Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 8 (2026) 3097

Journal of the Nigerian Society of Physical Sciences

Bridging context gaps in low-resource language chatbots through multilevel attention and hybrid embedding approaches

Godson Chetachi Uzoaru*, Ikechukwu Ignatius Ayogu, Juliet Nnenna Odii, Aloysius Chijioke Onyeka

Department of Computer Science, Federal University of Technology, Owerri, Imo State, Nigeria

Abstract

Conversational agents for low-resource languages (LRLs), such as Igbo, face major challenges, including limited annotated data, code-switching, and weak contextual coherence in multi-turn dialogue. This study proposes a multilevel-attention and hybrid-embedding framework that integrates FastText subword representations with multilingual BERT (mBERT) to improve semantic understanding and context retention. The architecture applies hierarchical attention at the word, utterance, and dialogue levels, enabling effective modeling of conversational dependencies and reducing context drift. The model was evaluated on a curated Igbo–English conversational dataset and benchmarked against long short-term memory (LSTM), Transformer, FastText, mBERT, and XLM-R baselines. For response generation, the proposed framework achieved a bilingual evaluation understudy (BLEU) score of 44.1%, a longest-common-subsequence recall-oriented understudy for gisting evaluation (ROUGE-L) score of 60.3%, and a context-retention accuracy (CRA) of 81.5%. For intent classification, it attained an F1-score of 87.3% and an area under the receiver operating characteristic curve (ROC-AUC) of 0.91; for context-dependency detection, it achieved an F1-score of 84.3%. The framework also reduced inference latency and was robust to code-switching and noisy conversational input. Human evaluation confirmed improvements in response clarity, cultural relevance, and multi-turn coherence. The findings show that hybrid embeddings combined with multilevel attention provide an effective and scalable approach to conversational AI for LRLs, with potential applicability to other African languages.

DOI: 10.46481/jnsps.2026.3097

Keywords: Low-resource languages, Igbo conversational agent, Multilevel attention, Hybrid embeddings, Context-aware dialogue systems

Article History:
Received: 18 August 2025
Received in revised form: 21 June 2026
Accepted for publication: 02 July 2026
Available online: 17 August 2026

Communicated by: B. J. Falaye

1. Introduction

In recent years, conversational agents (chatbots) have evolved from rule-based systems to highly sophisticated neural models powered by transformer architectures [1]. These advancements have enabled chatbots to exhibit human-like dialogue abilities, improved context retention, and adaptive reasoning [2]. However, most of these developments have been concentrated on high-resource languages [3] such as English [4], Mandarin, and Spanish [5], leaving many African languages including Igbo, significantly underrepresented in natural language processing (NLP) research.

Low-resource languages (LRLs) face two interconnected challenges in conversational AI [6]: the scarcity of annotated datasets [7] and the lack of pretrained models fine-tuned for their linguistic structures [8]. Igbo, a tonal and morphologically rich language spoken by over 27 million people in Nigeria and the diaspora, presents additional complexities such as

*Corresponding Author Tel. No.: +2348068475798.
Email address: uzoarugc@cliffordUni.edu.ng (Godson Chetachi Uzoaru)

code-switching between Igbo and English, morphological inflections, and limited digital presence [9]. These limitations result in poor dialogue relevance [10], incoherent responses, and cultural inappropriateness when using state-of-the-art chatbots originally trained on high-resource data [11].

Recent transformer-based models such as BERT, GPT-4, XLM-R, and mT5 have demonstrated significant improvements in contextual understanding and semantic alignment [12]. Nevertheless, their direct application to Igbo suffers from weak handling of out-of-vocabulary (OOV) words, insufficient cultural grounding, and reduced performance in multi-turn conversations [13]. To address these gaps, research has explored hybrid embedding strategies (e.g., combining FastText's subword modeling with BERT's contextual encoding) and multilevel attention mechanisms that dynamically allocate focus across word, sentence, and dialogue levels. These approaches promise better semantic alignment, context retention, and robustness against linguistic variability in low-resource settings.

This study introduces a multilevel attention and hybrid embedding framework for Igbo conversational agents. The model integrates word-, utterance-, and dialogue-level attention with combined FastText and mBERT embeddings to capture both subword and contextual information. This design improves dialogue coherence, supports code-switching, and enhances cultural relevance in low-resource settings without relying on large annotated datasets.

The key contributions of this paper are summarized as follows:

- Model Framework: We introduce a multilevel attention architecture integrated with hybrid embeddings (FastText + mBERT) to enhance contextual understanding, semantic continuity, and out-of-vocabulary (OOV) handling in Igbo conversational systems.
- System Implementation: We develop a functional end-to-end chatbot prototype with an interactive interface, enabling real-time user interaction and practical evaluation.
- Comprehensive Evaluation: The proposed model is rigorously evaluated against multiple baselines (LSTM, Transformer, FastText, mBERT, and XLM-R) using linguistic, classification, and system-level metrics, including BLEU, ROUGE-L, F1, CRA, and latency.
- Low-Resource NLP Advancement: The framework is specifically designed for low-resource settings and demonstrates adaptability to other African languages, contributing to the development of inclusive and culturally aware conversational AI.

The remainder of this paper is organized as follows: Section 2 reviews related work on low-resource language NLP, hybrid embeddings, and attention-based models. Section 3 presents the proposed methodology, including system architecture, dataset design, and mathematical formulation. Section 4 describes the experimental setup, deployment evaluation, and results. Section 5 discusses the findings, while Sections 6–8 outline the conclusion, limitations, and future work. The final sections include data availability, acknowledgements, conflict of interest statements, and references.

2. Literature review

2.1. Distributional semantics for low-resource languages

Distributional word representations such as Word2Vec, GloVe, and FastText have been widely used in natural language processing to capture semantic relationships between words [14]. Word2Vec learns word representations based on local context, while GloVe incorporates global co-occurrence statistics [15]. FastText extends these approaches by modeling subword information, making it particularly effective for morphologically rich and low-resource languages [16].

In African low-resource languages, FastText has shown better performance due to its ability to handle out-of-vocabulary words and morphological variations [24]. However, these models generate static embeddings, meaning that word meanings remain fixed regardless of context. This limitation reduces their effectiveness in conversational systems, where meaning depends on dialogue flow and user intent [14]. Table 1 summarizes the core distributional models used in NLP and highlights their suitability for low-resource languages. It shows that while Word2Vec and GloVe capture semantic relationships effectively, they struggle with morphology and rare words, whereas FastText provides better support for low-resource settings due to its subword modeling capability.

2.2. Transformer models and contextual representation

Transformer-based models such as BERT, GPT, XLM-R, and mT5 have significantly improved natural language understanding by introducing self-attention mechanisms that capture contextual relationships [25]. Unlike static embeddings, these models generate context-aware representations, allowing words to adapt meaning based on surrounding text [16].

Multilingual models such as mBERT and XLM-R enable cross-lingual transfer and have shown promising results in low-resource scenarios. However, their effectiveness is limited when applied to African languages due to insufficient training data, complex morphology, and frequent code-switching. In conversational settings, these models may also struggle with maintaining coherence across multiple dialogue turns.

Table 2 presents commonly used transformer-based models and their performance characteristics in low-resource contexts. Although these models provide strong contextual understanding, their effectiveness is reduced in African languages due to data scarcity, morphological complexity, and limited adaptation to local linguistic features.

2.3. Hybrid embeddings and attention mechanisms

To overcome the limitations of both static and contextual models, recent research has explored hybrid embedding approaches that combine subword-aware representations (e.g., FastText) with contextual embeddings (e.g., BERT or mBERT). This combination aims to improve robustness in low-resource

Table 1: Core distributional models for low-resource languages.

Model	Core Mechanism	Strengths	Limitations	Suitability for low-resource languages	Sources
Word2Vec	CBOW / Skip-gram	Efficient, captures local semantics	Poor OOV handling, ignores morphology	Low	[14]
GloVe	Global co-occurrence matrix	Captures global relationships	Weak on rare words and tonal variation	Moderate	[15]
FastText	Subword n-grams	Handles morphology and OOV effectively	Still static, lacks contextual adaptation	High	[16]

Table 2: Transformer models and limitations in low-resource contexts.

Model	Key capability	Strengths	Limitations in low-resource languages	Typical use cases	Sources
BERT	Bidirectional encoding	Strong contextual understanding	Limited cross-lingual generalization	Classification, named-entity recognition	[17]
GPT	Autoregressive generation	Fluent text generation	Data-intensive, weak in LRL adaptation	Dialogue, text generation	[18]
XLM-R	Multilingual pretraining	Cross-lingual transfer	Underperforms in tonal/morph-rich languages	Translation, NLU	[19]
mT5	Text-to-text framework	Flexible across tasks	Requires large-scale data	Multilingual NLP	[20]

Table 3: Hybrid embedding and attention approaches.

Approach	Description	Strengths	Limitations	Application scope	Sources
FastText + BERT	Combines subword + contextual embeddings	Improves OOV and semantic understanding	Limited dialogue integration	Classification, named-entity recognition	[21]
Hierarchical Attention	Multi-level attention over text structure	Captures long-range dependencies	computationally expensive	Document modeling	[22]
Hybrid + Attention models	Fusion of embeddings with attention layers	Better representation learning	Rarely applied to conversational systems	General NLP tasks	[23]

settings by capturing both morphological and contextual information.

Similarly, attention mechanisms particularly hierarchical and multilevel attention have been used to model relationships at different levels of text, including word-level, sentence-level, and document-level dependencies. These techniques have shown strong performance in tasks such as text classification and machine translation [27].

However, most existing studies apply hybrid embeddings and attention mechanisms to isolated NLP tasks rather than integrating them into end-to-end conversational systems. As a result, their effectiveness in multi-turn dialogue, context retention, and user interaction remains limited.

Table 3 outlines existing hybrid embedding and attention-based approaches. It indicates that while combining embeddings and attention improves representation learning, most prior work focuses on isolated NLP tasks and does not fully address conversational or multi-turn dialogue scenarios.

2.4. Conversational AI in African low-resource languages

Conversational AI research in African languages such as Igbo, Yorùbá, and Twi is still emerging. Existing systems are largely limited to translation, speech processing, or basic intent classification. Few studies address multi-turn dialogue, contextual coherence, and cultural relevance.

These languages present unique challenges, including limited annotated datasets, tonal variations, morphological complexity, and frequent code-switching with English. As a result, models trained on high-resource languages often produce irrelevant or incoherent responses when applied directly to African conversational contexts [28].

Table 4 highlights the major challenges affecting conversational AI in African low-resource languages. These include

data scarcity, morphological richness, tonal variation, and code-switching, all of which contribute to reduced model performance and poor dialogue quality.

2.5. Positioning and contribution of the proposed approach

While previous studies have explored hybrid embeddings and attention mechanisms, their application to conversational systems in African low-resource languages remains limited. This study differs from existing work in several key ways.

First, it introduces a task-specific integration of FastText and mBERT within a unified conversational architecture, rather than applying these techniques to isolated NLP tasks. Second, it employs multilevel attention tailored for dialogue modeling, enabling the system to capture dependencies across word-level, utterance-level, and conversation-level contexts.

Third, the framework is designed specifically for Igbo conversational scenarios, addressing challenges such as code-switching, morphological variation, and limited data availability. Finally, the study provides a comprehensive evaluation strategy, combining automatic metrics, error analysis, and human evaluation, along with a functional prototype for real-world interaction.

Table 5 compares the proposed framework with closely related prior approaches. It shows that the main contribution of this study lies in integrating hybrid embeddings and multilevel attention within a unified conversational system, supported by comprehensive evaluation and real-world applicability.

3. Methodology and model architecture

3.1. Overview of the framework

The proposed Igbo Conversational Agent (IgboCA) integrates hybrid embeddings (FastText + mBERT) with a multilevel attention mechanism to enhance dialogue relevance, context retention, and code-switching handling in low-resource settings. The model processes incoming user utterances through four main stages:

i. Data Preprocessing – Tokenization, normalization, tone marking, and code-switch detection.
ii. Hybrid Embedding Layer – Combination of FastText subword embeddings and mBERT contextual embeddings.
iii. Multilevel Attention Mechanism – Sequential attention layers for word, sentence, and utterance levels.
iv. Response Generation – Context-aware decoding using a Transformer decoder.

