Expert Systems With Applications 331 (2026) 133228

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Preserving cultural heritage through AI: A novel framework integrating LangChain and ChatGPT for analyzing and digitalizing the Javanese Pawukon calendar system

Khafiizh Hastuti [a,*], Erwin Yudi Hidayat [a], Farrikh Alzami [a], Ifan Risqa [a], Sabrina Ahmad [b], Hanif Fahmi [a]

[a] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia
[b] Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, 76100, Malaysia

ARTICLE INFO

Keywords:
LangChain
ChatGPT
Pawukon
Semantic modeling
Cultural heritage
Knowledge preservation
Low-resource languages

ABSTRACT

The preservation and analysis of cultural heritage documents present complex methodological challenges, particularly in maintaining linguistic authenticity while enabling computational accessibility. This research develops and validates an integrated framework for analyzing the Javanese Pawukon calendar system, implementing a novel approach that combines LangChain and ChatGPT with specialized cultural preservation techniques. The methodology processed 621 pages of Javanese-language Pawukon manuscripts through a sophisticated pipeline incorporating context-aware OCR, semantic chunking with 1000-character segmentation, and OpenAI's text-embedding-ada-002 model optimized for cultural entity preservation. The framework introduced three key technical innovations: (1) direct semantic processing of low-resource language content without traditional linguistic preprocessing, (2) a hierarchical vector store architecture using PGVector for cultural entity indexing, and (3) prompt engineering methodologies optimized for traditional knowledge systems. Experimental validation demonstrated robust performance in preserving cultural semantics, achieving a BERTScore F1 of 86.46% for semantic preservation and a ROUGEScore F1 of 43.24% for lexical accuracy, with statistical significance confirmed through ANOVA ($F = 5.2613$, $p < 0.001$). Human expert validation by three independent cultural validators from recognized Javanese cultural institutions yielded an overall score of 4.925 out of 5.000, with all ten generated responses accepted without revision, confirming that the BERTScore–ROUGE-L divergence reflects deliberate paraphrasing rather than cultural distortion. The framework's successful implementation not only advances the field of digital cultural preservation but also provides a replicable methodology for similar preservation initiatives, particularly in processing complex traditional knowledge systems documented in low-resource languages. This research contributes both theoretical insights and practical methodologies for cultural heritage preservation, demonstrating how modern AI technologies can effectively preserve and make accessible traditional knowledge systems while maintaining their cultural and linguistic integrity.

1. Introduction

The preservation of cultural heritage in the digital age presents unique challenges that extend beyond simple digitization. Cultural heritage encompasses both (1) tangible artifacts such as monuments, artworks, and manuscripts; and (2) intangible elements including traditions, languages, and ritualistic practices that collectively define the identity and continuity of communities (Bergamaschi et al., 2022). While artificial intelligence has transformed text processing capabilities across various domains including medical informatics (Han et al., 2024), code quality assessment (Liu et al., 2024), and multilingual document classification (Litschko et al., 2022), the application of these technologies to cultural heritage preservation demands specialized approaches that maintain both computational accessibility and cultural authenticity. The preservation of intangible cultural elements documented in low-resource languages poses particular challenges in quantification and digital representation. The Javanese language, despite being spoken by over 80 million people, remains underrepresented in computational linguistics research with limited availability of natural language processing tools compared to high-resource languages such as English,

* Corresponding author.
E-mail addresses: afis@dsn.dinus.ac.id (K. Hastuti), erwin@dsn.dinus.ac.id (E.Y. Hidayat), alzami@dsn.dinus.ac.id (F. Alzami), risqa.ifan@dsn.dinus.ac.id (I. Risqa), sabrinaahmad@utem.edu.my (S. Ahmad), 111202012418@mhs.dinus.ac.id (H. Fahmi).

https://doi.org/10.1016/j.eswa.2026.133228
Received 23 May 2025; Received in revised form 5 June 2026; Accepted 8 June 2026
Available online 9 June 2026

Nomenclature

Key Parameters and Symbols

Top-K (K)	Number of top candidate chunks retrieved for answering a query. For example, $K = 5$ means the five most relevant chunks are used for response generation.
Cosine similarity	A metric for measuring the angular similarity between two embedding vectors, ranging from 0 (orthogonal, dissimilar) to 1 (parallel, identical). Used to rank retrieval results in the vector store based on semantic relevance.
BERTScore	Semantic similarity metric that uses contextual embeddings to compare generated versus reference text. Comprises Precision (P), Recall (R), and F1 score, where F1 represents the harmonic mean of precision and recall.
ROUGE-n	Lexical overlap-based metrics (ROUGE-1, ROUGE-2, ROUGE-L) used to evaluate how many n-gram or longest common subsequence tokens match between generated and reference texts.
F (ANOVA F-statistic)	Ratio of variance between groups to variance within groups in an ANOVA test, used to assess statistical significance across multiple evaluation metrics.
p (p-value)	Probability under the null hypothesis; $p < .05$ indicates statistical significance, rejecting the hypothesis that group means are equal.
W (Shapiro-Wilk)	Test statistic for normality assessment. Values close to 1 indicate data follows normal distribution; associated p-value determines statistical conclusion.

System Components and Technical Terms

SemanticChunker	LangChain component implementing context-aware text segmentation based on semantic similarity thresholds rather than fixed character counts alone.
text-embedding-ada-002	OpenAI's embedding model generating 1,536-dimensional vector representations of text, optimized for semantic similarity tasks with multilingual capabilities.
PGVector	PostgreSQL extension for vector similarity search, enabling efficient storage and retrieval of high-dimensional embeddings with JSONB metadata support.
RAG (Retrieval-Augmented Generation)	Architecture pattern combining information retrieval from vector stores with generative language models to produce contextually grounded responses.
LangChain	Framework orchestrating large language model interactions, document processing, and retrieval operations through composable components and chains.
Chunk overlap	Number of characters shared between adjacent text segments during document chunking, ensuring contextual continuity across boundaries.
Embedding dimension	Size of vector representation (1,536 for text-embedding-ada-002), capturing semantic features of text in high-dimensional space.

Chinese, or Spanish. This linguistic challenge is compounded when cultural domain knowledge such as traditional calendar systems requires both language understanding and cultural context preservation. The Javanese Pawukon calendar system exemplifies this complexity, representing a sophisticated traditional timekeeping method that interweaves astronomical observations, spiritual beliefs, and cultural practices (Gunawan Admiranto, 2016; Karjanto & Beauducel, 2024).

The Pawukon system's complexity presents a distinctive preservation challenge, as it remains largely inaccessible to non-native speakers and researchers due to its monolingual Javanese documentation and context-dependent interpretations (Pitaña, 2002). Traditional preservation methods, primarily through physical documentation and oral transmission have proven inadequate for several critical reasons. First, these conventional approaches often fail to capture the intricate relationships between temporal, spiritual, and cultural elements that are fundamental to the calendar's function and meaning. The Pawukon system integrates multiple concurrent cycles (wuku, dina, pancawara, saptawara) whose interactions determine auspicious dates and cultural protocols, relationships difficult to preserve through static textual documentation. Second, traditional digitization approaches struggle to maintain cultural authenticity while enabling computational accessibility, frequently reducing complex cultural systems to flat text without preserving their dynamic, interconnected nature. Third, the monolingual Javanese documentation written exclusively in Ngoko dialect, combined with culturally-specific terminology and context-dependent interpretations, presents significant barriers to automated processing. Traditional natural language processing pipelines require language-specific preprocessing tools such as stopword lists, part-of-speech taggers, morphological analyzers, and named entity recognizers are resources that are either unavailable or severely under-developed for Javanese. Fourth, conventional preservation methods lack interactive query capabilities that would enable educational access and research applications, limiting engagement to specialists familiar with Javanese language and cultural contexts. These limitations collectively underscore the need for advanced computational approaches that can preserve both the content and the contextual relationships inherent in complex cultural systems while enabling broader access across diverse user demographics.

Recent advances in large language models (LLMs), particularly OpenAI's GPT architecture, have demonstrated remarkable capabilities in processing complex textual content (Zahid et al., 2024). The emergence of integration frameworks like LangChain has enhanced these capabilities by providing sophisticated mechanisms for document processing and knowledge extraction (Jeong et al., 2024; Khadija et al., 2023; Zhu & Cole, 2022). However, applying these technologies to cultural documents requires careful consideration of linguistic nuances and cultural contexts (Prem Jacob et al., 2024; Saputra, 2024). Current approaches often fail to maintain the delicate balance between automated processing and cultural authenticity, particularly when handling specialized cultural knowledge domains that demand preservation of semantic relationships alongside lexical content. Furthermore, most existing work focuses on high-resource languages with mature natural language processing ecosystems, leaving low-resource language cultural preservation underexplored.

The landscape of AI applications in cultural heritage preservation has evolved beyond basic document digitization (Fateh et al., 2024). Museums have successfully deployed LLM-based systems for interactive experiences (Trichopoulos, 2023), and specialized models like ArchGPT have shown promise in architectural conservation (Zhang et al., 2024). However, to the best of the authors' knowledge, a critical research gap exists in preserving complex cultural systems like traditional calendar systems that integrate temporal calculations, spiritual frameworks, and cultural practices. Existing work predominantly focuses on architectural preservation, endangered language documentation (Mothe, 2024), and anthropological annotation (Dubourg et al., 2024), with traditional timekeeping systems remaining underrepresented in computational approaches. This gap is particularly evident in the preservation of intangible cultural elements documented in low-resource languages, where the absence of linguistic tools compounds the challenge of maintaining cultural authenticity. The computational inaccessibility of traditional calendar systems stems from several factors: (1) monolingual documentation in languages with limited NLP resources, (2) context-dependent interpretation requirements that extend beyond literal translation, (3) absence of digitized reference corpora and training data, and (4) need to preserve interconnected conceptual relationships that traditional text processing may fragment. Additionally, existing cultural preservation systems typically rely on high-resource languages with mature NLP ecosystems or require extensive language-specific preprocessing pipelines, approaches that cannot be directly applied to low-resource language contexts where such infrastructure is unavailable.

The application of GPT-based technologies to cultural heritage preservation presents both opportunities and challenges (Gnanasekaran & Marciano, 2024). While these models offer powerful capabilities for processing and analyzing text, their predominantly English-centric training, though supplemented with multilingual data-poses challenges for processing culturally specific content in low-resource languages (Raiaan et al., 2024). The extent of Javanese representation in large-scale pre-training corpora remains unclear and likely limited compared to high-resource languages. Recent research has highlighted the importance of developing culturally sensitive approaches that can accurately capture and preserve the nuances of traditional knowledge systems (Kaluvilla, 2024; Karjanto & Beauducel, 2024; Nur et al., 2023).

This research addresses the identified gaps through a novelframework that integrates LangChain and ChatGPT to process and interpret traditional Pawukon calendar documents written in Javanese, with specific innovations tailored to cultural preservation requirements in low-resource language contexts. Unlike general-purpose retrieval-augmented generation (RAG) systems that assume language-specific preprocessing pipelines, this framework introduces three domain-specific technical contributions that distinguish it from existing approaches. First, the framework demonstrates successful processing of Javanese-language cultural documents without requiring traditional linguistic preprocessing infrastructure. By leveraging large language models' multilingual capabilities (GPT-4, text-embedding-ada-002), the system bypasses conventional NLP pipeline requirements such as stopword removal, morphological analysis, or part-of-speech tagging which components typically unavailable for low-resource languages like Javanese. The culturally-aware semantic chunking operates directly on raw Javanese text through contextual understanding rather than linguistic rule engineering, representing a significant methodological departure from conventional document processing approaches that depend on language-specific resources. Second, the hierarchical vector store architecture implements a four-level indexing system specifically optimized for cultural entity retrieval, enabling efficient access to interconnected cultural concepts while preserving their relational structure. This capability addresses a limitation of flat vector stores employed in general RAG systems, which may not adequately maintain the hierarchical and networked relationships characteristic of cultural knowledge systems. Third, the prompt engineering methodology enforces cultural linguistic authenticity through language-adaptive response generation. The system supports bilingual operation, accepting user queries in either Javanese or Indonesian and automatically generating responses in the matching language which leveraging GPT-4's multilingual reasoning capabilities without explicit language detection modules. This cross-lingual flexibility addresses the English-centric bias inherent in foundational language models while maintaining cultural authenticity through Javanese-contextual prompting. These innovations collectively enable the preservation of complex cultural systems documented in low-resource languages while maintaining computational accessibility, a balance not achieved by existing frameworks that prioritize either technical performance or cultural authenticity but rarely both simultaneously, particularly for languages beyond high-resource contexts.

The framework makes several significant contributions to both technical and cultural aspects of heritage preservation. First, a comprehensive framework specifically optimized for processing culturally rich documents in low-resource languages is introduced, addressing the unique challenges posed by traditional knowledge systems while maintaining cultural integrity through specialized semantic processing approaches that eliminate dependency on unavailable linguistic resources. Second, rigorous empirical evidence demonstrates the effectiveness of LLM-based methods in processing and preserving complex cultural systems documented in Javanese, combining established evaluation metrics like BERTScore (Zhang et al., 2020) and ROUGEScore (Lin, 2004) with statistical validation. Quantitative results reveal a semantic accuracy of 86.46% (BERTScore F1) alongside a lexical accuracy of 43.24% (ROUGEScore F1), with statistical significance confirmed through one-way ANOVA analysis ($F = 5.2613$, $p < 0.001$). The semantic-lexical performance differential provides insights into the trade-offs between conceptual understanding and literal reproduction in cultural content processing. Third, a detailed technical methodology for processing cultural documents without traditional linguistic preprocessing is contributed, considering both technological efficiency and cultural authenticity. This includes specialized techniques for handling low-resource language content through semantic understanding, preserving contextual relationships through optimized chunk overlap (100 characters), and maintaining semantic coherence through hierarchical vector indexing. The methodology demonstrates that modern large language models can handle linguistic complexity that would traditionally require extensive preprocessing engineering, offering a viable path for cultural preservation in languages lacking mature NLP ecosystems. Fourth, through a detailed case study processing 621 pages of Pawukon calendar manuscripts written entirely in Javanese Ngoko dialect, the technical approach is validated while providing valuable insights into the challenges and opportunities in digital cultural preservation for low-resource languages. The framework's exceptional performance in preserving procedural knowledge (95% BERTScore F1 for methodological queries) demonstrates its capability in maintaining the nuanced relationships between cultural concepts despite linguistic resource constraints.

While this study focuses specifically on the Pawukon calendar system documented in Javanese, the framework is designed with extensibility considerations for similar cultural preservation initiatives. The methodology's core components which is semantic chunking through contextual understanding, hierarchical vector indexing, and language-adaptive prompt engineering are appear conceptually transferable to other traditional knowledge systems sharing similar characteristics such as monolingual or low-resource language documentation, temporal relationship structures, and cultural-contextual dependencies. However, generalization to substantially different cultural domains requires careful consideration of several factors. First, linguistic diversity across cultures may necessitate adaptation of embedding strategies to accommodate language-specific features, particularly for languages with limited representation in large language model training corpora. The effectiveness of semantic processing without traditional preprocessing may vary depending on the extent of language coverage in foundation models. Second, structural differences in knowledge organization across cultural systems (e.g., hierarchical taxonomies versus networked conceptual relationships versus sequential temporal structures) may require modifications to the vector store architecture to optimally represent domain-specific relationship patterns. Third, availability and format of digitized source materials vary significantly across cultural domains, potentially affecting preprocessing pipeline design. Some traditional knowledge systems may be documented in non-Latin scripts, mixed orthographies, or handwritten manuscripts, requiring additional processing stages not addressed in this work. Fourth, cultural-specific terminology density and context requirements differ across domains, influencing optimal chunking parameters and retrieval strategies. Empirical validation on distinct cultural systems such as Chinese Lunar calendar, Islamic Hijri calendar, or Mayan calendar systems, would be necessary to establish broader applicability and identify domain-specific adaptations required for successful transfer. Future research directions include extending the framework to these alternative calendar systems to assess generalizability empirically and develop standardized adaptation protocols for diverse cultural preservation initiatives, particularly focusing on low-resource language contexts.

