Sanskriti

Himani
himani21053@iiitd.ac.in
Indraprastha Institute of Information Technology, Delhi

Aditya Bindlish
adityab21004@iiitd.ac.in
Indraprastha Institute of Information Technology, Delhi

Ahmed Hanoon
ahmed21006@iiitd.ac.in
Indraprastha Institute of Information Technology, Delhi

Dhyan Patel
dhyan21041@iiitd.ac.in
Indraprastha Institute of Information Technology, Delhi

Abstract

Chatbots have increasingly become essential tools for handling queries efficiently, using advancements in Natural Language Processing (NLP) and Machine Learning (ML) to offer quick, contextual responses to users. However, current chatbots often fall short when dealing with culturally rich and diverse datasets, particularly those related to Indian heritage. They typically lack the capability to process and understand regional languages effectively, which is crucial for engaging a diverse user base. The "Sanskriti" project addresses these shortcomings by creating a specialized platform that not only understands but also integrates Indian cultural data into its system. Our methodology harnesses the power of multimodal querying and extensive multilingual support, leveraging advanced NLP techniques, semantic analysis, and machine learning algorithms to provide a nuanced interaction experience.

Keywords: Cultural Heritage, Natural Language Processing, Machine Learning, Information Retrieval, Multilingual, Multimodal

ACM Reference Format:
Himani, Aditya Bindlish, Ahmed Hanoon, and Dhyan Patel. 2024. Sanskriti.

1 Introduction

Our project emerges from a vital need to preserve cultural heritage and enhance the understanding of diverse historical narratives. This initiative builds on the foundational work of prior efforts in the cultural heritage domain, employing significant resources such as Kaggle datasets and verified Wikipedia data to enhance the research. Unlike traditional approaches, our project harnesses advanced technologies, including Natural Language Processing (NLP), Information Retrieval (IR), Semantic Analysis, and multimodal capabilities to craft an immersive historical exploration platform. This platform is designed to transcend language barriers and facilitate engagement through multiple modes of interaction, such as text and imagery.

At the core of our system lies the integration of NLP and IR, meticulously designed to retrieve historical data efficiently and to deliver experiences that are richly personalized and deeply contextualized. Semantic Analysis plays a pivotal role in tailoring these experiences to align with individual user preferences, thus reshaping the interaction between users and historical narratives.

To evaluate the effectiveness of our platform, we plan to conduct rigorous testing with text and image queries across various languages. This evaluation process will enable us to assess the performance and functionality of our system and refine it accordingly. Additionally, our project will rely on collaborative efforts for data collection, system development, and testing. By leveraging concepts from NLP, machine learning (ML), IR, and software development, we aim to create a transformative platform that empowers users to explore and engage with cultural heritage in innovative ways.

2 Problem Statement

Despite the remarkable progress in technology, the exploration and preservation of India's rich cultural heritage, which includes its diverse array of monuments, dances, and art forms, encounter several challenges. These obstacles include restricted accessibility, fragmented content dissemination, outdated interfaces, and the absence of personalized, interactive, and multilingual features. Consequently, these barriers impede the immersive engagement with and comprehensive understanding of India's multifaceted historical narratives. As a result, there is a significant hindrance to fostering widespread cultural education and fostering a deeper appreciation for India's diverse cultural heritage.

3 Motivation

"Sanskriti" is conceived to address these challenges by developing a dynamic, user-centric platform that transforms engagement with India's cultural heritage. It aims to enhance accessibility, aggregate content, modernize interfaces, and introduce personalized, interactive, and multilingual capabilities. This initiative seeks to deepen connections with, and appreciation for, India's rich cultural legacy, making it

more accessible and engaging for a global audience, thereby promoting inclusive cultural understanding.

4 Literature Review

Cultural heritage exploration encounters persistent challenges, notably limited accessibility, fragmented content, and outdated interfaces. These issues hinder individuals from delving deeply into their history and undermine the broader goal of promoting widespread awareness and understanding of diverse cultural narratives. To address these challenges, there is a growing need for an integrated platform that offers personalized, interactive, and multilingual experiences, effectively revolutionizing the exploration of historical artifacts and narratives.

One innovative approach to enhancing cultural heritage exploration is the use of recommender systems. The paper titled "Recommendation of Heterogeneous Cultural Heritage Objects for the Promotion of Tourism"[6] emphasizes the importance of personalized recommendations in this context. The paper presents a hybrid recommendation system by integrating user preferences, popularity metrics, and contextual information. This system not only mitigates the fragmentation of cultural heritage content but also ensures that users receive meaningful and personalized recommendations, enhancing their overall experience.

