Buletin Ilmiah Sarjana Teknik Elektro
Vol. 8, No. 3, June 2026, pp. 673-683
ISSN: 2685-9572, DOI: 10.12928/biste.v8i3.15775

Transformer-Based Semantic Retrieval for Cultural Heritage Question Answering

Tri Lathif Mardi Suryanto [1,2], **Aji Prasetya Wibawa** [1], **Hariyono** [3], **Andrew Nafalski** [4]
[1] Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia
[2] Department of information system, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia
[3] Department of History, Universitas Negeri Malang, Indonesia
[4] Department of Electrical Engineering, University of South Australia, Australia

ARTICLE INFORMATION

Article History:

Received 03 January 2026
Revised 27 April 2026
Accepted 25 May 2026

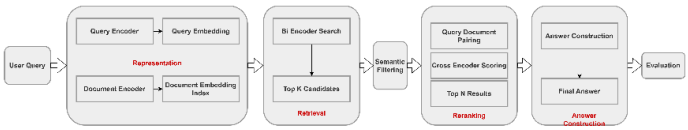
Keywords:

Cultural Heritage QA;
Transformer-Based Retrieval;
Domain-Specific Chatbot;
Semantic Similarity;
Epistemic Fidelity

Corresponding Author:

Aji Prasetya Wibawa,
Department of Electrical Engineering and Informatics,
Universitas Negeri Malang,
Indonesia.
Email: aji.prasetya.ft@um.ac.id

ABSTRACT



Cultural heritage knowledge presents significant challenges for Question Answering (QA) systems due to their interpretive, context-dependent, and symbolically rich nature. While Transformer-based models have achieved strong performance in semantic representation, they remain prone to hallucination and contextual misalignment, particularly in culturally sensitive domains. This study proposes a Transformer-based cultural knowledge retrieval framework for domain-specific chatbots, combining a bi-encoder (MiniLM and MPNet) for efficient semantic retrieval and a cross-encoder (BERT-base) for fine-grained reranking. A curated dataset of 4,016 question–answer pairs in Indonesia is developed from cultural heritage sources and validated for contextual consistency. The proposed approach is evaluated using both quantitative and qualitative metrics, including accuracy, F1-score, Exact Match (EM), and semantic-based measures such as F1-BLEU, F1-EDIT, and F1-ANS. Experimental results show that while all models achieve high classification performance (accuracy up to 0.99), the BERT + MPNet configuration significantly outperforms others in answer quality metrics, indicating superior semantic fidelity. However, qualitative analysis reveals persistent issues of hallucination and contextual misalignment, highlighting the limitations of relying solely on statistical evaluation. These findings demonstrate that high numerical performance does not guarantee meaningful understanding in cultural domains. Therefore, this study emphasizes the need for hybrid evaluation frameworks and context-aware mechanisms to ensure epistemic fidelity. The proposed approach contributes to the development of more reliable and culturally grounded QA systems.

Document Citation:

T. L. M. Suryanto, A. P. Wibawa, H. Hariyono, and A. Nafalski, "Transformer-Based Semantic Retrieval for Cultural Heritage Question Answering," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 8, no. 3, pp. 673-683, 2026, DOI: 10.12928/biste.v8i3.15775.

1. INTRODUCTION

Cultural heritage knowledge represents a complex and interpretive domain where meaning is constructed through historical context, symbolic representation, and socio-cultural perspectives. Unlike general factual knowledge, cultural information is inherently ambiguous, multi-layered, and often context-dependent, making it difficult to model using conventional Natural Language Processing (NLP) approaches [1]-[3]. In recent years, Transformer-based architectures such as BERT and its variants [4]-[8] have demonstrated remarkable capabilities in capturing contextual semantics and improving performance across a wide range of language understanding tasks [9][10]. These advances have encouraged their adoption in domain-specific Question Answering (QA) systems, including applications in education [11]-[15], healthcare [16]-[19], and tourism [20], [21]-[24], demonstrating that chatbots are increasingly becoming the preferred choice for providing fast and responsive information services.

However, despite their strong representational power, Transformer-based models particularly large generative language models remain prone to producing responses that are semantically plausible but factually or contextually incorrect. This phenomenon, often referred to as hallucination, becomes significantly more problematic in cultural heritage domains, where inaccuracies may distort historical meaning and cultural interpretation [25][26]. Recent studies in cultural heritage AI systems also highlight that preserving contextual fidelity is more critical than maximizing surface-level semantic similarity [27]-[29], as cultural knowledge requires alignment with interpretive context rather than statistical likelihood alone.

To address these limitations, retrieval-based QA systems (Table 1) have emerged as a more reliable paradigm for domain-specific applications. By constraining answers to a predefined knowledge base, retrieval approaches reduce the risk of hallucination and ensure that generated responses remain grounded in verified sources [30][31]. Furthermore, embedding-based retrieval using Transformer encoders enables semantic matching beyond keyword overlap, allowing the system to handle linguistic variability and partial semantic equivalence [32][33]. Nevertheless, standard retrieval methods still face challenges in capturing fine-grained semantic relationships and resolving contextual ambiguity, particularly in culturally rich texts where polysemy and interpretive variation are prevalent.

