LoFTI: Localization and Factuality Transfer to Indian Locales

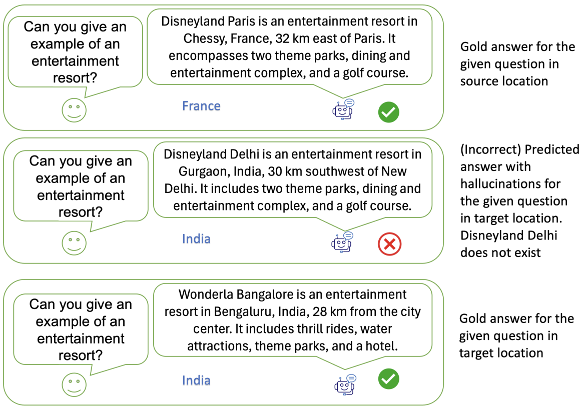
Anonymous ACL submission

Abstract

Large language models (LLMs) encode vast amounts of world knowledge acquired via training on large web-scale datasets crawled from the internet. However, the datasets used to train the LLMs typically exhibit a geographical bias towards English-speaking Western countries. This results in LLMs producing biased or hallucinated responses to queries that require answers localized to other geographical regions. In this work, we introduce a new benchmark named LoFTI (Localization and Factuality Transfer to Indian Locales) that can be used to evaluate an LLM's contextual localization and factual text transfer capabilities. LoFTI consists of factual statements about entities in source and target locations; the source locations are spread across the globe and the target locations are all within India with varying degrees of hyperlocality (country, states, cities). The entities span a wide variety of categories. We use LoFTI to evaluate Mixtral, Llama3.3-70B, GPT-4 and two other Mixtral-based approaches well-suited to the task of localized factual transfer. We demonstrate that LoFTI is a high-quality evaluation benchmark and all the models, including GPT-4, produce skewed results across varying levels of hyperlocality.

1 Introduction

Large language models (LLMs) are proficient in text generation and are also extensive repositories of world knowledge, owing to their pretraining and fine-tuning on vast and diverse internet data. This suggests that LLMs might be effective at associating and transferring factual knowledge across geographical locations. They can generate localized text in a given target location by transferring from a reference text in a source location. However, *there is no existing benchmark that helps assess this specific form of localization and fact-driven transfer*. Benchmarks that measure LLMs' ability to understand cultural concepts and their transference across geographical regions are steadily emerging in recent work (Li et al., 2024a,c,b; Rao et al., 2024). We argue that it is also important to evaluate the ability of models to transfer factual knowledge across geographical regions. Figure 1 illustrates this point by showing two use-cases: 1) Generating a localized response given a common question asked across locations and, 2) accurate factuality transfer from one locale to another.



Can you give an example of an entertainment resort?

Disneyland Paris is an entertainment resort in Chessy, France, 32 km east of Paris. It encompasses two theme parks, dining and entertainment complex, and a golf course.

Gold answer for the given question in source location

Can you give an example of an entertainment resort?

Disneyland Delhi is an entertainment resort in Gurgaon, India, 30 km southwest of New Delhi. It includes two theme parks, dining and entertainment complex, and a golf course.

(Incorrect) Predicted answer with hallucinations for the given question in target location. Disneyland Delhi does not exist

Can you give an example of an entertainment resort?

Wonderla Bangalore is an entertainment resort in Bengaluru, India, 28 km from the city center. It includes thrill rides, water attractions, theme parks, and a hotel.

Gold answer for the given question in target location

Figure 1: Illustrating localization and factual transfer.

In this work, we introduce a new evaluation benchmark called **LoFTI (Localization and Factuality Transfer for Indian Locales)**:

- It contains factual statements in source and target locations involving source and target entities. The entities span diverse categories such as food, sports, etc.
- The statements are grounded in various source locations across the globe, while all the target locations are in India.
- The target locations are at different levels of hyperlocality namely specific to India as a whole, or specific to states and cities within India.
- Each parallel pair is accompanied by common questions that can be answered at any location.

Motivation. A key motivation behind the LoFTI benchmark arises from the need to evaluate how well LLMs understand and transfer entity-specific knowledge across different cultural and regional contexts. Training datasets of large LLMs are typically dominated by Western entities and facts (Manvi et al., 2024). The challenge lies in how well these models can adapt their understanding of entities and factual statements to align with localized contexts, even when such entities are not explicitly represented in their context. For example, consider a dataset with instructions involving a customer and a cafe owner with the answer containing references to breakfast items such as waffles or pancakes. An LLM trained on such instructions many not accurately localize to regions in India where the appropriate answer might involve dosas or parathas. This is an example of the inability of an LLM to contextualize entity-specific behaviour appropriately. The issue becomes even more pronounced in factual queries. For instance, if a user asks about an "entertainment resort," the model might default to referencing globally prominent examples like Disneyland, even when the regional context strongly implies the need for a local reference, such as Ramoji Film City in India.

India provides a particularly valuable testbed for exploring these challenges due to its linguistic and cultural diversity. Our benchmark focuses on English as the medium for evaluation: our goal is to test how well English-centric LLMs can understand and transfer contextual relationships between entities localized to India, before addressing additional challenges posed by multilingual inputs.

Another important aspect of factual entity transfer is recognizing which characteristics of an entity should remain invariant and which require adaptation during localization. If a model's pre-training lacks sufficient representation of localized entities, it struggles to make these distinctions. Prompt engineering or fine-tuning cannot fully compensate for such gaps because the model lacks a foundational understanding of localized entities and their contextual relevance. External tools, such as APIs or retrieval-augmented generation (RAG), can partly address this issue by fetching relevant information from localized databases or the Web. However, such solutions introduce dependencies on the availability and accuracy of external data sources. More importantly, they do not address the core issue: the internal understanding of localized entities by the model and their contextual significance. If a model fundamentally misunderstands what an entity represents in a cultural or regional context, even accurate external information can be misinterpreted or poorly integrated into the response.

This underscores the importance of creating benchmarks that evaluate the usage ability of a model or tool-based technique to transfer entity-specific factual knowledge across localized contexts. By focusing on Indian entities within an English framework, our LoFTI benchmark addresses critical gaps in the understanding of regional nuances of current models. While our current focus is limited to India, we release all the details of our data creation pipeline to encourage similar efforts for other regions, enabling a broader evaluation of how LLMs localize and contextualize entity-specific information globally.

We define two tasks using LoFTI: Localized text transfer and localized question answering. Both tasks involve prompt-driven factual transfer to target locations with access to factual statements from source locations. We define three different metrics to evaluate the quality of both localization and factuality transfer on LoFTI. We evaluate the performance of powerful open-source (Mixtral, Llama3.3-70B) and closed-source models (GPT-4) on LoFTI. We also develop **two new retrieval-augmented approaches (using Mixtral)** that leverage external sources of evidence to significantly improve performance on all three metrics. While GPT-4 is expectedly superior in performance compared to Llama3.3-70B and all Mixtral variants, it shows degradation in performance across target locations of varying hyperlocality, thus revealing clear gaps in coverage across geographical regions. We publicly release LoFTI under the Apache 2.0 license.[1]

2 Methodology for Dataset Creation

Figure 2 illustrates the overall dataset creation pipeline with the help of an example.

Generation of Entity-pairs. For dataset creation, we compile pairs of entities $(e_{\text{ref}}, e_{\text{tar}})$, where e_{ref} is an entity from a reference location outside India and e_{tar} is an entity from India that serves as a suitable substitute for e_{ref}. These pairs are curated by human annotators to cover diverse categories and hyperlocal regions.

Reference Text Generation. Given the reference entity e_{ref}, a fact-based reference text T_{ref} is obtained from the entity's description derived using either the Google API Client or Wikipedia. If no en-

[1] The LoFTI dataset and codebase are available at: https://anonymous.4open.science/r/LoFTI-FC32

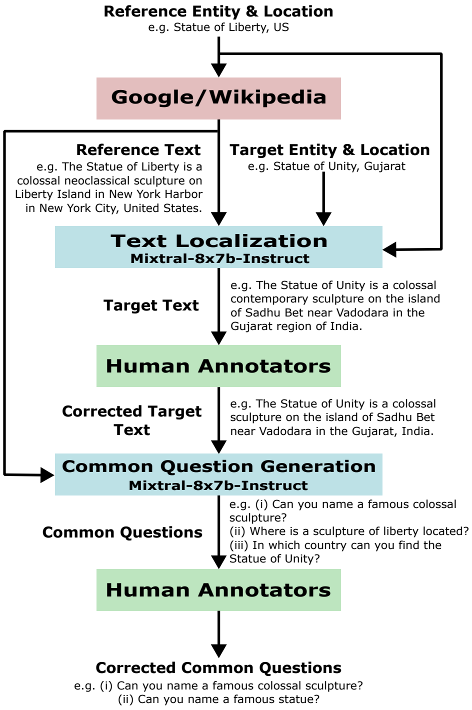

Reference Entity & Location
e.g. Statue of Liberty, US
Google/Wikipedia
Reference Text
e.g. The Statue of Liberty is a colossal neoclassical sculpture on Liberty Island in New York Harbor in New York City, United States.
Target Entity & Location
e.g. Statue of Unity, Gujarat
Text Localization
Mixtral-8x7b-Instruct
Target Text
e.g. The Statue of Unity is a colossal contemporary sculpture on the island of Sadhu Bet near Vadodara in the Gujarat region of India.
Human Annotators
Corrected Target Text
e.g. The Statue of Unity is a colossal sculpture on the island of Sadhu Bet near Vadodara in the Gujarat, India.
Common Question Generation
Mixtral-8x7b-Instruct
Common Questions
e.g. (i) Can you name a famous colossal sculpture?
(ii) Where is a sculpture of liberty located?
(iii) In which country can you find the Statue of Unity?
Human Annotators
Corrected Common Questions
e.g. (i) Can you name a famous colossal sculpture?
(ii) Can you name a famous statue?