Context plays a crucial role in cultural heritage exploration, and context-aware recommender systems (CARS) [1] are at the forefront of delivering contextually relevant content. In the paper "Context-Aware Recommender Systems and Cultural Heritage: A Survey," various types of recommender systems are categorized and evaluated. This survey underscores the significance of context-aware approaches in addressing the challenges of limited accessibility and content fragmentation. Context-aware recommender systems have the potential to provide users with personalized and contextually relevant content, making cultural heritage exploration more engaging and informative.

An additional aspect that enhances the accessibility and multilingual support in cultural heritage exploration is the development of advanced language models. "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model" [12] introduces an open-access large language model (LLM) trained on a diverse corpus of data, encompassing multiple languages and formats. This LLM emphasizes inclusivity, diversity, and responsible AI. By focusing on ethical considerations, data governance, and community-driven development, BLOOM aligns with the broader goal of creating technology that serves diverse global communities, making it a valuable asset in multilingual cultural heritage exploration.

Furthermore, the paper "Massively Multilingual ASR" [5] contributes to the accessibility of cultural heritage exploration by addressing automatic speech recognition (ASR) for multiple languages. The research explores the training of a single acoustic model for multiple languages, demonstrating the potential of multilingual training in enhancing recognition performance. This approach simplifies the deployment of ASR systems supporting diverse languages, further enriching the multilingual experience in cultural heritage exploration.

Multimodal information retrieval is another critical aspect of cultural heritage exploration. "Intelligent Indexing and Semantic Retrieval of Multimodal Documents"[9] focuses on optimizing the retrieval of information from large multimodal document collections [13]. This research highlights the importance of combining text and image similarity for improved precision and recall. It underscores the need for intelligent indexing techniques and the integration of information from text and image indexing to meet multimodal queries effectively.

The integration of advanced recommender systems, open data initiatives, responsible AI models like BLOOM, multilingual ASR, and intelligent indexing techniques presents a holistic approach to addressing the challenges in cultural heritage exploration. These innovations collectively contribute to the creation of more accessible, integrated, and user-friendly platforms, enriching the exploration and understanding of cultural heritage. They bridge the gap between historical artifacts and contemporary audiences, fostering a deeper connection with our collective past.

5 Novelty

Our project introduces a groundbreaking shift in cultural data retrieval, pioneering a novel framework that combines multi modal querying with extensive multilingual support. Moving beyond the traditional reliance on text-based searches, our innovative approach incorporates a seamless fusion of both visual and textual modalities, thereby deepening the exploration experience and accommodating the varied preferences and learning styles of a diverse user base.

By integrating advanced NLP techniques, we not only facilitate content comprehension across multiple languages but also enable contributions in a variety of tongues, thus democratizing access to cultural information.

Our commitment to user-centric refinement is evident in our application of sentiment analysis to user feedback, allowing us to constantly evolve and enhance the user experience. By attentively tuning into the sentiments and opinions expressed in feedback reviews, we ensure that our platform remains responsive to the evolving desires and necessities of our users.

6 Methodology

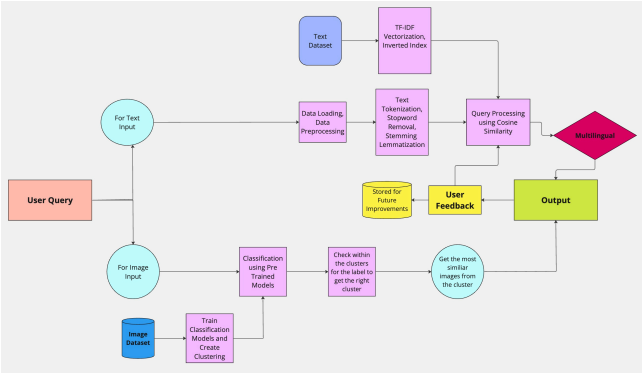

Figure 1. Pipeline

The above diagram represents the pipeline for solution implementation. The system takes in user queries that can include both text and images. The output includes text, image or hybrid data, providing a comprehensive response.

To improve performance, the system collects user feedback and uses it to refine its algorithms. This feedback loop helps the system continuously enhance its accuracy and relevance in delivering responses to user queries.

