linguistic community determines whether those outputs are acceptable for incorporation into the governed corpus.

Dataset Splitting and Leakage Prevention

Sovereign corpus annotation should also incorporate experimental controls required for reliable machine-learning evaluation. Speech data should be separated into training, validation, and test partitions in a way that minimizes speaker leakage. If recordings from the same speaker appear across all partitions, a model may appear more accurate than it would be on genuinely unseen speakers.

Where possible, speaker-independent and context-aware partitioning should be used. Duplicate recordings, near-duplicate narratives, repeated prompts, and derivative versions of the same source should also be identified. This is particularly important when community members contribute multiple recordings of the same lexical items or texts [45].

Sovereignty and Reproducible Research

Community governance and scientific reproducibility are not inherently contradictory. A restricted corpus can still support reproducible research if researchers are provided with sufficient

metadata, documented preprocessing procedures, annotation guidelines, dataset version identifiers, evaluation protocols, and controlled access mechanisms.

Where raw data cannot be publicly released, researchers can publish dataset statistics, annotation schemas, code, model configurations, evaluation procedures, and controlled-access instructions. This permits scientific scrutiny without requiring unrestricted publication of sensitive or community-governed data [47].

Proposed Community-to-AI Data Lifecycle

The complete lifecycle can therefore be summarized as a governed transformation from community knowledge into validated computational resources and, ultimately, back into community-facing applications:

The cycle is intentionally iterative. Language changes, new speakers contribute data, previously uncertain analyses may become clearer, and AI models reveal new categories of error. A sovereign corpus should therefore be treated as a living research and cultural infrastructure with controlled evolution rather than as a static dataset [77].

Expected Contribution to the Gadé Language Digitization Model Within the broader proposed architecture, community-led ingestion and sovereign annotation provide the data-governance foundation on which the AI components depend. The Digital Text Corpus requires validated texts; the Speech Corpus requires trusted recordings and annotations; the Lexical Database requires community and linguistic verification; and the Tone-Aware ASR requires accurate phonemic and tonal labels. Consequently, the reliability of the AI Language Intelligence Layer is structurally dependent on the reliability of the underlying governed corpus. The contribution can be expressed conceptually as.

Community Authority + Linguistic Expertise + Technical Infrastructure → Governed Gadé Corpus → Reliable AI Language Resources

This framework also strengthens the distinction between technological innovation and linguistic authority. Artificial intelligence supplies mechanisms for pattern recognition, prediction, search, classification, and generation. The community and qualified language experts provide the linguistic evidence, contextual interpretation, and governance decisions required to determine what should be represented as validated Gadé knowledge.

Research and Implementation Implications

From a research perspective, the proposed approach introduces a methodological bridge between indigenous-language documentation and computational language technology. It allows subsequent experiments to identify precisely which portion of the AI system depends on which type of linguistic resource. For example, ASR performance can be evaluated against validated speech and tone annotations, lexical retrieval can be assessed against the governed dictionary, and translation can be evaluated against human-reviewed parallel material.

From an implementation perspective, the approach suggests that the corpus platform should not be developed as a conventional database alone. It should combine data storage with annotation tools, contributor management, provenance tracking, validation workflows, access control, audit logging, dataset versioning, and APIs through which approved resources can be consumed by downstream AI systems.

The architecture therefore positions the corpus as an active computational layer within the Gadé language ecosystem. Rather than merely storing historical records, it becomes a continuously validated foundation for speech recognition, language modelling, search, translation, digital dictionaries, educational applications, and future language technologies.

Tone-Aware Automatic Speech Recognition (ASR) Pipeline

The proposed Tone-Aware Automatic Speech Recognition (ASR) Pipeline constitutes a core component of the Gadé Language Digitization Model. Its primary objective is to convert spoken Gadé language into machine-readable and orthographically appropriate text while preserving tonal information that may contribute to lexical and semantic distinctions. Conventional ASR systems are predominantly optimized for high-resource languages and may perform poorly on indigenous languages because of limited training corpora, dialectal variation, inadequate linguistic resources, and insufficient representation of prosodic features. A tone-aware architecture is therefore proposed to address these limitations.