Figure 1 illustrates the architecture of the proposed Multilevel Attention Igbo Chatbot, designed to address the linguistic and technical challenges of low-resource conversational AI. The architecture integrates hybrid embeddings, hierarchical attention mechanisms, and dialogue-level contextual memory to ensure accurate semantic interpretation and coherent multi-turn dialogues.

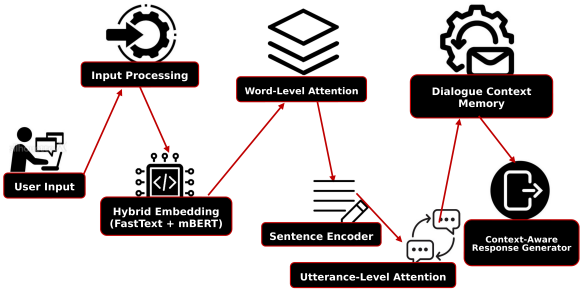

Figure 1: Proposed system architecture for the multilevel-attention Igbo chatbot.

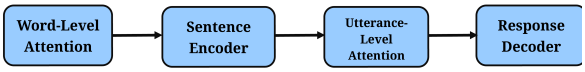

Figure 2: Hierarchical attention in dialogue processing.

The framework is optimized for the Igbo language, characterized by tonal variations, agglutinative morphology, and frequent code-switching with English. To meet these challenges, the model applies mathematical optimization techniques that improve semantic dependency modeling, context retention, and response generation accuracy under low-resource constraints.

3.1.1. Application in Igbo conversational agents

The proposed Igbo chatbot applies multilevel attention to dynamically track discourse flow and semantic alignment across conversation layers. For instance, if a user says:

Turn 1: “Ị gara ahịa?” (Did you go to the market?)
Turn 2: “Ee, m zụtara akwa.” (Yes, I bought some clothes.)
Turn 3: “Olee mgbe i rutere?” (When did you get home?)

A basic model may treat Turn 3 in isolation. However, hierarchical attention allows the system to track "ị gara" (you went) from Turn 1 and "zụtara akwa" (bought clothes) in Turn 2, providing a coherent and logically consistent response in Turn 3.

Figure 2 illustrates the architecture used in the proposed system. It begins with word-level attention, focusing on important tokens within an utterance [29]. These are processed through a sentence encoder that structures the syntax and aggregates intra-sentence meaning. At the next stage, utterance-level attention identifies which prior sentences are most relevant to the current user input [30]. The dialogue context memory captures conversational history and maintains cross-turn coherence. Finally, a response decoder generates fluent, context-aware replies [31]. This layered design significantly enhances the system’s ability to manage long-term dependencies, resolve ambiguous references, and respond naturally in low-resource language conversations.

3.1.2. Hybrid architecture integration

Table 6 shows that the hybrid FastText + mBERT approach outperforms all other embedding strategies for Igbo

Table 4: Challenges in African low-resource conversational AI.

Challenge	Description	Impact on chatbots
Data scarcity	Limited annotated corpora	Poor model training and generalization
Morphological Complexity	Rich word formation and inflection	Incorrect tokenization and semantics
Tonal Variation	Meaning depends on tone	Misinterpretation of intent
Code-Switching	Mixing local languages with English	Reduced contextual coherence
Cultural Context	Language tied to local usage patterns	Inappropriate or irrelevant responses

Table 5: Comparative analysis of the proposed framework and existing hybrid-attention models.

Feature / aspect	Prior hybrid NLP models	Proposed framework
Embedding strategy	hybrid (static + contextual)	Hybrid (FastText + mBERT, task-specific)
Attention mechanism	General attention	Multilevel dialogue-aware attention
Application domain	Classification / machine translation	Conversational AI (multi-turn dialogue)
Language focus	General multilingual	Igbo (framework designed for adaptation to other African low-resource languages)
Code-switch handling	Limited	Explicitly considered
Evaluation	Mostly automatic metrics	Automatic + human + error analysis
System implementation	Model-level only	End-to-end conversational prototype

chatbot development, offering very high context awareness, excellent OOV handling, and the highest overall suitability, while Word2Vec ranks lowest across all criteria.

3.1.3. Analysis of embedding performance

As shown in Figure 3, Word2Vec performs the weakest across all dimensions due to its lack of subword modeling and context encoding [32]. FastText improves OOV handling significantly but falls short on context alignment in longer sequences [33]. mBERT, while strong in capturing syntactic [34] and semantic relations, struggles with rare tokens [35], tone-sensitive words [36], and low-resource code-switching contexts [37]. The hybrid approach which combines FastText's subword-level detail with mBERT's deep contextual understanding achieves: Very high context sensitivity across long and multi-turn utterances. Excellent robustness to unseen or inflected words. Strong applicability in tonal, agglutinative, and multilingual dialogue systems, this performance is shown in Figure 3.

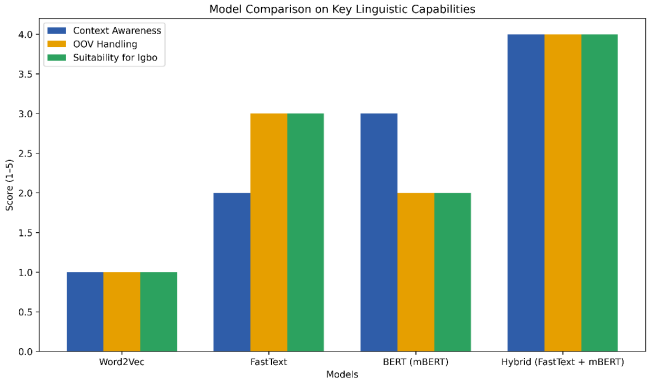

Figure 3: Model comparison on key linguistic capabilities.

3.2. Dataset description and reproducibility

The dataset used in this study is composed of two complementary components:

a) large monolingual Igbo corpus for representation learning, and
b) a curated Igbo–English conversational dataset for dialogue modeling and evaluation.

This separation ensures methodological clarity and addresses reproducibility concerns by distinguishing between pretraining data and task-specific conversational data.

It should be noted that the 383,449 monolingual Igbo sentences were used solely for representation learning and embedding construction, whereas the conversational modeling experiments were conducted using the curated dataset of 21,000 Igbo–English dialogue pairs. All quantitative results reported in Section 4 are based on the conversational dataset and its corresponding train, validation, and test partitions.

3.2.1. Data sources and composition

Table 7 presents diverse data sources combining formal text, lexical resources, and conversational datasets, ensuring rich language coverage and realistic dialogue modeling.

3.2.2. Conversational dataset structure

The conversational dataset consists of Igbo–English dialogue pairs derived from crowdsourced and manually constructed interactions (Table 8). The dataset captures realistic conversational patterns, including multi-turn dialogue flow, contextual dependencies, and code-switching behavior commonly observed in Igbo communication.

3.2.3. Data collection criteria and ethics

Data were included based on the following criteria:

- Valid Igbo or Igbo–English linguistic content
- Natural conversational or informational structure
- Removal of duplicated, noisy, or incomplete entries

All crowdsourced data were collected with informed consent, anonymized, and stripped of personally identifiable information. Public datasets were used in compliance with their respective licenses, while copyrighted materials were either licensed or paraphrased.

3.2.4. Preprocessing pipeline

Table 9 outlines the preprocessing pipeline, including cleaning, normalization, diacritic restoration, subword tokenization, and filtering, ensuring high-quality and consistent input for model training.

3.2.5. Annotation schema and procedure

Annotation was conducted by two native Igbo experts following a predefined guideline. Inter-annotator agreement was evaluated using Cohen's kappa ($\kappa = 0.82$), indicating strong consistency as depicted in Table 10.

3.2.6. Dataset splitting strategy

The conversational dataset was partitioned into training, validation, and test subsets using a stratified sampling approach to ensure balanced representation of intents, linguistic features, and dialogue structures as shown in Table 11.

3.2.7. Reproducibility and availability

To ensure reproducibility, the following resources are provided:

- Preprocessed conversational dataset (21,000 pairs)
- Monolingual corpus references and access instructions
- Annotation guidelines and label definitions
- Preprocessing and tokenization scripts

All materials are available in a public repository (see Data Availability section) to enable independent validation and reuse.

3.3. Optimization objectives

The proposed multilevel attention chatbot framework is designed to achieve two core objectives: linguistic performance and computational efficiency. These objectives guide the model design to ensure accurate, context-aware dialogue generation while maintaining efficient operation in low-resource environments.

3.3.1. Linguistic performance objectives

The computational component of the framework is designed to improve low latency, memory efficiency, and scalable training, as illustrated in Figure 4. The model employs a lightweight architecture to reduce inference time and support near real-time response generation. Memory efficiency is achieved through compact model representations and parameter-efficient optimization, thereby minimizing computational overhead. In addition, the framework incorporates transfer learning and data augmentation techniques to enable scalable training and maintain robust performance despite the limited availability of annotated Igbo language data.

As illustrated in Figure 4: Optimization Framework for Multilevel Attention-Based Igbo Conversational Agent, the linguistic component focuses on two key aspects: contextual continuity and hybrid embeddings.

The framework maintains contextual continuity across multi-turn conversations by leveraging attention mechanisms that capture dependencies between tokens, utterances, and dialogue history, thereby reducing contextual drift. In addition, hybrid embeddings combining FastText and mBERT are employed to enhance semantic representation. FastText captures subword and morphological patterns, while mBERT provides contextual understanding. This integration enables effective handling of out-of-vocabulary words, tonal variations, and Igbo–English code-switching, ensuring coherent and culturally relevant responses.

3.3.2. Computational efficiency objectives

The computational component focuses on low latency, memory efficiency, and scalable training, as shown in Figure 4.

The framework is optimized to achieve low inference latency, enabling near real-time responses. Memory efficiency is ensured through compact model representations and parameter-efficient design, reducing computational overhead. Furthermore, scalable training is supported through transfer learning and data augmentation strategies, allowing the model to perform effectively even with limited Igbo-specific data.

4. Mathematical formulation and implementation alignment

To ensure consistency between theory and implementation, the proposed framework is formulated using unified notation, and each component is explicitly mapped to its corresponding module in the architecture.

4.1. Notation and input representation

Let a dialogue be represented as

$$D = \{U_1, U_2, \ldots, U_T\}, \tag{1}$$

where D denotes a dialogue session, U_t represents the tth utterance, and T is the total number of utterances. Each utterance consists of a sequence of tokens:

$$U_t = \{x_1, x_2, \ldots, x_n\}. \tag{2}$$

Table 6: Comparison of embedding strategies for the Igbo language.

Embedding model	Context awareness	OOV handling	Suitability for Igbo
Word2Vec	Low	Poor	Limited
FastText	Moderate	Good	High
BERT (mBERT)	High	Moderate	Moderate
Hybrid (FastText + mBERT)	Very high	Excellent	Excellent

Table 7: Dataset sources and sizes.