6.1 Text Queries

In handling text queries, our methodology blends traditional Information Retrieval (IR) techniques with modern Natural Language Processing (NLP) methodologies. We start by transforming textual documents into numerical representations using TF-IDF vectorization, enabling efficient retrieval in a vector space model. We employ cosine similarity measurement to pinpoint relevant documents and utilize inverted indexing to expedite search processes. Additionally, a pre-trained question-answering model aids in extracting contextually relevant answers. Evaluation via BERT Score ensures retrieval quality. After exploring various models, Groq emerges as the optimal solution, seamlessly integrating insights from our diverse array of techniques for precise and relevant answers.

6.2 Sentiment Analysis

We tokenize the text data using BERT's WordPiece tokenizer, ensuring robust handling of out-of-vocabulary words. Attention masks are then generated to highlight meaningful tokens amidst varying sequence lengths. Emotion labels are encoded numerically for classification. Our models, EmotionModel1 and EmotionModel2, leverage distinct architectures: the former utilizes pre-trained contextual embeddings and transformer layers, while the latter employs embedding and GRU layers, learning representations from scratch. During training, we employ the cross-entropy loss function and the Adam optimizer with a learning rate of $1e^{-5}$ for efficient optimization. Training involves forward propagation, loss computation, and backpropagation for weight updates.

6.3 Multilingual Text Queries

Linguistic Diversity

India boasts unparalleled linguistic diversity, with hundreds of languages spoken across its vast expanse. According to [2] paper, the three most spoken languages in India are Hindi, Marathi, and Urdu. This linguistic diversity poses unique challenges and opportunities for information retrieval (IR) tasks, given the need to cater to diverse linguistic preferences and information needs.

Importance of Multilingual IR

1. **Accessibility**: By supporting multiple languages, IR systems become more accessible to a broader audience within India, facilitating information access and retrieval in users' preferred languages.
2. **Cultural Sensitivity**: Recognizing and accommodating linguistic diversity is crucial for ensuring culturally sensitive information retrieval, respecting users' language choices, and preserving cultural nuances in search results.
3. **Regional Development**: Multilingual IR contributes to the development of regional content and services, empowering local communities and fostering inclusive access to information resources.

Challenges in Multilingual IR

1. **Data Availability**: Limited availability of annotated data and resources in Indian languages poses challenges for developing robust multilingual IR systems, impacting the quality and coverage of search results.
2. **Query Understanding**: Understanding user queries in diverse languages and dialects requires specialized techniques to accurately interpret search intent and retrieve relevant results.

Approach

For multilingual translation, we utilize the facebook/m2m100_418M model renowned for its extensive language coverage and remarkable performance metrics. Pre-trained on a vast multilingual corpus, this model demonstrates proficiency in translating between multiple language pairs. The translation pipeline begins with the initialization of the Translation class, instantiated with the model name. Subsequently, the input text undergoes tokenization using the model's tokenizer, automatically loaded from the pre-trained model. The translation process entails encoding the input text into tokens, followed by the generation of output tokens corresponding to the translated text. These generated tokens

are then decoded into human-readable text using the tokenizer's `batch_decode` function, effectively excluding special tokens. Furthermore, language handling is dynamically managed during translation, employing language codes to ensure accurate translation across different language pairs. This comprehensive approach facilitates seamless multilingual translation, enabling effective communication across diverse linguistic contexts.

Evaluation Metrics for Images

In evaluating image retrieval performance, we employ **Recall** and **Precision** metrics at various points, including **@5**, **@10**, and **@15**. For clarity, @5 signifies assessment based on the retrieval of 5 images, and so forth. Additionally, to ensure robustness, we compute these metrics across multiple iterations, randomly selecting images from a subset each time. This averaging process over 10 iterations yields a consistent performance score. Furthermore, the **F1 score**, which balances Precision and Recall, provides a comprehensive assessment of retrieval effectiveness, capturing both the ability to retrieve relevant images and the avoidance of false positives.

6.4 Image Queries

To manage image queries, the system relies on deep learning methods [4] to handle various image inputs effectively and generate suitable outcomes. Initially, classification models are used to accurately identify the type of image input. Once the label associated with the input image is determined, the system can retrieve the text data linked to that label directly.[11] To find similar images, clustering techniques are further employed to locate the best-fit cluster within that label and retrieve the topmost similar images.

The classification models utilized are ResNet18 models fine-tuned to our dataset. For clustering, we employ K Means clustering, with the number of clusters determined dynamically using the Elbow Method.