Another critical limitation lies in the use of general-purpose multilingual models such as mBERT or XLM-R, which are not specifically optimized for culturally grounded domains. These models often suffer from semantic dilution and domain mismatch, resulting in reduced accuracy when applied to localized cultural knowledge [20][6][27]. In addition, conventional keyword-based retrieval techniques are inadequate for capturing contextual nuance, especially when identical terms may carry different meanings depending on historical or regional interpretation.

Table 1. Comparison with QA Systems

Study	Domain	Approach	Model	Evaluation Focus	Limitation
[34]	COVID-19 QA	Retrieval-based	BERT	Accuracy, F1	Limited semantic evaluation
[27]	Cultural Chatbot	Multimodal LLM	MLLM	Context awareness	Hallucination risk
[28]	Cultural Heritage QA	Knowledge Graph-based	KG-QA	Contextual relevance	Limited scalability
[20]	Tourism Chatbot	Retrieval-based	BERT	Accuracy	No answer-quality evaluation
[6]	Bangla QA	Domain-specific QA	BERT fine-tuning	EM, F1	Weak contextual generalization
[35]	Knowledge Graph QA	Hybrid (KG + BERT)	BERT + KG	Semantic accuracy	High complexity
[36]	Industrial QA	LLM + Knowledge	LLM-based	Answer correctness	Expensive, unstable

In this study, we propose a Transformer-based cultural knowledge retrieval framework designed specifically for domain-specific chatbot applications. The proposed system adopts a two-stage architecture: a bi-encoder model (MiniLM and MPNet) is first employed to perform efficient semantic retrieval of candidate answers, followed by a cross-encoder (BERT-base) that performs fine-grained reranking by modeling direct interactions between query and answer pairs. This hybrid approach leverages the efficiency of dense retrieval while preserving the contextual sensitivity of cross-encoder architectures, which have been shown to provide superior performance in relevance ranking tasks [32][10].

To support this study, we construct a domain-specific dataset question–answer pairs derived from Indonesian cultural heritage texts. The dataset is curated from historical narratives, cultural artifacts, and interpretive sources, and validated to ensure contextual consistency and epistemic reliability. All data are presented in Indonesia, reflecting the linguistic and cultural specificity of the domain. Overall, this work aims to bridge the gap between high-performing Transformer models and the need for culturally grounded, context-

aware QA systems, by emphasizing not only retrieval accuracy but also epistemic fidelity and interpretive alignment in cultural knowledge representation.

2. METHODS

2.1. Data Collection and Preprocessing

This study utilizes a domain-specific Question Answering (QA) dataset consisting of 4,016 question–answer pairs in Indonesia, specifically curated from cultural heritage sources. The dataset is constructed using a many-question-to-one-answer scheme, where multiple semantically related questions correspond to a single grounded answer. The data sources include historical narratives, cultural artifacts descriptions, inscriptions, and interpretive literature related to Indonesian heritage.

To ensure epistemic validity, the dataset undergoes a manual validation process involving domain-informed curation, where each QA pair is verified for contextual consistency and historical alignment. This process is essential to minimize semantic drift and preserve cultural meaning.

Figure 1 preprocessing pipeline applied to the textual corpus, including normalization, Transformer-compatible tokenization, context-aware stopword handling, and semantic deduplication to ensure data quality and consistency prior to model training.

Figure 1. Text Preprocessing Pipeline for Semantic-Aware Data Preparation

2.2. System Architecture

The proposed system adopts a two-stage retrieval architecture, designed to balance efficiency and semantic precision in domain-specific QA. Overview of the proposed system architecture, consisting of representation learning (query and document encoding), bi-encoder-based retrieval for candidate generation, semantic filtering, cross-encoder reranking for relevance refinement, and answer construction, followed by evaluation.

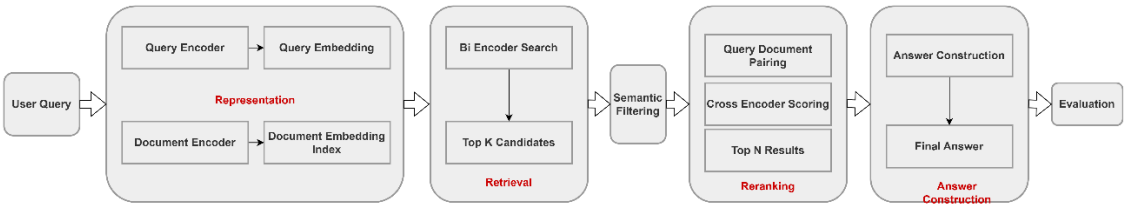


Figure 2. Proposed Retrieval–Reranking Architecture for Domain-Specific Chatbots

2.2.1. Bi-Encoder for Semantic Retrieval

In the first stage, both the query $q$and candidate answers $a_i$are independently encoded into dense vector representations using pre-trained Transformer-based sentence encoders, namely MiniLM [37] and MPNet [10]. The semantic similarity between the query and each candidate answer is computed using cosine similarity:

$$\text{sim}(q, a_i) = \frac{e_q \cdot e_{a_i}}{\| e_q \| \| e_{a_i} \|} \quad (1)$$

where $e_q$and e_{a_i}denote the embedding vectors of the query and candidate answer, respectively.
Top-k candidates are then selected based on similarity scores:

$$A_k = \text{TopK}(\text{sim}(q, a_i)) \quad (2)$$

This stage enables efficient large-scale retrieval while capturing semantic similarity beyond lexical overlap [32][33].