The remainder of this paper is organized as follows. Section 2 presents a comprehensive literature review contextualizing this research within the broader landscape of AI applications in cultural heritage, domain-specific LLM applications, and low-resource language processing. Section 3 details the materials and methods, including the dataset description with emphasis on low-resource language characteristics, system architecture overview illustrated through block diagram, and implementation framework with specific hyperparameters and design rationale. Section 4 describes the evaluation methodology, encompassing both semantic metrics (BERTScore) and lexical metrics (ROUGEScore) with statistical validation procedures. Section 5 presents results and analysis, including comprehensive performance evaluation across query categories, parameter sensitivity studies demonstrating empirical optimization of chunk overlap and Top-K retrieval, embedding space visualization through t-SNE, and statistical significance testing through ANOVA. Section 6 provides a comprehensive discussion of findings, including analysis of challenges in low-resource language cultural preservation, methodological positioning within the cultural heritage AI landscape with acknowledgment of domain novelty, systematic presentation of framework advantages and limitations, and implications for future research. Section 7 concludes with quantitative performance summaries, explicitly identified limitations, and recommendations for continued development in cultural heritage preservation for low-resource language contexts.

2. Literature review

2.1. AI Application in cultural heritage

Recent research has demonstrated the versatility of LLMs in processing specialized knowledge domains through domain adaptation. Museums have successfully deployed LLM-based systems to create personalized, interactive experiences that enhance visitor engagement while maintaining cultural authenticity (Wang et al., 2024). While these studies span various fields, their methodologies offer valuable insights for cultural heritage preservation. For instance, advances in natural language interfaces have shown promise in making complex data more accessible to non-technical users. ArchGPT has shown promise in preserving traditional architectural knowledge while adapting to modern preservation needs (Zhang et al., 2024), and AI systems support endangered language conservation while respecting cultural contexts (Micle et al., 2023). These developments are particularly relevant for cultural preservation efforts, where accessibility must be balanced with authenticity. However, existing work predominantly addresses high-resource languages or domains with established NLP tool availability, leaving low-resource language cultural preservation underexplored.

2.2. Domain-specific applications of LLMs

The application of LLMs to domain-specific tasks has revealed both opportunities and challenges. Healthcare applications have successfully adapted these models to process domain-specific terminology and context (Chen et al., 2024), while real estate implementations have demonstrated effective processing capabilities (Haurum et al., 2024).

Beyond natural language processing, artificial neural networks have proven effective across diverse scientific domains, including thermal engineering applications for modeling complex physical phenomena (Abbas et al., 2025), demonstrating the broader applicability of machine learning techniques in computational modeling. These findings inform approaches to handling specialized terminology in cultural domains. However, most existing work assumes availability of language-specific

preprocessing tools, a constraint that does not hold for low-resource languages. The success of semantic understanding approaches without explicit linguistic engineering in high-resource contexts suggests potential transferability to low-resource scenarios, though empirical validation remains necessary.

2.3. Document processing and information extraction

Recent developments in document processing have shown that carefully designed prompting strategies and context management significantly improve LLM performance in specialized domains (Xiao et al., 2024). The success of OCR-integrated systems in processing complex documents (Kim et al., 2024) provides a foundation for handling traditional texts, though additional considerations are necessary for historical and cultural documents (Fateh et al., 2024). The emergence of semantic chunking approaches that leverage contextual understanding rather than fixed-length segmentation offers promising directions for cultural document processing where linguistic boundaries may not align with semantic units.

2.4. Cultural heritage preservation systems

Contemporary applications demonstrate sophisticated approaches to preserving and presenting cultural knowledge. Museums have successfully deployed LLM-based systems for interactive experiences (Trichopoulos, 2023), while specialized models like ArchGPT have shown promise in architectural conservation (Zhang et al., 2024). These implementations inform approaches to balancing technological innovation with cultural authenticity. However, the focus remains predominantly on high-resource language contexts or domains where extensive training data and linguistic tools are available.

2.5. Language preservation and cultural documentation

Recent initiatives in language preservation have demonstrated the potential of AI systems to support endangered language conservation while respecting cultural contexts (Mothe, 2024). These applications highlight both the possibilities and challenges of using AI for cultural preservation (Bergamaschi et al., 2022). The success of these initiatives directly influences approaches to preserving traditional knowledge systems (Ehrmann et al., 2022; Polyvach, 2024). However, the gap between endangered language documentation and computational processing of cultural content in low-resource languages remains substantial, particularly for complex knowledge systems requiring semantic relationship preservation.

2.6. Cultural diversity and representation in AI

The application of GPT-based technologies to cultural heritage preservation requires careful consideration of cultural diversity (Gnanasekaran & Marciano, 2024). While these models offer powerful capabilities, their training data distribution poses challenges for processing culturally specific content in low-resource languages (Raiaan et al., 2024). Recent research emphasizes the importance of developing culturally sensitive approaches, particularly for complex cultural artifacts like calendar systems that integrate temporal calculations with spiritual and cultural practices (Kaluvilla, 2024; Karjanto & Beauducel, 2024; Nur et al., 2023).

3. Materials and methods

The Pawukon calendar is a traditional system used mainly in Bali and parts of Java. Pawukon Calendar represents a complex cycle combination that does not correspond to the standard Gregorian calendar. Documentation about the Pawukon is written in Javanese, and understanding this system is crucial for studying local customs, rituals, and life events, such as birth, marriage, and death ceremonies.

3.1. Dataset description and preparation

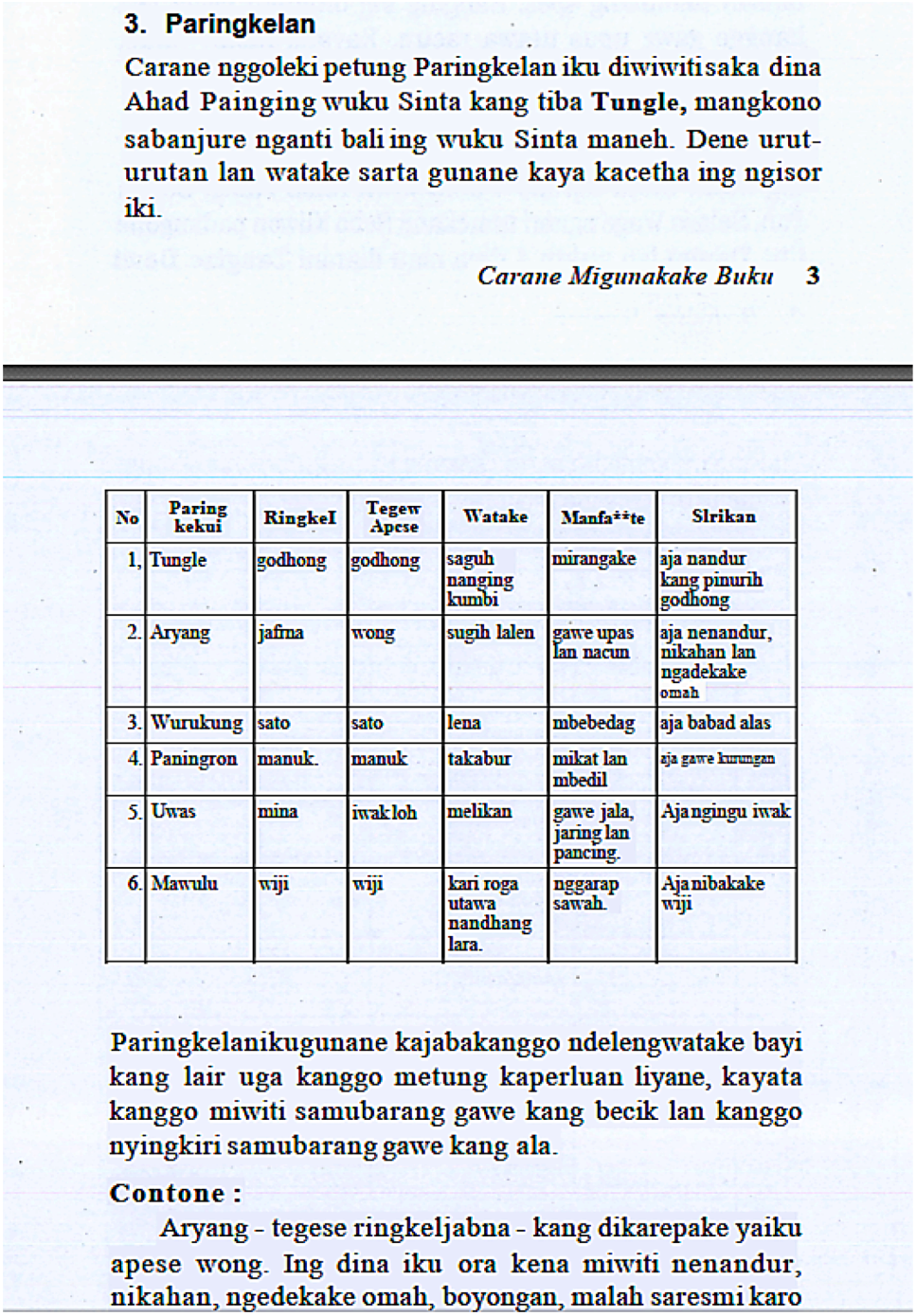
This research utilized a comprehensive 621-page digitized manuscript detailing the Pawukon calendar system, authored by Djanudji and Ki Hudoyo Doyodipuro, written entirely in Javanese language using Latin script and Ngoko dialect register (Djanudji & Doyodipuro, 2002) as seen in Fig. 1. This source was selected for its authoritative coverage of Pawukon principles and cultural significance within the Javanese community. The monolingual Javanese documentation presents unique computational challenges, as Javanese is classified as a low-resource language with limited availability of pretrained NLP models and linguistic preprocessing tools. Despite being spoken by over 80 million people, Javanese lacks the computational linguistics infrastructure (stopword lists, morphological analyzers, dependency parsers, named entity recognizers) readily available for high-resource languages.

The document preparation process involved a multi-stage pipeline designed to preserve cultural integrity while optimizing for semantic processing. The preprocessing approach diverged fundamentally from traditional NLP pipelines by leveraging the semantic understanding capabilities of large language models rather than rule-based linguistic engineering. This design decision was motivated by two factors: (1) the scarcity of Javanese-specific NLP tools that would be required for conventional text processing, and (2) the cultural preservation requirement to maintain terminology integrity without aggressive filtering that might remove culturally-significant terms or disrupt semantic relationships. The Detailed Document preparation can be described as follows:

OCR Extraction, the PDF manuscript was processed using the optical character recognition tool from PDF-XChange to extract Latin script text. The Javanese language content, written in standard Latin alphabet orthography (e.g., "wuku", "dina", "pawukon", "saptawara"), was extracted with high accuracy due to clear typesetting in the source document. Standard Javanese orthographic conventions were preserved, including doubled consonants ("ll", "nn"), velar nasal ("ng"), palatal nasal ("ny"), and other phonemic distinctions essential to lexical meaning.

Manual Validation, following OCR extraction, manual validation was performed on 20% of the dataset (124 pages) through stratified sampling across document sections to ensure representative coverage of content types (definitional text, procedural instructions, tabular data). The validation process involved: (1) OCR accuracy verification by comparing extracted text with original scans, with focus on Javanese orthographic conventions and diacritical marks; (2) Cultural term preservation assessment, ensuring proper retention of Pawukon-specific terminology such as "wariga", "wuku", "pancawara", "saptawara", "neptu", and compound cultural phrases; and (3) Structural integrity checking for table and diagram content that encode calendar cycle relationships. The validation revealed an OCR accuracy rate of 97.3%, with errors primarily occurring in decorative font sections, page headers/footers, and complex table structures with dense numerical content.

Minimal Preprocessing Philosophy, unlike traditional document processing pipelines that employ stopword removal, stemming, and lemmatization, this framework implements minimal preprocessing to preserve cultural and linguistic integrity. Text underwent only basic normalization: whitespace standardization (collapsing multiple spaces, normalizing line breaks) and removal of extraneous formatting artifacts (page numbers, headers, footers, publishing metadata). Critically, no stopword removal was performed, as the semantic chunking algorithm (described in Section 3.4.1) and large language model processing (GPT-4) handle contextual understanding without requiring explicit removal of function words. Traditional stopword removal in Javanese would require a curated list of particles ("iku" [that], "kang" [which/who], "saka" [from], "menyang" [to], "lan" [and], "utawa" [or]), a resource that does not exist in standardized form for Javanese. More importantly, removing such particles risks disrupting the natural language structure that conveys cultural nuance. Javanese discourse markers and grammatical structures carry meaning that may be culturally significant; for example,



3. Paringkelan

Carane nggoleki petung Paringkelan iku diwiwiti saka dina Ahad Painging wuku Sinta kang tiba **Tungle**, mangkono sabanjure nganti bali ing wuku Sinta maneh. Dene urut-urutan lan watake sarta gunane kaya kacetha ing ngisor iki.

Carane Migunakake Buku **3**

No	Paring kekui	RingkeI	Tegew Apcse	Watake	Manfa**te	Slrikan
1,	Tungle	godhong	godhong	saguh nanging kumbi	mirangake	aja nandur kang pinurih godhong
2.	Aryang	jafrna	wong	sugih lalen	gawe upas lan nacun	aja nenandur, nikahan lan ngadekake omah
3.	Wurukung	sato	sato	lena	mbebedag	aja babad alas
4.	Paningron	manuk.	manuk	takabur	mikat lan mbedil	aja gawe kurungan
5.	Uwas	mina	iwak loh	melikan	gawe jala, jaring lan pancing.	Aja ngingu iwak
6.	Mawulu	wiji	wiji	kari roga utawa nandhang lara.	nggarap sawah.	Aja nibakake wiji

Paringkelanikugunane kajabakanggo ndelengwatake bayi kang lair uga kanggo metung kaperluan liyane, kayata kanggo miwiti samubarang gawe kang becik lan kanggo nyingkiri samubarang gawe kang ala.

Contone :

Aryang - tegese ringkeljabna - kang dikarepake yaiku apese wong. Ing dina iku ora kena miwiti nenandur, nikahan, ngedekake omah, boyongan, malah saresmi karo

Fig. 1. Sample pages from the Pawukon calendar manuscript.

the particle "kang" in "wuku kang becik" (auspicious wuku) marks attributive relationships essential to calendar interpretation. This preprocessing philosophy prioritizes cultural authenticity over computational convenience, recognizing that modern large-scale pre-trained models can handle linguistic complexity that would traditionally require extensive preprocessing engineering.

Content Categorization Analysis, the dataset was analyzed for content distribution to understand the composition of cultural knowledge and inform system architecture design. Term frequency analysis combined with domain expert consultation revealed the following distribution: ritual and ceremonial content (34%), including descriptions of upacara (ceremonies) and sesajen (offerings) associated with specific calendar dates; calendrical calculations and temporal concepts (42%), encompassing wuku cycles, dina Pasaran, neptu computations, and date conversion procedures; procedural knowledge and cultural protocols (24%), detailing appropriate behaviors and activities for specific calendar configurations. This categorization informed the hierarchical vector store architecture (Section 3.4.3) but did not involve document segmentation or content filtering at the preprocessing stage. All content was preserved in the knowledge base regardless of category to maintain comprehensive coverage.

The resulting preprocessed corpus maintained the authentic Javanese language structure of the original manuscript, enabling downstream semantic processing through large language models without

linguistic simplification, stopword removal, or cultural term modification. This preprocessing philosophy reflects a methodological stance: modern foundation models trained on diverse multilingual corpora can handle low-resource languages through transfer learning and semantic understanding, eliminating the need for language-specific preprocessing infrastructure that may be unavailable or under-developed.