Figure 2: Illustration of our dataset creation pipeline.

tity description is found from these sources, human annotators provide the reference text.

Text Localization. Given a reference text T_{ref} and a target entity e_{tar} (paired with e_{ref}) from a target location L_{tar}, text localization aims to generate a target text T_{tar} accurately localized to L_{tar} that retains stylistic and semantic features of T_{ref}. Given L_{tar}, e_{tar}, and T_{ref}, we prompt the *Mixtral-8x7b-instruct-v0.1.Q4_K_M* model (using the prompt shown in Appendix A.2) to generate the localized target text T_{tar}.

Common Question Generation. In addition to the reference and target text pairs, LoFTI also contains questions that capture common aspects shared by T_{ref} and T_{tar}. These common questions are ones a user would ask the model from a particular location for which answers should be appropriately localized. Given a pair of text $(T_{\text{ref}}, T_{\text{tar}})$, we generate these questions by identifying shared properties or descriptions of the entities mentioned in the text pairs. We use few-shot prompting on *Mixtral-8x7b-instruct-v0.1.Q4_K_M* model for common question generation, using the prompt in Appendix A.2.

Human Annotators. All the generations at each stage were carefully verified by three human annotators employed by an annotation company. The annotators represent diverse demographics within India and are knowledgeable about different geographic and hyperlocal regions. Guidelines used by the human annotators, the annotation process, inter-annotator agreement and overall costs are detailed in Appendix A.1.

3 Properties of LoFTI Dataset

Figure 3 shows the distribution of entities across reference and target locations.[2] Table 1 presents salient statistics of LoFTI and an example with all its metadata detailed below.

- **Region**: The region of the reference location.
- **Category**: The category of the entity.
- **Reference Location**: A non-Indian location.
- **Reference Entity**: An entity specific to the reference location.
- **Reference Text**: Factual statement about the reference entity.
- **Target Location**: A location in India.
- **True Target Entity**: An example of a correct localization of the reference entity in the target location.
- **True Target Text**: A localized factual statement about the true target entity.
- **Hyperlocal Score**: The degree of hyperlocality within India. LoFTI includes three hyperlocality scores: 1, 2, and 3 that correspond to the target locations 'India,' 'any state in India,' and 'any city in India,' respectively.
- **High Cardinality**: This is a yes/no field denoting the potential count of entities that can be localized to the target location. "Yes" means there are more than 3 replaceable entities.
- **Common Questions**: Questions extracted from the reference and target texts. An appropriate answer to this question will depend on the location where it is asked.

Category Distribution. The dataset consists of 99 unique categories grouped into 10 domains namely Entertainment, Buildings/Monuments/-Companies, Food & Lifestyle, Professions, Nature, Finance & Economy, Sports, Incidents, Places & Landmarks, and Others. More details about the category distribution are in Appendix A.3.

[2] Our source entities are mainly from US/Europe, as we want the reference entities to be drawn from countries that are fairly well-represented by popular pretraining corpora (e.g. CommonCrawl).

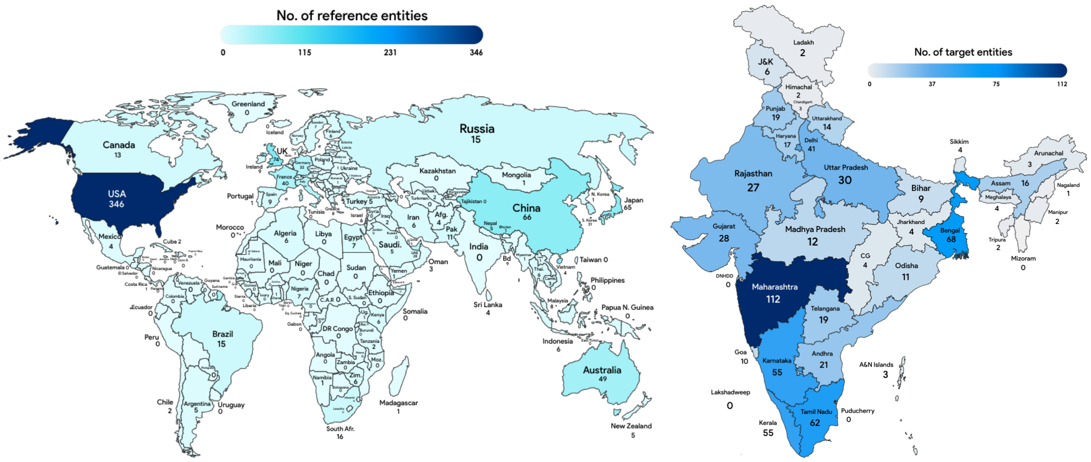


Figure 3: Illustrates the global distribution of the reference entities and the spread of target entities in India.

LoFTI Dataset Details					
No. of entity pairs	1100	No. of entities from US/Europe	651	No. entities with high cardinality	835
No. of categories	99	No. of entity pairs from other places	449	No. entities with low cardinality	265
No. of entities with hyperlocal score = 1	369	No. of entities with hyperlocal score = 2	391	No. of entities with hyperlocal score = 3	340

Example:

Region	**Category**	**Reference Location**	**Reference Entity**	**Reference Text**	**Target Location**	**High Cardinality**	**Hyperlocal Score**	**Target Entity**	**Target Text**	**Common Questions**
US/ Europe	Monument	US	Statue of Liberty	The Statue of Liberty is a colossal neoclassical sculpture on Liberty Island in New York Harbor in New York City, United States.	Gujarat	Yes	2	Statue of Unity	The Statue of Unity is a colossal sculpture on the island of Sadhu Bet near Vadodara in the Gujarat, India.	(i) Can you name a famous colossal sculpture? (ii) Can you name a famous statue?

Table 1: The statistics of LoFTI dataset and an example with all its metadata.

4 Evaluation Metrics

4.1 Entity Correctness

To evaluate entity correctness of a generated target text, the human annotator checks if the entity present in the target text is correctly localized to a target location given the reference entity in the reference text. Note that there can be multiple correct localized entities for a given target location. If the entity localization is correct, then a score of 1 is assigned, else it is 0. If this binary score assigned to each target text T^i is denoted as E_i, then the entity correctness metric is computed as $\mathbf{EC} = \frac{1}{N}\sum_{i=1}^{N} E^i$.

4.2 Common Question Correctness

For each target text with EC $= 1$, common questions are further used to evaluate the localization capability of the model. Human evaluators check if the target text correctly answers the common questions given the target location. Each question is evaluated separately and they assign a binary score of 1 if answered correctly, else it is 0.

Let the binary scores for all m_i common questions of a target text T^i be denoted as $\{C_j^i\}_{j=1}^{m_i}$. The common question correctness metric across N texts will be $\mathbf{CQ} = \frac{1}{\sum_{i=1}^{N} m_i}\sum_{i=1}^{N}\sum_{j=1}^{m_i} C_j^i$.

4.3 Factual Correctness

For each target text T^i with EC $= 1$, the human annotator checks if every detail in the text is factually correct and provides a binary score $F^i = 1$ if every fact is correct, else $F^i = 0$. The factual correctness metric across N texts is calculated as $\mathbf{FC} = \frac{1}{N}\sum_{i=1}^{N} F^i$.

Other Evaluations. More sophisticated metrics like FActScore (Min et al., 2023b) have challenges with localization as elaborated in Appendix A.9. We also show fluency of generations using UniEval (Zhong et al., 2022) in Appendix A.9.

5 Models and Approaches

5.1 Models

We evaluate the performance of three state-of-the-art LLMs on LoFTI: Mixtral (Jiang et al., 2024), Llama3.3-70B (Meta, 2024) and GPT-4 (OpenAI, 2023). The Mixtral-8x7B LLM is a pre-trained decoder-only sparse mixture-of-experts model. For our analysis, we utilize the quantized Mixtral model *Mixtral-8x7b-instruct-v0.1.Q4_K_M*[3] with zero-shot prompting. Interestingly, we observed that few-shot prompting did not improve performance compared to the zero-shot setting and adding more localization examples appeared to confuse the model. Similarly, we also experiment on the latest Llama3.3-70B model using its quantized version (*Llama-3.3-70B-Instruct-Q4_K_M*). Llama3.3-70B is an autoregressive language model with an optimized transformer architecture, pretrained and instruction-tuned for generative tasks. We also evaluate the performance of the state-of-the-art GPT-4 (OpenAI, *gpt-4-turbo*) model on LoFTI. We use the same prompt for both Mixtral and GPT-4 (detailed in Appendix A.4).

5.2 Our Approaches

Mixtral + RARR. LLM generations, while fluent, are known to be prone to hallucinations and factual inaccuracies. To address this, Gao et al. (2023) proposed RARR (Retrofit Attribution using Research and Revision), an attribution mechanism that leverages external evidence from the web to validate and edit LLM-generated text while aiming to maintain the original style of the output. RARR consists of three modules: (i) Question Generation Module, (ii) Evidence Retrieval Module, and (iii) Editor Module. The Question Generation Module formulates questions from the text to be edited and the Evidence Retrieval Module queries these questions on the web for factual evidence. While querying, the target location of the text is appended to the start of each question to extract evidence relevant to that location. The retrieval module also checks if the text to be edited disagrees with the evidence. The Editor Module then utilizes all the disagreed evidence to make factual edits to the text. We employ the *Mixtral-8x7b-instruct-v0.1.Q4_K_M* model in both the Question Generation and Editor Modules. As in the original RARR pipeline, we utilize Microsoft Bing for evidence retrieval. We adhere to the RARR pipeline, except for one detail. We aggregate all the evidence obtained for all the generated questions and make a single edit, whereas RARR makes edits for each question individually. We found that sequential editing increased the text context and disrupted the style. Making a single edit helped maintain the text length and style better.