Source	Type	Size	Role	Reference
BBC Igbo news	Sentence corpus	3,190 sentences	Formal language patterns	[38]
Igbo Radio & Kaoditaa	Lexical items	8,000 words	Vocabulary enrichment	-
IGBONLP + Literature (Ezeani *et al.*, Teta!, UloNche!)	Monolingual corpus	383,449 sentences	Embedding training	[37, 40]
Crowdsourcing & Forums	Conversational data	12,600 dialogue pairs	Real-world interactions	[39]
Manually constructed dialogues	Multi-turn conversations	8,400 dialogue pairs	Contextual control & Annotation	-

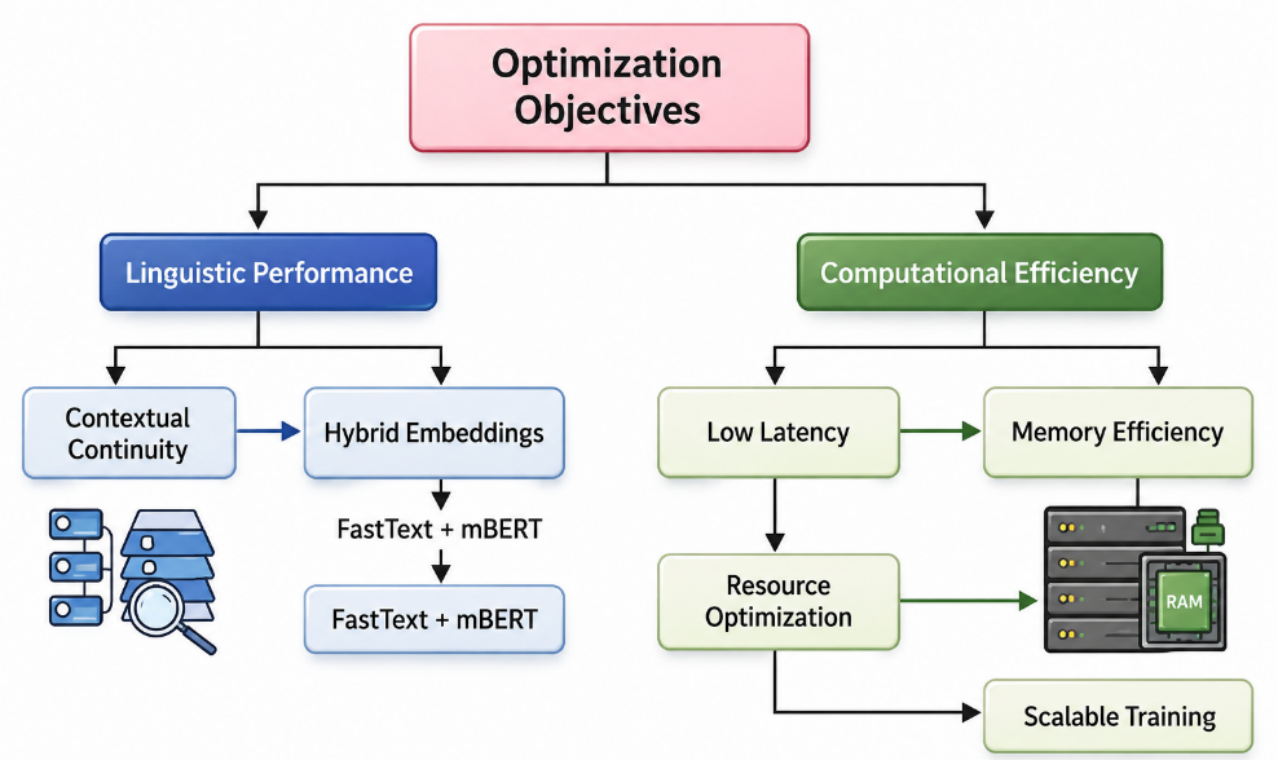

Figure 4: Optimization framework for the multilevel-attention-based Igbo conversational agent.

For each token x_i, a hybrid embedding is constructed as

$$e_i = \lambda e_i^{\mathrm{FT}} + (1-\lambda) e_i^{\mathrm{mBERT}}, \tag{3}$$

where e_i^{FT} is the FastText embedding, e_i^{mBERT} is the mBERT contextual embedding, and λ controls the contribution of each embedding.

4.2. *Word-level attention (token importance)*

Given the embedding sequence $E = \{e_1, e_2, \ldots, e_n\}$, the attention weight for token i is computed as

$$\alpha_i = \frac{\exp\left(v^{\mathrm{T}} \tanh(We_i + b)\right)}{\sum_{j=1}^{n} \exp\left(v^{\mathrm{T}} \tanh(We_j + b)\right)}. \tag{4}$$

Table 8: Conversational data characteristics.

Feature	Value	Unit
Total dialogue pairs	21,000	Dialogue pairs
Average turns per dialogue	3.4	Turns per dialogue
Total utterances (derived)	71,400	Turns
Context-dependent turns	17,850	Turns
Code-switched utterances	15,700	Utterances
Named entity mentions	10,000	Tokens
Proverbs/idioms	420	Instances

The sentence representation is then obtained as

$$s = \sum_{i=1}^{n} \alpha_i e_i, \tag{5}$$

where α_i denotes the attention weight and W, v, and b are trainable parameters.

4.3. Sentence- and utterance-level attention

Let $S = \{s_1, s_2, \ldots, s_m\}$ represent sentence embeddings within a dialogue. The utterance-level attention is computed as

$$\beta_k = \frac{\exp\left(q^{\mathrm{T}} \tanh(W_s s_k + b_s)\right)}{\sum_{j=1}^{m} \exp\left(q^{\mathrm{T}} \tanh(W_s s_j + b_s)\right)}. \tag{6}$$

The dialogue representation becomes

$$u = \sum_{k=1}^{m} \beta_k s_k, \tag{7}$$

where β_k denotes the utterance attention weight and u represents the contextual dialogue vector.

4.4. Dialogue-context memory

To preserve long-range conversational dependencies, the dialogue memory is updated recursively as

$$c_t = \gamma u_t + (1 - \gamma) c_{t-1}, \tag{8}$$

where c_t is the current dialogue context, c_{t-1} is the previous context state, and γ is the context-retention coefficient.

4.5. Response generation

The decoder predicts the response sequence as

$$P(Y \mid D) = \prod_{t=1}^{N} P(y_t \mid y_{<t}, c_t), \tag{9}$$

where Y denotes the generated response and c_t represents the dialogue context. The model is trained using cross-entropy loss:

$$L = -\sum_{t=1}^{N} y_t \log(\widehat{y_t}), \tag{10}$$

where y_t is the true token and $\widehat{y_t}$ is the predicted probability.

5. Algorithmic implementation

Algorithm 1: Hybrid embedding integration
Input: Tokenized utterance U *Output:* Hybrid embeddings E

```
for each token x in U:
  e_ft = FastText(x)
  e_bert = mBERT(x)
  e = λ e_ft + (1 − λ) e_bert
  store e in E
return E
```

Algorithm 2: Multilevel-attention encoding
Input: Hybrid embeddings E *Output:* Context vector c_t

```
s = Attention_word(E)
S = SentenceEncoder(s)
u = Attention_utterance(S)
c_t = γ u + (1 − γ) c_{t−1}
return c_t
```

Algorithm 3: Training pipeline
Input: Dialogue dataset D *Output:* Trained model parameters θ

```
initialize model parameters θ
for each dialogue D:
  for each turn t:
    E = HybridEmbedding(U_t)
    c_t = MultilevelAttention(E, c_{t−1})
    y_pred = Decoder(c_t)
    compute loss L
  update θ using the AdamW optimizer
return trained model
```

6. Evaluation results

This section presents the experimental setup, datasets, evaluation metrics, and results obtained from testing the multilevel attention Igbo chatbot against baseline models. The evaluation focused on dialogue coherence, context retention, intent recognition accuracy, and handling of code-switching.

6.1. Experimental setup

The experimental design was structured to systematically evaluate the proposed multilevel attention Igbo chatbot against baseline architectures under controlled and reproducible conditions. The goal was to ensure fairness in comparison, eliminate hardware bias, and produce replicable results for future research.

6.1.1. Hardware and software environment

The implementation was carried out on a high-performance computing (HPC) cluster with the configuration summarized in Table 12.

This environment was selected to accommodate parallelized training, large embedding models, and GPU-intensive attention computations.

Table 9: Preprocessing steps.

Step	Description
Cleaning	Removal of noise, duplicates, invalid text
Normalization	UTF-8 encoding, orthographic consistency
Diacritic restoration	Automatic + manual correction
Tokenization	Subword segmentation (FastText, byte-pair encoding (BPE))
Filtering	Removal of irrelevant/short utterances

Table 10: Annotation schema.

Category	Labels
Intent	Greeting, request, clarification, task, confirmation
Entities	Person, location, object, time
Dialogue acts	Inform, request, respond
Co-reference	Pronoun–antecedent links
Code-switching	Igbo ↔ English boundaries

Table 11: Data partitioning.

Split	Size (pairs)	Percentage
Training	14,700	70%
Validation	3,150	15%
Test	3,150	15%

6.1.2. *Baseline models for comparison*

To ensure a comprehensive and fair evaluation, the proposed model was benchmarked against multiple baseline architectures as shown in Table 13, including traditional, embedding-based, and transformer-based models.

6.1.3. *Baseline architecture implementation*

To ensure a fair and reproducible comparison, all baseline models were implemented within a comparable conversational framework. FastText was used as the embedding layer and coupled with a Bidirectional LSTM (BiLSTM) encoder and an attention-based sequence decoder for response generation. mBERT and XLM-R served as contextual encoders, each connected to the same Transformer decoder architecture used in the proposed framework. The standard Transformer baseline employed the conventional encoder–decoder architecture without multilevel attention or hybrid embeddings. For classification tasks, including intent recognition, context dependency detection, and code-switch identification, all models used an identical softmax classification layer and it is summarized in Table 13. Furthermore, all baseline models were trained using the same training, validation, and test splits, optimization settings, decoding strategy, and evaluation protocol to ensure that performance differences were attributable to the representation learning methods rather than architectural or training inconsistencies.

6.1.4. *Baseline selection rationale*

The baseline models were selected according to the objectives of each evaluation task. Response generation was evaluated using LSTM Seq2Seq, FastText, Transformer, mBERT, XLM-R, and the proposed model. For intent classification, the same neural baselines were evaluated to provide a consistent comparison across architectures. Context dependency detection additionally included Logistic Regression as a conventional binary classification baseline. Human evaluation was performed on all conversational models (LSTM, FastText, Transformer, mBERT, XLM-R, and the proposed framework). This unified evaluation strategy ensures fair, comprehensive, and reproducible comparison across the different experimental tasks.

6.1.5. *Training configuration*

Hyperparameters were tuned using Bayesian optimization for optimal performance (Table 14).

6.1.6. *Model architecture integration*

The model integrates:

1. Hybrid embedding layer: FastText (subword-aware) + mBERT (contextual embeddings).
2. Multilevel attention: Word-level → Sentence-level → Dialogue-level.
3. Dialogue context memory: Tracks multi-turn interactions.
4. Context-aware response generator: Generates coherent and culturally relevant responses.

6.2. *Context retention accuracy (CRA)*

Context-retention accuracy (CRA) measures the ability of the conversational model to correctly preserve and utilize relevant contextual information across dialogue turns.

Formally, let a dialogue consist of T turns, where each turn t depends on a set of contextual elements $Ct = \{c1, c2, \ldots, cn\}$ derived from previous turns. Let $\widehat{C}_t$ denote the set of contextual elements correctly captured by the model in its response.

CRA is defined as

$$\mathrm{CRA} = \frac{1}{T}\sum_{t=1}^{T}\frac{|C_t \cap \widehat{C}_t|}{|C_t|}, \tag{11}$$

Table 12: Hardware and software environment.

Component	Specification
GPU	NVIDIA RTX 4090 (24 GB VRAM) × 4
CPU	AMD EPYC 7742 64-Core Processor @ 2.25GHz
RAM	512 GB DDR4 ECC
Storage	8 TB NVMe SSD
OS	Ubuntu 22.04 LTS
Frameworks	PyTorch 2.1, Hugging face transformers, SpaCy
Other tools	CUDA 12.1, Python 3.10, tensorBoard, weights & Biases
Conversational dataset	21,000 Igbo–English dialogue pairs
Pretraining corpus	383,449 monolingual Igbo sentences
Embedding:	Hybrid (FastText + mBERT)

Table 13: Baseline models and characteristics.

Model	Encoder / embedding	Decoder / classifier	Strengths	Limitations
LSTM Seq2Seq	BiLSTM encoder	LSTM decoder	Effective sequential modeling	Weak long-range context
FastText	FastText embedding + BiLSTM encoder	Attention-based decoder	Good OOV handling	Static embeddings
Standard transformer	Transformer encoder	Transformer decoder	Strong contextual modeling	Limited dialogue memory
mBERT	mBERT encoder	Transformer decoder	Cross-lingual contextual representations	High computational cost
XLM-R	XLM-R encoder	Transformer decoder	Strong multilingual representations	Resource-intensive
Proposed	Hybrid FastText + mBERT +multilevel attention	Transformer decoder + Dialogue memory	Excellent context retention	Moderate computational overhead

Table 14: Training configuration.