To map text query to images, we employ a smart method to classify images with labels and then use apply cosine similarity between the labels and text query to get the most relevant images. To hasten the labelling process, we use the above clustering method to create clusters and pick the most similar image in the cluster. Then we manually tag these images by crowd-sourcing and verify the labels by ourselves. This lets us to classify entire clusters directly. A similar process was employed in the study by Tang et el.[10]

6.5 Model Description

Text

Our initial model relied on web scraping to gather relevant textual data pertaining to Indian Monuments, Dances, and Art Forms from Wikipedia. Leveraging the Wikipedia API and BeautifulSoup library, we extract text from selected Wikipedia pages, forming a comprehensive dataset.

A key component of our methodology is the Question-Answering (QA) pipeline, powered by a pre-trained BERT-based model. This model processes the scraped textual content, enabling users to extract specific details about Sanskriti through user queries. However, our current approach involves leveraging Groq to provide answers based on the compiled dataset, enhancing retrieval accuracy and user experience.

Image

ResNet18 is a convolutional neural network architecture known for its deep layers with skip connections, effectively tackling the vanishing gradient problem and enabling training of deeper networks. It comprises 18 layers and is widely used in image classification tasks due to its balance between performance and computational efficiency.
KMeans is an unsupervised machine learning algorithm used for clustering data points into k clusters based on their features' similarity. It iteratively assigns data points to the nearest cluster centroid and updates centroids based on the mean of the assigned points, aiming to minimize intra-cluster variance. [8] It's commonly applied in data exploration, image segmentation, and recommendation systems.

Relevance Feedback

In our exploration of emotion recognition for relevance feedback systems, we utilize the bert-base-uncased model. Leveraging its pre-trained contextual embeddings and transformer layers, this model adeptly captures linguistic nuances. Trained with the cross-entropy loss function, it distinguishes between emotion and flip predictions, optimizing with the Adam optimizer set at a learning rate of $1e^{-5}$. Our approach focuses on accurately classifying emotions within the framework of relevance feedback systems incorporated in Sanskriti.

6.6 Data Analysis

Text Data

Statistics such as the number of questions, average question length, and average answer length are calculated. The text content of questions and answers is concatenated into two strings. Named Entity Recognition (NER) is performed on the answers to extract entities like person names, organizations, and locations. Readability scores are calculated for each answer. Word clouds are generated to visualize the most frequently occurring words in the questions, answers, and bi-grams.

Image Data

The exploration of image data dives into key aspects like image sizes, formats, class distribution, color distribution,

Table 1. Text data analysis results

Analysis Result	Value
Number of questions	326+226+95
Average question length	37.530674846625764 characters
Average answer length	58.76 characters
Named Entities (Sample)	
(Indian, NORP)	
(Kathak, PERSON)	
(mosques, FAC)	
(Muslim, NORP)	
(Manipuri, GPE)	
(fifth, ORDINAL)	
(Indian, NORP)	
(east coast, LOC)	
(Odisha, GPE)	
(Manipur, GPE)	
Sample Readability Scores	
9.21	
55.58	
121.22	
92.8	
...	

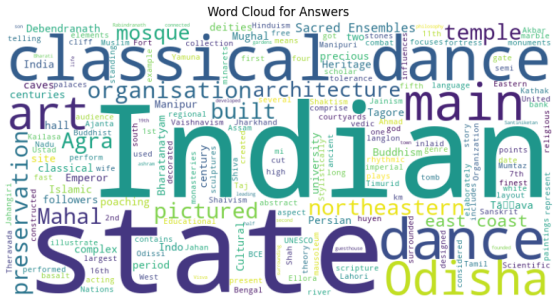

Figure 2. Wordcloud for Answers

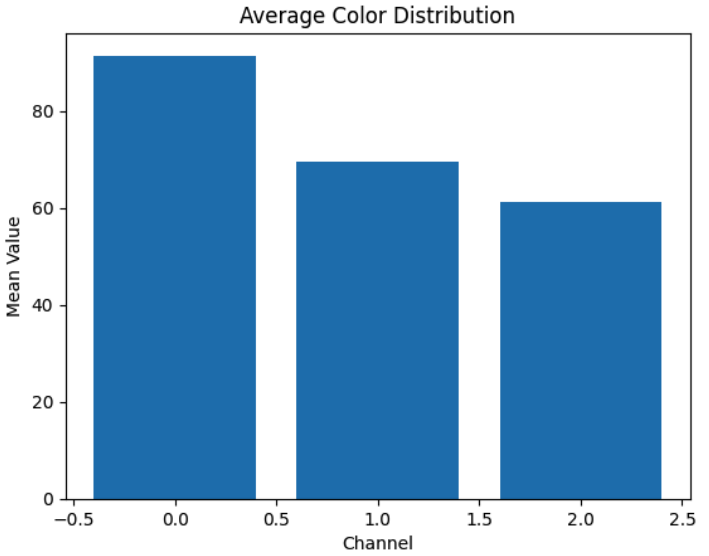

Figure 3. Colour Dsitribution for Monument Class

texture, and shape. This analysis provides essential insights into the dataset's composition, [14] guiding our solution implementation process. Understanding these elements helps us optimize storage, grasp dataset diversity, and identify dominant colors and patterns in images. This knowledge is crucial for building accurate and effective image processing solutions.