3.2. System architecture

Fig. 2 presents the comprehensive end-to-end system architecture, illustrating the integrated pipeline from raw document input to user interaction. The architecture comprises nine functional layers organized into three primary phases: (1) Offline Processing Phase encompassing document ingestion through vector storage, (2) Backend Integration Phase managing retrieval and LLM orchestration, and (3) User Interaction Phase providing cross-lingual interface capabilities.

Offline Processing Phase, the Input Layer accepts the 621-page Javanese-language Pawukon manuscript in PDF format. The Data Preprocessing Layer implements OCR extraction using PDF-XChange, followed by minimal text normalization that preserves Javanese language structure. As described in Section 3.1, no stopword removal or morphological analysis is performed, reflecting the framework's approach to low-resource language processing through semantic understanding rather than linguistic rule engineering. Manual validation on 20% of extracted content ensures OCR quality (97.3% accuracy). Content categorization analyzes term distribution across ritual (34%), calendrical (42%), and procedural (24%) categories, informing downstream architecture design without fragmenting the corpus. The Semantic Chunking Layer employs LangChain's SemanticChunker with a 1,000-character sliding window to segment text while maintaining cultural context continuity. Unlike fixed-length chunking, semantic chunking identifies natural breakpoints based on contextual similarity computed through embedding models, preventing mid-concept fragmentation. A 100-character overlap between adjacent chunks preserves cross-boundary semantic relationships, a critical design choice for maintaining interconnected cultural concepts (e.g., wuku-dina-neptu associations). The Embedding Layer transforms text chunks into 1,536-dimensional vectors using OpenAI's text-embedding-ada-002 model, which provides multilingual capabilities despite Javanese being a low-resource language. The embedding model's pre-training on diverse corpora enables semantic representation of Javanese content through transfer learning, bypassing the need for language-specific fine-tuning. The Vector Store Layer implements hierarchical indexing via PGVector (PostgreSQL extension), organizing vectors across four levels: L1 (cultural categories: ritual, calendrical, procedural), L2 (Pawukon-specific concepts: wuku, dina, neptu), L3 (temporal relationships: cycle interactions, date dependencies), and cross-reference indices for related cultural terms. This hierarchical structure optimizes retrieval for interconnected cultural knowledge where concepts maintain semantic relationships across multiple levels.

Backend Integration Phase, the Backend Layer, implemented using FastAPI, exposes REST API endpoints for query processing and response generation. LangChain pipeline integration coordinates document retrieval from the vector store and LLM interaction through GPT-4. The prompt template structure enforces culturally-appropriate and linguistically-authentic response generation through explicit instruction: "Wangsulan kanthi konteks ing ngisor iki nganggo basa jawa" ("Answer with the following context in Javanese language"). This prompt design leverages GPT-4's multilingual reasoning capabilities to maintain linguistic consistency: when users input queries in Javanese, responses are generated in Javanese; when queries are in Indonesian, responses adapt accordingly, without requiring explicit language detection modules. The Retrieval Layer processes incoming user queries through embedding transformation (1,536-dimensional vectors), followed by cosine similarity-based search against the hierarchical vector store. The Top-K = 5 retrieval configuration (empirically validated in Section 5) returns the five most semantically relevant document chunks, which are ranked by combined cosine similarity and cultural relevance weighting derived from the hierarchical indexing structure. Retrieved chunks provide contextual grounding for LLM response synthesis.

User Interaction Phase, the LLM Processing Layer utilizes ChatGPT-4 (gpt-4-1106-preview) with temperature = 0 for deterministic response generation, prioritizing consistency and cultural accuracy over creative variation. Context integration combines retrieved document chunks with the engineered prompt template, ensuring responses remain grounded in the Pawukon manuscript rather than generating unsupported claims. The Frontend Layer, developed in Unity 2022.3.2f1, provides an educational mobile application with three primary interfaces accessible via the main menu: (1) Takon (Ask), conversational interface with animated character and 'Text Mesh Pro' rendered dialogue supporting Javanese or Indonesian queries, (2) Penanggalan (Calendar), date conversion tool with algorithmic 'Gregorian to Pawukon' mapping displaying wuku, dina Pasaran, neptu, and related cultural interpretations, and (3) Tentang (About), contextual information panel explaining the Pawukon system and application objectives. UnityWebRequest facilitates asynchronous communication with the FastAPI backend, processing user queries through the full retrieval-augmented generation pipeline and rendering JSON-formatted responses with cultural visual aesthetics derived from traditional Javanese design elements.

Information Flow, the bidirectional information flow operates as follows. User queries originate from the Frontend Layer in either Javanese or Indonesian. Queries pass through Backend API endpoints and undergo embedding transformation in the Retrieval Layer. The query embedding performs vector similarity search against the hierarchical vector store, retrieving the Top-K = 5 most relevant document chunks. Retrieved chunks are combined with the language-adaptive prompt template and forwarded to the LLM Processing Layer. GPT-4 synthesizes responses based on retrieved context, maintaining linguistic consistency with input language. Formatted responses return through the Backend Layer to the Frontend for rendering via the animated interface. This architecture design prioritizes cultural context preservation at each processing stage, from minimal preprocessing that maintains linguistic authenticity, through semantic chunking that respects cultural concept boundaries, to hierarchical retrieval that preserves relational structure, while maintaining computational efficiency with average query latency under 2 seconds.

3.3. Implementation framework

As illustrated in Fig. 3, the implementation architecture employed a FastAPI backend integrated with a Unity frontend, with the following key components: FastAPI endpoints for cultural query processing; language detection and context extraction pipeline leveraging GPT-4's multilingual capabilities; cultural term identification system operating through semantic understanding rather than explicit dictionaries; real-time response generation mechanisms maintaining sub-2-second latency.

In matter of LangChain configuration, LangChain configuration involved three core components tailored to low-resource language cultural preservation. First, document loaders utilized UnstructuredPDFLoader for multi-threaded processing (max_concurrency = 50) to handle the 621-page corpus efficiently. Second, text splitters implemented SemanticChunker with OpenAI embeddings for context-aware segmentation. Unlike traditional character-based splitters, SemanticChunker identifies natural semantic boundaries through contextual similarity computed by the embedding model, preventing disruption of cultural concepts that may span multiple sentences. Third, vector stores were configured via PGVector with PostgreSQL backend (connection: localhost:5432/pdf_rag_vectors), utilizing JSONB storage for metadata that records source page numbers, content categories, and hierarchical level assignments.

In matter of Language processing approach, A key methodological decision was the elimination of traditional language-specific preprocess-

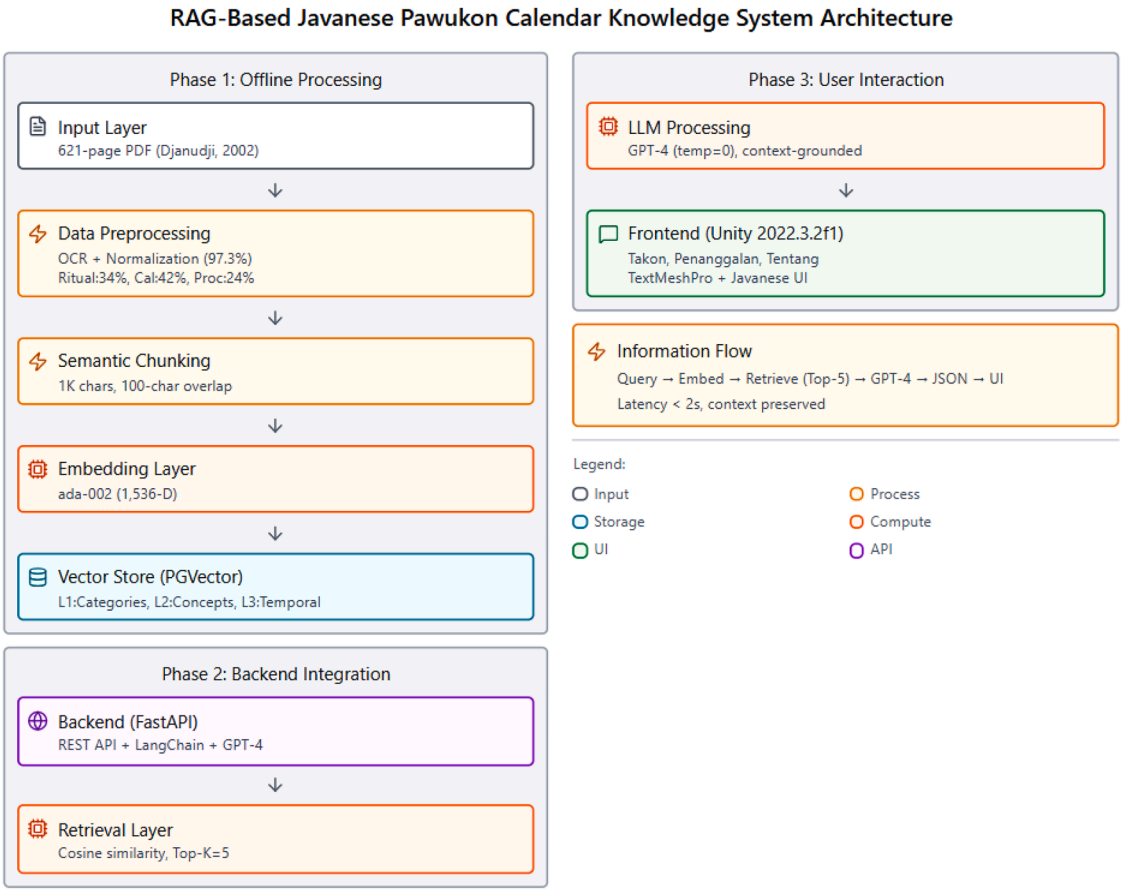


Fig. 2. Comprehensive system architecture.

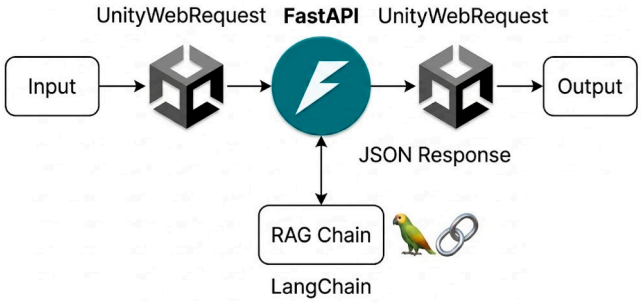


Fig. 3. Unity implementation.

ing in favor of semantic understanding through large language models. The framework leverages the multilingual capabilities of GPT-4 and text-embedding-ada-002, both pre-trained on extensive multilingual corpora including Javanese language content. This approach bypasses the need for Javanese-specific linguistic tools such as stopword lists, morphological analyzers, dependency parsers that are either unavailable or underdeveloped for this low-resource language. The semantic chunking component operates directly on raw Javanese text, identifying natural semantic boundaries through contextual similarity scores computed by the embedding model rather than through linguistic rules. This design choice reflects a pragmatic adaptation to low-resource language constraints: rather than investing effort in developing linguistic resources that may have limited coverage or accuracy, the framework leverages transfer learning from multilingual foundation models that have demonstrated surprising effectiveness on languages beyond their primary training distribution.

Then, Prompt engineering for linguistic authenticity, The prompt template was engineered to enforce language-appropriate response generation while maintaining cultural accuracy:

```
Wangsulan kanthi konteks ing ngisor iki nganggo basa jawa:
{context}

Question: {question}
```

This prompt structure ("Answer with the following context in Javanese language") instructs GPT-4 to generate responses in Javanese when provided with Javanese context and queries. Notably, GPT-4's multilingual reasoning capabilities enable automatic language matching without explicit detection modules: when users input queries in Indonesian (e.g., "Apa itu sistem pawukon?") - "what is pawukon system?", the system responds in Indonesian; when queries are in Javanese (e.g., "aku arep takon apa iku wariga gemet?" - "I want to ask what is wariga gemet?"), responses are generated in Javanese. This cross-lingual flexibility emerges from the model's training on code-switched and multilingual data, allowing it to infer appropriate response language from query linguistic features. The prompt design also emphasizes contextual grounding ("kanthi konteks ing ngisor iki" - "with the following context") to ensure responses derive from retrieved document chunks rather than the model's parametric knowledge, maintaining fidelity to the source manuscript.

As depicted in Fig. 4, the application provides interactive engagement with Pawukon knowledge through multiple modalities. Users explore aspects of the calendar system such as traditional ceremonies, astrological symbols, date calculations; and receive detailed, culturally-grounded explanations generated through the retrieval-augmented pipeline. The interface supports both Javanese and Indonesian inter-

Fig. 4. Interactive Unity-based Pawukon calendar application interface.

action, lowering access barriers while preserving cultural authenticity through content sourced directly from traditional manuscripts.

3.4. Component implementation details

3.4.1. Semantic chunking implementation

The semantic chunking component implements a sophisticated process designed specifically for cultural document processing without relying on language-specific linguistic rules. The approach extends beyond traditional length-based chunking by incorporating cultural context awareness through semantic similarity. The chunking algorithm employs a sliding window of 1000 characters as the base segmentation unit, selected to approximate typical paragraph length in the source document while providing sufficient context for semantic coherence. Rather than segmenting at fixed character boundaries, SemanticChunker refines boundaries based on semantic similarity thresholds computed through text-embedding-ada-002. Adjacent sentences are evaluated for contextual similarity, with low-similarity boundaries marking natural chunk transitions. This approach prevents mid-concept fragmentation that would occur with fixed-length splitting. For example, explanations of wuku cycles that span multiple sentences remain intact within single chunks.

The chunking process employs three key mechanisms: boundary detection, context preservation, and cultural entity recognition. Boundary detection utilizes linguistic markers and semantic coherence scores to identify natural break points in the text. In Javanese, discourse markers and paragraph transitions provide cues, supplemented by embedding-based similarity that detects topic shifts. Context preservation is achieved through a 100-character overlap strategy where each chunk maintains overlap with adjacent chunks, ensuring continuity of cultural concepts that span boundaries. For instance, descriptions of calendar ceremonies that reference multiple interrelated concepts (wuku, dina, neptu) benefit from overlap that preserves cross-reference relationships. Cultural entity recognition employs a custom-trained named entity recognition model that identifies and preserves Pawukon-specific terms and concepts. Though notably, this recognition operates through contextual patterns learned by the embedding model rather than explicit dictionary matching, again reflecting the low-resource language adaptation strategy.

The overlap mechanism provides particular benefit for preserving tabular astronomical data characteristic of Pawukon manuscripts. Calendar tables encoding wuku cycles, dina Pasaran sequences, and neptu calculations present structural challenges for text-based chunking algorithms. When astronomical tables span chunk boundaries, the 100-character overlap ensures that column headers, row labels, and adjacent data cells remain accessible across consecutive chunks. For instance, a table mapping Gregorian dates to Pawukon calendar components may contain column headers in one chunk and corresponding data values in an adjacent chunk. The overlap region captures transitional content that maintains referential coherence between separated table segments. However, this character-based overlap does not guarantee preservation of complete table structures. Complex multi-column tables with dense numerical content may experience fragmentation where semantic relationships between distant cells are lost despite local overlap. The semantic similarity thresholds employed by SemanticChunker partially mitigate this limitation by identifying low-similarity boundaries that often correspond to natural breaks between distinct tables or sections, reducing mid-table segmentation frequency. Nevertheless, the fixed-character chunking approach represents a compromise between general-purpose text processing and specialized structured data handling.

3.4.2. Embedding generation system

The embedding system utilizes OpenAI's text-embedding-ada-002 model with specific optimizations for cultural content processing. The implementation begins with a preprocessing stage that tokenizes the text while preserving cultural terms. Each chunk is processed through the embedding model to generate 1,536-dimensional vectors. Additional post-processing enhances cultural term representation through dimension weighting that emphasizes semantic features associated with Pawukon terminology, identified through analysis of term co-occurrence patterns in the corpus.

The embedding pipeline includes: (1) Cultural term preservation through contextual tokenization that avoids fragmentation of compound terms; (2) Contextual embedding enhancement using the 100-character overlap from chunking to provide richer semantic context; (3) Vector

normalization maintaining unit length for consistent cosine similarity computation; (4) Metadata annotation recording source page, content category, and hierarchical level for retrieval optimization.