The pipeline begins with the acquisition of authenticated Gadé speech recordings from native speakers representing relevant dialectal, demographic, and contextual variations. Audio preprocessing includes sampling-rate normalization, noise reduction, voice-activity detection, segmentation, and speaker normalization. The processed speech is subsequently transformed into acoustic representations, including Mel-frequency spectral features, Mel spectrograms, and fundamental-frequency (F0) contours. The explicit extraction of F0 is particularly important because pitch variation may encode lexical or grammatical information in tonal speech.

The architecture incorporates two complementary representation streams: an acoustic speech representation stream and a tone representation stream. The acoustic stream captures phonetic and spectral information, whereas the tone stream models pitch trajectory, F0 variation, voicing characteristics, and temporal tonal patterns. These representations are subsequently integrated through a neural feature-fusion mechanism before being passed to the speech recognition encoder.

A Transformer- or Conformer-based encoder may be employed to model long-range temporal dependencies within the speech signal. The recognition architecture can be formulated as a multi-task learning framework in which the system simultaneously learns speech transcription and tone classification (Scott, 1996). This can be expressed as:

Figure 19: Tone-Aware Automatic Speech Recognition (ASR) Pipeline

Originally compiled by the Author, GT Obadiah at August 9, 2026
The output of the acoustic encoder is therefore not limited to an unmarked phonetic sequence. Instead, the model attempts to establish a relationship between phonetic units, tonal patterns, lexical items, and contextual language information. A Gadé lexical database and language model can subsequently be incorporated into the decoding stage to improve recognition accuracy, particularly for low-frequency words and words that are acoustically similar but tonally distinct.

The proposed pipeline can be represented as:

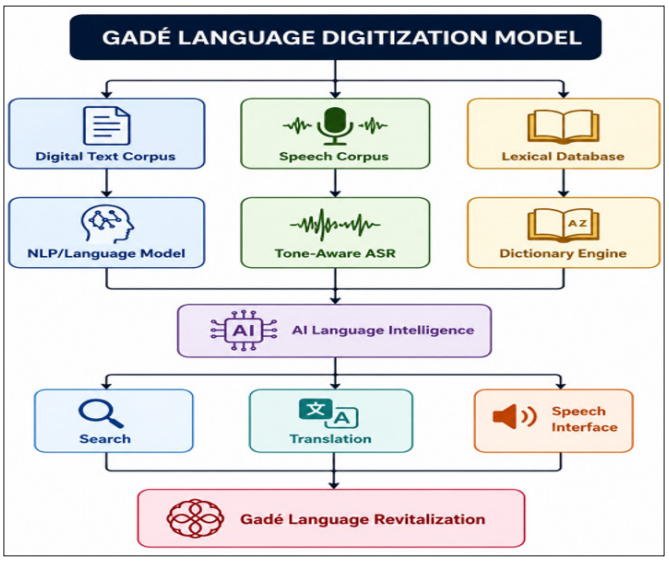


Figure 20: Gadé Language Digitalization Model. Originally compiled by the Author, GT Obadiah at August 9, 2026

Dataset Requirements
A major contribution of the proposed model is the construction of a Gadé Speech and Tone Corpus. Each recording should ideally be associated with:
I. native-speaker identification;
II. dialect/geographical classification;
III. orthographic transcription;
IV. phonemic transcription;
V. tonal annotation;
VI. word and sentence boundaries;
VII. timestamp information;
VIII. speaker demographic metadata where ethically appropriate;
IX. recording environment;
X. linguistic validation by competent Gadé speakers or linguists.
Particular attention should be given to tonal minimal pairs, because they provide controlled examples for determining whether the model can distinguish lexical items whose phonetic composition is similar but whose tonal patterns differ.