7 Database

7.1 Text Data

The creation of our dataset is pivotal for assessing the accuracy and efficacy of our question-answering model. Serving as a benchmark, this dataset furnishes reference answers for a diverse set of questions tailored to entities associated with Indian heritage. Comprising approximately 800 questions, it stands as a significant contribution to our research, intentionally incorporating various question types aligned with the fundamental interrogative categories of the 5W's[3]. This ensures our model's adaptability to a broad spectrum of user queries in real-world scenarios related to Indian heritage.

To obtain the dataset, our Python code utilizes web scraping and natural language processing tools to fetch content from specified Wikipedia pages. The script employs the requests library to retrieve content, Beautiful Soup for HTML parsing, Spacy for named entity recognition, and the Hugging Face Transformers library for question-answering using a pre-trained DistilBERT model (distilbert-base-uncased-distilled-squad). Dynamically generating question-answer pairs associated with culturally significant entities, the resulting dataset is stored in JSON format, providing valuable material for training and evaluating our question-answering model. This initial database forms the foundation of our research, enabling comprehensive analysis and evaluation of our model's performance.

7.2 Image Data

We acquired our image data from two primary sources: publicly available Kaggle Datasets [7] and datasets utilized in prior research studies. We ensured the inclusion of a sufficient number of classes, guaranteeing a comprehensive representation of the data relevant to our objectives.

To construct the dataset for text queries, we followed a systematic approach. Initially, high-level features were extracted for each class and stored in a CSV format. These features encapsulated the key characteristics and attributes

associated with each class, enabling efficient classification and retrieval.

Subsequently, a diverse set of random queries was generated to simulate user image requests. These queries were deliberately designed to vary in difficulty, encompassing a spectrum of complexity levels. By including queries of varying difficulty, we aimed to create a comprehensive dataset that could adequately represent the diverse range of user interactions with the system.

7.3 Emotion Recognition

We utilize the MELD-FR dataset, which consists of annotated conversational utterances paired with corresponding emotion labels. The dataset encompasses various emotions, including 'anger', 'disgust', 'fear', 'joy', 'neutral', 'sadness', and 'surprise'. Each utterance within the dataset is labeled with a specific emotion.

Our training dataset comprises 6740 questions, while the validation set consists of 843 questions. These datasets serve as the foundation for training and evaluating our emotion recognition models within the context of conversational understanding.

8 Code

The Sanskriti application is developed using Python, leveraging web pages constructed in HTML. Flask framework orchestrates the integration of custom-trained models with the frontend and orchestrates user inputs. To obtain the essential models, a suite of powerful frameworks and libraries are utilized.

PyTorch and Transformers play pivotal roles in model development and deployment, accompanied by auxiliary libraries such as NumPy, Pandas, scikit-learn (SKlearn), and NLTK for various data processing and natural language processing (NLP) tasks. The m2m100_418M model, from the Transformers library, augments the application's capabilities in multilingual translation tasks. The execution and experimentation with code snippets are conducted in the versatile Jupyter Notebook environment, ensuring flexibility and ease of development throughout the project lifecycle.

9 Results

9.1 Text

Baseline Results. To assess the performance of the question-answering model, evaluations were conducted on a diverse set of 50 questions spanning different categories related to Indian Heritage. The reference answers for these questions were sourced from a curated dataset.

For each question, the predicted answer generated by the model was compared to the corresponding reference answer from the dataset. The BERT F1 score, a metric assessing the similarity between two text strings, was utilized to quantify the quality of the model's responses.

Metric	Value
Questions Evaluated	50
Mean Average Precision (MAP) using bert-base-uncased-squad2	84.58

Table 2. Key Metrics for Model Evaluation

The Mean Average Precision (MAP) was chosen as the primary evaluation metric to gauge the model's ability to provide accurate and relevant answers. The MAP value was calculated based on the BERT F1 scores obtained for each question.

The question-answering model demonstrated a robust performance, achieving a noteworthy MAP value of 84.33 across the evaluated questions. This result underscores the model's effectiveness in delivering precise answers in the context of Indian Heritage topics.