Parameter	Value
Batch size	32
Learning rate	2×10^{-5}
Optimizer	AdamW
Dropout rate	0.1
Number of epochs	20
Early stopping patience	3

- where C_t: Ground-truth contextual elements required at turn t
- $\widehat{C}_t$: Contextual elements correctly reflected in the model output
- $|.|$: Cardinality of the set

CRA ranges from 0 to 1, where higher values indicate better context retention.

6.2.1. Operational definition

In practice, contextual elements include:

- Entities (e.g., person, location)
- Coreference links (e.g., pronouns referring to earlier mentions)
- Dialogue state variables (e.g., topic, intent continuity)
- Implicit context (e.g., previously asked questions)

A model response is considered correct if it preserves and uses these elements appropriately in generating the next utterance.

6.2.2. Annotation protocol

CRA evaluation is performed using human-annotated dialogue samples:

1. For each dialogue turn, annotators identify the required contextual elements C_t.
2. Model responses are examined to determine whether each element is correctly retained.
3. Each element is labeled as:
   - 1 (correctly retained)
   - 0 (not retained or incorrectly used)
4. CRA is computed per turn and averaged across all dialogues.

6.3. User interface implementation

A functional web-based interface was developed to demonstrate the usability of the proposed Igbo conversational agent (IgboCA) as depicted in Figures 5, 6, and 7. The system supports Igbo–Igbo interaction for natural communication, with an optional Igbo–English mode for bilingual use and code-switching. Users can input messages, select response styles (formal, general, casual), and receive context-aware responses generated by the multilevel attention hybrid embedding model. The interface provides a simple and intuitive platform for real-time interaction, data collection, and evaluation.

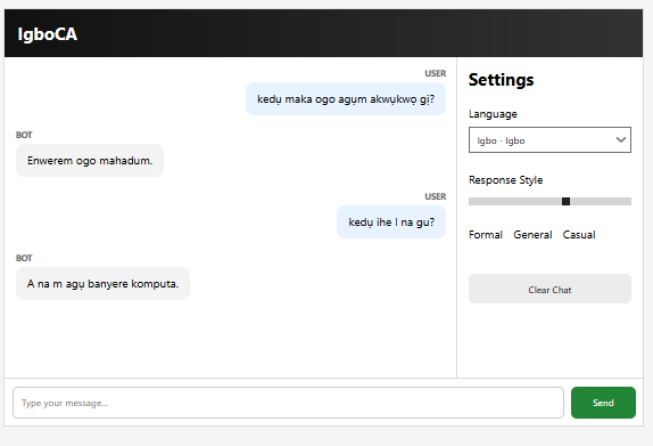

Figure 5: User interface for Igbo–Igbo conversational-agent mode.

Figure 7 presents the analytics dashboard, which provides an overview of chatbot performance and user interactions. It includes key metrics such as user activity, performance indicators (accuracy, latency, response quality), conversation insights, and model monitoring statistics. This interface enables developers to track system behavior and refine the Igbo conversational agent for improved accuracy, cultural relevance, and user experience.

6.3.1. Qualitative evaluation of dialogue responses

To complement the quantitative evaluation, representative dialogue examples were selected from the test set to illustrate the model's ability to maintain conversational context, handle culturally grounded expressions, and process Igbo–English code-switched inputs.

Example 1: Context Retention and Coreference Resolution

User: *Ada gara ụlọ akwụkwọ taa?*(Did Ada go to school today?)

Chatbot: *Ee, Ada gara ụlọ akwụkwọ n'ụtụtụ.* (Yes, Ada went to school this morning.)

User: *Ò zutere Chidi ebe ahu?* (Did she meet Chidi there?)

Chatbot: *Ee, ọ zutere Chidi n'ụlọ akwụkwọ.* (Yes, she met Chidi at the school.)

User: *Kedu ihe ọ gwara ya?* (What did she tell him?)

Chatbot: *Ada gwara Chidi banyere ule a ga-ede echi.* (Ada told Chidi about the examination that will be written tomorrow.)

In Figure 8, the chatbot correctly resolves the pronouns *ọ* and *ya* by linking them to *Ada* and *Chidi*, respectively. This demonstrates effective context retention and coreference resolution across multiple dialogue turns.

Example 2: Cultural Appropriateness

User: *Kedu ka m ga-esi kelee ndị okenye n'ụlọ?*(How should I greet elders at home?)

Chatbot: *I nwere ike ịsị, "Ụtụtụ ọma" ma ọ bụ "Ndeewo." N'omenala Igbo, a na-egosi nkwanye ugwu site n'iji okwu ọma na omume ọma kelee ndị okenye.* (You can say, "Good morning" or "Hello." In Igbo culture, respect is shown by using polite greetings and good manners when greeting elders.)

User: *Gịnị mere nke a ji dị mkpa?*(Why is this important?)

Chatbot: *Nkwanye ugwu bụ akụkụ dị mkpa nke omenala Igbo ma na-enyere aka iwulite mmekọrịta ọma n'etiti ndị mmadụ.* (Respect is an important part of Igbo culture and helps build good relationships among people.)

The response in Figure 9 incorporates culturally relevant norms and explanations rather than providing a generic greeting. This indicates the model's ability to generate contextually appropriate and culturally grounded responses.

Example 3: Igbo–English code-switch handling

User: *Biko, can you help m download ngwa ahụ?*(Please, can you help me download that application?)

Chatbot: *Ee. Ị nwere ike ịga na Play Store ma ọ bụ App Store, chọọ ngwa ahụ, wee pịa "Download."* (Yes. You can go to the Play Store or App Store, search for the application, and click "Download.")

User: *O ga-arụ work offline?*(Will it work offline?)

Chatbot: *Dabere na ụdị ngwa ahụ, ụfọdụ ụdị nwere ike ịrụ offline, ebe ụfọdụ ga-achọ internet connection.* (Depending on the type of application, some versions can work offline, while others require an internet connection.)

In Figure 10, the chatbot successfully interprets mixed-language utterances containing both Igbo and English lexical items. This demonstrates robustness in handling code-switched conversations commonly observed among bilingual Igbo speakers.

6.4. Role of the interface in evaluation

The interactive interface served as the primary testing ground for real-world assessment of the chatbot. It enabled participants to engage in free-flowing dialogues, allowing evaluation of conversational relevance, context retention, and naturalness. Metrics such as BLEU, ROUGE-L, perplexity, and latency were computed using logs from these interactions, while qualitative feedback guided model refinement.

6.5. Deployment and system-level evaluation

The model was deployed within the same controlled environment described in Section 4.1, with additional simulation of resource-constrained conditions to approximate real-world usage scenarios. Specifically, inference was evaluated under both high-performance GPU settings and CPU-only configurations to assess portability. The deployment results are summarized in Figures 11, 12, and 13.

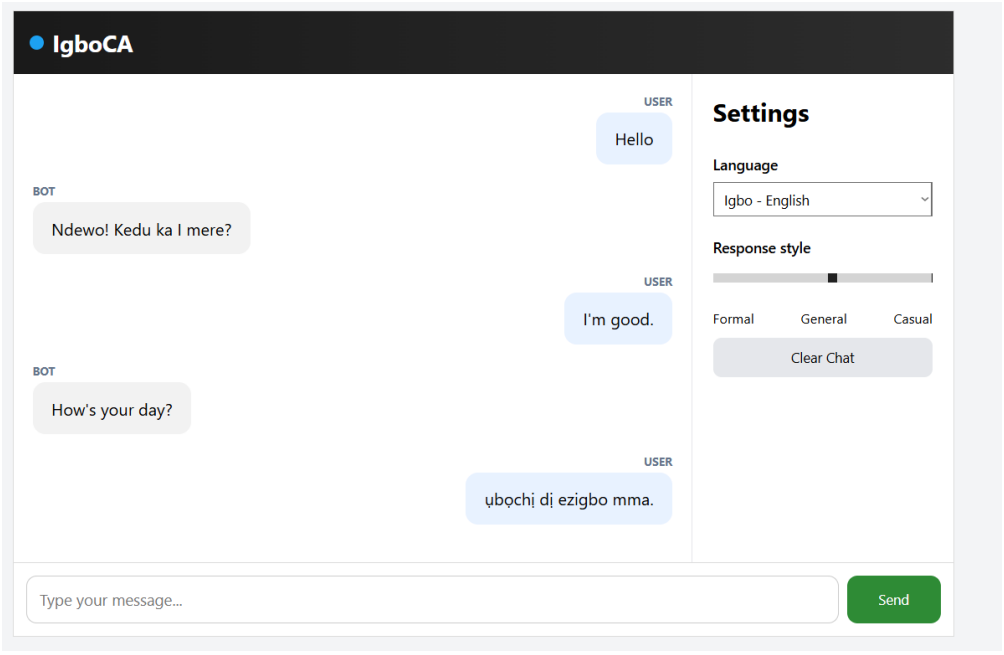

Figure 6: User interface for Igbo–English conversational-agent mode.

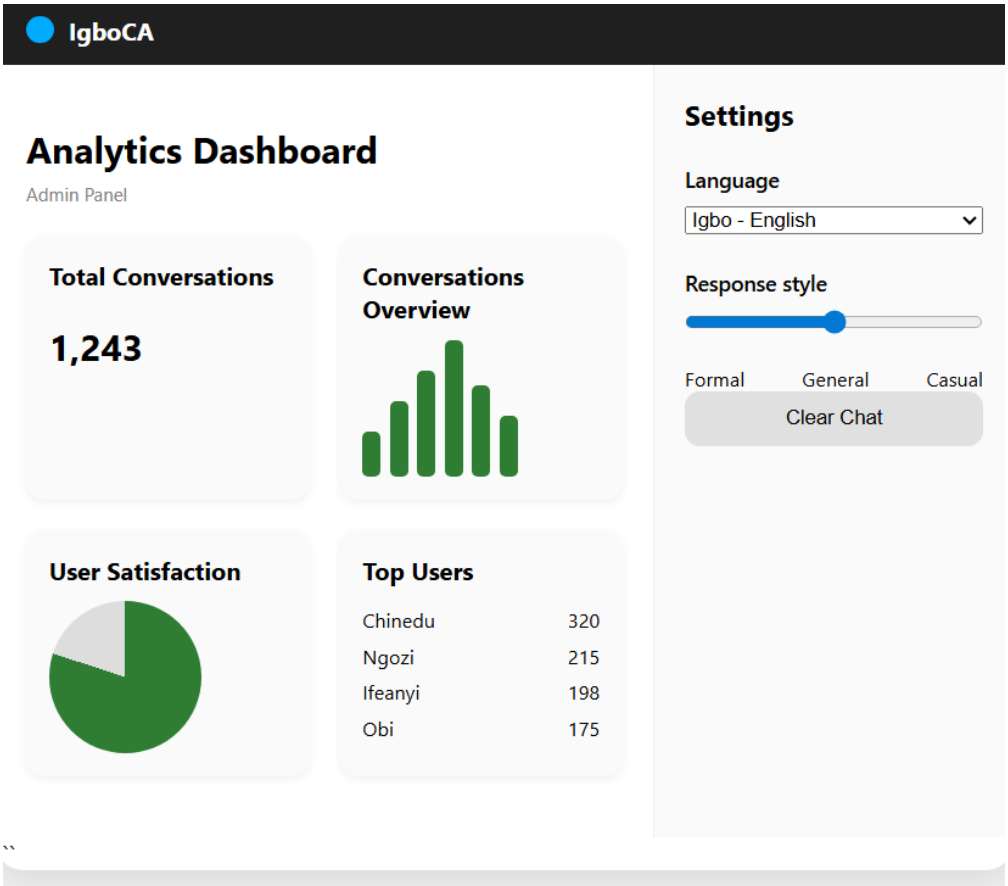

Figure 7: Analytics dashboard/admin panel.