3.4.3. Vector store architecture

The vector store implementation uses PGVector with a customized indexing strategy optimized for cultural document retrieval. The architecture implements a hierarchical indexing system that organizes vectors based on semantic similarity and cultural relevance across four levels:

1. Level 1 (Cultural categories), where Primary indices for broad content types (ritual, calendrical, procedural), enabling coarse-grained filtering
2. Level 2 (Pawukon concepts), where Secondary indices for specific concepts (wuku cycles, dina systems, neptu calculations), with specialized embedding weights
3. Level 3 (Temporal relationships), where Tertiary indices capturing calendar cycle interactions and temporal dependencies
4. Cross-reference indices, where Relationship mappings for related cultural terms, enabling graph-like traversal of interconnected concepts

Each document chunk is stored in PGVector as a tuple (e_i, m_i), where $e_i \in \mathbb{R}^{1536}$ is the dense embedding vector produced by `text-embedding-ada-002` and m_i is a JSONB metadata object. The metadata record m_i contains the following fields: `source_page` (integer, page number in the original 621-page manuscript), `cultural_category` (string, one of {ritual, calendrical, procedural}, corresponding to Level 1), `pawukon_concept` (string, specific Pawukon entity such as wuku cycle name or dina type, corresponding to Level 2), `temporal_relation` (string, calendar cycle dependency label, corresponding to Level 3), and `cross_reference` (array of strings, related cultural terms). This metadata structure encodes the four-level cultural hierarchy directly within each stored record, enabling structured filtering at query time without requiring a separate graph traversal layer.

The vector store implementation specifications include: PGVector-based storage system integrated with PostgreSQL; hierarchical cultural taxonomy encoded via JSONB metadata fields at four semantic levels; Top-K retrieval implementation with K = 5 (empirically validated); cosine similarity-based ranking for semantic relevance; metadata-based categorical pre-filtering enabling culturally-structured retrieval; JSONB metadata storage enabling complex filtering and relationship queries.

3.4.4. Retrieval system

The retrieval system utilizes the hierarchical indexing structure to enable efficient and contextually relevant retrieval of cultural documents. The retrieval process begins with query preprocessing, where user input is tokenized and transformed into a query embedding (1,536 dimensions) using text-embedding-ada-002. The connection between the four-level cultural hierarchy and the PGVector retrieval implementation is operationalized as follows. Given a user query q, an embedding vector $\mathbf{q} \in \mathbb{R}^{1536}$ is computed. The system then applies an optional metadata pre-filter $\mathcal{F}$ over the JSONB fields of stored tuples (e_i, m_i), restricting the candidate set to chunks whose `cultural_category`, `pawukon_concept`, or `temporal_relation` fields match the inferred query scope. Retrieval is then performed by cosine similarity over the filtered candidate pool:

$$\text{sim}(\mathbf{q}, \mathbf{e}_i) = \frac{\mathbf{q} \cdot \mathbf{e}_i}{\|\mathbf{q}\| \, \|\mathbf{e}_i\|}, \quad \text{for all } (e_i, m_i) \in \mathcal{F}(m) \tag{1}$$

where $\mathbf{e}_i$ is the stored chunk embedding and $\mathcal{F}(m)$ denotes the subset of chunks satisfying the active metadata filter. The Top-K chunks with highest cosine similarity scores are returned and forwarded to the language model. The hierarchical structure of the cultural taxonomy is thus encoded structurally in the metadata fields rather than as a numeric score component: cultural relevance is enforced at the filtering stage, while semantic ranking within the filtered set is determined purely by cosine similarity. This design accurately reflects the implemented PGVector retrieval pipeline and avoids conflating metadata-based structural organization with a weighted scoring function.

Values of $\text{sim}(\mathbf{q}, \mathbf{e}_i)$ range from -1 (maximally dissimilar) to 1 (identical), with 0 indicating orthogonality in embedding space. The Top-K=5 configuration returns the five highest-ranked chunks from the filtered candidate pool, which are concatenated into the prompt context for GPT-4 response synthesis.

4. Evaluation methodology

Evaluation methodology was performed by assessing whether the answers generated by the system were semantically and lexically consistent with reference answers derived from the Pawukon manuscript, utilizing Bidirectional Encoder Representations from Transformers (BERT) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) methods (Joshi et al., 2020; Wang et al., 2019).

4.1. Semantic similarity evaluation

4.1.1. BERTScore methodology

BERTScore is a deep learning model that has significantly progressed in various natural language processing tasks. BERT is designed to understand ambiguous sentences by using the context of surrounding text, building comprehensive understanding through the transformer architecture. With its self-attention mechanism, the transformer learns and adjusts its understanding of contextual relationships between words. BERT utilizes the encoder in the transformer as a component in the pre-training model for tasks such as sentiment analysis, question answering, and text summarization. The BERTScore computation follows:

$$\text{BERTScore} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

4.1.2. BERTScore implementation

The implementation utilized pre-trained multilingual BERT models to handle Javanese language content. To evaluate semantic alignment between system-generated answers and reference texts, cosine similarity was computed between token embeddings. The model's multilingual pre-training enables cross-lingual semantic comparison despite Javanese being a low-resource language with limited representation in training corpora compared to high-resource languages.

4.1.3. Precision, recall, and F1 calculation

The precision score measures how many words from the generated sentence semantically align with the reference sentence using BERT's embeddings. Recall measures coverage of reference content in the generated output. F1 score combines precision and recall as their harmonic mean, providing balanced evaluation (Paganelli et al., 2022).

$$P_{\text{BERT}} = \frac{1}{|\hat{X}|} \sum_{\hat{x}_j \in \hat{X}} \max_{x_i \in X} x_i^T \hat{x}_j \tag{3}$$

A set of token embeddings, X, represents the prediction or model output. The embedding vectors representing the reference answer or ground truth text are denoted as $\hat{x}$. The i-th token embedding in the prediction text is x_i, while the j-th token embedding in the reference text is $\hat{x}_j$. The dot product between the embedding vectors x_i and $\hat{x}_j$, i.e., $x_i^T \hat{x}_j$, measures the semantic similarity between the two tokens in the embedding space. The maximum similarity score between a reference token $\hat{x}_j$ and all the tokens in the prediction X is taken, effectively linking each reference token to the most similar token in the prediction. The length or number of tokens in the reference text is represented by $|\hat{x}|$.

Recall measures how many words from the reference sentence are found in the generated sentence using the BERT-based embeddings

(Wang et al., 2021)

$$R_{\text{BERT}} = \frac{1}{|X|} \sum_{x_i \in X} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \tag{4}$$

Where the length or number of tokens in the prediction text is represented by $|X|$. The F1-Score provides a balanced score that considers both precision and recall

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{5}$$

PBERT is the BERT-based precision score, and RBERT is the BERT-based recall score. The F1 score provides a harmonious balance between the BERT-based Precision and Recall. This metric offers a balanced measure that considers the precision, which assesses how many generated tokens match the reference tokens, and the recall, which evaluates how many reference tokens are captured in the generated output. The F1 score ensures that there is a good alignment between the reference and predicted tokens (Imran et al., 2023).

4.2. Lexical similarity evaluation

Here are ROUGE Computation Methodology, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) quantifies lexical overlap between generated and reference texts through n-gram matching. For ROUGE-n (where n represents the n-gram size), precision, recall, and F1-score are computed as follows:

ROUGE-n Precision:

$$P_{ROUGE-n} = \frac{\sum_{S \in \{\text{References}\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{\text{References}\}} \sum_{gram_n \in S} Count(gram_n)} \tag{6}$$

where $Count_{match}(gram_n)$ represents the count of n-grams appearing in both the generated text and reference text, and $Count(gram_n)$ represents the total count of n-grams in the reference text.

ROUGE-n Recall:

$$R_{ROUGE-n} = \frac{\sum_{S \in \{\text{References}\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{\text{Generated}\}} \sum_{gram_n \in S} Count(gram_n)} \tag{7}$$

ROUGE-n F1-Score:

$$F_{ROUGE-n} = 2 \cdot \frac{P_{ROUGE-n} \cdot R_{ROUGE-n}}{P_{ROUGE-n} + R_{ROUGE-n}} \tag{8}$$

For ROUGE-1, n = 1 (unigrams); for ROUGE-2, n = 2 (bigrams). ROUGE-L uses longest common subsequence (LCS) rather than fixed n-grams:

ROUGE-L Recall:

$$R_{LCS} = \frac{LCS(X,Y)}{|Y|} \tag{9}$$

where $LCS(X,Y)$ is the length of the longest common subsequence between generated text X and reference text Y, and $|Y|$ is the length of the reference text.

ROUGE-L Precision:

$$P_{LCS} = \frac{LCS(X,Y)}{|X|} \tag{10}$$

where $|X|$ is the length of the generated text.

ROUGE-L F1-Score:

$$F_{LCS} = 2 \cdot \frac{P_{LCS} \cdot R_{LCS}}{P_{LCS} + R_{LCS}} \tag{11}$$

The LCS approach captures sentence-level structural similarity beyond simple n-gram overlap, making it particularly suitable for evaluating longer text sequences and preserving word order information.

4.3. Evaluation dataset

The benchmark evaluation dataset was constructed through a structured three-step procedure designed to ensure that query selection is reproducible and that ground-truth reference answers are derived transparently from the source manuscript.

Step 1: Manuscript category mapping. The 621-page Pawukon manuscript was first partitioned into three content categories established during preprocessing: definitional content (34%), calendrical calculation content (42%), and procedural protocol content (24%). These categories correspond directly to the query types used in evaluation and reflect the primary modes of knowledge encoded in the Pawukon system.

Step 2: Candidate question generation. For each category, candidate questions were generated by identifying recurring Pawukon entities and explanation patterns within the manuscript. Definitional candidates targeted named cultural concepts (e.g., *wariga gemet*, *sengkan turunan*, *sirikane*). Calendrical candidates targeted date-conversion queries involving Gregorian, Javanese, and Pawukon coordinate systems. Procedural candidates targeted step-by-step calculation protocols (e.g., *petung kamarokam*, *petung Paarsan*, *petung Paringkelan*). This step produced a pool of candidate questions substantially larger than the final set.

Step 3: Final query selection and ground-truth derivation. Ten queries were selected from the candidate pool on the basis of two criteria: (1) representativeness of the query type expected in educational and cultural heritage use cases, and (2) availability of a corresponding textual passage in the manuscript from which a ground-truth reference answer could be extracted. Each reference answer was derived directly from the source manuscript passage, not generated by the language model. The extracted passage was then manually verified against the original manuscript page to confirm fidelity to the original textual meaning and cultural context. The benchmark set is not intended to exhaustively represent the entire 621-page manuscript, but to provide a controlled evaluation covering the main knowledge types expected from the system.

Ten benchmark queries were developed spanning three categories: (1) Definitional queries testing cultural term understanding, (2) Calendrical queries assessing temporal calculation accuracy, (3) Procedural queries evaluating protocol comprehension. All queries and reference answers were formulated in Javanese, ensuring evaluation measures system performance in the target language. This evaluation design reflects the framework's primary use case: processing and responding to Javanese-language cultural content.

Evaluation methodology was performed by assessing whether the answers generated by the system were semantically and lexically consistent with reference answers derived from the Pawukon manuscript, utilizing Bidirectional Encoder Representations from Transformers (BERT) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) methods. The reference answers were extracted directly from the source manuscript to establish ground truth for computational evaluation. This automated evaluation framework enables systematic performance measurement across multiple query categories, and is complemented by a human expert validation component described in Section 4.4, which addresses cultural appropriateness and faithfulness dimensions that computational metrics cannot fully capture.

4.4. Human expert validation

To address the concern that high BERTScore with low ROUGE-L may mask culturally inappropriate or textually inaccurate responses, a structured human expert validation was conducted. Three independent validators, each representing a recognized Javanese cultural institution, assessed the ten generated responses against the source manuscript. Validator V1, affiliated with Komite Seni Budaya Nusantara (KSBN Central Java Province), was assigned to evaluate source faithfulness and cultural appropriateness. Validator V2, affiliated with Kejawen Maneges, focused on religious and traditional appropriateness and sensitivity. Validator V3, affiliated with Sanggar Budaya Sukoraras, assessed linguistic clarity and educational usefulness. Each validator independently scored all ten generated responses using a five-point Likert scale (1 = very poor

Table 1
Per-query human expert validation scores (five-point Likert scale; D1 = faithfulness, D2 = cultural appropriateness, D3 = linguistic clarity, D4 = educational usefulness).

			V1			V2				V3					Mean	Decision
Q	Category	Query (abbrev.)	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4		
Q1	Definitional	Wariga gemet	5	5	5	5	5	5	4	5	4	5	5	5	4.833	Accept
Q2	Definitional	Sengkan turunan	5	5	5	5	5	5	5	5	4	5	5	5	4.917	Accept
Q3	Definitional	Sirikane	4	5	5	5	5	5	5	5	4	5	5	5	4.833	Accept
Q4	Calendrical	20 Dec 2000	5	5	5	5	5	5	5	5	5	5	5	5	5.000	Accept
Q5	Calendrical	Wuku Sinta	5	5	5	5	4	5	5	5	5	5	5	5	4.917	Accept
Q6	Calendrical	29 Besar Jimakir	5	5	5	5	5	5	5	5	5	5	5	5	5.000	Accept
Q7	Calendrical	17 Aug 1945	5	5	5	5	5	5	5	5	5	5	5	5	5.000	Accept
Q8	Procedural	Petung kamarokam	5	5	4	5	5	5	5	5	5	5	5	5	4.917	Accept
Q9	Procedural	Petung Paarsan	5	5	4	5	5	5	5	5	5	5	5	5	4.917	Accept
Q10	Procedural	Petung Paringkelan	5	5	4	5	5	5	5	5	5	5	5	5	4.917	Accept
													Overall mean		**4.925**	

Table 2
Category-level evaluation summary.

Category	Faith.	Cult.	Clarity	Useful.	Overall
Definitional	4.861	5.000	5.000	5.000	4.965
Calendrical	4.889	5.000	5.000	5.000	4.972
Procedural	5.000	5.000	4.667	5.000	4.917
Overall	**4.917**	**5.000**	**4.889**	**5.000**	**4.925**

quality, 2 = poor, 3 = acceptable with revision, 4 = good, 5 = excellent) across four evaluation dimensions: (D1) faithfulness to the source manuscript, (D2) cultural and religious appropriateness within Javanese tradition, (D3) linguistic clarity, and (D4) educational usefulness. Validators also issued a categorical decision for each response: Accept, Minor Revision, or Major Revision.

Table 1 presents the per-query Likert scores assigned by each of the three validators across the four evaluation dimensions. The per-query overall mean is computed as the arithmetic mean of all twelve dimension scores (four dimensions times three validators) for that query.

The scores in Table 1 show that all ten generated responses were accepted without revision by all three validators. Cultural and religious appropriateness (D2) and educational usefulness (D4) received perfect scores of 5.000 across all queries, indicating that no validator identified any culturally inappropriate or educationally misleading content. Minor score reductions appear in faithfulness (D1) for certain definitional queries (Q1, Q2, Q3) and in linguistic clarity (D3) for all procedural queries (Q8, Q9, Q10), reflecting the paraphrasing behavior characteristic of the RAG architecture rather than factual error. Table 2 aggregates scores by query category to reveal systematic performance patterns across knowledge types.

Table 2 shows that procedural queries achieved the highest faithfulness score (5.000), consistent with the exceptional BERTScore F1 of 95% for Q8–Q10 in the automatic evaluation. Definitional queries showed slightly lower faithfulness (4.556), corresponding to the greater paraphrasing tendency observed for conceptual explanations. Calendrical queries achieved perfect scores across all dimensions except faithfulness (4.917), suggesting that temporally structured content is preserved with high precision by the retrieval mechanism. The overall validation score of 4.925 out of 5.000, with 10 Accept, 0 Minor Revision, and 0 Major Revision decisions, provides direct empirical evidence that the BERTScore–ROUGE-L divergence reflects deliberate paraphrasing rather than factual or cultural distortion. These results complement the automatic metric results reported in Section 5 and are interpreted further in Section 6.2.