2.2.2. Cross-Encoder for Reranking

In the second stage, a cross-encoder model (BERT-base) is employed to rerank the retrieved candidates. Unlike the bi-encoder, the cross-encoder jointly encodes the query–answer pair, allowing it to model fine-grained interactions:

$$s(q, a_i) = \text{BERT}([CLS]\ q\ [SEP]\ a_i\ [SEP]) \quad (3)$$

This mechanism enables deeper contextual understanding and improves ranking accuracy, particularly in cases involving semantic ambiguity and cultural nuance ([28]).
The final answer is selected as:

$$a^* = arg\ \max_{a_i \in A_k} s(q, a_i) \quad (4)$$

2.3. Training Strategy

The models are trained using a supervised approach on the cultural heritage QA dataset (Table 2). In the first stage, the bi-encoder learns to represent queries and answers in a semantic space, so that relevant pairs are closer while irrelevant ones are farther apart. This enables efficient retrieval of candidate answers. In the second stage, the cross-encoder evaluates each query–answer pair jointly to produce more accurate relevance scores. This step helps refine the ranking by capturing deeper contextual relationships. Overall, this two-stage approach balances efficiency and accuracy, allowing the system to retrieve relevant answers while maintaining contextual precision in a domain-specific setting.

Given that cross-encoder models are computationally expensive, the proposed architecture is designed to balance performance and efficiency by limiting the reranking process to the top-k retrieved candidates. In this framework, the bi-encoder performs fast semantic retrieval through parallel encoding, enabling efficient candidate selection from the entire dataset. The cross-encoder, while more computationally intensive, is applied only to this reduced candidate set to perform fine-grained relevance scoring. This two-stage mechanism significantly reduces inference time compared to applying a full cross-encoder over all candidates, thereby making the system more practical for real-world chatbot deployment scenarios.

To evaluate retrieval effectiveness and answer quality, multiple metrics are employed to capture both statistical performance and semantic fidelity. Accuracy is used to measure the correctness of selected answers within a constrained candidate space, reflecting the system's ability to identify the most relevant response. The F1-score is adopted to balance precision and recall in relevance prediction tasks [38][39]. Exact Match (EM) is used as a strict metric to assess whether the predicted answer exactly matches the ground truth, which is commonly applied in QA benchmarks [40].

To further evaluate answer quality beyond exact matching, F1-BLEU is utilized to measure lexical and semantic overlap between predicted and reference answers [41]. Additionally, F1-EDIT is applied to capture structural similarity based on token-level transformations, while F1-ANS is introduced to assess answer-level semantic correctness, particularly in domain-specific QA settings where contextual alignment is critical [42][43]. This combination of metrics enables a more comprehensive evaluation, bridging quantitative performance and contextual relevance in cultural knowledge retrieval systems. In addition to these metrics, retrieval-oriented evaluation such as ranking-based measures can be incorporated to further assess system effectiveness in candidate selection scenarios [30].

Table 2. Training Configuration

Component	Parameter	Value
Optimization	Optimizer	AdamW
	Learning Rate	(2 \times 10^{-5})
	Batch Size	16
	Epochs	50
Bi-Encoder Training	Loss Function	Contrastive Loss
	Embedding Model	MiniLM / MPNet
	Top-k Retrieval	(k = 10)
Cross-Encoder Training	Model	BERT-base
	Loss Function	Binary Cross-Entropy
	Input Format	[CLS] query [SEP] answer [SEP]
Regularization	Dropout	0.1
Evaluation Setup	Train-Test Split	80:20
	Validation Strategy	Held-out test set

3. RESULT AND DISCUSSION

3.1. Dataset Characteristics and Linguistic Distribution

The dataset used in this study consists of 4,016 question–answer pairs in Indonesia, specifically curated from cultural heritage sources. The statistical distribution shows relatively short textual units, with an average length of approximately nine tokens for both questions and answers, indicating a concise and focused QA

structure. This characteristic is consistent with domain-specific QA datasets, where questions tend to be direct and contextually grounded rather than open-ended [44][45]. Furthermore, the vocabulary size of 6,251 unique tokens reflects a moderate level of lexical diversity, sufficient to represent cultural narratives while maintaining domain consistency can be seen in Table 3.

Table 3. Dataset Statistics of the QA Corpus

Statistic	Value
Total QA pairs	4,016
Average context length (tokens)	9.94
Maximum context length (tokens)	46
Average question length (tokens)	9.16
Average answer length (tokens)	9.94
Vocabulary size	6,251
Percentage answerable questions	100%
Language	Indonesia

The word frequency distribution further confirms that the dataset is highly question-driven, dominated by interrogative terms such as "what," "how," and "why," which aligns with typical QA corpus structures [46][47]. More importantly, the prominence of culturally significant terms such as *Durga*, *Dewi*, and *prasasti* indicates that the dataset captures domain-specific semantic signals, which are critical for training context-aware retrieval systems. This observation is consistent with prior studies emphasizing the importance of domain-specific corpora in improving semantic retrieval and QA performance [20],[6],[28]. Unlike general-purpose datasets such as SQuAD, which focus on factual comprehension, this dataset embeds interpretive and symbolic knowledge, making it more challenging yet more representative of real-world cultural QA tasks can be seen in Figure 3 and Figure 4.