Importantly, this ASR pipeline should not exist in isolation. In your paper, it can be positioned as one layer of a broader architecture:

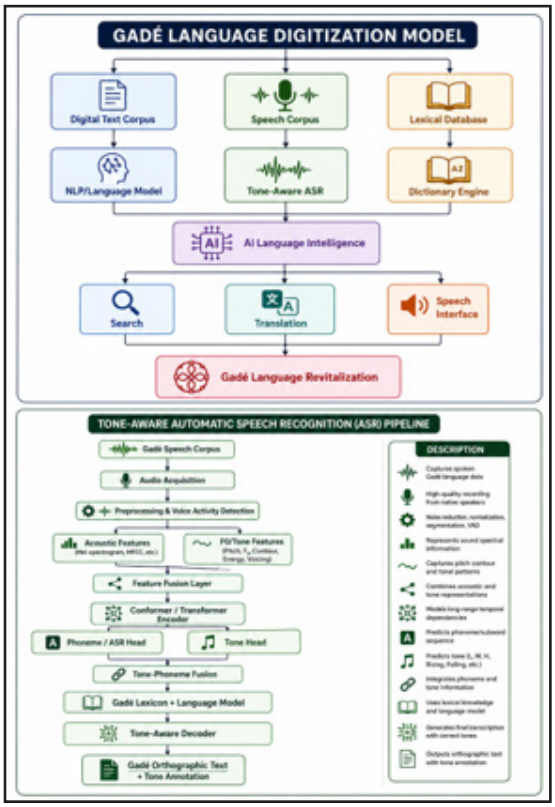


Figure 21: Position within the Overall Digitization Model. Originally compiled by the Author, GT Obadiah at August 9, 2026

Theoretical Architecture: The proposed Tone-Aware Automatic Speech Recognition (ASR) component is conceptualized as a low-resource, tone-sensitive speech-processing architecture specifically adapted to the linguistic characteristics of Gadé. The architecture combines established speech-processing technologies with a Gadé-specific linguistic layer. Established technologies include speech preprocessing, acoustic-feature extraction, fundamental-frequency (F0) analysis, Transformer/Conformer encoders, phoneme recognition, language modelling, and neural sequence decoding. Their integration within a tonal indigenous-language framework constitutes the proposed architecture.

The architecture consists of five principal stages: (i) speech acquisition and preprocessing, (ii) acoustic and tonal feature extraction, (iii) multimodal feature fusion and neural encoding, (iv) phoneme-tone recognition and linguistic decoding, and (v) orthographic transcription with tonal annotation.

The first stage receives speech recordings from the Gadé Speech Corpus. Voice Activity Detection (VAD), noise reduction, normalization, segmentation, and quality control are applied before feature extraction. The second stage generates two complementary representations. Acoustic features capture spectral and phonetic characteristics, while the tonal stream extracts F0, pitch trajectory, energy, voicing characteristics, and temporal tonal patterns.

The two streams are subsequently combined through a feature-fusion layer and processed by a Conformer or Transformer encoder. Two recognition heads are proposed: a Phoneme/ASR Head, responsible for recognizing the underlying speech sequence, and a Tone Head, responsible for identifying the corresponding tonal characteristics. Their outputs are integrated before being passed to a Gadé lexicon and language model. The final decoder produces orthographic Gadé text and, where applicable, explicit tone annotation.

Thus, the architecture does not treat pitch merely as an incidental acoustic property. It treats tone as a linguistic representation that participates directly in recognition and decoding.

Gadé Speech Dataset Construction: A major component of the proposed model is the development of a Gadé Speech and Tone Corpus. The scarcity of computational resources for indigenous Nigerian languages presents a significant limitation to the direct application of existing high-resource ASR systems. Consequently, dataset development is considered a foundational component of the digitization model.