Model	# Samples	EC	CQ	FC
Mixtral	1100	**0.63**	**0.50**	**0.35**
Hyperlocal 1	369	0.72	0.58	0.41
Hyperlocal 2	391	0.63	0.49	0.38
Hyperlocal 3	340	0.54	0.43	0.25
GPT-4	1100	**0.80**	**0.64**	**0.62**
Hyperlocal 1	369	0.85	0.71	0.67
Hyperlocal 2	391	0.80	0.63	0.63
Hyperlocal 3	340	0.75	0.59	0.56

Table 2: Comparison of Mixtral and GPT-4 performance for localized Text Transfer using human evaluators.

Mixtral Revised. To improve the factual accuracy of Mixtral generations, we propose a revised version (henceforth referred to as Mixtral Revised). We use the Question Generation and Evidence Retrieval Modules as discussed in Section 5.2(Mixtral + RARR). However, we replace the Editor module with a Re-generation module which filters the evidence and re-generates the text using *Mixtral-8x7b-instruct-v0.1.Q4_K_M* model. The evidences retrieved from the Evidence Retrieval module are filtered to assess their relevance to the context and are added to the localized text transfer prompt to obtain more factually correct re-generation. This approach focuses on improving the factual correctness of the entity generated by Mixtral while preserving the style.

All the prompts used in the two above-mentioned approaches are detailed in Appendix A.4.

6 Experiments and Results

6.1 Comparison of Mixtral and GPT-4 for Localized Text Transfer

Table 2 compares the performance of state-of-the-art models GPT-4 and Mixtral on LoFTI using all three metrics that collectively help evaluate localization and factual transfer. We observe that GPT-4 significantly outperforms Mixtral in EC, CQ, and FC by 17%, 14%, and 27%, respectively. As hyperlocal scores increase, both models show a decline in accuracy, indicating the difficulty in recalling knowledge about these locales. GPT-4 surpasses Mixtral by achieving superior localization even at higher hyperlocal scores, resulting in more accurate outputs.

Table 3 shows examples highlighting the limitations of both Mixtral and GPT-4 at different hyperlocality values. For the example localized to India, we observe that both models localize well

[3] The quantized Mixtral-8x7b models Q6 and Q8 gave similar performance to Q4.

Reference Sentence	Target Location	Mixtral	GPT-4
The Reichenbach Falls are a waterfall cascade of seven steps on the stream called Rychenbach in the Bernese Oberland region of Switzerland. They drop over a total height of about 250 meters (820 ft).	India	The Dudhsagar Falls is a multi-step waterfall cascade on the Mandovi river in the Indian state of Goa, dropping over a total height of about **310 meters (1017 ft)**.	The Jog Falls are a segmented waterfall on the Sharavathi River located in Karnataka, India. They drop from a height of about 253 meters (830 ft).
Masashi Kishimoto is a Japanese manga artist. His manga series, Naruto, which was in serialization from 1999 to 2014, has sold over 250 million copies worldwide in 46 countries as of May 2019.	Maharashtra	**Padmashri Rahul Aware** is a renowned Marathi cartoonist. His comic series, 'Toofan Tai', which was in serialization from 2005 to 2016, has sold over 3 million copies within Maharashtra as of May 2021.	Anant Pai was an Indian comic book creator. His comic book series, Amar Chitra Katha, which was launched in 1967, has sold over 100 million copies worldwide in multiple languages.
Eric Otto Valdemar Lemming was a Swedish track and field athlete who competed at the 1900, 1906, 1908, and 1912 Olympics in a wide variety of events, which mostly involved throwing and jumping.	Khandra	**Surendra Kumar Singh** was a renowned Indian athlete who competed in several Athletics competitions at the 1956, 1960, and 1964 Olympics, primarily excelling in events such as shot put and javelin throw.	**Milkha Singh** was an Indian track and field athlete from Khandra who competed in the 1956, 1960, and 1964 Olympics, specializing in the 400 meters and other sprint events.

Text in **bold** represents the error in the text.

Table 3: Examples from different hyperlocal levels to illustrate the limitations of Mixtral and GPT-4 for localized text transfer.

Models	Mixtral				Mixtral + RARR				Mixtral Revised				Llama3.3-70B				GPT-4			
Hyperlocal	Overall	1	2	3	Overall	1	2	3	Overall	1	2	3	Overall	1	2	3	Overall	1	2	3
Human Evaluation																				
EC	0.60	0.70	0.63	0.45	0.65	0.77	0.65	0.48	0.66	0.78	0.65	0.52	0.66	0.79	0.65	0.50	0.81	0.83	0.83	0.75
CQ	0.48	0.54	0.53	0.36	0.57	0.68	0.61	0.39	0.56	0.64	0.57	0.45	0.63	0.80	0.60	0.46	0.65	0.68	0.66	0.60
FC	0.35	0.38	0.44	0.20	0.48	0.60	0.49	0.31	0.51	0.61	0.53	0.34	0.50	0.61	0.44	0.43	0.63	0.63	0.66	0.58
GPT-4 Evaluation																				
EC	0.72	0.79	0.71	0.62	0.76	0.85	0.76	0.63	0.76	0.82	0.73	0.72	0.79	0.79	0.80	0.78	0.81	0.80	0.80	0.84
CQ	0.61	0.68	0.60	0.52	0.64	0.78	0.61	0.49	0.61	0.68	0.59	0.54	0.69	0.75	0.67	0.64	0.66	0.70	0.63	0.63
FC	0.44	0.54	0.46	0.28	0.47	0.57	0.49	0.30	0.53	0.63	0.48	0.44	0.50	0.55	0.47	0.44	0.62	0.62	0.61	0.61

Table 4: Performance of Mixtral, Mixtral + RARR, Mixtral Revised, Llama3.3-70B and GPT-4 models for localized text generation on (a subset of) LoFTI using both human and GPT-4 evaluations.

but Mixtral tends to make errors in the factual details (e.g., the height of the waterfall). For Maharashtra (hyperlocal score of 2), Mixtral hallucinates and creates an imaginary entity ("Padmashri Rahul Aware") while GPT-4 localizes correctly. For Khandra (hyperlocal score of 3), both models fail to localize "Eric Otto Valdemar Lemming" correctly. Mixtral returns an entity from a different category and location ("Surendra Kumar Singh" is a politician from Madhya Pradesh), while GPT-4 returns an entity from the correct category but a different location ("Milkha Singh" is a track-and-field athlete from Chandigarh). The category-wise performance analysis of the models is given in Appendix A.5.

6.2 Comparison of Models/Approaches for Localized Text Transfer

In Table 4, we compare the performance of Mixtral, Mixtral + RARR, Mixtral Revised, Llama3.3-70B, and GPT-4 on a subset of 250 randomly chosen samples from LoFTI using human evaluators[4]. Attribution using factual evidence helps Mixtral + RARR in improving Mixtral generations, especially in the CQ and FC metrics, where the scores improve by 9% and 13%, respectively. However, the length of text obtained by RARR attribution is usually more than the original length of the text, and it fails to preserve the style.

Mixtral Revised utilizes factual evidence similar to RARR but regenerates the text instead of editing it. Including factual evidence in the prompt enhances the Mixtral outputs and results in improvements in both FC and CQ. The approach focuses mainly on revising the factual correctness of the text while largely retaining the entity present in it. However, we still see an enhancement in EC as factual evidence provides a richer context for the effective localization of the entity. While both Mixtral Revised and Mixtral + RARR use evidence, the former re-generates the text and the latter edits the text by retaining the entity. Re-generation helps in obtaining a factually correct entity. Improvements using Mixtral Revised are statistically significant at $p < 0.01$ using a Wilcoxon signed-rank test (details in Appendix A.10).

The latest Llama3.3-70B demonstrates superior performance compared to the Mixtral approaches, with its larger model size (70B vs. 8B parameters)

[4] We restrict this evaluation to a 250-sample subset due to annotation costs. However, we point to the Mixtral and GPT-4 scores in Tables 2 and 4 which are very similar, thus affirming that the 250-sample subset of LoFTI is representative of the full set.

Reference Text	Avenged Sevenfold is an American heavy metal band from Huntington Beach, California, formed in 1999. The band's current lineup consists of vocalist M. Shadows.			
Target Location	Kolkata			
Common Questions	(i) Name a music band which was formed many years ago? (ii) Which band's current lineup consists of a vocalist?			
Model	**Text Generated**	**EC**	**CQ**	**FC**
Mixtral	Cactus is a popular Indian heavy metal band from Kolkata, formed in **2006**. The band's current lineup consists of vocalist **Bipul Das**, lead guitarist **Sidharth Roy**, rhythm guitarist **Swarnendu Biswas**, and drummer **Shayan Ghosh**.	1	[0, 0]	0
Mixtral + RARR	Cactus is a popular Indian heavy metal band from Kolkata, formed in 1992. The current lineup consists of vocalist **Bipul Das**, lead guitarist **Sidharth Roy**, rhythm guitarist **Swarnendu Biswas**, and drummer **Shayan Ghosh**. *The band has named Deep Purple, Pink Floyd, Led Zeppelin, Iron Maiden and many others as their inspiration. Cactus is one of the first professional Bangla rock bands, with their professional debut occurring on 6th March 1993 at Aban Mahal (CLT), Kolkata.*	1	[1, 0]	0
Mixtral Revised	Cactus is a renowned Kolkata heavy metal band, formed in 1992. The band's current lineup includes vocalists Sidhartha Sankar "Sidhu" Ray and Abhijit "Pata" Barman, guitarists Boidurjyo Chowdhury and Samrat Banerjee, keyboardist Sayantan Chatterjee, bassist Proshanto Mahato, and drummer Arnab Tabla Dasgupta.	1	[1, 1]	1
Llama3.3-70B	Nicotine is an Indian heavy metal band from **Kolkata**, **West Bengal**, formed in 2006. The band's current lineup consists of vocalist **Rajdeep Chowdhury**.	0	[0, 0]	0
GPT-4	Fossils is an Indian rock band from Kolkata, West Bengal, formed in 1998. The band's current lineup consists of vocalist Rupam Islam.	1	[1, 1]	1

Bold represents the error in the text, underline represents the correct edits and *italics* represents the extra text generated.