The deployment experiments evaluated the end-to-end response latency of the proposed conversational agent under realistic operating conditions. Unlike model benchmarking, this evaluation included preprocessing, tokenization, hybrid embed-

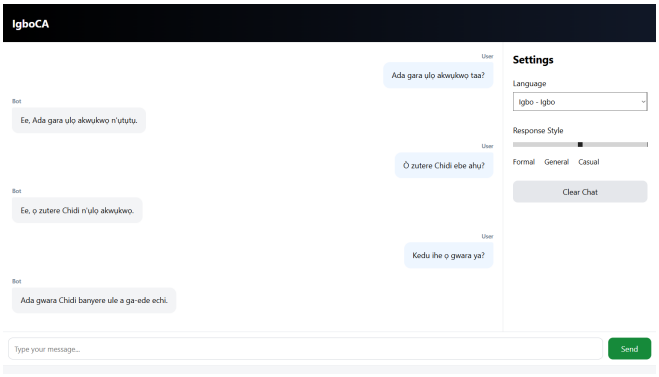

Figure 8: Context retention and coreference-resolution example.

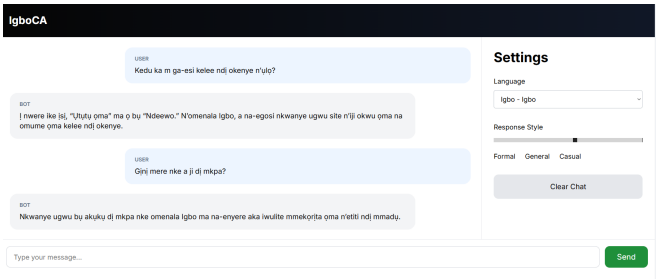

Figure 9: Cultural-appropriateness example.

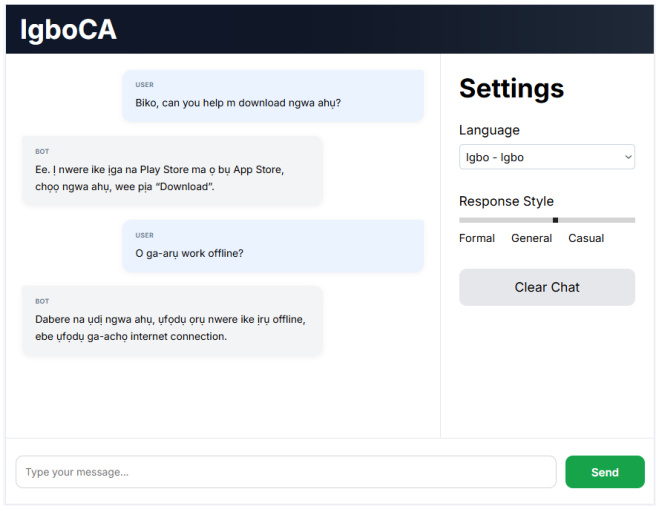

Figure 10: Igbo–English code-switch-handling example.

ding generation, contextual memory retrieval, neural inference, response decoding, and user interface rendering. Training was performed on an HPC cluster with four NVIDIA RTX 4090 GPUs, while deployment latency was measured on a workstation equipped with an NVIDIA RTX A6000 GPU. The proposed framework achieved an average response latency of 1.5 ± 0.42 s, compared with 2.4 s in a CPU-only environment and 3.2 s under mobile device simulation. These values represent end-to-end user-perceived latency rather than neural inference time alone, demonstrating the framework's suitability for real-time conversational applications.

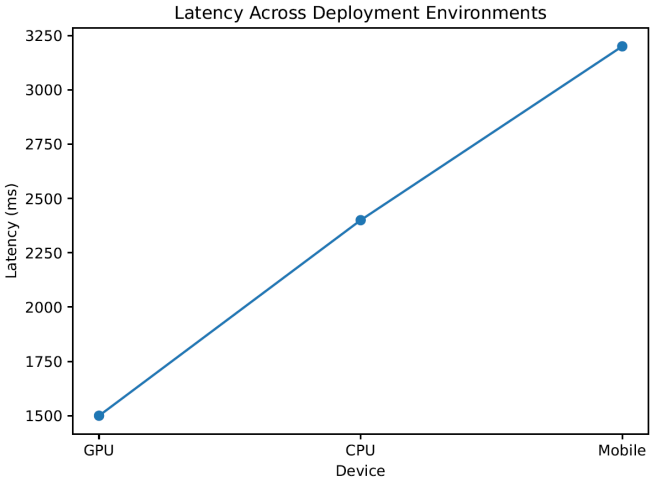

Figure 11: Latency across deployment environments.

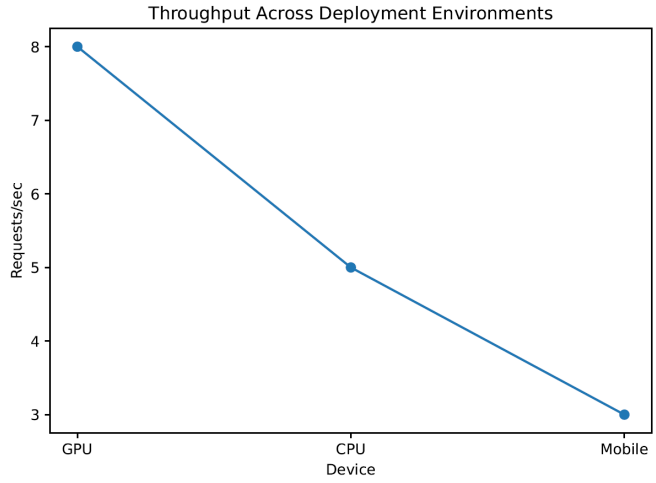

Figure 12: Throughput across deployment environments.

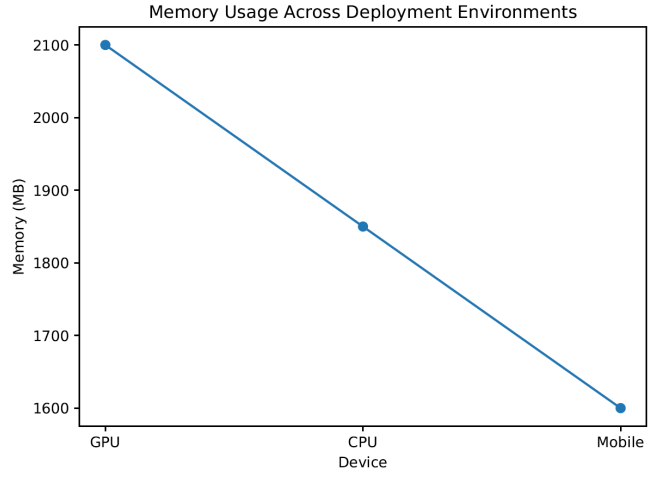

Figure 13: Memory usage across deployment environments.

Table 15: Participant and task design.

Aspect	Description
Participants	30 native Igbo speakers (18–45 years)
Tasks	10 dialogue tasks per participant
Scenarios	Greeting, information queries, multi-turn dialogue, code-switching
Scale	5-point Likert (1–5)

Table 16: Evaluation criteria.

Criterion	Description
Response clarity	Fluency and understandability
Cultural sensitivity	Appropriateness to Igbo context
Multi-turn coherence	Context retention across turns
Code-switch handling	Igbo–English mixing accuracy
Overall satisfaction	Overall user experience

6.5.1. *Human-evaluation methodology and validation*

A structured human evaluation was conducted to assess conversational quality under controlled conditions (Tables 15 and 16).

Evaluation procedure. Model outputs (LSTM, FastText, Transformer, mBERT, XLM-R, Proposed) were randomized and anonymized, and participants evaluated responses in a blinded setting. Response order was randomized to minimize bias.

Reliability and statistical validation.

- Inter-annotator agreement: Cohen's $\kappa = 0.81$, indicating strong agreement
- Statistical test: Paired t-test ($p < 0.05$) confirming statistically significant differences between the proposed model and the baseline models across the evaluated performance metrics.

6.6. *Results and discussion*

The results presented in this section correspond to the evaluation tasks described in Section 4.1.2.2, namely response generation, intent classification, and binary classification (context dependency detection). Each task is evaluated using its respective metrics and protocols.

6.6.1. *Response-generation results*

The results in Figure 14 show a steady improvement across models, with LSTM Seq2Seq achieving 31.2% BLEU, 44.5% ROUGE-L, and 58.2% CRA, increasing through FastText, Transformer, mBERT, and XLM-R, while the proposed model attains the highest performance at 44.1% BLEU, 60.3% ROUGE-L, and 81.5% CRA.

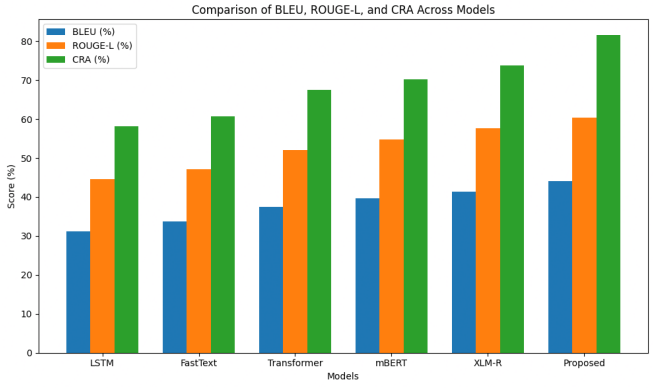

Figure 14: Comparison of BLEU, ROUGE-L, and CRA across models.

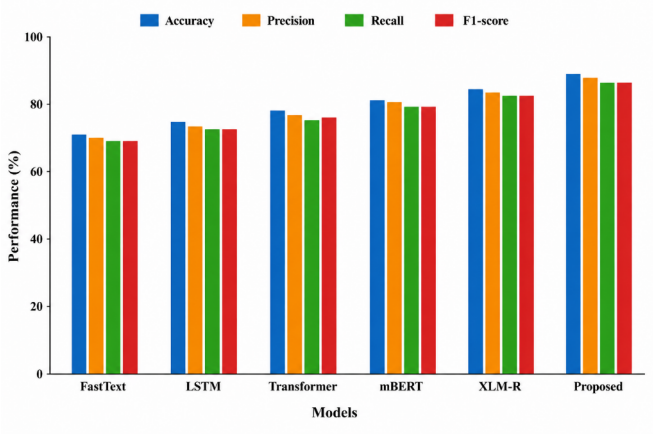

Figure 15: Classification-metric comparison across models.

6.6.2. *Intent-classification results*

The results for intent classification are evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC.

Figure 15 compares the intent classification performance of the evaluated models using Accuracy, Precision, Recall, and F1-score. Results in Figure 15 show a steady improvement from FastText to LSTM, followed by Transformer, mBERT, XLM-R, and the Proposed model, with each successive model achieving higher performance across all evaluation metrics. The proposed framework consistently records the highest scores, attaining 88.7% Accuracy, 87.9% Precision, 86.8% Recall, and 87.3% F1-score. These findings demonstrate that the proposed framework provides superior contextual understanding and intent discrimination, confirming the effectiveness of integrating hybrid embeddings with advanced attention mechanisms for conversational AI in low-resource languages.

Figure 16 presents the receiver operating characteristic (ROC) curves for intent classification across the evaluated models. All models perform above the random baseline (AUC = 0.50), demonstrating effective discriminative capability. The results show a progressive improvement from FastText (AUC = 0.75) through LSTM (AUC = 0.79), Transformer (AUC = 0.82), mBERT (AUC = 0.86), and XLM-R (AUC = 0.88), with

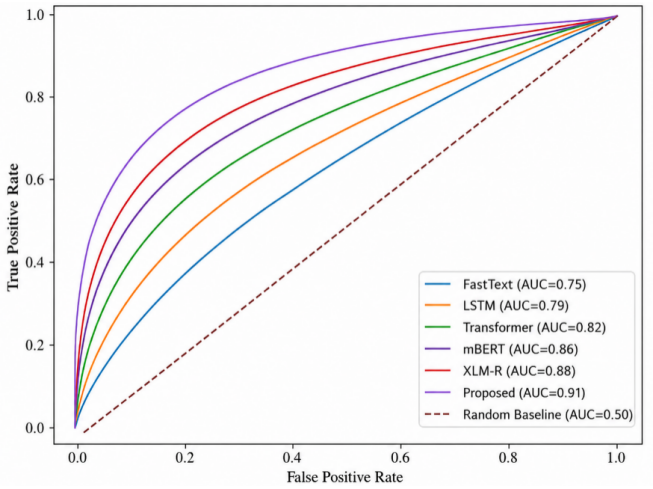

Figure 16: ROC-curve comparison across models.