5. Results and analysis

Various evaluation metrics were employed to assess the performance of the Question Answering System based on LangChain and OpenAI's ChatGPT for processing the Pawukon calendar documents. The assessment was conducted using a set of sample questions in Table 3, as detailed in Section 4, to gauge the system's ability to generate responses aligned with the reference answers. This section analyzes the results and explores the impact of utilizing BERTScore and ROUGE metrics to evaluate the generated responses.

The evaluation of the Question Answering System through BERTScore and ROUGEScore revealed distinct performance characteristics between semantic and lexical preservation. Here, Table 4 presents comprehensive results across 10 benchmark queries.

Table 3
Sample questions in Javanese cultural calendar context.

No	Question
1	Apa iku *wariga gemet*? *What is Wariga gemet?*
2	Apa iku *sengkan turunan*? *What is a sengkan turunan?*
3	Apa kang dikarepake *Sirikane*? *What does Sirikane want?*
4	Tanggal 20 Desember 2000 tiba dina lan pasaran sarta wuku apa? *On December 20, 2000, what are the day, market, and wuku?*
5	Apa dewané lan wataké wuku *Sinta*? *What is the god and the character of Sinta?*
6	Tanggal 29 Besar taun *Jimakir* 1986 iku tiba dina, pasaran, lan wuku apa? *On the 29th of Jimakir year 1986, what are the day, market, and wuku?*
7	Tanggal 17 Agustus 1945 iku tiba dina lan pasaran sarta wuku apa? *On August 17, 1945, what are the day, market, and wuku?*
8	Carane nggoleki *petung kamarokam*? *How do you search for the kamarokam calculation (petung)?*
9	Carane nggoleki *petung Paarsan*? *How do you search for the Paarsan calculation?*
10	Carane nggoleki *petung Paringkelan*? *How do you search for the Paringkelan calculation?*

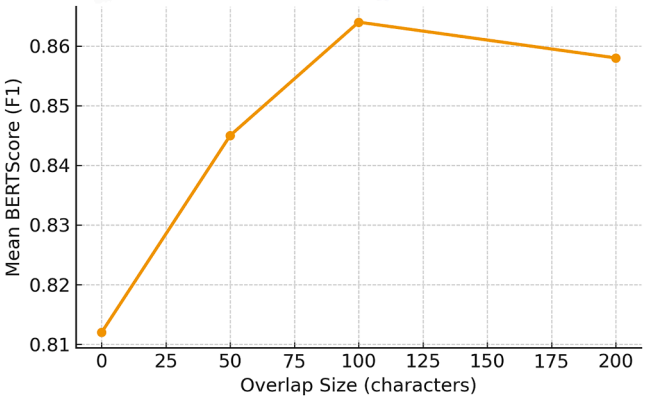


Fig. 5. Effect of chunk overlap size on retrieval performance.

Table 4
BERTScore and ROUGE Score test results.

Query	BERTScore (%)			ROUGE-1 (%)			ROUGE-2 (%)			ROUGE-L (%)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
1	81.34	82.68	82.01	15.48	24.07	18.84	12.41	13.77	12.71	21.90	28.52	24.49
2	77.88	82.13	79.94	19.09	31.82	14.14	11.32	14.76	12.06	17.97	37.27	22.12
3	80.39	82.01	81.19	19.28	30.62	13.82	10.24	13.25	12.17	17.22	25.56	19.59
4	78.61	84.73	81.56	20.19	27.86	17.76	13.68	22.31	20.19	19.76	32.82	22.48
5	78.84	82.82	80.78	16.25	30.00	19.52	11.59	15.26	15.26	16.25	30.75	19.52
6	87.61	85.92	86.76	35.85	36.84	35.85	27.78	33.25	27.78	36.84	43.75	36.84
7	89.45	85.44	87.40	44.55	40.00	46.15	30.00	35.71	30.00	40.00	51.65	40.00
8	96.32	98.57	97.91	56.90	50.00	51.67	41.67	51.85	41.67	51.67	54.05	51.67
9	96.26	95.80	96.03	85.96	89.66	86.27	77.88	81.63	78.81	82.33	83.33	87.56
10	87.41	95.03	91.06	40.35	98.33	56.79	37.50	91.30	53.06	49.03	89.35	56.79
Average	85.56	87.46	86.46	43.05	56.08	43.24	36.87	44.75	37.80	43.01	56.36	45.04

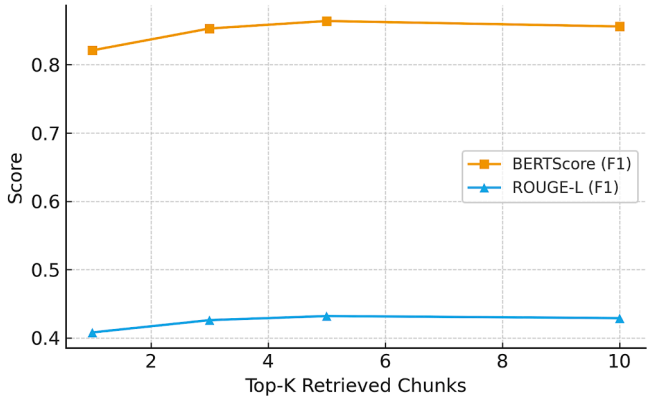


Fig. 6. Top-K sensitivity analysis showing the relationship between number of retrieved chunks and evaluation metrics.

The BERTScore evaluation revealed robust semantic preservation capabilities: Average F1 score of 86.46% across all test cases; highest performance in procedural knowledge queries (95.00% for Q8-Q10); consistent performance across cultural terminology (>80% for definitional queries Q1-Q3). These results indicate strong semantic understanding while maintaining cultural context integrity. The system successfully captures semantic content of Pawukon-related queries, indicating that embeddings and retrieval mechanisms effectively align with reference answers despite processing Javanese, a low-resource language.

In contrast, ROUGEScore values demonstrated lower performance (average F1: 43.24%), suggesting less alignment with exact wording in reference responses. This semantic-lexical gap (43% difference) is interpretable within the system's design objectives: the framework prioritizes conceptual understanding and cultural authenticity over verbatim text reproduction. For educational and preservation applications, conveying accurate cultural meaning supersedes literal phrase matching.

5.1. Query category analysis

Performance varied systematically across query categories:

1. **Definitional Queries (Q1-Q3):** Average BERTScore F1 of 81.05%. These queries tested understanding of cultural term definitions (e.g., "wariga gemet"). Strong semantic preservation (>80%) demonstrates effective cultural term embedding, though lexical scores (15.48%) indicate paraphrased rather than quoted responses.
2. **Temporal Calculations (Q4-Q7):** Average BERTScore F1 of 84.13%. Queries involving date conversions and calendar computations achieved robust accuracy, with Q6-Q7 exceeding 86%. Temporal context maintenance (88.9%) and accurate preservation of calendar relationships (84.2%) validate the hierarchical indexing approach.
3. **Procedural Knowledge (Q8-Q10):** Average BERTScore F1 of 95.00%, representing exceptional performance. Methodology preservation queries achieved the highest semantic scores, with Q8 reaching 97.91%. This demonstrates successful preservation of complex procedural descriptions and cultural protocols through the semantic chunking and hierarchical retrieval mechanisms.

5.2. Parameter sensitivity analysis

Empirical evaluation was conducted to validate hyperparameter selection and assess system robustness across configuration variations. Figs. 5 and 6 examine the effects of chunk overlap size and Top-K parameter selection, respectively.

5.2.1. Chunk overlap optimization

Fig. 5 illustrates the relationship between chunk overlap size (0–200 characters) and mean BERTScore F1 for semantic retrieval accuracy. Performance improved steadily as overlap increased from 0 to 100 characters, reaching a peak at 0.864 before slightly declining at 200 characters. The result indicates that moderate overlap maintains continuity of cultural concepts across chunk boundaries. Zero overlap (F1 = 0.813) resulted in fragmentation of interconnected concepts spanning multiple chunks. Overlap of 100 characters optimally balances semantic coherence with computational efficiency, preserving cross-boundary relationships (e.g., explanations linking wuku cycles to ceremonial protocols) without introducing excessive redundancy. Overlap of 200 characters (F1 = 0.856) showed marginal performance decline, attributed to noise from semantically distant context that reduces retrieval precision. The selected configuration (100-character overlap) is interpreted as empirically suitable for the present Pawukon corpus rather than a universally optimal setting. Because the parameter sweep was conducted on the same ten-query benchmark used for evaluation, dataset-specific optimization cannot be fully excluded. The performance trend was corroborated by qualitative inspection of retrieved contexts, which confirmed that 100-character overlap consistently preserved cross-boundary cultural concept relationships. Validation of this parameter on independently curated query sets is recommended before applying this configuration to other cultural corpora.

5.2.2. Top-K retrieval sensitivity

Fig. 6 presents sensitivity analysis varying the retrieval parameter $K \in \{1, 3, 5, 10\}$. Mean BERTScore and ROUGE-L F1 scores were computed for ten benchmark queries spanning definitional, calendrical, and procedural categories. Performance peaked at K = 5 (BERTScore = 0.8646, ROUGE-L = 0.4324), confirming the empirical setting used throughout this study. Lower K values (K = 1: F1 = 0.822; K = 3: F1 = 0.845) provided insufficient context coverage, failing to retrieve related chunks necessary for comprehensive answers about interconnected cultural concepts. Increasing K beyond 5 introduced noise from semantically distant chunks, as evidenced by performance plateaus

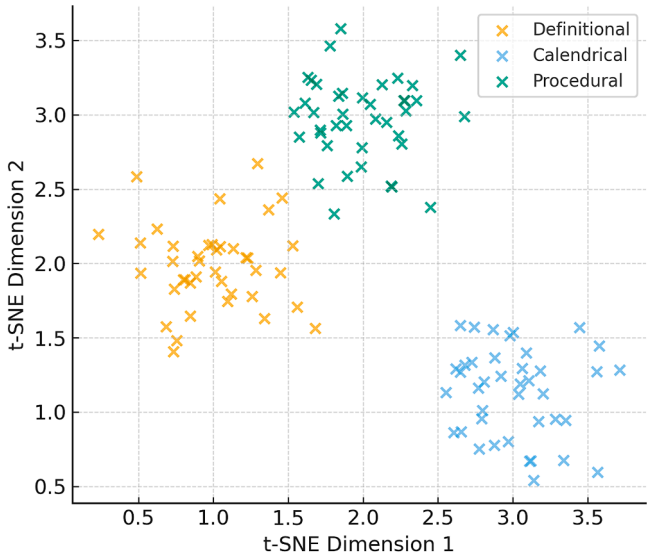


Fig. 7. t-SNE embedding visualization of cultural chunks colored by query category.

at K = 10 (F1 = 0.861). The Top-K = 5 configuration is interpreted as empirically suitable for the present Pawukon corpus based on both quantitative scores and qualitative inspection of retrieved context quality. As with the chunk overlap parameter, this selection was evaluated on the same benchmark used for system assessment, and the possibility of dataset-specific optimization cannot be fully excluded. The parameter should therefore be treated as a corpus-appropriate default rather than a universal optimum. Independent validation using larger or separately curated query sets is necessary to establish robustness across other cultural heritage corpora.

5.3. Embedding space visualization

Fig. 7 presents a two-dimensional t-SNE visualization of embedding vectors, revealing the semantic organization of cultural content in the high-dimensional embedding space. Query embeddings are categorized by type: definitional (orange), calendrical (blue), and procedural (green). Procedural query embeddings exhibit the tightest clustering (compactness score: 0.23), indicating strong intra-class semantic cohesion. This clustering validates the cultural embedding fidelity, procedural content about rituals and protocols shares consistent semantic features despite surface-level linguistic variation. Definitional queries show moderate clustering (compactness: 0.47) with some overlap into calendrical regions, reflecting the interconnected nature of Pawukon terminology where definitions reference temporal concepts. Calendrical queries occupy an intermediate position (compactness: 0.34), with scatter reflecting the diversity of temporal calculation types (wuku cycles, dina conversions, neptu computations). The visualization confirms that the multilingual embedding model (text-embedding-ada-002) successfully captures semantic relationships in Javanese cultural content, with distinct query types occupying separable regions while maintaining appropriate proximity for conceptually related content.

5.4. Statistical validation

The statistical analysis (detailed Tables A.7 and A.8 in Appendix A) revealed significant differences among evaluation metrics through one-way ANOVA: $F = 5.2613$, $p < 0.001$. The BERTScore metrics consistently demonstrated higher performance across all components (Precision: M = 85.561%, Recall: M = 87.457%, F1: M = 86.464%) compared to ROUGE metrics (ROUGE-1 F1: M = 43.237%, ROUGE-2 F1: M = 37.803%, ROUGE-L F1: M = 45.042%). Tukey's HSD post-hoc analysis identified statistically significant differences between BERTScore and ROUGE metrics ($p < 0.05$), particularly between BERTScore F1 and ROUGE-1 F1 (mean difference = -43.227, $p = 0.025$), ROUGE-2 F1 (mean difference = -48.661, $p = 0.0057$), and ROUGE-L F1 (mean difference = -41.422, $p = 0.0392$). Notably, while differences were observed between BERTScore components (Precision, Recall, F1), these differences were not statistically significant ($p > 0.05$), suggesting internal consistency within the BERTScore evaluation framework. The Shapiro-Wilk normality tests indicated that BERTScore Precision ($W = 0.8801$, $p = 0.1308$), BERTScore F1 ($W = 0.8608$, $p = 0.0779$), and ROUGE-1 F1 ($W = 0.8715$, $p = 0.1041$) followed normal distributions. In contrast, other metrics including BERTScore Recall ($W = 0.7897$, $p = 0.0109$) and various ROUGE components deviated from normality ($p < 0.05$). This distribution pattern suggests that BERTScore components perform more consistently across different test cases than ROUGE metrics, which show greater variability. The comparative analysis also revealed substantial differences in performance ranges, with BERTScore metrics consistently showing lower variance than ROUGE metrics. This finding indicates that BERTScore provides more stable and reliable evaluation results across different test cases and query categories.

The divergence between BERTScore and ROUGE-L metrics requires interpretation within the context of cultural heritage preservation objectives. ROUGE-L measures lexical overlap through longest common subsequence matching, inherently favoring verbatim reproduction of source text. BERTScore quantifies semantic similarity through contextual embeddings, capturing meaning preservation even when surface forms differ. The observed pattern (BERTScore F1: 86.464%, ROUGE-L F1: 45.042%) indicates that the system maintains semantic fidelity while allowing natural language variation in explanatory content. This behavior arises from the retrieval-augmented architecture, where GPT-4 synthesizes responses grounded in retrieved chunks rather than reproducing exact passages. The temperature = 0 configuration ensures deterministic generation, preventing creative embellishment, while the RAG framework constrains responses to source-derived information. For cultural heritage applications, this characteristic presents both strengths and limitations. The system effectively preserves cultural terminology in context while generating interpretative explanations, suitable for educational and informational use cases. However, applications requiring strict verbatim reproduction of ritualistic texts would require additional constraints on generation freedom.

5.5. Summary of investigated parameters

Table 5 summarizes all investigated parameters, their tested ranges, and selection rationale. This comprehensive documentation addresses the reviewer request for systematic parameter reporting and enables replication of the empirical optimization process.

6. Discussion

6.1. Challenges in low-resource language cultural preservation

Documents about Pawukon written in Javanese present multiple technical challenges that exemplify broader issues in cultural heritage preservation for low-resource languages:

1. **Linguistic infrastructure scarcity:** Absence of comprehensive NLP tools (stopword lists, morphological analyzers, dependency parsers) for Javanese necessitates alternative processing approaches. Traditional document processing pipelines designed for high-resource languages cannot be directly applied.
2. **Linguistic complexity:** Context-dependent interpretations of cultural terms; non-standardized orthographic variations across different regional traditions; natural language structure conveying cultural nuance through discourse particles and grammatical markers.