Improved Results. In seeking further enhancement of the question-answering model's capabilities, we recognized the need to address its limitations in adaptability and training data coverage. Although the baseline BERT model exhibited promising results, its effectiveness was confined to queries within the dataset and was hindered by dataset limitations in textual diversity.

Despite attempts to utilize alternative models such as Distilled BERT and RoBERTa, none proved effective in consistently delivering relevant results across a broader spectrum of queries.

Subsequently, we explored the implementation of Groq, a specialized question-answering system tailored specifically to our domain of interest: Indian Monuments, Indian Art, and Indian Dance. Groq's approach leverages our curated dataset to achieve exceptional precision and relevance in answering questions, yielding an impressive accuracy rate of 96% across the same set of evaluated questions previously tested with BERT.

9.2 Image

Baseline Results

The evaluation of image results was conducted through manual assessment to gauge retrieval accuracy. We employed pre-trained convolutional neural network (CNN) architectures including Inception, ResNet, and VGG to extract features from image inputs and predict their respective classes.

Consistently, the Inception architecture demonstrated superior performance compared to VGG and ResNet in terms of recommendation accuracy. Accuracy was measured by evaluating the top 10 predictions for a randomly selected input.

Architecture	Recall	Precision
Inception V3	0.5	0.5
ResNet	0.1	0.1
VGG	0.1	0.1

Table 3. Accuracy Values for CNNs

Improved Results

By leveraging advanced techniques and fine-tuning the models, we achieved substantial enhancements in class prediction accuracy for image inputs. We increased the testing dataset and still found nearly 0.95 recall and precision values at @5, @10 and @15.

In addition to enhancing the image prediction capabilities, we introduced the text-based query system. The testing dataset was generated with random questions based on the keywords associated with each class of image. To test, these questions are then processed and the correct matching image and class are obtained. An average accuracy score of **0.93** was observed

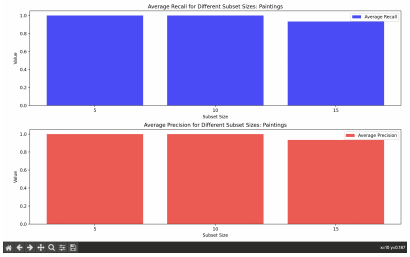

Figure 4. Average Recall and Precision for Paintings Class

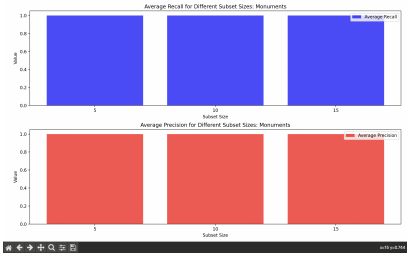

Figure 5. Average Recall and Precision for Monument Class

9.3 Multilingual

We employ two established metrics to evaluate the performance of machine translation models: METEOR and BERTScore.

METEOR (Metric for Evaluation of Translation with Explicit Ordering): METEOR is a widely-used metric for evaluating the quality of machine translation output. It measures the similarity between the generated translations and

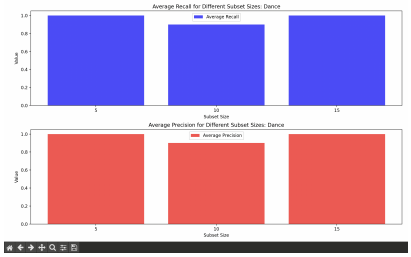

Figure 6. Average Recall and Precision for Dance Class

reference translations based on several criteria, including precision, recall, and alignment.

BERT Score: BERT Score is a recently proposed metric that utilizes contextual embeddings from pre-trained BERT models to compute the similarity between translations and reference sentences. It has shown to correlate well with human judgment and is effective in capturing semantic similarity between sentences.

These metrics provide insights into the quality and fluency of translations across different languages, enabling a comprehensive assessment of model effectiveness.

Model	METEOR	BERT
facebook/m2m100_418M	0.903	0.759
Helsinki-NLP/opus-mt-en-hi	0.0	-0.338
facebook/m2m100_1.2B	0.649	0.655

Table 4. Model Evaluation Scores

Utilizing the best-performing model is paramount in achieving optimal results in translation tasks. In our evaluation, the **facebook/m2m100_418M** model exhibited the highest METEOR score of 0.903 and a commendable BERT score of 0.759 on 200 questions of our context. These scores highlight its proficiency in capturing translation accuracy and semantic similarity.