Table 5: An example to illustrate the various localization approaches: Mixtral, Mixtral + RARR, Mixtral Revised, Llama3.3-70B and GPT-4.

and diverse training enabling better localization. However, even Llama3.3-70B struggles at higher hyperlocal levels. We note that the revision step in Mixtral Revised significantly improves the scores across all metrics, particularly for regions with a hyperlocal score of 3, and matches the performance of the much larger Llama3.3-70B model.

GPT-4[5] surpasses all the models due to its extensive training and diverse world knowledge. With increasing hyperlocal scores, performance degrades even with GPT-4. We also analyze the capability of GPT-4 (*gpt-4-turbo*) as an evaluator (prompts in Appendix A.4) for the task of localized text transfer. Compared to humans, GPT-4 is prone to false positives resulting in a 0.1-0.2 increase across all metrics and models. This shows that GPT-4 may not be reliable for absolute numbers. However, we observe similar *overall trends* in both human and GPT-4 evaluations. This shows that GPT-4 could be used as an LLM evaluator for localized text transfer to study the trends across models. There are minimal performance changes across hyperlocal levels in GPT-4 based on its own evaluations. However, human evaluations reveal a performance drop at the hyperlocal score of 3, indicating a bias in GPT-4 towards its own outputs. The limitations of GPT-4 as an evaluator are elaborated further in Appendix A.8. Table 5 shows a detailed example for all the models discussed.

6.3 LoFTI for Localized Question Answering

LoFTI can also be used as a benchmark to evaluate localized question answering. Given a target location and a question, the model has to generate text that answers the question while being correctly localized to the given target location. To aid this task, we also provide the reference location and the reference text as an example to guide localization and the style of generation. Simple prompt modifications, such as appending the target location to the question, do not improve generalization and often underperform compared to our prompt template.

From Table 6, we see that Mixtral obtains accuracies of 64% and 59% on the EC and FC metrics, respectively. We discuss some examples of Mixtral generations in Appendix A.7. As noted previously, performance degrades as hyperlocal scores increase. We also show the performance of GPT-4 as an evaluator; it nearly matches human evaluation when targeting India as a whole (hyperlocal score = 1), but highly overestimates scores for regions with hyperlocal scores of 2 and 3. The overall trends

Mixtral	**Overall**	**1**	**2**	**3**
# Samples	250	96	83	71
# Questions	447	168	145	134
Human Evaluation				
EC	0.64	0.81	0.63	0.45
FC	0.59	0.77	0.58	0.37
GPT-4 Evaluation				
EC	0.77	0.83	0.77	0.69
FC	0.61	0.74	0.53	0.52

Table 6: Performance of Mixtral model for localized question answering on (a subset of) LoFTI using both human and GPT-4 evaluations.

[5] We also evaluated the performance of GPT-3.5 (gpt-3.5-turbo) and it performed similarly to Mixtral with EC = 0.70, CQ = 0.50, and FC = 0.48 on the 250-sample subset.

of human evaluation are maintained by GPT-4.

7 Discussion

Errors. Two types of errors were mainly observed in the LLM generations. (i) Hallucinations: The LLM generates fictional entities or events that do not exist in order to create a localized sentence for the target location. This was mainly observed with Mixtral models for hyperlocal scores 2 and 3. (ii) Lack of fine-grained knowledge: Due to insufficient knowledge about specific categories and entities, the model produces sentences involving entities from closely related categories or with lower hyperlocal scores. This pattern was observed with both Mixtral and GPT-4 models for regions with hyperlocal scores 2 and 3.

GPT-4 evaluations. From Table 4, we observe that human and GPT-4 evaluations are most similar for GPT-4 generations. For all other model generations, GPT-4 gives inflated scores for all metrics (particularly EC) compared to the human evaluations. However, the trends in GPT-4 evaluations across models for both EC and FC mimic the trends observed in human evaluations. (This is not as clear for the CQ metric.) This suggests that one could use GPT-4 evaluations instead of expensive human evaluations *to observe trends in scores across multiple models*.

8 Related Work

Factual Correction, Transfer and Localization. Improving factual accuracy of LM generations is a very important problem that has gathered recent interest. Evidence integration, LLM post-editing modules, Rank-One Model Editing (ROME) are some of the recent techniques used to correct factual errors but they all struggle with consistency, specificity and generalizability (Thorne and Vlachos, 2021; Cao et al., 2021; Meng et al., 2023). Evaluating factual accuracy is another important problem for which fine-grained measures such as FActScore (Min et al., 2023a; Shafayat et al., 2024) have been developed. However, such measures are prone to biases across language and regions (Mirza et al., 2024) as we show in Appendix A.9.

In factual transfer, we also want the text style and intent of the reference text to be preserved (Jin et al., 2021). ModQGA is a framework that transfers facts without altering style (Balepur et al., 2023). Techniques like inverse prompting (Zou et al., 2021) have been used to improve the generation quality of LLMs for factual transfer. The RARR system improves reliability and attribution by correcting unsupported content using external evidence (Gao et al., 2023). Hence, we utilize RARR with LOFTI.

Cultural Adaptability and Diversity. LLMs tend to be geographically biased on various dimensions such as culture, race, language, politics due to its training being dominated by Western/English-centric datasets (Manvi et al., 2024). Recent works such as CultureLLM (Li et al., 2024a,c,b) and culture-sensitive evaluations (Rao et al., 2024) focus on cultural adaptability. In our work, we focus on factual transfer across geographical regions.

Existing Benchmarks. Existing benchmarks explore bias in various dimensions: (i) Social bias: CrowS-Pairs (Nangia et al., 2020) and BBQ (Bias Benchmark for QA) (Parrish et al., 2022) highlight social biases in the U.S. BOLD (Dataset and Metrics for Measuring Biases in Open-Ended Language Generation) (Dhamala et al., 2021) consists of prompts that evaluate bias in categories such as profession, gender, race, religion, and political ideology. (ii) Factual bias: X-FACT (A New Benchmark Dataset for Multilingual Fact Checking) (Gupta and Srikumar, 2021) consists of multilingual factual statements. FEVER (Fact Extraction and VERification) (Thorne et al., 2018) is a similar factual dataset in English. (iii) Geographical bias: WorldBench (Moayeri et al., 2024) consists of numerical questions to evaluate the factual recall abilities of LLMs on a per-country basis. None of the above-mentioned benchmarks capture the challenges in accurately localizing entities across different locales and maintaining factual accuracy.

9 Conclusion

This work introduces a new evaluation benchmark LOFTI to test the localization and factual transfer capabilities of LLMs. We attempt to localize factual statements from across the globe to multiple target locations within India spanning different levels of hyperlocality. We establish various baselines (Mixtral, GPT-4, etc.) and multiple benchmark tasks for the different models. We find that GPT-4 struggles with localization at higher levels of hyperlocality (i.e., when localizing to Indian cities), so much so that it cannot be reliably used as an automatic evaluator. We hope LOFTI helps the research community in designing improved localization and factual transfer techniques.

Limitations

The main limitations of the current benchmark are detailed below:

- GPT-4 is not good at identifying hyperlocal entities and facts about them. Hence, it cannot be used to reliably evaluate whether or not the localization produced is correct. Thus, there is still a need for human evaluators to check whether the localization produced is correct or not. We hope that expanding the possible target entities will help eventually mitigate the need for human evaluators to check for correctness. This is something that we plan to eventually add to our dataset in the near future.

- There can be several correct target entities localized to a target location which we refer to as high cardinality. High cardinality can make it hard to make the resulting evaluations precise, especially since some entities can be added in the future with respect to localization.

- This dataset consists only of factual data. However, localization can take place with respect to actions as well. For example, suppose we are localizing a conversation between a human and a shopkeeper about a special dinner. In the west, this typically would include conversations about buying steaks, lobsters etc. while in India, the conversation would likely be more about buying spices, rice and chicken. This is a broader style of localization that we intend to explore further as future work.

- The dataset is designed for localization from different locations in the world to regions in India. In order to perform localization to regions other than in India, we will need additional annotations. This is also reserved for a future release.

- LOFTI is entirely in English and does not contain any multilingual localizations. It is possible to use simple translation models to translate the data but it is not robust. This is a significant extension that we also intend to explore as future work.

References

Nishant Balepur, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Text fact transfer. *Preprint*, arXiv:2310.14486.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2021. Factual error correction for abstractive summarization models. *Preprint*, arXiv:2010.08712.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. Rarr: Researching and revising what language models say, using language models. *Preprint*, arXiv:2210.08726.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2021. Deep learning for text style transfer: A survey. *Preprint*, arXiv:2011.00416.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Huihan Li, Liwei Jiang, Jena D. Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024c. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *Preprint*, arXiv:2404.10199.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *Preprint*, arXiv:2402.02680.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023a. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Shujaat Mirza, Bruno Coelho, Yuyuan Cui, Christina Pöpper, and Damon McCoy. 2024. Global-liar: Factuality of llms over time and geographic regions. *Preprint*, arXiv:2401.17839.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1211–1228, New York, NY, USA. Association for Computing Machinery.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

OpenAI. 2023. Opengpt-4.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *Preprint*, arXiv:2404.12464.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore. *Preprint*, arXiv:2402.18045.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. *Preprint*, arXiv:2012.15788.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. 2021. Controllable generation from pre-trained language models via inverse prompting. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

A Appendix

A.1 Annotation Process

The LoFTI dataset was annotated by humans at various stages of its generation. The annotation was performed by an annotation company in India. The annotators were from diverse locations, occupations, age groups (21-40 yrs), and gender. Annotators were selected based on their performance and annotation quality on a sample set.