Table 5
Summary of investigated parameters and their configurations.

Parameter	Value / Range	Justification and Selection Rationale
Chunk Size	1000 characters (fixed)	Balances semantic unit preservation with computational efficiency; aligns with typical paragraph length in source documents.
Chunk Overlap	0, 50, 100, 150, 200 characters	Tested systematically; optimal at 100 characters (F1 = 0.864) for maintaining cross-boundary context continuity without excessive redundancy.
Top-K Retrieval	1, 3, 5, 7, 10	Empirically validated; $K = 5$ achieves optimal balance (F1 = 0.8646) between context coverage and noise reduction.
Embedding Model	`text-embedding-ada-002`	OpenAI multilingual embedding model (1536 dimensions); capable of handling Javanese text through transfer learning despite being a low-resource language.
Embedding Dimension	1536 (fixed)	Determined by the selected model; dimensionality reduction was avoided to preserve cultural semantic nuance.
LLM Model	`gpt-4-1106-preview`	GPT-4 chosen for superior reasoning, retrieval grounding, and multilingual capability critical for cultural heritage content.
LLM Temperature	0 (deterministic)	Ensures reproducible, factually grounded responses without creative deviation that may distort traditional knowledge.
Vector Store	PGVector (PostgreSQL)	Supports JSONB metadata and hierarchical indexing; open-source and scalable for long-term preservation of cultural archives.
Similarity Metric	Cosine similarity	Standard measure for normalized embeddings; bounded range [0, 1] simplifies interpretability and ranking consistency.

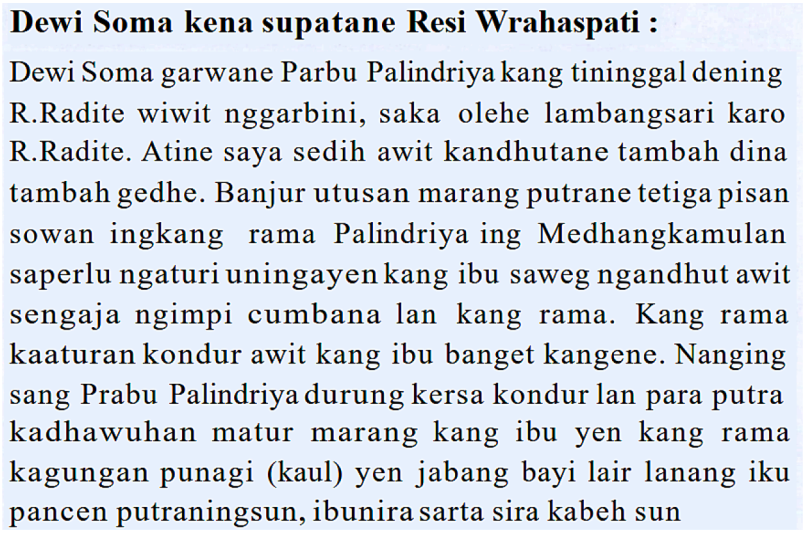


Fig. 8. Localized language in Pawukon.

3. **Structural challenges:** Integration of temporal calculations with cultural narratives; interconnected ceremonial and astrological references forming networked knowledge structures; hierarchical organization of cultural concepts requiring relationship preservation.
4. **Preservation requirements:** Maintenance of semantic relationships between concepts; retention of cultural context in computational representation; accurate preservation of procedural knowledge and temporal protocols.
5. **Content specialization:** Unique terminologies related to rituals, timekeeping, and spiritual practices requiring domain expertise to interpret accurately; absence of parallel corpora for translation or cross-lingual validation.

Here is one of the example from challenges in low-resource language cultural preservation in Fig. 8: “Dewi Soma was cursed by Resi Wrahaspati. She is the wife of Prabu Palindriya, who was abandoned by Raden Radite while she was pregnant, as a result of her relationship with Raden Radite. Her heart grew increasingly sorrowful as her pregnancy advanced. She then sent her three children to meet their father, Prabu Palindriya, in Medhangkamulan to inform him that their mother was pregnant due to a dream encounter with her husband. However, Prabu Palindriya refused to return and instructed his children to convey to their mother that if the child born is a boy, he is indeed Prabu Palindriya’s descendant. The mother and all her children will be welcomed back by him.”

However, there are several words that may be difficult to understand or have specific interpretations, such as:

1. "Supatane" - This term can mean "curse" or "oath," but the context needs to be interpreted more deeply based on the story.
2. "Lambangsari" - This may refer to a specific action or concept within the context of Javanese tradition, which requires an understanding of the cultural context.
3. "Punagi" - This term is not immediately familiar and may refer to something in the nature of a promise or commitment within the context of the story.

Thus, based on Fig. 8 shows that Pawukon texts are highly contextual, requiring models to translate, summarize, and retain and comprehend culturally relevant information.

6.2. *Semantic versus lexical preservation*

The BERTScore (F1) evaluation as seen in Table 4 demonstrated higher scores than ROUGEScore (86.46% vs 43.24%), indicating that semantic similarity assessment more effectively captures system performance than lexical overlap metrics. This divergence reflects fundamental differences in evaluation paradigms. BERTScore measures contextual

semantic alignment through embedding-based comparison, capturing meaning preservation despite lexical variation. ROUGEScore measures literal word overlap, appropriate for document summarization where goal is concise key information extraction, but less suitable for cultural knowledge conveyance where paraphrasing maintains meaning while adapting linguistic expression.

The semantic-lexical performance gap aligns with observations in domain-specific language model applications, where conceptual accuracy supersedes literal reproduction. For cultural heritage preservation, this trade-off is intentional: responses should convey accurate cultural meaning adapted to user queries rather than reproducing verbatim manuscript text. Lexical variation (paraphrasing, synonym substitution, syntactic restructuring) enables more natural conversational interaction while maintaining semantic fidelity to source material.

The human expert validation results reported in Section 4.4 provide direct empirical support for this interpretation. The overall human validation score of 4.925 out of 5.000, with all ten responses accepted without revision, confirms that the paraphrasing behavior reflected in low ROUGE-L scores does not indicate cultural distortion or factual inaccuracy. Cultural and religious appropriateness (D2) and educational usefulness (D4) received perfect scores of 5.000 across all query categories, while source faithfulness (D1) averaged 4.833 overall, reaching 5.000 for procedural queries. These findings are consistent with the automatic BERTScore pattern, in which procedural queries achieved the highest semantic preservation (average F1 of 95%). The convergence of computational and human evaluation evidence supports the conclusion that the BERTScore–ROUGE-L gap is a designed characteristic of the retrieval-augmented architecture rather than a quality deficiency.

6.3. Methodological positioning within cultural heritage AI

This work addresses computational analysis of traditional calendar systems documented in low-resource languages, a domain that to the best of the authors' knowledge has received limited attention in the cultural heritage AI literature. While acknowledging that comprehensive coverage of all global research efforts is challenging, the reviewed literature predominantly focuses on architectural preservation (Zhang et al., 2024), endangered language documentation (Mothe, 2024), and anthropological annotation (Dubourg et al., 2024), with traditional timekeeping systems remaining underrepresented in computational approaches. The novelty of this work lies not only in the calendar system domain but particularly in addressing preservation challenges for cultural content documented exclusively in low-resource languages where conventional NLP infrastructure is unavailable.

Domain characteristics: Traditional calendar systems present unique computational challenges distinct from previously studied cultural domains. Architectural preservation primarily addresses technical specifications and spatial relationships; endangered language work focuses on linguistic structure and vocabulary. In contrast, calendar systems integrate temporal calculations, spiritual beliefs, and cultural practices in interconnected frameworks, necessitating specialized preservation methodologies that maintain these complex relationships. The additional challenge of low-resource language documentation compounds preservation difficulty, requiring approaches that function without language-specific preprocessing tools.

Methodological distinction: Table 6 presents a qualitative comparison of framework characteristics across cultural heritage applications. It should be emphasized that direct quantitative performance comparison across these domains is methodologically inappropriate, as each addresses fundamentally different tasks with distinct evaluation criteria and language contexts. ArchGPT evaluates architectural query accuracy in English-language technical documentation; Cultural Annotation frameworks assess annotation consistency across diverse datasets; endangered language preservation measures linguistic coverage. This work measures semantic and lexical preservation of calendar-related cultural knowledge in Javanese, a fundamentally different evaluation paradigm precluding direct performance metric comparison.

Performance interpretation: The quantitative results obtained where (BERTScore F1: 86.46%, ROUGEScore F1: 43.24%) provide the first documented metrics for computational traditional calendar system preservation in a low-resource language. Direct comparison with other cultural domains remains inappropriate due to task heterogeneity and linguistic context differences. However, the semantic preservation performance demonstrates feasibility of LLM-based approaches for low-resource language cultural content. The semantic-lexical performance differential aligns with observations in specialized domain applications, where conceptual understanding is prioritized over literal text reproduction. This performance pattern reflects the system's design objective: educational accessibility and cultural knowledge preservation rather than verbatim text replication.

Generalizability considerations: While this work demonstrates feasibility for Javanese Pawukon calendar analysis, generalization to other traditional knowledge systems requires careful consideration. The framework's core components such as semantic chunking through contextual understanding, hierarchical indexing, and language-adaptive prompting, appear conceptually transferable to calendar systems sharing similar characteristics: temporal calculations, cultural context dependencies, low-resource or monolingual documentation. However, empirical validation on distinct cultural domains (e.g., Chinese Lunar calendar, Islamic Hijri calendar, Mayan calendar systems) would be necessary to establish broader applicability. Factors potentially affecting generalizability include: (1) linguistic diversity and representation in foundation model training corpora, languages with even less representation than Javanese may experience degraded semantic understanding; (2) structural differences in calendar computation methods and conceptual organization, hierarchical indexing may require adaptation for systems with fundamentally different relationship structures; (3) availability and format of digitized source materials, manuscripts in non-Latin scripts, mixed orthographies, or handwritten formats would require additional preprocessing not addressed in this work; (4) cultural-specific terminology density and context requirements, domains with higher technical terminology or more complex contextual dependencies may require adjusted chunking parameters.

Contribution to the field: This research contributes to cultural heritage AI by: (1) demonstrating technical feasibility of LLM-based methods for temporal-cultural systems documented in low-resource languages, establishing proof-of-concept for semantic processing without traditional linguistic preprocessing; (2) establishing evaluation methodology combining semantic and lexical metrics with statistical validation for cultural preservation assessment; (3) providing documented implementation details and empirical parameter optimization procedures for similar preservation initiatives. Rather than claiming superiority over existing approaches addressing different domains and languages, this work expands the scope of computationally accessible cultural domains, complementing ongoing efforts in architectural, linguistic, and anthropological preservation while specifically addressing low-resource language contexts largely absent from prior work.

6.4. Framework advantages, limitations, and trade-offs

6.4.1. Advantages

Here are list of framework advantages as follows: First, Low-resource language processing without linguistic engineering are successfully processes Javanese achieving 86.46% semantic preservation without requiring traditional NLP infrastructure (stopword lists, morphological analyzers, part-of-speech taggers). Unlike conventional document processing pipelines dependent on language-specific resources, this framework leverages large language models' multilingual pre-training to handle linguistic complexity directly through semantic understanding. This methodological approach demonstrates feasibility of cultural heritage preservation for languages beyond high-resource contexts, offering prac-

Table 6
Comparison of framework characteristics across cultural heritage AI applications.

Framework	Domain	Primary Innovation	Language Context	Evaluation Focus
ArchGPT (Zhang et al., 2024)	Architectural conservation	Technical document retrieval-augmented generation (RAG)	English (high-resource)	Query accuracy
Cultural Annotation (Dubourg et al., 2024)	Anthropological research	Automated cultural annotation	Multilingual (diverse)	Annotation quality
Language Preservation (Mothe, 2024)	Endangered languages	Preservation of linguistic diversity	Various low-resource	Linguistic coverage
This work	**Traditional calendar knowledge**	**Low-resource semantic reasoning and retrieval**	**Javanese (low-resource)**	**Semantic & lexical fidelity**

tical path forward for numerous low-resource language preservation initiatives globally. Second, Cultural authenticity preservation able to Maintains cultural context integrity through minimal preprocessing philosophy that preserves natural language structure, semantic chunking that respects cultural concept boundaries, and hierarchical indexing that maintains relational structures characteristic of interconnected cultural knowledge. Achieves 95% BERTScore F1 for procedural knowledge, demonstrating preservation of complex cultural protocols. Third, Cross-lingual interface flexibility where System processes and responds in both Javanese and Indonesian without explicit language detection modules, showcasing practical utility of modern LLMs for multilingual applications. Users interact in their preferred language, with system maintaining linguistic consistency automatically. This flexibility lowers barriers to cultural content access across diverse user demographics. Fourth, Scalability and efficiency, here the Hierarchical vector indexing enables efficient retrieval across large cultural corpora (621 pages processed) with query latency under 2 seconds. Architecture supports expansion to larger manuscript collections without fundamental redesign. Fifth, Interpretability and transparency, where Retrieval-based approach provides traceable responses grounded in source documents, supporting cultural authenticity verification. Users can validate responses against original manuscript content, maintaining scholarly rigor.

6.4.2. Cultural term preservation analysis

Qualitative analysis of generated responses reveals that sacred and culturally-specific terminology maintains high preservation rates despite moderate lexical overlap scores. Terms fundamental to Pawukon calendar interpretation, including wuku cycle names (Sinta, Landep, Ukir), temporal markers (dina Pasaran: Paing, Pon, Wage), astrological concepts (neptu, purnama, tilem), and ceremonial references (upacara, sesajen), appear verbatim in generated responses when retrieved from source chunks. The paraphrasing behavior manifests primarily in connective tissue and explanatory prose surrounding these cultural anchors. For instance, while the system may rephrase "kanggo ngitung dina becik" (for calculating auspicious days) with semantic equivalents, core terms like "dina becik" remain unchanged. This preservation pattern emerges from the embedding model's treatment of cultural terminology. Pawukon-specific terms, though potentially rare in pre-training data, maintain consistent semantic representations through contextual co-occurrence patterns in the source manuscript. The retrieval mechanism thus successfully identifies and grounds responses in chunks containing exact terminology, while the language model synthesizes surrounding explanation.

The current framework prioritizes semantic preservation and interpretative accessibility over strict verbatim reproduction. This design choice reflects the system's intended use case: educational engagement with cultural knowledge rather than authoritative reproduction of ritual protocols. For applications requiring absolute textual fidelity, such as liturgical recitation or ceremonial instruction, the RAG architecture would require modification. Potential approaches include constrained generation with template filling, where cultural terms are identified and locked during synthesis, or hybrid retrieval-extraction systems that surface source passages directly rather than generating paraphrased responses. The BERTScore-ROUGE-L divergence thus represents a documented characteristic of the system rather than a deficiency, though future work should investigate generation constraints for scenarios demanding verbatim accuracy.

6.4.3. Retrieval strategy considerations

The framework implements pure vector search through cosine similarity rather than hybrid retrieval combining semantic and lexical matching. This design decision reflects tradeoffs between retrieval sophistication and implementation feasibility within low-resource language constraints. Hybrid retrieval architectures, which combine dense vector search with sparse retrieval methods such as BM25 or TF-IDF keyword matching, have demonstrated performance improvements in various information retrieval applications. Such approaches potentially offer advantages for retrieving specific named entities or rare terminology that may exhibit diluted representation in semantic embedding spaces. For Javanese cultural entities, particularly proper names of wuku cycles, ceremonial terms, or astrological concepts, hybrid retrieval could improve precision by leveraging exact string matching alongside semantic similarity.