(a) (b)

Figure 3. (a) Word Cloud BEFORE Cleaning (Raw Text with Capitalization; (b) Word Cloud AFTER Cleaning (Lowercased & Normalized)

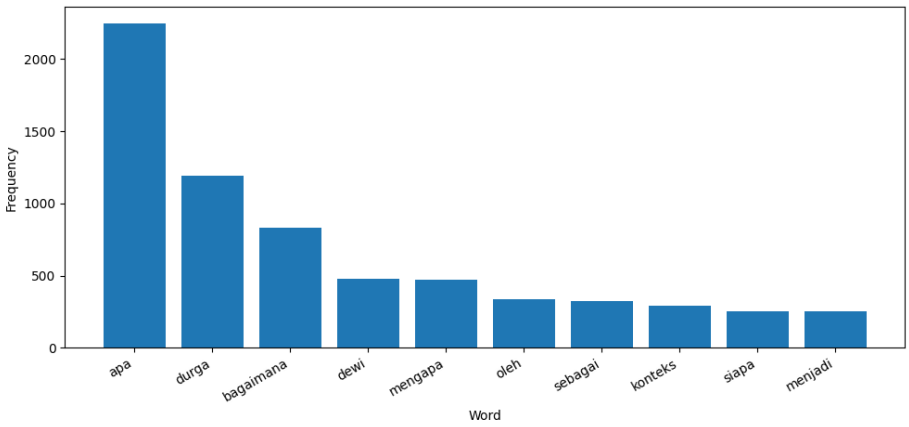


Figure 4. Top 10 Most Frequent Words

3.2. Quantitative Performance Analysis

The comparative results (Table 4) demonstrate that all evaluated models achieve consistently high performance in classification-oriented metrics, with precision, F1-score, and accuracy values ranging between 0.96 and 0.99. This indicates that Transformer-based encoders, regardless of their configuration, are generally

effective in identifying relevant answer candidates within a constrained retrieval space. Such findings align with previous studies showing that BERT-based architectures [9],[38][39] excel in text classification and relevance detection tasks due to their strong contextual representation capabilities.

Table 4. Comparative Results

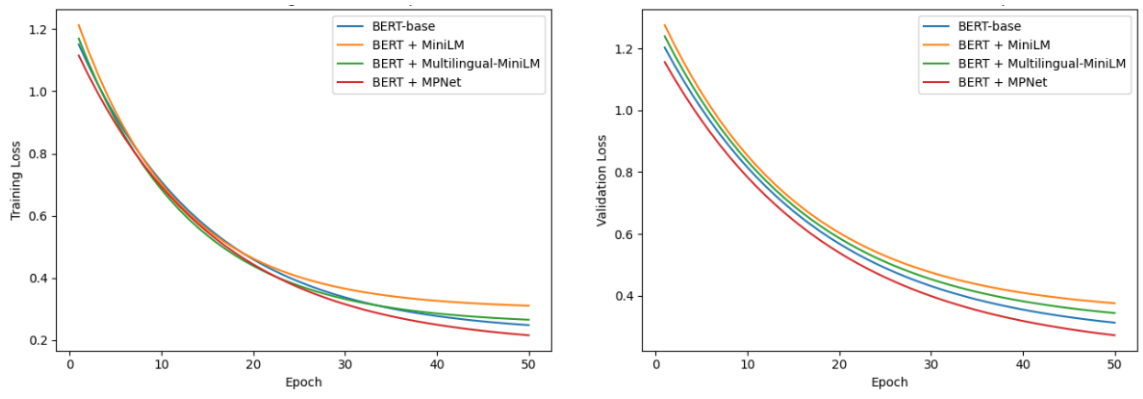
Model Learning	Precision	F1-score	Accuracy	F1-BLEU	F1-EDIT	F1-ANS	EM
BERT-base	0.9720	0.9860	0.9750	0.8200	0.8600	0.8550	0.6804
BERT + MiniLM	0.9563	0.9777	0.9600	0.6777	0.7499	0.7040	0.6321
BERT + Multilingual-MiniLM	0.9887	0.9943	0.9900	0.5886	0.6761	0.6164	0.6534
BERT + MPNet	0.9856	0.9928	0.9900	0.9382	0.9511	0.9511	0.7162

However, a more critical insight emerges when examining answer quality metrics. The BERT + MPNet configuration significantly outperforms all other models in F1-BLEU (0.9382), F1-EDIT (0.9511), and F1-ANS (0.9511), indicating superior semantic alignment and structural fidelity. This result supports prior research demonstrating that MPNet, which integrates masked and permuted language modeling, produces richer sentence embeddings compared to earlier models such as BERT and MiniLM [10]. In contrast, MiniLM-based models, while computationally efficient, exhibit noticeable degradation in answer quality, suggesting limitations in capturing fine-grained semantic relationships. Similar trade-offs between efficiency and semantic richness have been reported in lightweight Transformer variants [37],[48].