A.1.1 Inter-annotator Agreement

The dataset was divided into subsets, each of which was assigned to two annotators based on their expertise with the entities. The annotators independently annotated their assigned subsets. A third annotator then reviewed their work, selecting the best annotations and making necessary adjustments if required. Additionally, the company's in-house team conducted random reviews to ensure quality.

A.1.2 Annotation Cost

The total annotation cost was approximately ₹1 lakh, covering both data annotation and evaluation. In total, 13 annotators were involved in the process, with each receiving an average monetary reward of ₹7,600.

A.1.3 Guidelines for Annotators

The following guidelines were provided to the human annotators.

- Generation of Entities
  - Entities should cover a diverse set of 99 categories. Examples of categories: Politician, Music Band, Historical Monument, Airline, Web Series, etc.
  - On average 10 entity-pairs under each category. Note: Reference entity can be repeated, but do not repeat target entity.
  - Ensure the target entity is sufficiently similar to the reference entity selected. For example, refer row 1 of Table A1.
  - Ensure the new entities are spread over India and have different hyperlocal scores. For example, refer row 2-4 of Table A1.
  - The reference entities of the dataset should be spread across different countries, with 60% from the US/Europe and the remaining 40% from other parts of the world.
- Correction of Target Sentences
  - Check if the target sentences are factually correct and localized correctly.
  - Altering multiple elements within the target sentence might be necessary to guarantee factual accuracy within the specific domain.
  - Check for fluency, grammar, and vocabulary accuracy in the sentences while eliminating unnecessary symbols or words.
  - Align the structure of the target sentence with that of the reference sentence. Remove or add any additional or missing content/information present in the reference sentence. For example, refer to Table A2.
- Correction of Target Sentences
  - Check if the target sentences are factually correct and localized correctly.
  - Altering multiple elements within the target sentence might be necessary to guarantee factual accuracy within the specific domain.
  - Check for fluency, grammar, and vocabulary accuracy in the sentences while eliminating unnecessary symbols or words.
  - Align the structure of the target sentence with that of the reference sentence. Remove or add any additional or missing content/information present in the reference sentence. For example, refer to Table A2.
- Common questions
  - It should be generated based on the common description of the entities in the pairs of text provided.
  - It should be of the type such that it can be asked in any target location and still be valid.
  - It should be free from specific details such as locations, timings, or unique identifiers connected to either event.
  - Remove or correct any incorrect questions present. There should be a minimum of one correct common question for each sample sentence pair. Add more questions if needed.
  - For common question correction, refer to the example in Table A3.

Human Evaluation Guidelines The outputs generated by the models were evaluated by humans to assess Entity Correctness, Common Question Correctness, and Factual Correctness. The following guidelines were provided to the human annotators.

- Entity Correctness (EC)
  - The entity detected from the sentence should be from the target location.
  - Check if the entity is a correct localization of the reference entity provided.
  - If the entity is an exact match to the true target entity, please mention “Exact match" in the reason.
  - Always provide a reason when the score is 0.
- Common Question Correctness (CQ)
  - Each sample will have multiple questions, evaluate each (sample, question) pair separately.
  - For each sample, return the score as a list of 0’s and 1’s with the scores indexed at the question number.
  - Common Question Correctness for all questions should be given a score of 0 if that sample’s entity correctness (EC) is 0.
  - Check if the sentence correctly answers the question for the “target location".
  - Ensure factual correctness in these answers.
  - Always provide a reason when the score is 0.
- Factual Correctness (FC)
  - Factual correctness should be given a score of 0, if that sample’s entity correctness (EC) is 0.
  - Assign a score of 1, if the sentence is fully factually correct, else assign a score of 0.
  - If the sentence contains any information that lacks factual evidence online, assign a score of 0.
  - Always provide a reason when the score is 0.

Refer to Table A4, for examples of human evaluation.

A.2 Prompts used for Dataset Creation

For dataset creation we used the *mixtral-8x7b-instruct-v0.1.Q4_K_M.gguf* model from HuggingFace. The prompt used for text localization is given in Figure A1. And the prompt used for the generating of common questions from the reference and target text is given in Figure A2.

A.3 Category Distribution of LoFTI

The LoFTI dataset contains 99 diverse categories like Movies, Accidents, Currency, Sports, etc. The number of entities under each category is uniformly distributed with an average of 10 entities in each category. Figure A3 shows the distribution of entities across the categories. As shown in Table A5, categories can be grouped mainly into 10 clusters namely Entertainment, Professions, Buildings/-Monuments/Companies, Food & Lifestyle, Places & Landmarks, Nature, Sports, Incidents, Finance & Economy and Others.

A.4 Implementation Details

We ran our experiments on a single NVIDIA DGX A100 GPU. For the Mixtral and Llama3.3-70B experiments, we used the *mixtral-8x7b-instruct-v0.1.Q4_K_M.gguf* and *Llama-3.3-70B-Instruct-Q4_K_M* models respectively from HuggingFace. A maximum sequence length of 32768 was used. For GPT-4 generations and evaluations, we used the *gpt-4-turbo* model using OpenAI API. For evidence extraction from the web, we used Bing Search API from Microsoft Azure.

A.4.1 Prompt for localized text transfer

The prompt used for localized text transfer is given in Figure A4. We use the same prompt for Mixtral, Llama3.3-70B and GPT-4 models.

A.4.2 Prompt for localized question answering

The prompt used for localized question answering is given in Figure A5. We use the same prompt for both Mixtral and GPT-4 models.

A.4.3 Mixtral + RARR Prompts

The prompts used in the Question Generation module, Evidence Retrieval module (to check whether the evidence agrees/disagrees with the text to be edited), and the Editor module are given in Figure A6, A7 and A8 respectively.

Category	Reference Location	Reference Entity	Target Location	Target Entity	Hyperlocal Score
Singer	US	Taylor Swift	India	~~Neha Kakkar~~ Ravi Shankar	1
Educational Institution	Australia	The University of Melbourne	India	Indian Institute of Technology, Bombay	1
Educational Institution	Florida	University of Central Florida	Kerala	Central University of Kerala	2
Educational Institution	Miami	University of Miami	Tiruchirappali	Bharathidasan University	3

~~Strikethrough text~~ is the incorrect entity.

Table A1: Example to illustrate how to create correct entity pairs for LoFTI dataset.

Category	Reference Location	Reference Entity	Target Location	Target Entity	Reference sentence	Target sentence
Automotive company	US	Ford Motor	India	Tata Motors	Ford Motor Company is an American multinational automobile manufacturer headquartered in Dearborn, Michigan, United States. It was founded by Henry Ford and incorporated on June 16, 1903.	Tata Motors Limited is an Indian multinational automotive ~~manufacturing company.~~ **[manufacturer headquartered in Mumbai, Maharashtra, India]**. It was founded by J. R. D. Tata and incorporated on September 1, 1945. ~~The company sells passenger cars, trucks, vans, coaches, buses, sports cars, construction equipment and military vehicles under the Tata brand. Tata Motors is the largest automobile manufacturer in India with a revenue of over 470 billion Indian rupees.~~

[Text in square brackets] is the additional content added and ~~strikethrough text~~ is the additional content that has to be removed.

Table A2: An example to illustrate the annotation process for the target sentence generated for the LoFTI dataset.

A.4.4 Mixtral Revised Prompts

The prompt used for verifying the relevance of the evidence for the target context is given in Figure A9. The text re-generation prompt of the Mixtral Revised model is given in Figure A10.

A.4.5 GPT-4 Evaluation

We used the GPT-4 model to evaluate LoFTI on the EC, CQ, and FC metrics. For EC, we first identify the entity in the target sentence using the prompt provided in Figure A11. Then, we check if the identified entity is an exact match to the true target entity or a possible correct localization using the prompt given in Figure A12. An EC score of 1 is given if it's an exact match or a possible correct localization; otherwise, a score of 0 is assigned.

For CQ, we compute a binary score for each sample against each common question in the LoFTI dataset corresponding to that sample, using the prompt given in Figure A13. CQ for a sample was computed only if EC $= 1$; otherwise, a CQ score of 0 was assigned. For FC, if the entity detected during EC was an exact match to the true target entity, then factual correctness was computed by comparing it with the true target sentence in LoFTI using the prompt given in Figure A14. If it is not an exact match, then the true target entity and the true target sentence in LoFTI are given as examples to the model to evaluate the target sentence using the prompt given in Figure A15.

A.5 Localized Text Transfer: Category-wise Performance Analysis of Models using Human Evaluators

In this section, we compare the performance of Mixtral and GPT-4 outputs across different categories. The LoFTI has 99 unique categories and we have grouped them into 10 category clusters for our analysis.

Table A6 shows that the performance varies across categories 'Professions', 'Entertainment', and 'Incidents' obtain the lowest scores by Mixtral and GPT-4 models due to the presence of diverse entities like Web Series, Movies, YouTubers, Motivational speakers, Accidents, etc. that have higher cardinality and lack of factual evidence. Both Mixtral and GPT-4 perform well in categories like 'Buildings/Monuments/Companies', 'Places & Landmarks', and 'Nature' due to the sufficient amounts of factual evidence available during training.