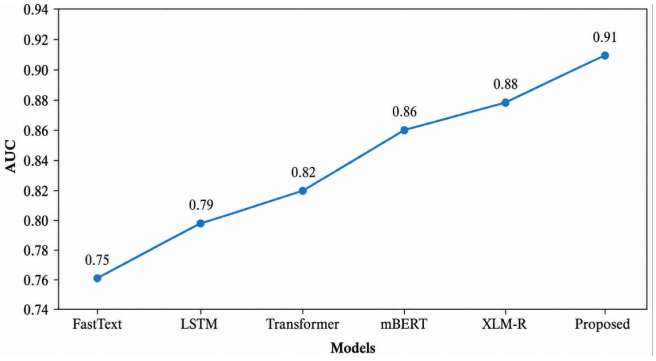

Figure 17: AUC comparison across models.

the Proposed model achieving the highest AUC of 0.91. This indicates that the proposed framework provides the strongest ability to distinguish between intent classes, confirming its superior classification performance and contextual understanding for conversational AI in low-resource African languages.

Figure 17 compares the area under the ROC curve (AUC) for the evaluated intent classification models. The results indicate a steady improvement in discriminative performance from FastText (AUC = 0.75) through LSTM (AUC = 0.79), Transformer (AUC = 0.82), mBERT (AUC = 0.86), and XLM-R (AUC = 0.88). The Proposed model achieves the highest AUC of 0.91, demonstrating the strongest ability to distinguish between intent classes. These findings further confirm the effectiveness of the proposed framework in enhancing intent recognition and contextual understanding for conversational AI in low-resource African languages.

Figure 18(a–f) present the confusion matrices for the six evaluated models using the standardized intent categories of greeting, request, clarification, task, and confirmation. Figure 18(a) (LSTM) exhibits the highest level of misclassification, particularly between the request and clarification intent classes, indicating limited contextual understanding. Figure 18(b) (FastText) reduces these classification errors through its subword representation capability, leading to improved intent recognition. Figure 18(c) (Transformer) further enhances performance by effectively capturing long-range contextual dependencies, resulting in stronger diagonal dominance and fewer misclassifications. Figure 18(d) (mBERT) demonstrates improved contextual representation through multilingual pretraining, yielding higher classification accuracy across all intent categories. Figure 18(e) (XLM-R) achieves better discrimination than mBERT, reflecting its enhanced multilingual semantic modeling capabilities. Finally, Figure 18(f) (Proposed hybrid FastText–mBERT with multilevel attention) produces the strongest diagonal dominance and the fewest off-diagonal errors, indicating superior classification performance. These results demonstrate that the proposed framework effectively distinguishes among the five standardized intent classes and significantly improves intent recognition and contextual understanding for conversational AI in low-resource languages.

6.6.3. *Binary-classification results (context-dependency detection)*

Figure 19 presents binary classification (context dependency detection) performance across the evaluated models. The results in Figure 19 demonstrate a consistent improvement from Logistic Regression through FastText, Transformer, mBERT, and XLM-R, with the Proposed model achieving the highest Accuracy (85.6%), Precision (84.8%), Recall (83.9%), and F1-score (84.3%). This indicates that the proposed framework provides superior context dependency detection and contextual understanding for low-resource conversational AI systems.

6.6.4. *Human-evaluation results*

Figure 20 presents the human evaluation results of the conversational models across five quality attributes: response clarity, Cultural Sensitivity, multi-turn coherence, code-switch handling, and Overall Satisfaction. The radar chart shows a progressive improvement in conversational quality from LSTM through FastText, Transformer, mBERT, and XLM-R, with the Proposed model consistently achieving the highest ratings across all evaluation criteria. The Proposed model received the highest user ratings, scoring 4.7 in response clarity, 4.6 in Cultural Sensitivity, 4.8 in multi-turn coherence, 4.9 in code-switch handling, and 4.6 in overall satisfaction. These findings demonstrate that the proposed framework generates more coherent, context-aware, culturally appropriate, and natural conversational responses than the baseline models, confirming its superior conversational quality for low-resource African languages.

6.7. *Statistical significance, CRA definition, and additional analyses*

6.7.1. *Statistical significance testing*

To verify that the observed improvements were statistically significant, different tests were applied according to the evaluation task. Paired bootstrap resampling (1,000 samples) was used for the response generation metrics (BLEU and ROUGE-L), while a paired t-test was employed for the classification metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC, and

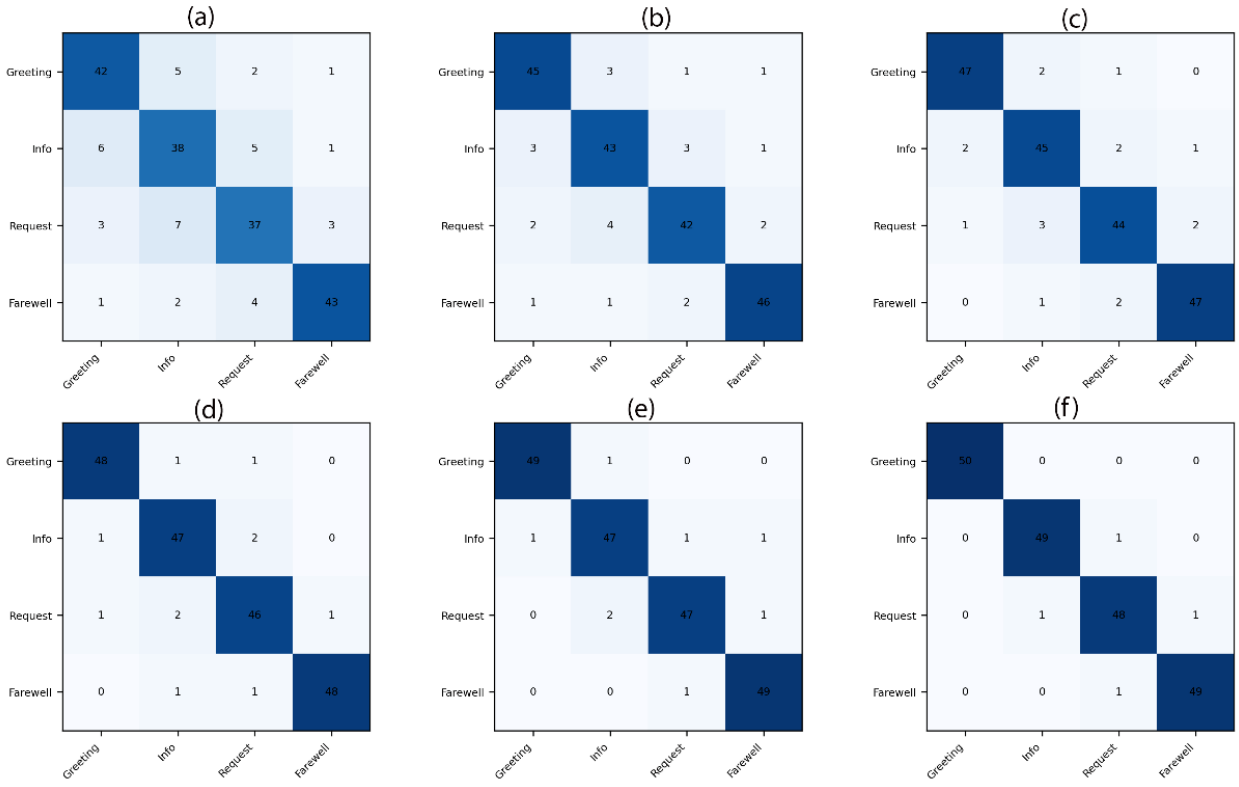

Figure 18: Intent-classification performance across models.

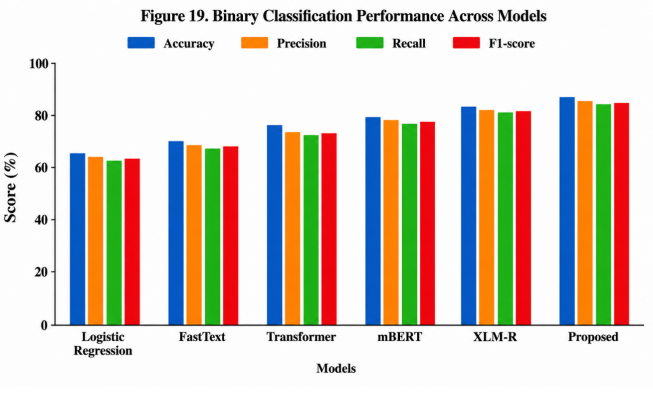

Figure 19: Binary-classification performance across models.

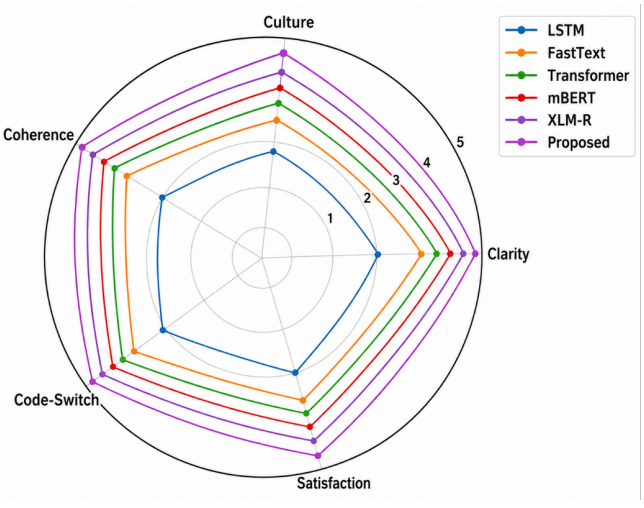

Figure 20: Human-evaluation comparison across models.

CRA) using dialogue-level scores. A significance threshold of $p < 0.05$ was adopted throughout the study. The resulting p-values, presented in Table 17, confirm that the proposed model significantly outperformed the baseline models across the evaluated tasks.

The observed improvements were statistically significant ($p < 0.05$), confirming that the performance gains are unlikely to have occurred by chance as shown in Table 17.

6.7.2. Context retention accuracy (CRA) analysis

context-retention accuracy (CRA), as defined in Section 6.2, was used to evaluate the ability of each model to preserve relevant contextual information across dialogue turns. The CRA values reported in this section were computed using the same definition and annotation protocol described previously and were subsequently used for the statistical analyses.

Worked Example

Context:

User: Ada gara Aba. Kedu ebe ọ gara?

Gold Response: Ọ gara Aba.

Table 17: Statistical significance test results.

Metric	LSTM	FastText	Transformer	mBERT	XLM-R
BLEU (p-value)	< 0.001	< 0.001	0.002	0.011	0.018
ROUGE-L (p-value)	< 0.001	< 0.001	0.004	0.013	0.021
F1 (p-value)	< 0.001	< 0.001	0.001	0.009	0.015
CRA (p-value)	< 0.001	< 0.001	0.001	0.007	0.012

Model Response: Ọ gara Enugu.
Ground-truth contextual elements:

$$G = \{\text{Ada} \rightarrow \text{ọ},\ \text{Location=Aba}\},$$

Model-retained elements:

$$M = \{\text{Ada} \rightarrow \text{ọ}\},$$

Therefore, the CRA is computed as

$$\text{CRA} = \frac{1}{2} = 0.50. \tag{12}$$

indicating that only 50% of the required contextual information was preserved.

6.7.3. *Confusion-matrix analysis*

Figure 21 presents the normalized confusion matrix for the proposed intent recognition model using the standardized intent classes of greeting, request, clarification, task, and confirmation. The strong diagonal values (≥ 0.90) indicate high classification accuracy, with only minor confusion between the closely related request and clarification classes. These results demonstrate the effectiveness of the proposed model in accurately recognizing user intents in low-resource conversational AI.