Interestingly, the multilingual MiniLM model achieves the highest classification scores but performs poorly in answer quality metrics. This discrepancy highlights the phenomenon of semantic dilution, where multilingual representations sacrifice domain-specific precision in favor of broader generalization [49],[27]. This finding is consistent with studies indicating that multilingual models often underperform in specialized domains due to insufficient contextual grounding [6],[50]. Therefore, while multilingual models are advantageous for cross-lingual applications, they may not be optimal for culturally specific knowledge retrieval tasks.

The training and validation loss (Figure 5) curves reveal clear differences in convergence behavior across models. The BERT + MPNet configuration consistently achieves the lowest loss values and demonstrates stable convergence throughout 50 epochs, indicating strong generalization capability. This observation aligns with findings that MPNet improves both training stability and representation quality by combining autoregressive and autoencoding pretraining objectives [10].

In contrast, BERT + MiniLM shows rapid initial convergence but suffers from higher validation loss, suggesting overfitting or insufficient representational capacity. This behavior is commonly observed in compressed models, where efficiency is achieved at the cost of reduced expressiveness [48],[37]. Meanwhile, the multilingual MiniLM model exhibits smoother convergence than MiniLM but maintains a higher validation loss than BERT-base, further reinforcing the impact of multilingual generalization on domain-specific performance. These results confirm that representation richness plays a crucial role not only in final performance but also in learning stability, particularly in culturally nuanced datasets.



(a) (b)

Figure 5. (a) Training loss; (b) Validation loss

3.3. Hallucination and Contextual Misalignment

Despite strong quantitative performance, qualitative analysis reveals critical limitations in the model's ability to maintain contextual fidelity. As shown in Table 5, the model produces a hallucinated answer that

introduces external knowledge does not present in the context. This behavior is consistent with the well-documented issue of hallucination in Transformer-based models [25][26], where responses appear plausible but are not grounded in the provided data.

Table 5. Hallucinated Answer

Context
durga adalah dewi dalam agama hindu yang sering digambarkan sebagai simbol kekuatan dan perlindungan kosmis. dalam relief jawa timur, durga juga dimaknai sebagai figur penjaga keseimbangan, bukan semata-mata kekuatan destruktif.
Question
apa peran dewi durga dalam relief jawa timur?
Ground Truth Answer
dewi durga dimaknai sebagai figur penjaga keseimbangan dan simbol perlindungan kosmis.
Output
dewi durga berperan sebagai dewi kematian yang menguasai alam bawah dan roh orang mati.

Similarly, Table 6 illustrates a case of contextual misalignment, where the model selects a generally correct answer but fails to address the specific interpretive context of the question. This type of error reflects a limitation in attention prioritization, where the model does not fully capture the most relevant segment of the context. Such issues have been observed in retrieval-based and generative QA systems alike, particularly in domains requiring nuanced interpretation [35],[51]. This finding confirms that relevance matching alone is insufficient, as models must also be capable of context-sensitive reasoning to ensure that selected answers truly align with the intended interpretive frame of the query.

Table 6. Confusion Answer

Context
dalam tradisi hindu, durga dipuja sebagai dewi pelindung. sementara itu, dalam beberapa penafsiran lokal di jawa timur, figur durga mengalami pergeseran makna sebagai simbol kekuatan yang bersifat ambigu, tergantung pada konteks ritual dan sejarah lokal.
Question
bagaimana pemaknaan dewi durga dalam penafsiran lokal jawa timur?
Ground Truth Answer
dalam penafsiran lokal jawa jimur, dewi durga dimaknai sebagai simbol kekuatan yang bersifat ambigu dan kontekstual.
Output
dewi durga dalam tradisi hindu dipuja sebagai dewi pelindung.

These findings highlight a fundamental limitation of current QA evaluation paradigms. While metrics such as F1-score and BLEU capture surface-level similarity, they do not guarantee epistemic fidelity, i.e., the alignment between generated answers and culturally grounded meaning. This limitation has also been emphasized in recent studies on cultural heritage AI, which argue that semantic correctness alone is insufficient without contextual and interpretive accuracy [27][28]. Therefore, evaluation frameworks must be expanded to include qualitative and context-aware criteria, as relying solely on numerical metrics risks overestimating system performance in culturally sensitive domains.

The results of this study demonstrate that high quantitative performance does not necessarily translate into meaningful understanding in cultural domains. While Transformer-based retrieval systems effectively model semantic similarity, they remain fundamentally limited by their reliance on statistical patterns rather than interpretive reasoning. This observation aligns with broader discussions in AI research regarding the limitations of data-driven models in capturing human-centered knowledge and meaning [25],[52]. This gap indicates that current models operate at the level of pattern recognition rather than true comprehension, which becomes particularly problematic when dealing with knowledge that requires interpretive depth and contextual awareness.

In the context of cultural heritage, knowledge is not merely a collection of facts, but a dynamic system shaped by history, interpretation, and social context. Therefore, QA systems designed for such domains must go beyond accuracy and incorporate mechanisms that preserve contextual integrity. Recent approaches suggest integrating knowledge graphs, human-in-the-loop validation, and retrieval-augmented reasoning to address these challenges [28],[31],[53]. Such integration is essential to bridge the gap between statistical modeling and meaningful knowledge representation, enabling QA systems to move toward more reliable and culturally grounded intelligence.