A.6 Localized Text Transfer: GPT-4 Evaluation of Mixtral on the full LoFTI Dataset

We also analyze the performance of GPT-4 as an evaluator for localized text transfer on the full LoFTI dataset. In Table A7, we compare the human and GPT-4 evaluations on the Mixtral model for the full dataset. Similar to our observation on the 250 subset (Table 4), GPT-4 closely aligns with

Reference sentence	Target sentence	Common questions
Rishi Sunak is a British politician who has served as Prime Minister of the United Kingdom and Leader of the Conservative Party since 2022.	Narendra Modi is an Indian politician who has served as Prime Minister of India and President of the Bharatiya Janata Party since 2014.	(i) **Can you name a current Prime Minister?** (ii) **Who is a well-known politician serving as the head of a major political party?**
Poshmark is a social commerce marketplace where users can buy and sell new and secondhand fashion, home goods, and electronics. The platform has over 80 million users, with over 200M available listings. The company is headquartered in Redwood City, California, with offices in Canada, Australia, and India.	Meesho is a social commerce marketplace based in India where users can buy and sell new and secondhand fashion, home goods, and electronics. The platform has over 60 million users, with millions of available listings. The company is headquartered in Bengaluru, India, and operates independently.	(i) **Name a social commerce marketplace ?** (ii) **Tell me about a company in the social commerce space?** (iii) ~~Name a social commerce marketplace in California?~~ (iv) ~~Who operates Poshmark as an independent subsidiary since January 2023?~~ (v) ~~Where is Meesho headquartered, and do they have any connections to Naver Corporation or headquarters outside of India?~~
KLVE is a commercial radio station licensed to Los Angeles, California with a Spanish AC format. The station is owned by TelevisaUnivision, and is the flagship station for the Uforia Audio Network.	Radio Mango 91.9 FM is a private radio station licensed to Kochi, Kerala with a Malayalam language format. The station is owned by the Malayala Manorama Group and serves as the flagship station for their radio network.	(i) ~~What kind of radio station is KLVE?~~ (ii) ~~Who owns the radio station "Radio Mango 91.9 FM"?~~ (iii) **Can you mention a radio station that is a flagship station for a network?**
Mindhunter is an American psychological crime thriller television series created by Joe Penhall, which debuted in 2017, based on the 1995 true-crime book Mindhunter: Inside the FBI's Elite Serial Crime Unit by John E. Douglas and Mark Olshaker.	Kerala Crime Files is a Malayalam-language psychological crime drama web series directed Ahammed khabeer, which debuted in 2023.	(i) ~~Can you name an American psychological crime thriller television series that debuted in 2017?~~ (ii) ~~Is there a television series based on a true-crime book that released recently?~~ (iii) **Mention a series focusing on criminal psychology.**

Bold text represents the correct questions and ~~strikethrough text~~ represents the incorrect questions.

Table A3: Examples to illustrate the annotation process for the common question generated for the LoFTI dataset.

human evaluation for regions with a hyperlocal score of 1 but significantly overestimates scores for regions with hyperlocal scores of 2 and 3. Despite this, GPT-4 maintains the overall trends observed in human evaluation.

A.7 Localized Question Answering: Examples of Mixtral Generation

In Table A8, we illustrate the performance of Mixtral on the benchmark task of localized question answering using some examples.

A.8 Limitations of GPT-4 as an Evaluator for LoFTI

Table A9 illustrates the limitation using GPT-4 as an evaluator with an example. On text localization, Mixtral hallucinates and returns the entity "Mystic Moods". GPT-4 incorrectly claims it to be a factually correct localization and assigns a score of 1 for all the metrics. This shows that GPT-4 evaluations result in false positives and it not a reliable evaluator for absolute numbers.

A.9 Other Evaluation Metrics

Metrics to measure factuality. FActScore (Min et al., 2023b) is a more sophisticated evaluation metric for LLMs that represents the percentage of atomic facts (pieces of information) supported by a given knowledge source. It breaks down the generations from an LLM into atomic facts and assigns a binary score to evaluate factual precision, using GPT models for scoring these atomic facts. We compared the performance of GPT-4 (*gpt-4-turbo*) and human annotators in assigning scores to these atomic facts. Mixtral generations gave a FActScore of 0.70 and 0.56 using GPT-4 and human annotators respectively on 250 samples. However, FActScore has two main drawbacks: (i) the generated atomic facts lack sufficient context for accurate evaluation, and (ii) GPT-4 produces many false positives.

Metrics to measure fluency. Regarding fluency, unsurprisingly, models like Mixtral and GPT-4 are very good at generating coherent and fluent text. We computed the reference-free fluency scores using UniEval (Zhong et al., 2022) and observed that these scores are equally high for both Mixtral(0.93) and GPT-4(0.94) on the full LoFTI dataset. They do not, however, capture the nuances of localization and factuality transfer (that our proposed metrics do).

A.10 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test results for Mixtral + RARR and Mixtral Revised in comparison to Mixtral across all three metrics are given in Table A10.

Claim ID	Category	Reference Location	Reference Entity	Target Location	True Target Entity	Reference sentence	True Target sentence	Questions	Model Generation
114	Books	London	A Tale of Two Cities	Cheruvathur	The House of Blue Mangoes	A Tale of Two Cities is a historical novel published in 1859 by Charles Dickens, set in London and Paris before and during the French Revolution.	The House of Blue Mangoes is a historical novel published in 2002 by David Davidar, set in Chevathar during the historical events like caste conflicts, World War II, and India's struggle for independence.	(i) What is an example of a historical novel? (ii) Can you name a novel that is set during the political change happening in the country?	"Kaalam Maarunnu Kaalangal" is a historical novel written in 1916 by C. V. Raman Pillai, set in Kilimanoor (Kerala) and Thiruvithamkoor (Travancore) during the late 18th and early 19th centuries.
255	Educational institution	US	Johns Hopkins University	India	Indian Institute of Science	Johns Hopkins University is a private research university in Baltimore, Maryland. Founded in 1876, Johns Hopkins was the first U.S. university based on the European research institution model.	The Indian Institute of Science is a public research university in Bangalore, Karnataka, India. Established in 1909, the Indian Institute of Science was the first to introduce a master's programs in engineering.	(i) Can you give an example of a research university? (ii) Can you provide name of a university which introduced something new?	The Indian Institute of Science (IISc) is a public research university in Bangalore, Karnataka. Established in 1909, IISc was modeled after the European research institution and is one of the first institutions of its kind in India.

Evaluation

Claim ID	EC	Reason	FC	Reason	CQ	Reason
114	0	wrong target location	0	EC = 0	[0, 0]	EC = 0
255	1	Exact match	0	some details like European are incorrect	[1, 0]	(ii)The sentence does not answer the question.

Table A4: Examples to illustrate human evaluation.

You are a localization assistant. Convert the reference entity sentence from English to the Indian domain by replacing the source entity with the target entity. Make the needed modifications in the sentence to make it factually correct for the target entity. Output answers in English using multi-entity localization.
Reference entity: **Rishi Sunak**
Reference entity location: **UK**
Reference entity sentence: **Rishi Sunak is a British politician who has served as Prime Minister of the United Kingdom and Leader of the Conservative Party since 2022.**
Target entity: **Narendra Modi**
Target entity location: **India**

Figure A1: Prompt for text localization on Mixtral

You are tasked with generating basic questions from common property or common description of the entities in pairs of sentences provided. The goal is to create 2 or more questions such that they can be asked in any location and still be valid. The questions should not have any entity or location mentioned in it. Example:
Given the following pair of sentences:
(1) Poshmark is a social commerce marketplace where users can buy and sell new and secondhand fashion, home goods, and electronics. The platform has over 80 million users, with over 200M available listings. The company is headquartered in Redwood City, California, with offices in Canada, Australia, and India;
(2) Meesho is a social commerce marketplace based in India where users can buy and sell new and secondhand fashion, home goods, and electronics. The platform has over 60 million users, with millions of available listings. The company is headquartered in Bengaluru, India, and operates independently.
The correct questions are:
(i) Name a social commerce marketplace.
(ii) Tell me about a company in the social commerce space.
The wrong questions are:
(i) Name a social commerce marketplace in California.
(ii) Who operates Poshmark as an independent subsidiary since January 2023?
(iii) Where is Meesho headquartered, and do they have any connections to Naver Corporation or headquarters outside of India?
As shown in the examples, the correct questions should be free from specific details such as locations, timings, or unique identifiers connected to either event. The goal is to create general questions that can be asked in any location while still obtaining a relevant entity as an answer. Keep the questions simple.
Now generate only correct questions for the following pair:
Sentence 1: **Rishi Sunak is a British politician who has served as Prime Minister of the United Kingdom and Leader of the Conservative Party since 2022.**
Sentence 2: **Narendra Modi is an Indian politician who has served as Prime Minister of India, since 2014 and is a member of the Bharatiya Janata Party.**