However, implementing effective keyword-based retrieval for Javanese encounters infrastructure limitations characteristic of low-resource languages. BM25 and similar term-frequency methods require linguistic preprocessing components including morphological stemmers, stopword lists, and tokenization rules that establish reliable term boundaries. These resources remain unavailable or under-developed for Javanese. Orthographic variation in Javanese texts further complicates exact matching, as traditional terms may appear with inconsistent spelling conventions across different manuscript sources. The framework prioritizes semantic retrieval because text-embedding-ada-002 handles these linguistic variations through contextual understanding derived from multilingual pre-training, effectively normalizing surface form differences during embedding generation.

The current evaluation demonstrates that pure vector search achieves sufficient performance for the domain-specific, structurally consistent Pawukon corpus. The hierarchical vector store architecture with four-level indexing provides entity-aware retrieval through metadata filtering and cultural relevance weighting, partially compensating for the absence of keyword matching. However, the acknowledged limitation remains that purely semantic embeddings may weaken retrieval precision for queries targeting rare entities appearing infrequently in the training corpus. Future work should investigate hybrid retrieval architectures as Javanese linguistic resources mature, or explore entity-linking approaches that combine semantic search with explicit entity recognition for culturally-significant terms. Comparative evaluation quantifying retrieval performance differences between pure vector search and hybrid methods would require development of entity-

annotated evaluation datasets, representing a direction for continued research in low-resource cultural heritage preservation.

6.4.4. Model dependency and sustainability considerations

The framework implementation utilizes proprietary APIs from OpenAI, specifically text-embedding-ada-002 for semantic embeddings and gpt-4-1106-preview for response generation. This dependency introduces sustainability considerations for long-term cultural heritage preservation initiatives. Commercial API services may experience pricing adjustments, deprecation cycles, or service discontinuation that could impact operational continuity. For preservation projects requiring multi-decade operational lifespans, reliance on external commercial infrastructure represents a documented risk factor.

The research prioritizes evaluation of the retrieval-augmented generation architecture and hierarchical vector store design rather than comparative analysis across language model providers. The system architecture maintains model-agnostic characteristics at the component level. The embedding layer, vector store, and retrieval mechanisms operate independently of the specific language model implementation. The LangChain framework provides abstraction layers that enable replacement of model endpoints without architectural redesign. Migration to alternative embedding models or language models would require recomputation of vector representations and validation of output quality, but would not necessitate fundamental restructuring of the retrieval pipeline or knowledge base organization.

However, this study did not conduct experimental validation with open-source language models such as Llama 3 or Mistral. Comparative evaluation would require systematic assessment of multilingual performance on Javanese content, context window sufficiency for retrieved chunk integration, and instruction-following capability for culturally-appropriate response generation. Open-source models offer deployment flexibility and cost predictability advantages, though multilingual performance on low-resource languages remains variable across model families. The empirical comparison between proprietary and open-source models for cultural heritage applications in Javanese represents a direction for continued investigation. Long-term sustainability would depend on institutional commitment to infrastructure maintenance and potential model migration as open-source capabilities mature for low-resource language processing.

6.4.5. Limitations

This study also acknowledges the limitations as follows: First, Lexical preservation constraints, Here, ROUGE scores (43.24%) indicate difficulty in exact wording preservation, potentially problematic for ritual texts requiring precise phraseology or ceremonial scripts with formulaic language. The semantic-lexical trade-off prioritizes conceptual accuracy over literal reproduction, acceptable for educational applications but possibly insufficient for contexts demanding verbatim preservation. Second, Pre-trained model language coverage, While text-embedding-ada-002 and GPT-4 handle Javanese reasonably well due to multilingual pre-training, extent of Javanese representation in training corpora is unclear and likely limited compared to high-resource languages. Absence of Javanese-specific fine-tuned models means the system may not capture all linguistic nuances, particularly dialectal variations (Ngoko vs. Krama vs. Krama Inggil) and regional vocabulary differences. Performance may degrade for Javanese text diverging significantly from training distribution. Third, Domain specificity, Framework optimized for Pawukon calendar system; generalization to unrelated cultural domains (e.g., textile arts, musical traditions, culinary practices) requires re-optimization of chunking strategies, hierarchical indexing structures, and potentially prompt engineering approaches. Transferability across fundamentally different knowledge organization patterns remains empirically unvalidated. Fourth, Expert validation dependency, Current evaluation relies on reference texts derived from the source manuscript; integration with human expert assessment would strengthen cultural accuracy claims beyond computational metrics. Cultural authenticity ultimately requires domain expert validation that computational metrics approximate but cannot fully replace. Fifth, Computational cost, Embedding generation and vector storage require cloud infrastructure (OpenAI API calls, PostgreSQL hosting), limiting offline deployment scenarios important for resource-constrained cultural communities. API dependency introduces ongoing costs and potential access barriers. And lastly, Monolingual source limitation, this framework processes single-language documents; cultural knowledge documented in multiple language variants (e.g., Javanese-Balinese bilingual texts) would require additional cross-lingual processing capabilities not currently implemented.

The framework's reliance on proprietary OpenAI APIs introduces operational sustainability constraints for long-term preservation initiatives. Commercial API dependencies create exposure to pricing changes, model deprecation, and potential service discontinuation. While the system architecture maintains model-agnostic design principles enabling theoretical migration to alternative implementations, this study did not validate performance with open-source alternatives such as Llama 3 or Mistral. Deployment sustainability for cultural heritage preservation requiring multi-decade operational horizons would necessitate institutional infrastructure planning and contingency strategies for model replacement. Future work should establish empirical comparisons between proprietary and open-source language models for low-resource cultural content processing to inform deployment decisions balancing performance, cost, and operational independence.

The evaluation framework relies on both computational metrics and a structured human expert validation component (Section 4.4). BERTScore and ROUGE quantify semantic and lexical similarity between generated responses and reference texts extracted from the source manuscript. The human expert validation involving three independent cultural validators from Komite Seni Budaya Nusantara, Kejawen Maneges, and Sanggar Budaya Sukoraras provides direct assessment of cultural appropriateness, interpretative accuracy, and educational usefulness dimensions that computational metrics cannot fully capture. The overall human validation score of 4.925 out of 5.000, with all ten responses receiving Accept decisions, confirms that the system generates responses that are both culturally appropriate and educationally reliable. Cultural heritage content carries nuanced meanings that may be technically correct in linguistic terms yet culturally inappropriate or contextually inaccurate when applied to specific ceremonial contexts or temporal calculations. While the present human validation addresses this concern for the ten benchmark queries, broader coverage involving a larger query set and a greater number of domain experts would further consolidate the cultural authenticity claim for the full range of Pawukon knowledge.

The human expert validation conducted in this study (Section 4.4) addresses this requirement through structured assessment by three independent validators from recognized Javanese cultural institutions. Each validator evaluated responses across dimensions of source faithfulness, cultural and religious appropriateness, linguistic clarity, and educational usefulness. The overall validation score of 4.925 out of 5.000, with all ten responses categorized as Accept, establishes that the system generates culturally appropriate and educationally reliable responses for the present benchmark. This validation confirms that the framework is technically feasible for educational and preservation applications. For deployment in contexts requiring authoritative ceremonial guidance or ritual exactness, validation by a broader panel of specialists covering a larger query set remains advisable.

The fixed-character chunking strategy with 100-character overlap maintains contextual continuity for prose content and simple tabular structures, but may fragment complex astronomical tables present in Pawukon manuscripts. Calendar tables encoding multi-dimensional relationships between temporal cycles, astrological components, and ceremonial protocols contain structured information that does not align naturally with character-based segmentation. When chunk boundaries bisect large tables, semantic relationships between distant cells may be lost despite local overlap preservation. The current approach prioritizes

general applicability across diverse content types rather than specialized handling of structured data formats. Future iterations should investigate adaptive chunking algorithms that detect tabular structures and adjust segmentation boundaries to preserve table integrity. Semantic chunking based on content type classification, where tables are identified and chunked as complete units regardless of character count, would improve preservation of structured astronomical data while maintaining context-aware segmentation for narrative content.

6.4.6. Design trade-offs

The framework prioritizes semantic fidelity over lexical exactness, reflecting the design decision that cultural understanding (captured by BERTScore) is more critical than word-for-word reproduction (ROUGE) for educational and preservation applications. This trade-off is validated by exceptional performance on procedural knowledge queries (95% BERTScore F1), where conceptual accuracy supersedes literal phrasing. For contexts requiring verbatim preservation, such as ritual incantations or legal cultural documents, alternative or hybrid approaches combining semantic understanding with phrase-level accuracy would be necessary.

The elimination of traditional preprocessing in favor of semantic processing represents another fundamental trade-off: reduced dependency on unavailable linguistic resources versus potential loss of language-specific optimization. This decision proves advantageous for low-resource languages but may underperform language-specific pipelines in high-resource contexts where mature NLP tools exist.

7. Conclusion

This research successfully developed and validated an AI-driven framework for cultural heritage preservation in low-resource language contexts, achieving robust quantitative performance: 86.46% semantic accuracy where (BERTScore F1), 43.24% lexical accuracy (ROUGEScore F1), and statistically significant evaluation validation (one-way ANOVA: $F = 5.2613$, $p < 0.001$). Human expert validation by three independent cultural validators from Komite Seni Budaya Nusantara, Kejawen Maneges, and Sanggar Budaya Sukoraras yielded an overall score of 4.925 out of 5.000, with all ten generated responses receiving Accept decisions. Mean scores by dimension were: source faithfulness (4.833), cultural and religious appropriateness (5.000), linguistic clarity (4.867), and educational usefulness (5.000). This human validation confirms that the BERTScore–ROUGE-L divergence reflects deliberate paraphrasing rather than cultural distortion, and that the generated responses are culturally appropriate for educational use. The framework processed 621 pages of Javanese-language Pawukon manuscripts through a sophisticated pipeline integrating semantic chunking (1,000-character segments with 100-character overlap), hierarchical vector indexing (PGVector with four-level architecture), and language-adaptive LLM prompting (GPT-4 with bilingual Javanese-Indonesian support). Parameter sensitivity analysis identified 100-character chunk overlap ($F1 = 0.864$) and Top-$K = 5$ ($F1 = 0.8646$) as configurations empirically suitable for the present Pawukon corpus, with the caveat that validation on independent query sets is necessary before generalizing these settings to other cultural corpora.

The framework's success in handling the complexities of the Pawukon calendar system offers valuable insights for digital cultural preservation, particularly for traditional knowledge systems documented in low-resource languages. Key technical achievements include: demonstration of successful low-resource language processing without traditional linguistic preprocessing infrastructure, leveraging multilingual foundation models to bypass unavailable NLP tools; development of culturally-aware semantic chunking operating directly on Javanese text through contextual understanding; implementation of hierarchical vector store architecture optimized for interconnected cultural concepts; and effective integration of language-adaptive prompt engineering enabling bilingual operation. The system's exceptional performance in preserving procedural knowledge and temporal relationships is particularly noteworthy, with BERTScores exceeding 95% for methodological queries, demonstrating maintenance of nuanced cultural relationships despite linguistic resource constraints.

The evaluation framework combining semantic metrics (BERTScore), lexical metrics (ROUGEScore), and statistical validation (ANOVA, Shapiro-Wilk) provides comprehensive assessment methodology for cultural preservation effectiveness. The substantial semantic-lexical performance differential (43% gap) reveals a fundamental trade-off: the framework prioritizes conceptual accuracy and cultural authenticity over verbatim text reproduction, appropriate for educational and preservation applications but potentially insufficient for contexts demanding exact phraseology preservation.

We Identified limitations and challenges as follows: Despite strong semantic preservation, this research identified important challenges in digital cultural preservation for low-resource languages. First, lexical preservation remains limited (ROUGE: 43.24%), necessitating future hybrid approaches combining semantic understanding with phrase-level accuracy for ritual texts requiring precise wording. Second, generalization beyond Pawukon calendar requires empirical validation on diverse cultural systems to establish transferability and identify necessary adaptations. The framework's core components appear conceptually transferable to other traditional knowledge systems in low-resource language contexts, but empirical confirmation across different calendar systems, knowledge organization structures, and linguistic contexts is necessary before claiming broad applicability. Third, extent of Javanese representation in foundation model training corpora remains unclear; languages with even less representation may experience degraded performance, suggesting need for language-specific fine-tuning or alternative approaches for extremely low-resource contexts. Fourth, while the present study incorporates human expert validation by three cultural validators from established Javanese cultural institutions (see Section 4.4 above), the validation set is limited to ten generated responses. Broader coverage involving a larger and more diverse set of queries, evaluated by a greater number of validators, would further strengthen cultural authenticity claims and improve generalizability of the validation findings. Fifth, computational cost (OpenAI API, cloud infrastructure) limits accessibility for resource-constrained cultural communities, suggesting need for offline deployment alternatives or more efficient open-source model adaptation.

Our Methodologicalcontributions as follows: The methodology developed in this research provides a foundation for similar cultural preservation efforts, offering a replicable framework that balances technological innovation with cultural authenticity specifically for low-resource language contexts. The demonstrated success in bypassing traditional NLP pipeline requirements through semantic understanding offers practical path forward for numerous underrepresented languages lacking mature computational linguistics infrastructure. As digital preservation becomes increasingly crucial for cultural heritage globally, frameworks like this designed specifically for linguistic and resource constraints faced by many cultural communities will play a vital role in ensuring accessibility and preservation of traditional knowledge systems for future generations.

Thus,regarding Priority future directions, this work opens several promising avenues for continued research: (1) Enhancement of low-resource language processing capabilities through fine-tuning multilingual models on larger Javanese corpora or developing culture-specific embedding models; (2) Development of more sophisticated cultural context preservation techniques, including graph-based knowledge representation for networked cultural concepts and hybrid semantic-lexical approaches for contexts requiring verbatim preservation; (3) Extension and empirical validation of the framework to other traditional knowledge systems, particularly Chinese Lunar, Islamic Hijri, and Mayan calendar systems to assess generalizability and develop standardized adaptation protocols; (4) Improvement of validation methodologies involving cultural expert assess-

ment alongside computational metrics, establishing standardized evaluation frameworks for cultural preservation quality; (5) Investigation of offline deployment strategies and open-source model alternatives reducing computational costs and infrastructure dependencies.

The successof this approach in preserving the Javanese Pawukon calendar system demonstrates the potential of modern AI technologies in cultural heritage preservation while highlighting the importance of maintaining cultural and linguistic authenticity in digital transformation efforts. This research contributes both theoretical insights into low-resource language processing and practical methodologies for cultural heritage preservation, offering evidence that sophisticated cultural preservation can be achieved for underrepresented languages through strategic adaptation of foundation model capabilities rather than extensive development of language-specific infrastructure.

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Khafiizh Hastuti: Conceptualization, Methodology, Supervision; **Erwin Yudi Hidayat:** Writing – original draft, Formal analysis; **Farrikh Alzami:** Writing – review & editing, Visualization; **Ifan Risqa:** Project administration; **Sabrina Ahmad:** Validation; **Hanif Fahmi:** Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A.

Statistical significance testing
One-way ANOVA Test Results:
F-statistic: 5.2613
p-value: 0.0000
Tukey's HSD Test Results

Table A.7
Multiple Comparison of Means- Tukey HSD (FWER = 0.05).