The corpus should contain recordings from competent native Gadé speakers representing relevant linguistic and geographical variation. Each speech sample should be paired with a validated orthographic transcription and, where possible, phonemic and tonal annotations.
The proposed dataset schema includes:

Table 3: Dataset Construction. Compiled originally by the Corresponding author, Dr GT Obadiah, August 10, 2026 at 11:30pm (Nigerian Time)

Dataset Element	Description
Audio ID	Unique identifier for each recording
Speaker ID	Anonymized speaker identifier
Speech recording	Original high-quality audio
Orthographic transcription	Standard Gadé written representation
Phonemic transcription	Phoneme-level representation
Tone annotation	Tone category or contour
Word segmentation	Word-level boundaries
Time alignment	Start/end timestamps
Dialect/variant	Relevant linguistic variety
Speech context	Word, phrase, sentence, narrative, etc.
Recording condition	Studio, field, indoor, outdoor, noisy, etc.
Validation status	Reviewed/verified linguistic annotation

Particular emphasis should be placed on tonal minimal pairs, homophonous forms, frequently occurring lexical items, culturally specific vocabulary, personal and geographical names, traditional expressions, and naturally occurring conversational speech.

The dataset should be divided into training, validation, and test partitions, with speaker-independent separation wherever feasible. This prevents the model from simply learning speaker-specific acoustic characteristics and provides a more reliable assessment of generalization.

Model Training Strategy: The proposed model employs a multi-task learning strategy in which transcription, phonemic recognition, and tonal recognition are jointly optimized.
The overall training objective can be represented as:

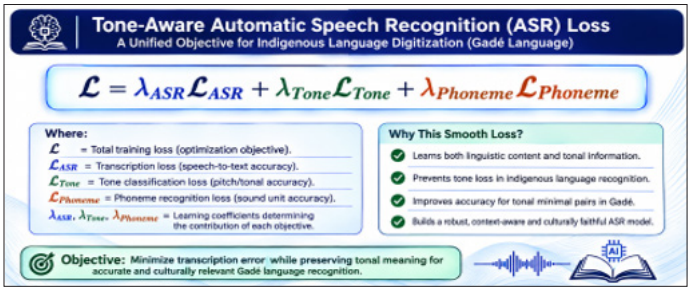


Figure 22: Tone-Aware Automatic Recognition. Originally compiled by the Author, GT Obadiah at August 9, 2026

The formulation allows the model to optimize speech recognition while simultaneously learning the phonological and tonal characteristics of Gadé. The weighting coefficients should not be arbitrarily fixed without experimentation. They should be determined through validation experiments, potentially using grid search, Bayesian optimization, or adaptive multi-task weighting.

For a low-resource language, transfer learning and self-supervised speech representation learning may substantially reduce the amount of labelled Gadé speech required. A pretrained multilingual speech encoder can serve as the acoustic foundation and subsequently be adapted using the Gadé corpus. However, the pretrained model should not be assumed to possess adequate Gadé phonological or tonal representations. Fine-tuning and explicit tone modelling remain necessary.

Data augmentation may additionally be employed through controlled variations in noise, speaking rate, amplitude, and other acoustic conditions. Care must be taken with pitch-shifting augmentation because uncontrolled modification of F0 could alter the linguistic tone and consequently introduce incorrect labels.

Evaluation Framework
Evaluation should extend beyond conventional ASR accuracy because a transcription can be orthographically close to the reference while still containing a linguistically significant tonal error. The proposed evaluation framework therefore contains four complementary dimensions. Automatic Speech.

Conflict of Interest
The author declares that there are no conflicts of interest regarding the publication of this study. The research was conducted independently, and there are no financial, institutional, or personal relationships that could have inappropriately influenced the content, interpretation, or conclusions presented in this paper.

Authour's Contribution
The author conceived and designed the study, conducted the research, wrote and revised the manuscript, and read and approved the fnal version for submission.

Acknowledgement
The author expresses sincere appreciation to scholars in anthropology, ethnology, and legal studies whose works provided valuable intellectual foundations for this research. Special recognition is given to the custodians of Gadé cultural knowledge,

whose traditions and worldview continue to inspire academic inquiry and preservation efforts.

The author also acknowledges the support and intellectual environment provided by the Institute of Gade and Classical Studies, which has contributed significantly to the conceptual development of the Gaboic Theory of Evolution.