Figure A2: Few-shot prompt for common question generation on Mixtral

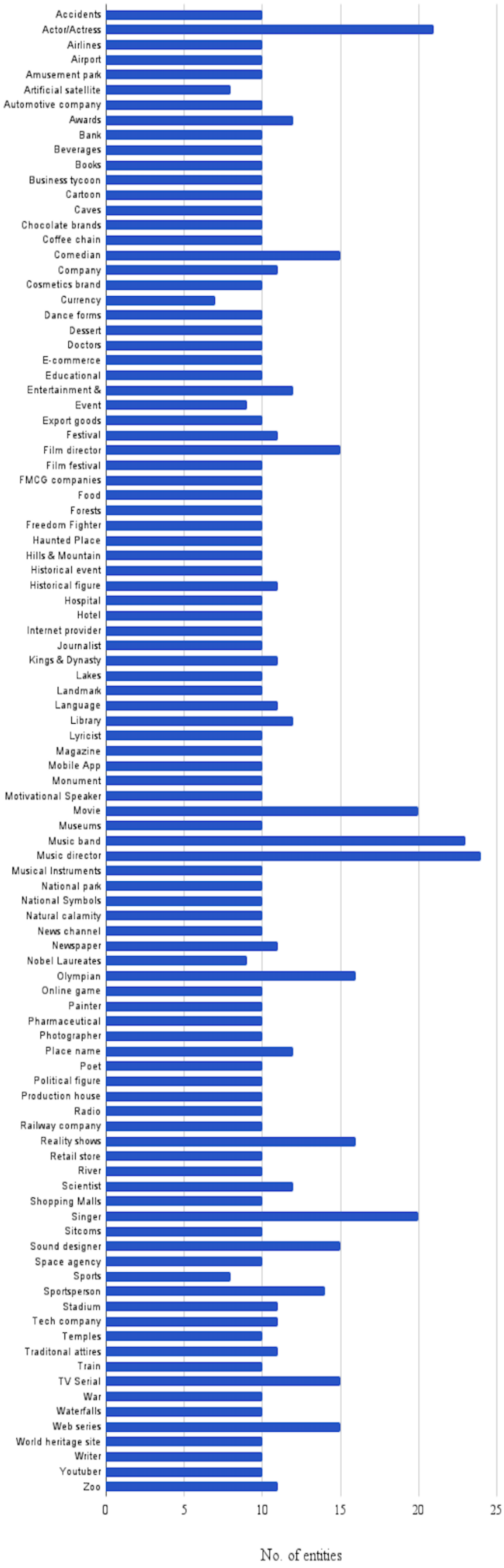


Figure A3: LoFTI dataset category distribution

Category Cluster	Categories	# Samples
Entertainment	Actor/Actress, TV Serial, Cartoon, Film Festival, Event, Magazine, Mobile App, Movie, Music Band, Radio, Sitcoms, Online Game, Web Series, News Channel, Newspaper, Production House, Awards, Books, Reality Shows, Dance Forms, Musical Instruments, Entertainment & Sports Channel	274
Professions	Business Tycoon, Comedian, Doctors, Film Director, Lyricist, Journalist, Motivational Speaker, Music Director, Poet, Writer, Singer, Scientist, Painter, Youtuber, Sound Designer, Photographer, Political Figure, Nobel Laureates	220
Buildings/ Monuments/ Companies	Automotive Company, Company, Airlines, Educational Institution, FMCG Companies, Hospital, Hotel, Library, Temples, Pharmaceutical Companies, Airport, Tech Company, Space Agency, Monument, Railway Company, Museums, Internet Provider	174
Others	Traditional Attires, Train, Language, Kings & Dynasty, National Symbols, Artificial Satellite, Historical Figure, Festival, Freedom Fighter	93
Food & Lifestyle	Beverages, Chocolate Brands, Coffee Chain, Cosmetics Brand, Food, Dessert, Shopping Malls, Retail store, E-commerce company	90
Places & Landmarks	Landmark, Place Name, Haunted Place, Zoo, Amusement Park, National Park, World Heritage Site	73
Nature	Caves, Forests, Hills & Mountains, Lakes, River, Waterfalls	60
Sports	Sports, Sportsperson, Olympian, Stadium	49
Incidents	Historical Event, War, Accidents, Natural Calamity	40
Finance & Economy	Bank, Currency, Export Goods	27

Table A5: Category Clusters and Categories in LoFTI dataset.

Category Cluster	#	Mixtral			GPT-4		
		EC	FC	CQ	EC	CQ	FC
Entertainment	274	0.57	0.27	0.44	0.75	0.57	0.56
Professions	220	0.56	0.25	0.45	0.71	0.55	0.54
Building/ Monument / Company	174	0.80	0.59	0.68	0.88	0.76	0.71
Others	93	0.71	0.52	0.56	0.88	0.75	0.79
Food & Lifestyle	90	0.67	0.35	0.51	0.83	0.70	0.65
Places & Landmarks	73	0.81	0.46	0.63	0.93	0.75	0.73
Nature	60	0.75	0.42	0.58	0.83	0.65	0.58
Sports	49	0.61	0.24	0.35	0.79	0.56	0.60
Incidents	40	0.57	0.30	0.45	0.75	0.63	0.50
Finance & Economy	27	0.85	0.63	0.67	0.74	0.66	0.63

Table A6: Category-wise performance analysis of Mixtral and GPT-4 model on localized text transfer using human annotators.

Mixtral	Overall	1	2	3
	Human Evaluation			
EC	0.63	0.72	0.63	0.54
CQ	0.50	0.58	0.49	0.43
FC	0.35	0.41	0.38	0.25
	GPT-4 Evaluation			
EC	0.71	0.76	0.71	0.67
CQ	0.58	0.65	0.56	0.52
FC	0.42	0.52	0.42	0.33

Table A7: Comparison of human and GPT-4 evaluation on Mixtral's localized text transfer outputs on the full LoFTI dataset.

Reference Text	Target Location	Question	Mixtral	EC	FC
Morgan Freeman is an American actor, producer, and narrator known for his distinctive, deep voice. Throughout a career spanning five decades and multiple film genres, he has received numerous accolades, including an Academy Award, a Golden Globe Award, and a Screen Actors Guild Award as well as a nomination for a Tony Award.	Thanjavur	Name an actor who has worked for more than two decades.	Kamal Haasan is a renowned Indian actor who has been active in the film industry for over four decades, featuring in more than 200 films across various languages.	0	0
Duronto TV is a Bangladeshi Bengali-language satellite and cable children's television channel, and the first of its kind in the country. It was initially licensed in 2013 as "Renaissance TV" but was later renamed to its current name.	Mumbai	Can you name a children's television channel that is the first of its kind in a country?	Nickelodeon Sonic is a children's television channel in India, and it was the first of its kind when it launched in 2001.	1	0

Table A8: Examples of Mixtral generations for the benchmark task of localized question answering.

Reference Text	The Free Spirits was an American band credited as the first jazz-rock group. The band also incorporated elements of pop and garage rock. Their first album Out of Sight and Sound was recorded in 1966 and released in 1967.						
Target Location	Delhi						
Common Questions	(i) Can you name a band credited as a jazz-rock band? (ii) Can you provide an example of a band which has released albums?						
True Target Entity	Indian Ocean						
True Target Sentence	The Indian Ocean is an Indian band credited as jazz-rock fusion group. The band also incorporated elements of folk and classical. Their first album Indian Ocean was recorded in 1992 and released in 1993.						
Mixtral	The Mystic Moods, a Delhi-based band, was credited as one of the first to experiment with fusion of jazz and rock in India. They also incorporated elements of Indian classical music and pop. Their first album "A New Dawn" was recorded in 1972 and released in 1973.						
Evaluator	**EC**	**Reason**	**CQ**	**Reason**	**FC**	**Reason**	
GPT-4	1	The Mystic Moods is also a music band based in India, and shares a similar category with Indian Ocean, although it is not the exact same band.	[1, 1]	It mentions that 'The Mystic Moods,' a Delhi-based jazz-rock band and have released an album titled "A New Dawn".	1	The Mystic Moods is a Delhi-based jazz-rock band. It accurately describes their origin, genre, and album details, making it relevant and truthful in the context of Delhi.	
Human	0	Entity does not exist.	[0, 0]	EC = 0	0	EC = 0	

Table A9: An example to illustrate the limitation of GPT-4 as an evaluator for text localization by comparing it with human evaluation.

Metric	Mixtral + RARR		Mixtral Revised	
	W Statistic	p-value	W Statistic	p-value
EC	40.0	0.026	20.0	0.001
CQ	6461.0	0.004	5814.0	0.009
FC	528.0	< 0.001	341.0	< 0.001

Table A10: Wilcoxon signed-rank test: Comparison of Mixtral + RARR and Mixtral Revised to Mixtral across all three evaluation metrics.

You are a localization assistant. Convert the reference entity sentence from English to the Indian domain by localizing the reference sentence with a similar entity from the target location. Make the needed modifications in the sentence to make it factually correct for the target location. Output answers in English using multi-entity localization. Use the below format.
My reference sentence: `<reference_claim>`
Target location: `<target_location>`
Target sentence: `<localized_target_sentence>`
Reason: `<reason_for_the_localization>`

My reference sentence: **Rishi Sunak is a British politician who has served as Prime Minister of the United Kingdom and Leader of the Conservative Party since 2022.**
Target location: **India**
Target sentence: `<fill_your_answer_here>`
Reason: `<fill_your_answer_here>`

Figure A4: The prompt used for localized text transfer in Mixtral and GPT-4 models.

Given a question and a target location, generate a factually correct sentence such that it answers the given question using an entity from the target location. A reference location and sentence is given as an example. Output the answers in English. Use the below format.
Question: `<question_to_be_answered>`
Reference location: `<example_reference_location>`
Reference sentence: `<example_reference_sentence>`
Target location: `<target_location>`
Target sentence: `<target_sentence>`
Reason: `<reason_for_the_localization>`

Question: **Can you name a current Prime Minister?**
Reference location: **UK**
Reference sentence: **Rishi Sunak is a British politician who has served as Prime Minister of the United Kingdom and Leader of the Conservative Party since 2022.**
Target location: **India**
Target sentence: `<fill_your_answer_here>`
Reason: `<fill_your_answer_here>`

Figure A5: The prompt used for localized question answering in Mixtral and GPT-4 models.