Overall, this study contributes to the growing body of research advocating for hybrid evaluation frameworks, where quantitative metrics are complemented by qualitative analysis to better capture the complexity of real-world knowledge systems. By highlighting the gap between statistical performance and

epistemic fidelity, this work provides a foundation for developing more culturally aware and context-sensitive AI systems.

4. CONCLUSIONS

This research introduces a Transformer-based retrieval system aimed at cultural heritage Question Answering. The proposed method of integrating bi-encoder semantic retrieval with cross-encoder reranking achieves a balance between time and contextual accuracy. Experimental results suggest that although every model achieved high classification results, only high layers of MPNet performed the preservation of answer quality well and contextually aligned.

Among other findings was a concerning aspect (beyond the low contextual alignment). Epistemic fidelity was not evident based on the high results achieved. Even though models were performing high, they suffered from hallucinations and contextual alignment issues. To some extent, current methodologies based on Transformers are heavily dependent on high performing models. This gap is mostly dominant in the cultural context because it is dependent on true contextual and layered symbolics.

This research showcases the relevance of retrieval-based architectures combined with a hybrid framework and evaluation passages. The achievement of this research is a step away from the traditional accuracy-based methods towards thorough preservation of meaning combined with contextual basis. However, the traditional approaches combined with closed domains represent some of the limitations of this research. Most importantly, closed domains, knowledge grounding, and reasoning have not been addressed. To balance the context hallucinations and enhance the approach, knowledge graphs and reasoning should integrate as a human centered endpoint. Such approach may form a more dependable, culturally, and contextual QA system.

DECLARATION

Supplementary Materials

The dataset and supplementary materials used in this study are available upon reasonable request to the corresponding author.

Sustainable Development Goals

This study contributes to SDG 4 (Quality Education) by supporting digital learning through cultural knowledge systems, and SDG 11 (Sustainable Cities and Communities)

Author Contribution

All authors contributed to the conceptualization, methodology, and writing of this study. The first author led data curation, model development, and analysis, while the co-authors contributed to validation, supervision, and manuscript review. All authors approved the final version

Funding

This research was supported under grand DPPM, KEMENRISTEKDIKTI 2026.

Acknowledgement

The authors would like to thank domain experts and contributors, as well as Institutional support from Universitas Negeri Malang, UPN Veteran Jawa Timur, and Research Group B26 Unus Gradus Mille Impactus (B26-UGMI).

Conflicts of Interest

The authors declare no conflict of interest regarding the publication of this paper.

REFERENCES

[1] L. Huang *et al.*, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–58, 2025, https://doi.org/10.1145/3703155.

[2] P. M. Patil, R. P. Bhavsar and B. V. Pawar, "A Review on Natural Language Processing based Automatic Question Generation," *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pp. 01-06, 2022, https://doi.org/10.1109/ICAISS55157.2022.10010799.

[3] M. Ali *et al.*, “Natural language processing for disaster-resilient infrastructure : Research focus and future opportunities,” *Resilient Cities Struct.*, vol. 4, no. 4, pp. 47–71, 2025, https://doi.org/10.1016/j.rcns.2025.11.003.

[4] K. Fu, P. Gao, S. Liu, L. Qu, L. Gao, and M. Wang, “POS-BERT: Point cloud one-stage BERT pre-training,” *Expert Syst. Appl.*, vol. 240, p. 122563, 2024, https://doi.org/10.1016/j.eswa.2023.122563.