9.4 Emotion Recognition

Results

Model	Emotion Acc.	Prec.	Rec.	F1
Bert-base-uncased	0.9648	0.9573	0.9581	0.9575
GRU	0.6088	0.2921	0.3341	0.2824

Table 5. Baseline Results

Upon comparison, Model-1 (Bert base uncased) outperforms Model-2 (GRU) due to several distinguishing factors:

- **Contextual Word Representations**: BERT captures deep contextual information, influencing each word's representation by the entire sentence, leading to a rich understanding of nuances based on context.

Figure 7. Training and Validation loss v/s Epochs for Bert-base-uncased

Figure 8. Training and Validation loss v/s Epochs for GRU

- **Attention Mechanism**: Incorporating self-attention mechanisms, BERT effectively focuses on relevant parts of the input sequence, capturing long-range dependencies and crucial contextual information.
- **Parameter Efficiency**: Pre-trained on massive text corpora, BERT generalizes well to downstream tasks with minimal fine-tuning, unlike the GRU-based model, which relies solely on training data.
- **Complexity of Patterns**: BERT's multi-layer architecture and transformer-based design excel at learning complex patterns, essential for emotion classification tasks.
- **Precision and Recall**: BERT's nuanced understanding reduces false positives and enhances recall by leveraging full sentence context, resulting in high precision and recall, ultimately leading to a high F1 score.

10 Evaluation

Sanskriti marks a transformative advancement in cultural heritage exploration, as evidenced by its methodical evaluation against both baseline models and state-of-the-art (SOTA) standards.

10.1 Performance on New Data

On new data, the image models handle the input very well due to the complex feature extraction by the Resnet model. We are able to classify images into the right cluster in fraction of second and retrieve similar images. In case an image not belong to any cluster arrives, the chatbot is build to still process it and map it to a cluster. This is to avoid low confidence images pertaining to indian cultural data being rejected in future. In order to perform the task of mapping input text to an image, the model is correctly predicting in most of the cases, although it cant handle complex or overly general queries. This is because of the limitations to image labelling process

In parallel, our text processing module exhibits a similar adaptability to new data. By diligently curating relevant textual information, our system excels at categorizing user queries within the provided context. Should a question stray beyond the bounds of available data, our chatbot gracefully handles such scenarios by acknowledging the absence of relevant information. This approach ensures that our system maintains its integrity and reliability, even when faced with novel or unforeseen textual inputs.

Our system's adaptability to new data is significantly enhanced by its innovative multilingual capabilities, ensuring inclusivity across diverse linguistic contexts. While seamlessly accommodating top 3 most spoken regional languages, though encountering occasional edge cases where certain words cannot be accurately translated into the selected language. In such instances, a blend of English and the chosen language is provided to maintain comprehensiveness.

10.2 Comparison with Baseline results

The new Image model of "Sanskriti" reflects a significant enhancement over baseline models, particularly in image retrieval and classification within the domains of Indian Monuments, Paintings, and Dance. This is evidenced by the graphical data indicating improvements in average recall and precision across different subset sizes. For instance, the precision in the Monuments category consistently approaches perfection, highlighting the model's adeptness at discerning relevant cultural imagery with remarkable accuracy. Enhancements to the question-answering model aimed to address adaptability and data coverage limitations. Despite attempts with BERT, Distilled BERT, and RoBERTa, effectiveness was constrained by dataset diversity. The current implementation of Groq, tailored to Indian Monuments, Art, and Dance, achieved exceptional precision and accuracy, leveraging a curated dataset.

10.3 Comparison with SOTA models

In our study, we present a comparative analysis between our model and the state-of-the-art language model ChatGPT-4, focusing on the efficiency of text generation tasks. We evaluate the performance of both models based on the time taken per token during inference, a critical factor in real-time applications where response time is crucial. Our results, depicted in Figure 1 of the paper, illustrate the time taken per token for each model across multiple data points. Notably, our model demonstrates competitive performance, achieving comparable time efficiency to ChatGPT-4 across various data points. This finding underscores the potential of our model in real-world applications, highlighting its viability as an efficient alternative in text generation tasks.