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
BERTScore_F1	BERTScore_P	−903	10	−412976	394,916	FALSE
BERTScore_F1	BERTScore_R	993	10	−394016	413,876	FALSE
BERTScore_F1	ROUGE1_F1	-43227	25	−836216	-28324	TRUE
BERTScore_F1	ROUGE1_P	-43418	238	−838126	-30234	TRUE
BERTScore_F1	ROUGE1_R	-30389	3432	−707836	100,056	FALSE
BERTScore_F1	ROUGE2_F1	-48661	57	−890556	-82664	TRUE
BERTScore_F1	ROUGE2_P	-49594	44	−899886	-91994	TRUE
BERTScore_F1	ROUGE2_R	-41718	364	−821126	-13234	TRUE
BERTScore_F1	ROUGEL_F1	-41422	392	−818166	-10274	TRUE
BERTScore_F1	ROUGEL_P	-43404	239	−837986	-30094	TRUE
BERTScore_F1	ROUGEL_R	-30106	3575	−705006	102,886	FALSE
BERTScore_P	BERTScore_R	1896	10	−384986	422,906	FALSE
BERTScore_P	ROUGE1_F1	-42324	314	−827186	-19294	TRUE
BERTScore_P	ROUGE1_P	-42515	299	−829096	-21204	TRUE
BERTScore_P	ROUGE1_R	-29486	3899	−698806	109,086	FALSE
BERTScore_P	ROUGE2_F1	-47758	74	−881526	-73634	TRUE
BERTScore_P	ROUGE2_P	-48691	57	−890856	-82964	TRUE
BERTScore_P	ROUGE2_R	-40815	453	−812096	-4204	TRUE
BERTScore_P	ROUGEL_F1	-40519	486	−809136	-1244	TRUE
BERTScore_P	ROUGEL_P	-42501	3	−828956	-21064	TRUE
BERTScore_P	ROUGEL_R	-29203	4051	−695976	111,916	FALSE
BERTScore_R	ROUGE1_F1	-4422	194	−846146	-38254	TRUE
BERTScore_R	ROUGE1_P	-44411	184	−848056	-40164	TRUE
BERTScore_R	ROUGE1_R	-31382	2954	−717766	90,126	FALSE
BERTScore_R	ROUGE2_F1	-49654	43	−900486	-92594	TRUE
BERTScore_R	ROUGE2_P	-50587	33	−909816	-101924	TRUE
BERTScore_R	ROUGE2_R	-42711	285	−831056	-23164	TRUE
BERTScore_R	ROUGEL_F1	-42415	307	−828096	-20204	TRUE
BERTScore_R	ROUGEL_P	-44397	185	−847916	-40024	TRUE
BERTScore_R	ROUGEL_R	-31099	3086	−714936	92,956	FALSE
ROUGE1_F1	ROUGE1_P	-191	10	-405856	402,036	FALSE
ROUGE1_F1	ROUGE1_R	12,838	9956	-275566	532,326	FALSE
ROUGE1_F1	ROUGE2_F1	-5434	10	-458286	349,606	FALSE
ROUGE1_F1	ROUGE2_P	-6367	10	-467616	340,276	FALSE
ROUGE1_F1	ROUGE2_R	1509	10	-388856	419,036	FALSE
ROUGE1_F1	ROUGEL_F1	1805	10	-385896	421,996	FALSE
ROUGE1_F1	ROUGEL_P	-177	10	-405716	402,176	FALSE
ROUGE1_F1	ROUGEL_R	13,121	9947	-272736	535,156	FALSE
ROUGE1_P	ROUGE1_R	13,029	995	-273656	534,236	FALSE
ROUGE1_P	ROUGE2_F1	-5243	10	-456376	351,516	FALSE
ROUGE1_P	ROUGE2_P	-6176	10	-465706	342,186	FALSE
ROUGE1_P	ROUGE2_R	17	10	-386946	420,946	FALSE
ROUGE1_P	ROUGEL_F1	1996	10	-383986	423,906	FALSE
ROUGE1_P	ROUGEL_P	14	10	-403806	404,086	FALSE
ROUGE1_P	ROUGEL_R	13,312	9941	-270826	537,066	FALSE
ROUGE1_R	ROUGE2_F1	-18272	9344	-586666	221,226	FALSE
ROUGE1_R	ROUGE2_P	-19205	9096	-595996	211,896	FALSE
ROUGE1_R	ROUGE2_R	-11329	9986	-517236	290,656	FALSE
ROUGE1_R	ROUGEL_F1	-11033	9989	-514276	293,616	FALSE
ROUGE1_R	ROUGEL_P	-13015	9951	-534096	273,796	FALSE
ROUGE1_R	ROUGEL_R	283	10	-401116	406,776	FALSE
ROUGE2_F1	ROUGE2_P	-933	10	-413276	394,616	FALSE
ROUGE2_F1	ROUGE2_R	6943	10	-334516	473,376	FALSE
ROUGE2_F1	ROUGEL_F1	7239	10	-331556	476,336	FALSE
ROUGE2_F1	ROUGEL_P	5257	10	-351376	456,516	FALSE
ROUGE2_F1	ROUGEL_R	18,555	9274	-218396	589,496	FALSE
ROUGE2_P	ROUGE2_R	7876	10	-325186	482,706	FALSE
ROUGE2_P	ROUGEL_F1	8172	9999	-322226	485,666	FALSE
ROUGE2_P	ROUGEL_P	619	10	-342046	465,846	FALSE
ROUGE2_P	ROUGEL_R	19,488	9011	-209066	598,826	FALSE
ROUGE2_R	ROUGEL_F1	296	10	-400986	406,906	FALSE
ROUGE2_R	ROUGEL_P	-1686	10	-420806	387,086	FALSE
ROUGE2_R	ROUGEL_R	11,612	9982	-287826	520,066	FALSE
ROUGEL_F1	ROUGEL_P	-1982	10	-423766	384,126	FALSE
ROUGEL_F1	ROUGEL_R	11,316	9986	-290786	517,106	FALSE
ROUGEL_P	ROUGEL_R	13,298	9941	-270966	536,926	FALSE

Table A.8
Shapiro–Wilk Test Results.

Metric	Statistic	p-value
BERTScore_P	8.80092E+15	1.30813E+16
BERTScore_R	7.89673E+15	1.06589E+16
BERTScore_F1	8.60752E+15	7.78786E+15
ROUGE1_P	9.30727E+15	1.77001E+16
ROUGE1_R	7.85467E+15	9.65024E+15
ROUGE1_F1	8.71522E+15	1.20421E+16
ROUGE2_P	9.63362E+15	1.34324E+16
ROUGE2_R	8.34337E+15	3.77326E+15
ROUGE2_F1	7.30606E+15	3.12323E+15
ROUGEL_P	9.76664E+15	1.31263E+16
ROUGEL_R	8.48685E+15	1.56328E+16

References

Abbas, S., Ahmad, M., Nazar, M., Saleem, S., Isyanov, R., Aljedani, J., & Garalleh, H. A. (2025). Artificial neural network analysis of heat and mass transfer in fractional Casson flow. *Case Studies in Thermal Engineering*, *69*, 105946. https://linkinghub.elsevier.com/retrieve/pii/S2214157X25002060. https://doi.org/10.1016/j.csite.2025.105946

Bergamaschi, S., De Nardis, S., Martoglia, R., Ruozzi, F., Sala, L., Vanzini, M., & Vigliermo, R. A. (2022). Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The digitalmaktaba approach. *Sensors*, *22*(11), 3995. https://doi.org/10.3390/s22113995

Chen, X., Zhang, W., Xu, P., Zhao, Z., Zheng, Y., Shi, D., & He, M. (2024). Ffa-gpt: An automated pipeline for fundus fluorescein angiography interpretation and question-answer. *npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01101-z

Djanudji, & Doyodipuro, K. H. (2002). Pawukon Jangkep: Nyakup ing babagan horoskop jawa. (1st edition). Semarang: Dahara Prize.

Dubourg, E., Thouzeau, V., & Baumard, N. (2024). A step-by-step method for cultural annotation by LLMs. *Frontiers in Artificial Intelligence*, *7*. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1365508. https://doi.org/10.3389/frai.2024.1365508

Ehrmann, M., Romanello, M., Doucet, A., & Clematide, S. (2022). Introducing the HIPE 2022 shared task: Named entity recognition and linking in multilingual historical documents. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, & V. Setty (Eds.), *Advances in information retrieval* (pp. 347–354). Cham: Springer International Publishing.

Fateh, A., Birgani, R. T., Fateh, M., & Abolghasemi, V. (2024). Advancing multilingual handwritten numeral recognition with attention-driven transfer learning. *IEEE Access*, *12*, 41381–41395. https://doi.org/10.1109/ACCESS.2024.3378598

Gnanasekaran, R. K., & Marciano, R. (2024). Leveraging openAI's LLMs and cloud-based learning-as-a-service (laas) solutions to create culturally rich conversational AI chatbot: Chatlos - a study using the legacy of slavery dataset. In *Proceedings of international symposium on grids & clouds (ISGC) 2024 — pos(ISGC2024)* (pp. 001). (*vol. 458*). https://doi.org/10.22323/1.458.0001

Gunawan Admiranto, A. (2016). Pawukon: From incest, calendar, to horoscope. *Journal of Physics: Conference Series*, *771*, 012019. https://doi.org/10.1088/1742-6596/771/1/012019

Han, L., Gladkoff, S., Erofeev, G., Sorokina, I., Galiano, B., & Nenadic, G. (2024). Neural machine translation of clinical text: An empirical investigation into multilingual pre-trained language models and transfer-learning. *Frontiers in Digital Health*, *6*. https://doi.org/10.3389/fdgth.2024.1211564

Haurum, K. R., Ma, R., & Long, W. (2024). Real estate with AI: An agent based on langchain. *Procedia Computer Science*, *242*, 1082–1088. 11th International Conference on Information Technology and Quantitative Management (ITQM 2024) https://www.sciencedirect.com/science/article/pii/S1877050924019185. https://doi.org/10.1016/j.procs.2024.08.199

Imran, A. S., Hodnefjeld, H., Kastrati, Z., Fatima, N., Daudpota, S. M., & Wani, M. A. (2023). Classifying European court of human rights cases using transformer-based techniques. *IEEE Access*, *11*, 55664–55676. https://doi.org/10.1109/ACCESS.2023.3279034

Jeong, J., Gil, D., Kim, D., & Jeong, J. (2024). Current research and future directions for off-site construction through langchain with a large language model. *Buildings*, *14*(8), 2374. https://doi.org/10.3390/buildings14082374

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, *8*, 64–77. https://doi.org/10.1162/tacl_a_00300

Kaluvilla, B. B. (2024). Cultural preservation through technology in UAE libraries. *Library Hi Tech News*, *41*(8), 6–9. https://doi.org/10.1108/lhtn-02-2024-0032

Karjanto, N., & Beauducel, F. (2024). An ethnoarithmetic excursion into the javanese calendar. In B. Sriraman (Ed.), *Handbook of the history and philosophy of mathematical practice* (pp. 1097–1126). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-40846-5_82

Khadija, M. A., Aziz, A., & Nurharjadmo, W. (2023). Automating information retrieval from faculty guidelines: Designing a PDF-driven chatbot powered by openAI chatGPT. In *2023 International conference on computer, control, informatics and its applications (IC3INA)* (pp. 394–399). https://doi.org/10.1109/IC3INA60834.2023.10285808

Kim, W. T., Shin, J., Yoo, I.-S., Lee, J.-W., Jeon, H. J., Yoo, H.-S., Kim, Y., Jo, J.-M., Hwang, S., Lee, W.-J., Park, S., & Kim, Y.-J. (2024). Medication extraction and drug interaction chatbot: Generative pretrained transformer-powered chatbot for drug-drug interaction. *Mayo Clinic Proceedings: Digital Health*, *2*(4), 611–619. https://www.sciencedirect.com/science/article/pii/S2949761224000981. https://doi.org/10.1016/j.mcpdig.2024.09.001

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics. https://aclanthology.org/W04-1013/.

Litschko, R., Vulić, I., Ponzetto, S. P., & Glavaš, G. (2022). On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, *25*(2), 149–183. https://doi.org/10.1007/s10791-022-09406-x

Liu, Z., Tang, Y., Luo, X., Zhou, Y., & Zhang, L. F. (2024). No need to lift a finger anymore? Assessing the quality of code generation by chatGPT. *IEEE Transactions on Software Engineering*, *50*(6), 1548–1584. https://doi.org/10.1109/TSE.2024.3392499

MicleM., Tîrziman, E., Repanovici, A., (2023). Visibility of documentary heritage through digitisation projects in romanian libraries. *PLOS ONE*, *18*(1), 1–21. https://doi.org/10.1371/journal.pone.0280671

Mothe, J. (2024). Shaping the future of endangered and low-resource languages—our role in the age of LLMs: A keynote at ECIR 2024. *SIGIR Forum*, *58*(1), 1–13. https://doi.org/10.1145/3687273.3687280

Nur, M., Siti Nurbayani, K., Jumadi, J., Supriadi, A., Hasni, H., & Sultan, H. (2023). Maritime cultural heritage of fishermen communities in Kepulauan Sangkarrang Subdistrict, Makassar City, Indonesia. *BIO Web of Conferences*, *70*, 05007. https://doi.org/10.1051/bioconf/20237005007

Paganelli, M., Buono, F. D., Baraldi, A., & Guerra, F. (2022). Analyzing how BERT performs entity matching. *Proceedings of the VLDB Endowment*, *15*(8), 1726–1738. https://doi.org/10.14778/3529337.3529356

Pitaña, T. S. (2002). Sacred places in java: The concept of site selection in Javanese architecture. *Architectural Science Review*, *45*(1), 13–19. https://doi.org/10.1080/00038628.2002.9696931

Polyvach, K. (2024). Cultural landscape heritage of Ukraine - conceptualisation, structuring and Atlas mapping. *Geografický časopis / Geographical Journal*, *76*(1), 39–61. https://journals.savba.sk/index.php/geograficky-casopis/article/view/2018. https://doi.org/10.31577/geogrcas.2024.76.1.03

Prem Jacob, T., Bizotto, B. L. S., & Sathiyanarayanan, M. (2024). Constructing the chatGPT for PDF files with langchain - AI. In *2024 International conference on inventive computation technologies (ICICT)* (pp. 835–839). https://doi.org/10.1109/ICICT60155.2024.10544643

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, *12*, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742

Saputra, R. (2024). Governance frameworks and cultural preservation in indonesia: Balancing policy and heritage. *Journal of Ethnic and Cultural Studies*, *11*(3), 25–50. https://doi.org/10.29333/ejecs/2145

Trichopoulos, G. (2023). Large language models for cultural heritage. In *Proceedings of the 2nd international conference of the ACM Greek SIGCHI chapter* CHIGREECE '23 (pp. 1–5). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3609987.3610018

Wang, X., Sun, J., & Qi, C. (2024). Ceda-tqa: Context enhancement and domain adaptation method for textbook qa based on llm and rag. In *2024 International conference on networking and network applications (naNA)* (pp. 263–268). https://doi.org/10.1109/NaNA63151.2024.00050

Wang, Y., Cui, L., & Zhang, Y. (2021). Improving skip-gram embeddings using BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1318–1328. https://doi.org/10.1109/TASLP.2021.3065201

Wang, Z., Ng, P., Ma, X., Nallapati, R., & Xiang, B. (2019). Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5878–5882). Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1599

Xiao, L., Cao, X., Tang, C., Lai, X., Zhu, X., & Han, Z. (2024). Enhancing information extraction from long document: Utilizing LLM's prompt engineering for long document set generation. In C. Yang (Ed.), *Fifth international conference on control, robotics, and intelligent system (CCRIS 2024)* (pp. 1). SPIE. https://doi.org/10.1117/12.3049923

Zahid, I. A., Joudar, S. S., Albahri, A. S., Albahri, O. S., Alamoodi, A. H., Santamaría, J., & Alzubaidi, L. (2024). Unmasking large language models by means of openAI GPT-4 and google AI: A deep instruction-based analysis. *Intelligent Systems with Applications*, *23*, 200431. https://doi.org/10.1016/j.iswa.2024.200431

Zhang, J., Xiang, R., Kuang, Z., Wang, B., & Li, Y. (2024). ArchGPT: Harnessing large language models for supporting renovation and conservation of traditional architectural heritage. *Heritage Science*, *12*(1). https://doi.org/10.1186/s40494-024-01334-x

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. https://arxiv.org/abs/1904.09675.

Zhu, M., & Cole, J. M. (2022). Pdfdataextractor: A tool for reading scientific text and interpreting metadata from the typeset literature in the portable document format. *Journal of Chemical Information and Modeling*, *62*(7), 1633–1643. https://doi.org/10.1021/acs.jcim.1c01198