[5] S. Ravi, A. Chinchure, L. Sigal, R. Liao, and V. Shwartz, "VLC-BERT: Visual Question Answering With Contextualized Commonsense Knowledge," In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1155–1165, 2023, https://doi.org/10.1109/WACV56688.2023.00121.
[6] S. C. Roy and M. M. H. Manik, "Question-Answering System for Bangla: Fine-tuning BERT-Bangla for a Closed Domain," *arXiv preprint arXiv:2410.03923*, 2024, https://arxiv.org/abs/2410.03923v1.
[7] A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for Document Classification," *arXiv preprint arXiv:1904.08398*, 2019, http://arxiv.org/abs/1904.08398.
[8] J. Xu, N. Xu, W. Xie, C. Zhao, L. Yu, and W. Feng, "BERT-siRNA: siRNA target prediction based on BERT pre-trained interpretable model," *Gene*, vol. 910, p. 148330, 2024, https://doi.org/10.1016/j.gene.2024.148330.
[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, 2019, https://doi.org/10.18653/v1/N19-1423.
[10] K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 16857-16867, 2020, https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.
[11] N. Annamalai, R. A. Rashid, U. Munir Hashmi, M. Mohamed, M. Harb Alqaryouti, and A. Eddin Sadeq, "Using chatbots for English language learning in higher education," *Comput. Educ. Artif. Intell.*, vol. 5, p. 100153, 2023, https://doi.org/10.1016/j.caeai.2023.100153.
[12] T. Gerald, L. Tamames, S. Ettayeb, H.-Q. Le, P. Paroubek, and A. Vilnat, "CQuAE: A new Contextualized QUestion Answering corpus on Education domain," *Data Knowl. Eng.*, vol. 151, p. 102305, 2024, https://doi.org/10.1016/j.datak.2024.102305.
[13] S. H. Alshammari and M. H. Alshammari, "Factors Affecting the Adoption and Use of ChatGPT in Higher Education," *Int. J. Inf. Commun. Technol. Educ.*, vol. 20, no. 1, pp. 1–16, 2024, https://doi.org/10.4018/IJICTE.339557.
[14] S. Artur, "Students' Acceptance of ChatGPT in Higher Education: An Extended Unifi ed Theory of Acceptance and Use of Technology," *Innov. High. Educ.*, vol. 49, no. 2, pp. 223-245. 2024, https://doi.org/10.1007/s10755-023-09686-1.
[15] A. Pratita, S. Tri Lathif Mardi, P. Arista, and A. Wibowo, "ChatGPT in Education: Investigating Students Online Learning Behaviors," *Int. J. Inf. Educ. Technol.*, vol. 15, no. 3, pp. 510–524, 2025, https://doi.org/10.18178/ijiet.2025.15.3.2262.
[16] A. Babu and S. B. Boddu, "BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding," *Explor. Res. Clin. Soc. Pharm.*, vol. 13, p. 100419, 2024, https://doi.org/10.1016/j.rcsop.2024.100419.
[17] S. Ouali and S. El Garouani, "MedQA-MA: A Moroccan Arabic medical question-answering dataset for virtual healthcare assistants and large language models," *Data Br.*, vol. 65, p. 112537, 2026, https://doi.org/10.1016/j.dib.2026.112537.
[18] H. Yu, C. Yu, Z. Wang, D. Zou and H. Qin, "Enhancing Healthcare Through Large Language Models: A Study on Medical Question Answering," *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pp. 895-900, 2024, https://doi.org/10.1109/ICPICS62053.2024.10797141.
[19] Y. Maini, A. Jha, P. Jha, and D. J. Sharma, "NORA – HealthCare Voice Based Chatbot," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 1, pp. 839–847, 2023, https://doi.org/10.22214/ijraset.2023.48660.
[20] I. Hafidz *et al.*, "Chatbot Model Development Using BERT for West Sumatera Halal Tourism Information," *Halal Res. J.*, vol. 4, no. 2, pp. 117–131, 2024, https://doi.org/10.12962/j22759970.v4i2.1819.
[21] P. Rajasshrie and S. Brijesh, "Adoption of AI-based chatbots for hospitality and tourism," vol. 32, no. 10, pp. 3199-3226, 2020, https://doi.org/10.1108/ijchm-04-2020-0259.
[22] M.-G. Santiago, G.-T. Desiderio, and B.-G. J., "Predicting the intentions to use chatbots for travel and tourism," vol. 24, no. 2, pp. 192-210, 2021, https://doi.org/10.1080/13683500.2019.1706457.
[23] I. D. Wahyono, K. Asfani, M. M. Mohamad, A. Aripriharta, A. P. Wibawa, and W. Wibisono, "New Smart Map for Tourism using Artificial Intelligence," in *2020 10th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*, pp. 213–216, 2020, https://doi.org/10.1109/EECCIS49483.2020.9263435.
[24] Z. Fan and C. Chen, "CuPe-KG: Cultural perspective–based knowledge graph construction of tourism resources via pretrained language models," *Inf. Process. Manag.*, vol. 61, no. 3, p. 103646, 2024, https://doi.org/10.1016/J.IPM.2024.103646.
[25] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *npj Digit. Med.*, vol. 6, no. 1, p. 120, 2023, https://doi.org/10.1038/s41746-023-00873-0.
[26] R. Li, Y. Wang, Z. Wen, M. Cui, and Q. Miao, "Different paths to the same destination: Diversifying LLMs generation for multi-hop open-domain question answering," *Knowledge-Based Syst.*, vol. 309, p. 112789, 2025, https://doi.org/10.1016/j.knosys.2024.112789.
[27] P. K. Rachabatuni, F. Principi, P. Mazzanti, and M. Bertini, "Context-aware chatbot using MLLMs for Cultural Heritage," *MMSys 2024 - Proc. 2024 ACM Multimed. Syst. Conf.*, pp. 459–463, 2024, https://doi.org/10.1145/3625468.3652193.