Finally, the author extends gratitude to colleagues, reviewers, chief editors, and board of International Journal of Anthropology and Enthology, and all those who offered constructive insights during the development of this work.

Funding
This research received no funding from any public, private, commercial, or non-profit funding agencies. The study was entirely birthed by the author as part of ongoing scholarly work on Language Digitalization Model using Artificial Intelligence and and Machine Learning with specific focus on Gadé Language and Gadé studies.

Data availability
No datasets were generated or analyzed during the current study. The next volume would focus on Data Pipeline and trainer on Indigenous and Minor Ethnic groups which are underdocumented and modernized.

Conclusion, and Further Discussion
This volume/article has successfully discussed its purpose

References

1. Zheng Z, Wu X, Srihari R (2004) Feature selection for text categorization on imbalanced data, ACM SIGKDD Explor. Newsl. 6: 80-89.
2. Dang NC, Moreno-García MN, De F, la Prieta (2020) Sentiment analysis based on deep learning: a comparative study, Electronics 9: 3.
3. Aldunate A, Maldonado S, Vairetti C, Armelini G (2022) Understanding customer satisfaction via deep learning and natural language processing, Expert Syst. Appl. 209.
4. Kaur G, Sharma A (2023) A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. J Big Data 10: 1.
5. Feldman R (2013) Techniques and applications for sentiment analysis: the main applications and challenges of one of the hottest research areas in computer science.
6. Chai DM, Moulitsas I, Bisandu DB (2024) Understanding the relevance of parallelising machine learning algorithms using CUDA for sentiment analysis, Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence: 28-38.
7. Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR (2021) ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis, Future Gener Comput Syst. 115: 279-294.
8. Mao Y, Liu Q, Zhang Y (2024) Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review, King Saud bin Abdulaziz University.
9. Mahendhiran PD, Kannimuthu S (2018) Deep learning techniques for polarity classification in multimodal sentiment analysis. Int J Inf Technol Decis Mak 17: 883-910.
10. Bharti SK, Gupta RK, Shukla PK, Hatamleh WA, Tarazi H, et al. (2022) Multimodal sarcasm detection: a deep learning approach. Wirel Commun Mob Comput.
11. Liu Y, Li P, Hu X (2022) Combining context-relevant features with multi-stage attention network for short text classification. Comput Speech Lang 71.
12. Diwan T, Tembhurne JV (2022) Sentiment analysis: a convolutional neural networks perspective, Multimed. Tools Appl 81: 44405-44429.
13. Mewada A, Dewang RK (2023) SA-ASBA: a hybrid model for aspect-based sentiment analysis using synthetic attention in pre-trained language BERT model with extreme gradient boosting J Supercomput 79: 5516-5551.
14. Lai S, Liu K, He S, Zhao J (2016) How to generate a good word embedding, IEEE Intell. Syst 31: 5-14.
15. Levy O, Goldberg Y (2014) Neural word embedding as implicit matrix factorization Adv. Neural Inf Process Syst 27: 2177-2185.
16. Li M, Lu Q, Long Y, Gui L (2017) Inferring affective meanings of words from word embedding. IEEE Trans Affect Comput 8: 443-456.
17. Song M, Park H, Shin KS (2019) Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean Inf Process Manag 56: 637-653.
18. Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation.
19. Lauriola I, Lavelli A, Aiolli F (2022) An introduction to deep learning in natural language processing: models, techniques, and tools, Neurocomputing 470: 443-456.
20. Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: a survey, Wiley Interdiscip. Rev Data Min Knowl. Discov 8: 4.
21. Acheampong FA, Nunoo-Mensah H, Chen W (2021) Transformer models for text-based emotion detection: a review of BERT-based approaches Artif Intell Rev 54: 5789-5829.
22. Sundermeyer M, Schlüter R, Ney H (2012) LSTM neural networks for language modeling. Proceedings of the 13th Annual Conference of the International Speech Communication Association: 194-197.
23. Mehta Y, Majumder N, Gelbukh A, Cambria E (2020) Recent trends in deep learning-based personality detection Artif. Intell Rev 53: 2313-2339.
24. Chatterjee A, Gupta U, Chinnakotla MK, Srikanth R, Galley M, Agrawal P (2019) Understanding emotions in text using deep learning and big data. Comput Hum Behav 93: 309-317.
25. Rezaeinia SM, Rahmani R, Ghodsi A, Veisi H (2019) Sentiment analysis based on improved pre-trained word embeddings. Expert Syst Appl 117: 139-147.
26. Onan A (2022) Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification J King Saud Univ Comput. Inf Sci 34: 2098-2117.
27. Tan KL, Lee CP, Lim KM (2023) RoBERTa-GRU: a hybrid deep learning model for enhanced sentiment analysis Appl Sci 13: 3915.
28. Gandhi UD, Kumar PM, Babu GC, Karthick G (2021) Sentiment analysis on Twitter data by using convolutional neural network (CNN) and long short-term memory (LSTM) Wirel Pers Commun.
29. Islam MdS, Ghani NA (2022) A novel BiGRU-BiLSTM model for multilevel sentiment analysis using deep neural network with BiGRU-BiLSTM. Lecture Notes in Electrical Engineering 730: 403-414.
30. Tam S, Ben Said R, Tanriover OO (2021) A ConvBiLSTM deep learning model-based approach for Twitter sentiment