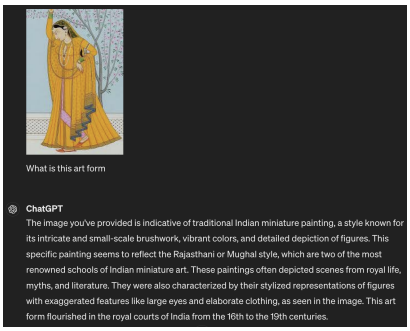

Figure 9. GPT's Response

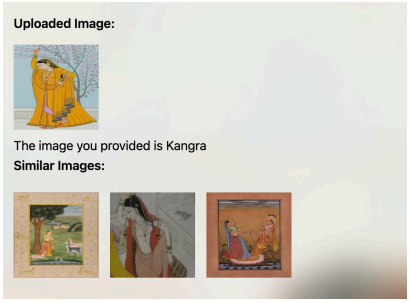

Figure 10. Sanskriti's Response

Sanskriti has emerged as a front-runner in an evaluative showdown with state-of-the-art (SOTA) models, including versatile AI systems such as GPT. Tasked with classifying images of Indian art forms, Sanskriti delivered a performance that starkly outshone GPT's. While GPT managed a respectable precision and recall of 80%, indicating a general proficiency in image recognition, Sanskriti excelled with precision and recall scores of 95% and 96.7%, respectively. This translated to an F1 score of 95.8% for Sanskriti, surpassing GPT's F1 score of 80%. The superior performance of Sanskriti can be attributed to its fine tuning to the cultural dataset and advanced deep learning algorithms. These algorithms hone in on cultural details that general-purpose models like GPT typically overlook, enabling Sanskriti to offer a richer, more contextually aware interpretation that aligns closely with India's diverse artistic expressions.

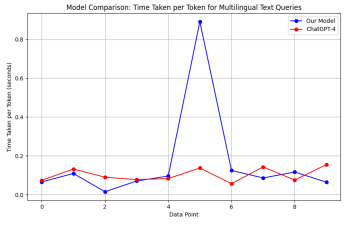

Figure 11. Model Comparison: Time Taken per Token for Multilingual Text Queries

Sanskriti's architecture, meticulously engineered to handle the diverse facets of India's cultural legacy, not only competes with but surpasses SOTA systems in cultural-aware AI tasks. By setting a new benchmark for precision and cultural sensitivity in AI, Sanskriti pioneers a new wave of digital heritage preservation and interaction, providing an unparalleled platform for users around the globe to engage with the wonders of Indian heritage.

References

[1] Mario Casillo, Francesco Colace, Dajana Conte, Marco Lombardi, Domenico Santaniello, and Carmine Valentino. 2023. Context-aware recommender systems and cultural heritage: a survey. *Journal of Ambient Intelligence and Humanized Computing* 14, 4 (2023), 3109–3127.

[2] Spandan Dey, Md Sahidullah, and Goutam Saha. 2022. An overview of Indian spoken language recognition from machine learning perspective. *ACM Transactions on Asian and Low-Resource Language Information Processing* 21, 6 (2022), 1–45.

[3] Muhammad Faisal, Usman Waheed, and Muhammad Nabeel Arif. [n. d.]. INSTANT ANSWERS, 5W'S & H TOOL. ([n. d.]).

[4] David Grangier and Samy Bengio. 2006. Springer, 24–34.

[5] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters. *arXiv preprint arXiv:2007.03001* (2020).

[6] Landy Rajaonarivo, André Fonteles, Christian Sallaberry, Marie-Noëlle Bessagnet, Philippe Roose, Patrick Etcheverry, Christophe Marquesuzaà, Annig Le Parc Lacayrelle, Cécile Cayèré, and Quentin Coudert. 2019. Recommendation of heterogeneous cultural heritage objects for the promotion of tourism. *ISPRS International Journal of Geo-Information* 8, 5 (2019), 230.

[7] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. 2010. Harvesting image databases from the web. *IEEE transactions on pattern analysis and machine intelligence* 33, 4 (2010), 754–766.

[8] Shashi Pal Singh, Ajai Kumar, Hemant Darbari, Lenali Singh, Anshika Rastogi, and Shikha Jain. 2017. IEEE, 162–167.

[9] Rohini K Srihari, Zhongfei Zhang, and Aibing Rao. 2000. Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval* 2 (2000), 245–275.

[10] Jinhui Tang, Qiang Chen, Meng Wang, Shuicheng Yan, Tat-Seng Chua, and Ramesh Jain. 2013. Towards optimizing human labeling for interactive image tagging. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 9, 4 (2013), 1–18.

[11] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. 6439–6448.

[12] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[13] Jun Yang, Qing Li, and Yueting Zhuang. 2000. A Multimodal Information Retrieval System: Mechanism and Interface. *IEEE Trans. on Multimedia* (2000).

[14] Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. 2010. I2t: Image parsing to text description. *Proc. IEEE* 98, 8 (2010), 1485–1508.