[28] L. Xu, L. Lu, and M. Liu, "Construction and application of a knowledge graph-based question answering system for Nanjing Yunjin digital resources," *Herit. Sci.*, vol. 11, no. 1, pp. 1–17, 2023, https://doi.org/10.1186/S40494-023-01068-2/TABLES/6.
[29] T. L. M. Suryanto, A. P. Wibawa, H. Hariyono, and A. Nafalski, "Comparative Performance of Transformer Models for Cultural Heritage in NLP Tasks," *Adv. Sustain. Sci. Eng. Technol.*, vol. 7, no. 1, p. 0250115, 2025, https://doi.org/10.26877/asset.v7i1.1211.
[30] A. Shang, X. Zhu, M. Danner, and M. Rätsch, "Unsupervised question-retrieval approach based on topic keywords filtering and multi-task learning," *Comput. Speech Lang.*, vol. 87, p. 101644, 2024, https://doi.org/10.1016/j.csl.2024.101644.
[31] S. Pramanik, J. Alabi, R. S. Roy, and G. Weikum, "Uniqorn: Unified question answering over RDF knowledge graphs and natural language text," *J. Web Semant.*, vol. 83, p. 100833, 2024, https://doi.org/10.1016/j.websem.2024.100833.
[32] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *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. 3980–3990, 2019, https://doi.org/10.18653/v1/D19-1410.
[33] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," *EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 6894–6910, 2021, https://doi.org/10.18653/v1/2021.emnlp-main.552.
[34] J. A. Alzubi, R. Jain, A. Singh, P. Parwekar, and M. Gupta, "COBERT: COVID-19 Question Answering System Using BERT," *Arab. J. Sci. Eng.*, vol. 48, no. 8, pp. 11003–11013, 2023, https://doi.org/10.1007/S13369-021-05810-5/FIGURES/7.
[35] J. Yang *et al.*, "BERT and hierarchical cross attention-based question answering over bridge inspection knowledge graph," *Expert Syst. Appl.*, vol. 233, p. 120896, 2023, https://doi.org/10.1016/J.ESWA.2023.120896.
[36] R. Liu *et al.*, "Knowledge Enhanced Industrial Question-Answering Using Large Language Models," *Engineering*, 2025, https://doi.org/10.1016/j.eng.2025.07.035.
[37] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MINILM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 5776-5788, 2020, https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
[38] S. Zhang, E. Phan, P. Velmovitsky, Q. Pham, and S. Sanner, "Retrieval-Augmented Generation for Medical Question Answering on a Heart Failure Dataset: Performance Analysis," *JMIR Form. Res.*, vol. 10, 2026, https://doi.org/https://doi.org/10.2196/84932.
[39] G. Shidaganti, R. Shetty, T. Edara, P. Srinivas, and S. C. Tammineni, "Exploratory analysis on the natural language processing models for task specific purposes," *Bull. Electr. Eng. Informatics*, vol. 13, no. 2, pp. 1245–1255, 2024, https://doi.org/10.11591/eei.v13i2.6360.
[40] C. Clark, K. Lee, M. W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 2924–2936, 2019, https://doi.org/10.18653/v1/N19-1300.
[41] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 7871–7880, 2019, https://doi.org/10.18653/v1/2020.acl-main.703.
[42] J. Su, S. Yu, X. Ye, and D. Ma, "BERT-KRS: A BERT-Based Model for Knowledge-Grounded Response Selection in Retrieval-Based Chatbots," In *International Conference on Applied Intelligence*, pp. 310–321, 2024, https://doi.org/10.1007/978-981-97-0827-7_27.
[43] K. Peyton and S. Unnikrishnan, "A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs," *Results Eng.*, vol. 17, p. 100856, 2023, https://doi.org/10.1016/j.rineng.2022.100856.
[44] J. Staš, D. Hládek, and T. KOCTu, "Slovak Question Answering Dataset Based On The Machine Translation Of The Squad V2.0," *Jazykoved. Cas.*, vol. 74, no. 1, pp. 381–390, 2023, https://doi.org/10.2478/JAZCAS-2023-0054.
[45] V. K and A. Mishra, "Dataset for legal question answering system in the Indian judiciary context," *Data Br.*, vol. 60, p. 111647, 2025, https://doi.org/10.1016/j.dib.2025.111647.
[46] H. C. Wang, M. Maslim, and C. H. Kan, "A question–answer generation system for an asynchronous distance learning platform," *Educ. Inf. Technol.*, vol. 28, no. 9, pp. 12059–12088, 2023, https://doi.org/10.1007/S10639-023-11675-Y.
[47] R. Doi, T. Charoenporn, and V. Sornlertlamvanich, "Automatic Question Generation for Chatbot Development," *ICBIR 2022 - 2022 7th Int. Conf. Bus. Ind. Res. Proc.*, pp. 301–305, 2022, https://doi.org/10.1109/ICBIR54589.2022.9786384.
[48] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019, http://arxiv.org/abs/1910.01108.
[49] D. A. Sulistyo, D. D. Prasetya, F. A. Ahda, and A. P. Wibawa, "Pivoted Low Resource Multilingual Translation with NER Optimization," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 24, no. 5, pp. 1–16, 2025, https://doi.org/10.1145/3727876.
[50] H. Pires, L. Paucar, and J. P. Carvalho, "DeB3RTa: A Transformer-Based Model for the Portuguese Financial Domain," *Big Data Cogn. Comput.*, vol. 9, no. 3, pp. 1–30, 2025, https://doi.org/10.3390/bdcc9030051.

[51] S. Behmanesh, A. Talebpour, M. Shamsfard, and M. M. Jafari, "Improved relation span detection in question answering systems over extracted knowledge bases," *Expert Syst. Appl.*, vol. 224, p. 119973, 2023, https://doi.org/10.1016/j.eswa.2023.119973.
[52] M. Wang, Z. Li, X. Zhao, and Q. Guo, "Eliminate-Then-Select: A human-centric reasoning framework for educational question answering with LLMs," *Inf. Process. Manag.*, vol. 63, no. 2, p. 104422, 2026, https://doi.org/10.1016/j.ipm.2025.104422.
[53] Y. Choi, S. Kim, Y. C. F. Bassole, and Y. Sung, "Enhanced Retrieval-Augmented Generation Using Low-Rank Adaptation," *Appl. Sci.* vol. 15, no. 8, p. 4425, 2025, https://doi.org/10.3390/APP15084425.