classification. IEEE Access 9: 41283-41293.
31. Tan KL, Lee CP, Anbananthen KSM, Lim KM (2022) RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access 10: 21517-21525.
32. Young T, Hazarika D, Poria S, Cambria E (2018) Recent Trends in Deep Learning Based Natural Language Processing, IEEE.
33. Xu J, Chen D, Qiu X, Huang X (2016) Cached long short-term memory neural networks for document-level sentiment classification, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: 1660-1669.
34. Moraes R, Valiati JF, Gavião Neto WP (2013) Document-level sentiment classification: an empirical comparison between SVM and ANN, Expert Syst. Appl. 40: 621-633.
35. Chatterjee A, Gupta U, Chinnakotla MK, Srikanth R, Galley M, Agrawal P (2019) Understanding emotions in text using deep learning and big data. Comput Hum Behav 93: 309-317.
36. Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing 1532-1543.
37. Tang D, Wei F, Yang N, Zhou M, Liu T, at all. (2014) Learning sentiment-specific word embedding for Twitter sentiment classification, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics 1555-1565.
38. Zhou X, Wan X, Xiao J (2016) Attention-based LSTM Network for cross-lingual sentiment classification, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing 247-256.
39. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification, Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics 1480-1489.
40. He R, Lee WS, Ng HT, Dahlmeier D (2018) Exploiting document knowledge for aspect-level sentiment classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics 579-585.
41. Basiri ME, Abdar M, Cifci MA, Nemati S, Acharya UR (2020) A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. Knowl Based Syst 198: 105949.
42. Wang J, Yu LC, Lai KR, Zhang X (2016) Dimensional sentiment analysis using a regional CNN-LSTM model. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics 225-230.
43. Wen S, Li J (2018) Recurrent convolutional neural network with attention for Twitter and Yelp sentiment classification arc model for sentiment classification, Proceedings of the ACM International Conference Proceeding Series.
44. Liu G, Guo J (2019) Bidirectional LSTM with attention mechanism and convolutional layer for text classification. Neurocomputing 337: 325-338.
45. Rezaeinia SM, Rahmani R, Ghodsi A, Veisi H (2019) Sentiment analysis based on improved pre-trained word embeddings Expert Syst Appl 117: 139-147.
46. Perumal T, Mustapha N, Mohamed R, Shiri FM (2024) A comprehensive overview and comparative analysis on deep learning models. J Artif. Intell 6: 301-360.
47. Aldhyani THH, Alkahtani H (2021) A bidirectional long short-term memory model algorithm for predicting COVID-19 in Gulf countries. Life 11.
48. Liciotti D, Bernardini M, Romeo L, Frontoni E (2020) A sequential deep learning application for recognising human activities in smart homes. Neurocomputing 396: 501-513.
49. Shiri FM, Perumal T, Mustapha N, Mohamed R, Ahmadon MAB, Yamaguchi S (2024) Recognition of student engagement and affective states using ConvNeXt-large and ensemble GRU in E-learning. Proceedings of the 2024 12th International Conference on Information and Education Technology 30-34.
50. Chai C et al. (2022) A multifeature fusion short-term traffic flow prediction model based on deep learnings. J Adv Transp 2022.
51. Tasdelen A, Sen B (2021) A hybrid CNN-LSTM model for pre-miRNA classification. Sci Rep 11: 1.
52. Qin L, Yu N, Zhao D (2018) Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Teh Vjesn 25: 528-535.
53. Babu BP, Narayanan SJ (2022) One-vs-all convolutional neural networks for synthetic aperture radar target recognition. Cybern Inf Technol 22: 179-197.
54. Li Z, Liu F, Yang W, Peng S, Zhou J (2022) A survey of convolutional neural networks: analysis, applications, and prospects IEEE Trans. Neural Netw Learn Syst 33: 6999-7019.
55. Lu W, Li J, Wang J, Qin L (2021) A CNN-BiLSTM-AM Method for Stock Price Prediction.
56. Chen L, Li S, Bai Q, Yang J, Jiang S, Miao Y (2021) Review of Image Classification Algorithms Based on Convolutional Neural Networks.
57. Liu Y (2019) RoBERTa: a robustly optimized BERT pretraining approach.
58. Sanh V, Debut L, Chaumond J, Wolf T (2020) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.
59. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using siamese BERT-networks.
60. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 Conference of the North 4171-4186.
61. Nguyen DQ, Vu T, Nguyen AT (2020) BERTweet: a pre-trained language model for English tweets. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing 9-14.
62. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Report Stanford 1: 12.
63. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique.
64. Vadicamo L (2017) Cross-Media learning for image sentiment analysis in the wild. Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops 308-317.
65. McAuley J, Targett C, Shi Q, Van Den Hengel A (2015) Image-based recommendations on styles and substitutes. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval: 43-52.
66. He R, McAuley J (2016) Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. Proceedings of the 25th International World Wide Web Conference: 507-517.
67. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance.
68. Demsar J (2006) Statistical comparisons of classifiers over