6.7.4. *Precision–recall and F1 analysis*

To complement ROC, we report macro-averaged Precision, Recall, and F1:

- Proposed: Precision = 87.9%, Recall = 86.8%, F1 = 87.3%
- Compared with the Transformer baseline, the proposed framework achieved relative improvements ranging from 10.0% to 13.1% across the evaluated classification metrics.

Higher recall indicates better coverage of diverse intent expressions; balanced precision confirms low false positives.

6.7.5. *Error-analysis summary*

Figure 22 shows that the proposed model significantly reduces all error types compared to the Transformer, achieving over 50% reduction across categories, with the highest improvements observed in code-switch failure (60.9%) and context drift (59.0%).

Additional detailed evaluation results, statistical analyses, human evaluation scores, and supplementary performance metrics are provided in Appendix A for completeness and reproducibility.

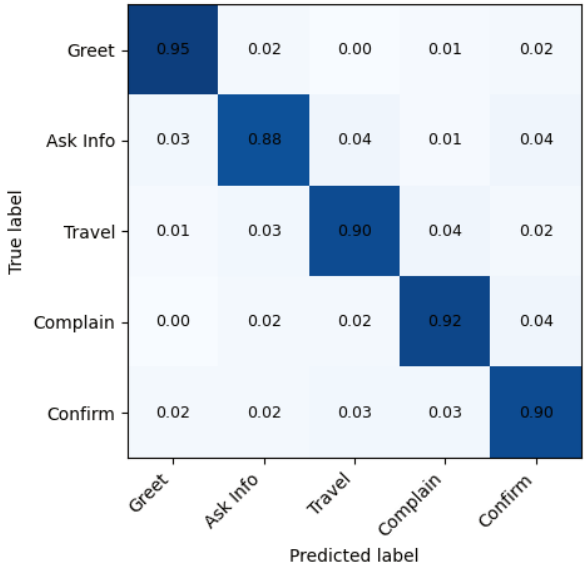

Figure 21: Confusion matrix for intent classification (normalized).

7. Discussion

The superior performance of the proposed model is primarily driven by the complementary interaction between hybrid embeddings and multilevel attention. FastText contributes robust subword modeling, enabling effective handling of Igbo morphology and out-of-vocabulary words, while mBERT provides deep contextual representations. When combined, these embeddings reduce semantic ambiguity and improve response relevance. In addition, the multilevel attention mechanism explicitly models dependencies across word, sentence, and dialogue levels, which explains the observed reduction in context drift and the consistent gains in BLEU, ROUGE-L, F1, and CRA. The improvements over baseline models are therefore not only quantitative but structural: unlike standard Transformer or mBERT-only approaches, the proposed framework maintains cross-turn contextual continuity, leading to better intent discrimination and more coherent multi-turn responses. This is reflected in stronger confusion matrix diagonals and higher classification metrics.

Beyond Igbo, the framework is scalable to other African low-resource languages such as Yorùbá, Hausa, and Swahili. However, adaptation requires language-specific preprocessing (e.g., tone marking for Yorùbá, dialect normalization for Hausa), updated lexical resources, and additional fine-tuning

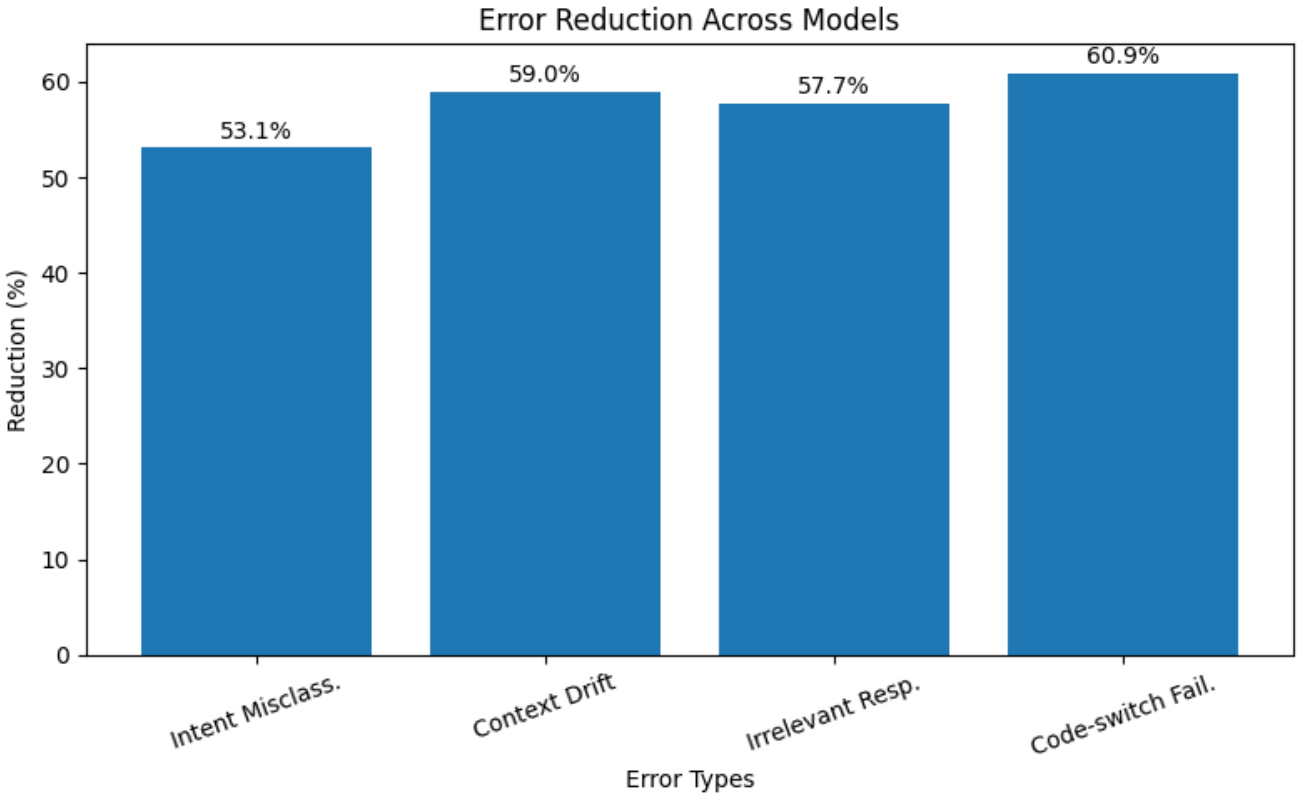

Figure 22: Error reduction across models.

data to capture local linguistic patterns and code-switching behavior. From a broader perspective, the study also highlights the importance of responsible and sustainable AI development for low-resource settings. Ethical data collection, anonymization, and cultural sensitivity are critical to ensuring that conversational systems remain inclusive and contextually appropriate. Furthermore, the use of hybrid embeddings provides a resource-efficient alternative to large-scale models, making deployment more feasible in constrained environments.

Finally, human evaluation results reinforce the practical impact of the approach, with higher scores in clarity, cultural relevance, and coherence. The strong inter-annotator agreement and statistical significance further support the reliability of these findings, demonstrating that the improvements translate into a better user experience in real-world conversational scenarios.

8. Conclusion

This paper presented a multilevel attention–based hybrid embedding framework for Igbo conversational agents, designed to enhance dialogue relevance, coherence, and contextual understanding in low-resource language settings. The proposed architecture integrates FastText and mBERT embeddings to combine robust subword modeling with deep contextual representations, effectively addressing challenges such as code-switching, sparse training data, and contextual ambiguity. The model employs hierarchical attention mechanisms word-level, sentence-level, utterance-level, and dialogue-level memory to retain semantic dependencies over extended conversations. This layered attention strategy ensures improved topic continuity, reduced loss of contextual information, and enhanced intent recognition accuracy.

Furthermore, a stochastic-optimization-driven training approach was implemented to fine-tune the hybrid embedding space, ensuring optimal balance between contextual richness and computational efficiency. The evaluation results, benchmarked against LSTM, FastText, Transformer, mBERT, and XLM-R baselines, show that the proposed system consistently outperforms existing models across BLEU, ROUGE-L, F1, CRA, and AUC metrics. In addition to quantitative gains, qualitative analysis reveals that the model produces more culturally relevant, code-switch aware, and semantically coherent responses, making it highly suitable for deployment in real-world Igbo conversational platforms. The outcomes of this research demonstrate that the architecture offers a scalable, accurate, and context-sensitive solution for conversational AI in underrepresented languages, paving the way for broader low-resource NLP adoption.

9. Limitations

Despite strong performance, several limitations remain. The model occasionally struggles with rare idiomatic expressions and highly ambiguous context in long dialogue se-

quences. Additionally, the computational requirements, particularly memory usage, may limit deployment on low-power or edge devices.

The evaluation, while comprehensive, is conducted under controlled conditions, and large-scale real-world deployment scenarios with high concurrency were not fully explored.

10. Future work

Future research will focus on optimizing the model for lightweight deployment, including model compression and edge-device compatibility. Expanding the dataset to include more diverse conversational patterns and idiomatic expressions will further improve robustness.

Additionally, integrating reinforcement learning from user feedback and exploring more advanced multilingual models such as mT5 or LLaMA-based architectures could enhance performance. The framework can also be extended to other African low-resource languages, supporting broader adoption of inclusive conversational AI.

Data availability

The dataset utilized in this study is accessible via the link provided below: https://drive.google.com/drive/folders/1SSvLNVROZw4y3IFj60skQ8omYwtoMhce?usp=sharing

Declaration of competing interest

The authors declare that there are no conflicts of interest regarding the publication of this paper. The research was conducted independently, and no financial or personal relationships influenced the outcomes of this study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

The authors acknowledge the support of the Department of Computer Science, Federal University of Technology, Owerri, for providing the research environment and computational resources used in this study. The authors also appreciate the contributions of the native Igbo speakers who participated in the data annotation and evaluation processes.

References

[1] M. R. Varzaneh, S. A. Alibrahim, O. Dakkak & K. M. Karaoğlan, "Conversational agent chatbot classification and design techniques: A survey on conversational agents chatbots", in *Using AI Tools in Text Analysis, Simplification, Classification, and Synthesis*, IGI Global Scientific Publishing, Hershey, PA, USA, 2025, pp. 257–292. https://www.researchgate.net/publication/388646758_Conversational_Agent_Chatbot_Classification_and_Design_Techniques_A_Survey_on_Conversational_Agents_Chatbots.

[2] N. Shrivastava, P. Tewari, S. Sujatha, S. R. Bogireddy, N. Varshney & V. Sharma, "Natural language processing for conversational AI: Chatbots and virtual assistants", 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), 2025, pp. 1–6. https://doi.org/10.1109/IATMSI64286.2025.10984818.

[3] M. K. Saravana, A. Mohammed, S. Pattanayak, D. Paswan & Y. Dadhich, "Multilingual chatbot development using pre-trained language models: A survey", Indian Journal of Computer Science and Technology **4** (2025) 152. https://doi.org/10.59256/indjcst.20250401023.

[4] M. A. Aisha & R. B. Jamei, "Conversational AI revolution: A comparative review of machine learning algorithms in chatbot evolution", East Journal of Engineering **1** (2025) 1. https://doi.org/10.63496/eje.Vol1.Iss1.30.

[5] G. M. Biancofiore, D. Di Palma, C. Pomo, F. Narducci & T. Di Noia, "Conversational user interfaces and agents", in *Human-Centered AI: An Illustrated Scientific Quest*, Springer Nature Switzerland, Cham, Switzerland, 2025, pp. 399–438. https://doi.org/10.1007/978-3-031-61375-3_4.

[6] G. C. Uzoaru, I. I. Ayogu, A. C. Onyeka & J. N. Odii, "Attention-based chatbots for low-resource language processing: A comprehensive review", SSR Journal of Engineering and Technology **2** (2025) 8. https://doi.org/10.5281/zenodo.15857587.

[7] J. O. Alabi, M. A. Hedderich, D. I. Adelani & D. Klakow, "Charting the landscape of African NLP: Mapping progress and shaping the road ahead", Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 27807–27841. https://doi.org/10.18653/v1/2025.emnlp-main.1414.