multiple data sets.
69. Shefani RG, Jeyalaksshmi S (2025) A Stacked Ensemble Approach for Robust Sentiment Classification on Twitter Data. IEEE: 1217-1224.
70. Sesmero MP, Ledezma AI, Sanchis A (2015) Generating ensembles of heterogeneous classifiers using stacked generalization. Wiley Interdiscip Rev Data Min Knowl Discov 5: 21-34.
71. Chai DM, Moulitsas I, Bisandu DB (2026) DAFN: A dual attention fusion network for textual data classification, Knowledge-Based Systems 351: 116791.
72. Abdelgaber N, Nikolopoulos C (2020) Overview on Quantum computing and its applications in artificial intelligence, Proceedings of the 2020 IEEE 3rd International Conference on Artificial Intelligence and Knowledge Engineering: 198-199.
73. Acemoglu D, Restrepo P (2018) The race between man and machine: Implications of technology for growth, factor shares, and employment, American Economic Review 108: 1488-1542.
74. Agarwal S, Bhardwaj G, Saraswat E, Singh N, Aggarwal R, et all. (2022) Insurtech Fostering Automated Insurance Process using Deep Learning Approach 386-392.
75. Arrazola JM, Bromley TR, Izaac J, Myers CR, Bradler K, Killoran N (2019) Machine learning method for state preparation and gate synthesis on photonic quantum computers. Quantum Science and Technology 4.
76. Baniata H (2024) SoK: Quantum computing methods for machine learning optimization. Quantum Machine Intelligence 6: 47.
77. Cai XD, Wu D, Su ZE, Chen MC, Wang XL, et all. (2015) Entanglement-based machine learning on a quantum computer, Physical Review Letters 114.
78. Castaldo D, Rosa M, Corni S (2021) Quantum optimal control with quantum computers. A hybrid algorithm featuring machine learning optimization Physical Review A 103.
79. Cherbal S, Zier A, Hebal S, Louail L, Annane B (2024) Security in internet of things: A review on approaches based on blockchain machine learning cryptography. and quantum computing J Supercomput 80: 3738-3816.
80. Dai Z, Nahum A (2020) Quantum criticality of loops with topologically constrained dynamics, Physical Review Research 2: 033051.
81. Damasevicius R, Misra S (2024) The rise of industry 6.0: 478-494.
82. Duarte LT, Deville Y (2024) Quantum-assisted machine learning by means of adiabatic quantum computing. 2024 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium 371-375.
83. Duggal AS, Malik PK, Gehlot A, Singh R, Gaba GS, et all. (2022) A sequential roadmap to industry 6.0: Exploring future manufacturing trends, IET Communications 16: 521-531.
84. Espuny M, da Motta Reis JS, Monteiro Diogo GMM, Reis Campos TL, de Mello Santos VH, et al. (2021) COVID-19 The importance of artificial intelligence and digital health during a pandemic 27-32.
85. Morgan W (1936b) The distribution of the Gadé people in Central Nigeria. British Colonial Administration.
86. Ngũgĩ wa Thiong'o (1986) Decolonising the mind: The politics of language in African literature, Heinemann.
87. Obadiah GT (2013) The Gabo religion of the Gade people. The etymological conception in ancient and modern African civilisation and industrialisation, Pan-African Institute of Paralegal Studies.
88. Obadiah GT (2015a) Gaboic theory of evolution, Brobadiah Printing and Publishing House.
89. Obadiah GT (2015b) Gaboic theory of evolution, Institute of Gade and Classical Studies.
90. Obadiah GT (2022) Gaboic theory of evolution, Institute of Gade and Classical Studies.
91. Obadiah GT (2023a) Gaboism: An introduction to Gadé philosophy and existential thought, Institute of Gade and Classical Studies.
92. Obadiah GT (2023b) Gade lexis and structure, Sahab Printing Press.
93. Obadiah GT (2023c) Gade tonal processes and grammatical melodies, Brobadiah Printing and Publishing House.
94. Obadiah GT (2024a) A linguistic and sociocultural analysis of Gade: Tonal phonology. orthographic evolution numeracy religion and education in pre-and-post-colonial African contexts Institute of Gade and Classical Studies.
95. Obadiah GT (2024b) GADE keyboard: Textual and architectural samples instructions and validity source code., National Association of Gade Students.
96. Obadiah GT (2025a) Gade people Origins, civilisation and industrialisation-An anthropological and historical study, GAS Journal of Arts Humanities and Social Sciences 3: 1-20.
97. Obadiah GT (2025b) Gaboism: An indigenous philosophy of mind and metaphysics, Zenodo.
98. Obadiah GT, Eling F (2025) The origin of democracy and republican systems of government of the Gade people in precolonial African contexts, SSR Journal of Arts, Humanities and Social Sciences 2: 35-53.
99. Obadiah GT (2026) Adakpụ Ancestral Guidance – The Unending Voice of Ancestors.
100. Obadiah GT (2026) Ancient AI vs Modern AI the Moral Machine Ethics Native AI Framework in Universal Technology Paradigm of the Gadé People. Journal of Data Analytics and Artificial Intelligence.
101. GWF (2026) GADÉ Worldwide Essential Guide Magazine: The Definitive Voice of Gadé History Culture Language and Global Identity January 10 2026 Miss Gade Nasarawa Second Edition, Government Day Secondary School, Uke, Karu LGA, Nasarawa State, Nigeria.