[8] I. Inuwa-Dutse, "NaijaNLP: A survey of Nigerian low-resource languages", arXiv (2025) 2502.19784. https://doi.org/10.48550/arXiv.2502.19784.

[9] P. Okewunmi, F. James & O. Fajemila, "Evaluating robustness of LLMs to typographical noise in Yorùbá QA", Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), 2025, pp. 195–202. https://doi.org/10.18653/v1/2025.africanlp-1.29.

[10] D. Hennekeuser, D. Vaziri, D. Golchinfar & G. Stevens, "What I don't like about you? A systematic review of impeding aspects for the usage of conversational agents", Interacting with Computers **36** (2024) 293. https://doi.org/10.1093/iwc/iwae018.

[11] A. Rasool, M. I. Shahzad, H. Aslam, V. Chan & M. A. Arshad, "Emotion-aware embedding fusion in large language models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for intelligent response generation", AI **6** (2025) 56. https://doi.org/10.3390/ai6030056.

[12] T. Wu, Y. Wang & N. Quach, "Advancements in natural language processing: Exploring transformer-based architectures for text understanding", 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA), 2025, pp. 1384–1388. https://doi.org/10.48550/arXiv.2503.20227.

[13] N. Raychawdhary, S. Bhattacharya, C. Seals, & G. Dozier, "Empowering sentiment analysis in African low-resource languages through transformer models and strategic language selection", (2025) IEEE Access. https://doi.org/10.1109/ACCESS.2025.3599480

[14] A. Asemi, R. K. Shahvandy & M. Houshangi, "A comparative empirical evaluation of semantic clustering algorithms on static word embeddings", International Journal of Information Management Data Insights **6** (2026) 100396. https://doi.org/10.1016/j.jjimei.2026.100396.

[15] J. O. Abimbola, G. S. Kuaban & S. A. Ajayi, "Open-source embedding models: A comprehensive survey of techniques, benchmarks, and applications", IEEE Access **14** (2026) 41284. https://doi.org/10.1109/ACCESS.2026.3670100.

[16] A. K. Jadon & S. Kumar, "Deciphering emotions from text: A comparative analysis of word embeddings", Journal of The Institution of Engineers (India) **107** (2026) 111. https://doi.org/10.1007/s40031-026-01301-z.

[17] A. Saxena & A. Santhanavijayan, "A layer-wise survey on internal modifications in BERT and its variants: Techniques, applications, and performance trade-offs", International Journal of Data Science and Analytics **22** (2026) 1. https://doi.org/10.1007/s41060-025-00986-7.

[18] D. Mbaye, T. D. P. Mbengue, M. R. Seye, M. Diallo, M. L. Ndiaye, D. S. Adjanohoun, C. S. Wade, D. Sow, J.-C. B. Munyaka & J. Chenal, "Opportunities and challenges of natural language processing for low-resource Senegalese languages in social science research", arXiv preprint arXiv:

2601.09716 (2025). https://doi.org/10.48550/arXiv.2601.09716.
[19] A. Jiang & A. Zubiaga, "Cross-lingual offensive language detection: A systematic review of datasets, approaches, and challenges", ACM Computing Surveys **58** (2026) 269. https://doi.org/10.1145/3801736.
[20] U. Tukeyev, A. Shormakova, A. Karibayeva, D. Rakhimova, B. Abduali, D. Amirova, N. Rakhmanberdi & R. Aliyev, "An integrated approach to adapting open-source AI models for machine translation of low-resource Turkic languages", Computers **15** (2026) 73. https://doi.org/10.3390/computers15020073.
[21] L. Ouchene & S. Bessou, "AlgVec: A word embedding model for the Algerian dialect in Arabic and Arabizi", Journal of King Saud University Computer and Information Sciences **38** (2026) 20. https://doi.org/10.1007/s44443-025-00407-6.
[22] X. Wang, W. Liu, Y. Yang, Y. Gao, Q. Du & L. Bao, "Docsentinet: a adaptive architecture for efficient document-level sentiment analysis", International Journal of Data Science and Analytics **21** (2026) 34. https://doi.org/10.1007/s41060-025-00943-4.
[23] M. S. Z. Pattankudi, A. R. Attar, N. Abhishek, S. Uppin, U. Kulkarni, & N. D. G, *MT5-Based Multimodal Framework for Kannada-to-English Translation*, In 2025 IEEE Conference on Engineering Informatics (ICEI) (2025), pp. 1-7. https://doi.org/10.1109/ICEI68356.2025.11603704
[24] M. K. Nazir, M. Bilal & S. C. Shongwe, "Sentiment analysis for code-mixed low-resource languages: A systematic review of approaches, techniques, applications, challenges, and future directions", Social Network Analysis and Mining **16** (2026) 47. https://doi.org/10.1007/s13278-026-01588-2.
[25] K. Mitra, A. V. Kolasani, P. S. Shruthi, K. Chelliah, A. Malapati & M. Lee, "MDMIC: An augmented Indic corpus and joint multitask attention-based fusion framework for cross-domain, multi-intent NLU in LoRes languages", IEEE Access **14** (2026) 28631. https://doi.org/10.1109/ACCESS.2026.3664703.
[26] J. Lee, & I. Joe, "A dialogue response generation model with latent variable modeling and LWE attention based on the VW-BLEU evaluation metric". IEEE Access **13** (2025), 49710.
[27] N. S. Mullah & W. M. N. W. Zainon, *AI and low-resource languages: Bridging the linguistic divide*, in *Reshaping Language and Cognition in Education Through AI*, IGI Global Scientific Publishing, Hershey, PA, USA, 2026, pp. 77–128. https://www.researchgate.net/publication/401236705_AI_and_Low-Resource_Languages_Bridging_the_Linguistic_Divide.
[28] J. Wang, B. Ma, Y. Nan, Y. He, H. Sun, C. Liu, S. Tao, Q. Qi & J. Liao, "Improving matching models with contextual attention for multi-turn response selection in retrieval-based chatbots", IEEE Transactions on Network Science and Engineering **12** (2025) 1497. https://doi.org/10.1109/TNSE.2025.3532663.
[29] B. M. Deeb, A. V. Savchenko & I. Makarov, "Enhancing emotion recognition in speech based on self-supervised learning: Cross-attention fusion of acoustic and semantic features", IEEE Access **13** (2025) 56283. https://doi.org/10.1109/ACCESS.2025.3554454.
[30] Y. Wang, X. Li, H. Yu, F. Hu, G. Wang & D. Lei, "Continuous entity reasoning for multi-turn medical dialogue generation", IEEE Transactions on Consumer Electronics **71** (2025) 10681. https://doi.org/10.1109/TCE.2025.3598217.
[31] P. Bojanowski, E. Grave, A. Joulin, & T. Mikolov, "Enriching word vectors with subword information. Transactions of the association for computational linguistics", **5** (2017) 135.
[32] A. Ba Alawi & F. Bozkurt, "Performance analysis of embedding methods for deep learning-based Turkish sentiment analysis models", Arabian Journal for Science and Engineering **50** (2025) 7299. https://doi.org/10.1007/s13369-024-09360-4.
[33] A. Somani, S. R. M. Kumar & A. Malapati, "Where does mBERT understand code-mixing? Layer-dependent performance on semantic tasks", IEEE Access **13** (2025) 135950. https://doi.org/10.1109/ACCESS.2025.3594135.
[34] S. Phani, A. Abdul, M. K. S. Prasad & V. D. Reddy, "MATSFT: User query-based multilingual abstractive text summarization for low-resource Indian languages by fine-tuning mT5", Alexandria Engineering Journal **127** (2025) 129. https://doi.org/10.1016/j.aej.2025.04.031.
[35] X. Li & K. Zhang, "Contrastive learning pre-training and quantum theory for cross-lingual aspect-based sentiment analysis", Entropy **27** (2025) 713. https://doi.org/10.3390/e27070713.
[36] L. B. Marques, S. Isotani, B. Pimentel, J. V. L. B. do Nascimento, M. Isotani, I. I. Bittencourt & C. E. Snow, "Structural and semantic analysis techniques for translation evaluation of educational materials", International Conference on Artificial Intelligence in Education, Cham, Switzerland, (2025), pp. 230. https://doi.org/10.1007/978-3-031-99261-2_21.
[37] I. Ezeani, M. Hepple, I. Onyenwe, & E. Chioma, *Igbo diacritic restoration using embedding models*. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, (2018) pp. 54.
[38] BBC News Igbo. https://www.bbc.com/igbo.
[39] Common Crawl Foundation, Common Crawl. https://commoncrawl.org.
[40] Teta!, "Jehovah's Witnesses (Awake!) Igbo Edition and Ulo Nche (Watchtower)". https://www.jw.org/ig.

11. Supplementary evaluation metrics

11.1. Extended response-generation results

The latency values reported in Table A1 represent the core neural inference time measured under controlled benchmarking conditions after data preprocessing and loading had been completed. Consequently, these values differ from those reported in Section 4.5, which represent end-to-end deployment latency, including preprocessing, contextual memory management, response generation, and user interface overhead. The two measurements evaluate different aspects of system performance and should therefore not be compared directly.

11.2. Detailed intent-classification results

The detailed intent-classification metrics are reported in Table A2.

11.3. Binary-classification results

The context-dependency detection results are reported in Table A3.

11.4. Human-evaluation results

The complete human-evaluation scores are reported in Table A4.

11.5. Statistical significance testing

The full significance-test results are reported in Table A5.

11.6. Error-analysis summary

The error-reduction summary is reported in Table A6.

This appendix provides extended quantitative results, detailed classification metrics, human evaluation outcomes, statistical significance analyses, and error analysis summaries that complement the main results presented in Section 4.

Table A1: Detailed response-generation performance

Model	BLEU (%)	ROUGE-L(%)	CRA (%)	Perplexity	Inference Latency (ms)
LSTM Seq2Seq	31.2	44.5	58.2	42.8	120
FastText	34.1	47.3	60.7	39.5	110
Standard Transformer	37.4	52.1	67.4	35.2	120
mBERT	40.2	54.8	70.5	32.4	115
XLM-R	42.0	57.3	74.1	30.6	105
Proposed Model	44.1	60.3	81.5	27.8	95

Table A2: Intent-classification metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC-AUC
FastText	71.5	70.2	69.8	70.0	0.75
LSTM	75.2	74.0	73.5	73.8	0.79
Transformer	78.4	77.3	76.9	77.1	0.82
mBERT	81.6	80.5	80.1	80.3	0.86
XLM-R	84.2	83.1	82.8	82.9	0.88
Proposed model	88.7	87.9	86.8	87.3	0.91

Table A3: Context-dependency detection performance.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic regression	65.3	64.1	63.5	63.8
FastText	70.2	69.1	68.4	68.7
Transformer	76.0	74.8	74.1	74.4
mBERT	80.3	79.1	78.5	78.8
XLM-R	82.6	81.7	81.0	81.3
Proposed model	85.6	84.8	83.9	84.3

Table A4: Human-evaluation scores (5-point Likert scale).

Attribute	LSTM	FastText	Transformer	mBERT	XLM-R	Proposed
Response clarity	3.2	3.6	3.9	4.2	4.4	4.7
Cultural sensitivity	2.9	3.4	3.5	4.0	4.2	4.6
Multi-turn coherence	3.1	3.8	4.0	4.3	4.5	4.8
Code-switch handling	3.3	4.0	4.1	4.4	4.6	4.9
Overall satisfaction	3.0	3.6	3.8	4.2	4.4	4.6

Table A5: Statistical significance of the proposed model against baselines.

Metric	LSTM	FastText	Transformer	mBERT	XLM-R
BLEU (p-value)	< 0.001	< 0.001	0.002	0.011	0.018
ROUGE-L (p-value)	< 0.001	< 0.001	0.004	0.013	0.021
F1 (p-value)	< 0.001	< 0.001	0.001	0.009	0.015
CRA (p-value)	< 0.001	< 0.001	0.001	0.007	0.012

Table A6: Error reduction achieved by the proposed model relative to the transformer.

Error category	Reduction (%)
Intent misclassification	53.1
Context drift	59.0
Irrelevant responses	57.7
Code-switch failure	60.9