Given a target sentence corresponding to a specific target location, your task is to ask questions about the target entity. Each question should be specific to the target entity and should not contain pronouns such as ’he,’ ’she,’ ’it,’ or ’they.’ The questions should seek relevant information about the target entity, its attributes, actions, or associations with the target location. Additionally, the questions should be structured in a way that the answer contains the target entity and/or the target location. Avoid general questions like ’Who is he?’ or ’Where does he live?’ Instead, focus on extracting detailed insights about the target entity. Ensure that the questions are clear, concise, and relevant to the context of the target sentence. Questions should be able to interrogate the factual information in the claim. Do not generate irrelevant questions based on other entities that have no relation with the target entity or target locations.
For example:
Target location: Delhi
Target sentence: The India Gate is a war memorial made of sandstone located in the heart of New Delhi, India. It is named after the engineer Sir Edwin Lutyens, who designed and built the monument in 1931 to honor the Indian soldiers who died during World War I and the Third Anglo-Afghan War.
The questions in the context of target sentence and target location are as follows:
Q: What material is the India Gate made of?
Q: In which city is the India Gate located?
Q: Who is the engineer credited with designing and building the India Gate?
Q: When was the India Gate constructed?
Target location: India
Target sentence: Narendra Modi is an Indian politician who has served as Prime Minister of India and leader of the Bharatiya Janata Party since 2014.
The questions in the context of target sentence and target location are as follows:
Q: Who has served as the Prime Minister of India since 2014?
Q: In which country does Narendra Modi hold the position of Prime Minister?
Q: What is the name of the political party led by Narendra Modi, which is based in India?
Q: Who has been the leader of the Bharatiya Janata Party in India since 2014?
Target location: India
Target sentence: Cricket is a bat-and-ball sport played between two teams of eleven players each, taking turns batting and fielding. The game occurs over the course of several overs, with each over consisting of six deliveries (pitches) generally made by a player on the fielding team, called the bowler, which a player on the batting team, called the batter, tries to hit with a bat.
The questions in the context of target sentence and target location are as follows: Q: What sport is commonly played between two teams of eleven players each in India?
Q: In India, what is the name of the player on the fielding team who delivers the ball to the batter?
Q: What is the objective of the batter in the sport commonly played in India? Q: In India, what is the term for a single set of deliveries made by a bowler in the sport?
Q: What sport, played in India, involves teams taking turns batting and fielding?
Target location: Telengana
Target sentence: S. S. Rajamouli is an Indian director and screenwriter, known for his work in Telugu industry based in Telengana, India. He is considered one of the leading filmmakers in the Indian film industry, having directed some of the highest-grossing Indian films of all time. His most notable works include the "Telugu-language fantasy action film series", Baahubali and RRR which broke several box office records and gained international recognition.
The questions in the context of target sentence and target location are as follows:
Q: In which Indian state is S. S. Rajamouli primarily associated with for his filmmaking?
Q: What are some of the notable works directed by S. S. Rajamouli in the Telugu film industry?
Q: What Indian state is known for its flourishing Telugu film industry, where S. S. Rajamouli has made significant contributions?
Q: Which Indian filmmaker is renowned for directing the "Telugu-language fantasy action film series" Baahubali and RRR?
Target location: Maharashtra
Target sentence: Sanjay Dutt is an Indian actor who works in Bollywood industry based in Maharashtra, India and whose career has seen highs and lows. He initially gained critical acclaim and popularity for his roles in Bollywood films during his youth. However, he also faced struggles with substance abuse and legal issues, including his involvement in the 1993 Bombay bombings case, which resulted in his arrest and imprisonment.
The questions in the context of target sentence and target location are as follows:
Q: In which Indian state is Sanjay Dutt primarily based for his work in the Bollywood film industry?
Q: What industry is Sanjay Dutt associated with, which is primarily based in Maharashtra, India?
Q: What are some challenges that Sanjay Dutt faced in his career, despite gaining critical acclaim and popularity for his roles in Bollywood films during his youth?
Q: What legal issue was Sanjay Dutt involved in, which led to his arrest and imprisonment in Maharashtra?
Q: Which Indian state is known for its Bollywood film industry, where Sanjay Dutt has worked throughout his career?
Target location: India
Target sentence: "Baahubali: The Conclusion" is an epic Indian film released in 2017, serving as the climax of the "Baahubali" film series. It brings together iconic characters like Baahubali, Bhallaladeva, and Devasena for a grand spectacle of action and drama. The film received widespread acclaim for its visual effects, emotional storytelling, and massive box office success, solidifying its place as one of India’s most beloved cinematic experiences.
The questions in the context of target sentence and target location are as follows:
Q: What is the name of the epic Indian film released in 2017, which is considered one of India’s most beloved cinematic experiences?
Q: What film series does "Baahubali: The Conclusion" serve as the climax for?
Q: Who are some of the iconic characters featured in "Baahubali: The Conclusion"?
Q: What aspects of "Baahubali: The Conclusion" contributed to its widespread acclaim and massive box office success in India?
Q: Which country is known for producing "Baahubali: The Conclusion," an epic Indian film released in 2017? Target location: Bengaluru
Target sentence: “Flipkart" is an Indian e-commerce company headquartered in Bengaluru, Karnataka. Founded by Sachin Bansal and Binny Bansal in 2007, it started as an online bookstore before diversifying into a wide range of product categories, including electronics, fashion, and home goods. With its user-friendly interface, extensive product offerings, and competitive pricing, Flipkart has emerged as one of India’s leading e-commerce platforms, revolutionizing the way millions of people shop online in the country.
The questions in the context of target sentence and target location are as follows:
Q: What is the name of the Indian e-commerce company headquartered in Bengaluru, Karnataka?
Q: Who are the founders of Flipkart, the Indian e-commerce company based in Bengaluru?
Q: In which Indian city is Flipkart headquartered?
Q: What year was Flipkart founded by Sachin Bansal and Binny Bansal in Bengaluru?
Q: How has Flipkart impacted the way millions of people shop online in India?
Target location: West Bengal
Target sentence: The Howrah Bridge is a cantilever bridge spanning the Hooghly River, the wide river that flows through West Bengal and connects the cities of Howrah and Kolkata.
The questions in the context of target sentence and target location are as follows:
Q: What is the name of the bridge spanning the Hooghly River in West Bengal?
Q: Which two cities does the Howrah Bridge connect in West Bengal?
Q: What type of bridge is the Howrah Bridge?
Q: Through which river does the Howrah Bridge span in West Bengal?
Q: What is the significance of the Howrah Bridge in connecting the cities of Howrah and Kolkata in West Bengal?

Target location: **West Bengal**
Target sentence: **Prosenjit Chatterjee is a renowned Bengali actor, whose career has been marked by critical acclaim in his early life, followed by personal challenges and a resurgence in popularity and commercial success in his later years.**
The questions in the context of target sentence and target location are as follows:

Figure A6: Mixtral + RARR: The prompt used for generating questions from the sentence and target location for evidence retrieval.

I will check some things you said.

1. You said: Your nose switches back and forth between nostrils. When you sleep, you switch about every 45 minutes. This is to prevent a buildup of mucus. It's called the nasal cycle.
2. I checked: How often do your nostrils switch?
3. I found this article: Although we don't usually notice it, during the nasal cycle one nostril becomes congested and thus contributes less to airflow, while the other becomes decongested. On average, the congestion pattern switches about every 2 hours, according to a small 2016 study published in the journal PLOS One.
4. Reasoning: The article said the nose's switching time is about every 2 hours, and you said the nose's switching time is about every 45 minutes.
5. Therefore: This disagrees with what you said.

1. You said: The Little House books were written by Laura Ingalls Wilder. The books were published by HarperCollins.
2. I checked: Who published the Little House books?
3. I found this article: These are the books that started it all – the stories that captured the hearts and imaginations of children and young adults worldwide. Written by Laura Ingalls Wilder and published by HarperCollins, these beloved books remain a favorite to this day.
4. Reasoning: The article said the Little House books were published by HarperCollins and you said the books were published by HarperCollins.
5. Therefore: This agrees with what you said.

1. You said: Real Chance of Love was an American reality TV show. Season 2 of the show was won by Cali, who chose to be with Chance.
2. I checked: Who won season 2 of Real Chance of Love?
3. I found this article: Real Chance of Love 2: Back in the Saddle is the second season of the VH1 reality television dating series Real Chance of Love. Ahmad Givens (Real) and Kamal Givens (Chance), former contestants on I Love New York are the central figures.
4. Reasoning: The article doesn't answer the question and you said that Cali won season 2 of Real Chance of Love.
5. Therefore: This is irrelevant to what you said.

1. You said: The Stanford Prison Experiment was conducted in the basement of Jordan Hall, Stanford's psychology building.
2. I checked: Where was Stanford Prison Experiment conducted?
3. I found this article: Carried out August 15-21, 1971 in the basement of Jordan Hall, the Stanford Prison Experiment set out to examine the psychological effects of authority and powerlessness in a prison environment.
4. Reasoning: The article said the Stanford Prison Experiment was conducted in Jordan Hall and you said the Stanford Prison Experiment was conducted in Jordan Hall.
5. Therefore: This agrees with what you said.

1. You said: Social work is a profession that is based in the philosophical tradition of humanism. It is an intellectual discipline that has its roots in the 1800s.
2. I checked: When did social work have its roots?
3. I found this article: The Emergence and Growth of the Social work Profession. Social work's roots were planted in the 1880s, when charity organization societies (COS) were created to organize municipal voluntary relief associations and settlement houses were established.
4. Reasoning: The article said social work has its roots planted in the 1880s and you said social work has its root in the 1800s.
5. Therefore: This disagrees with what you said.

1. You said: The Havel-Hakimi algorithm is an algorithm for converting the adjacency matrix of a graph into its adjacency list. It is named after Vaclav Havel and Samih Hakimi.
2. I checked: What is the Havel-Hakimi algorithm?
3. I found this article: The Havel-Hakimi algorithm constructs a special solution if a simple graph for the given degree sequence exists, or proves that one cannot find a positive answer. This construction is based on a recursive algorithm. The algorithm was published by Havel (1955), and later by Hakimi (1962).
4. Reasoning: The article said the Havel-Hakimi algorithm is for constructing a special solution if a simple graph for the given degree sequence exists and you said the Havel-Hakimi algorithm is for converting the adjacency matrix of a graph.
5. Therefore: This disagrees with what you said.

1. You said: "Time of My Life" is a song by American singer-songwriter Bill Medley from the soundtrack of the 1987 film Dirty Dancing. The song was produced by Michael Lloyd.
2. I checked: Who was the producer of "(I've Had) The Time of My Life"?
3. I found this article: On September 8, 2010, the original demo of this song, along with a remix by producer Michael Lloyd , was released as digital files in an effort to raise money for the Patrick Swayze Pancreas Cancer Resarch Foundation at Stanford University.
4. Reasoning: The article said that a demo was produced by Michael Lloyd and you said "Time of My Life" was produced by Michael Lloyd.
5. Therefore: This agrees with what you said.

1. You said: Tiger Woods is the only player who has won the most green jackets. He has won four times. The Green Jacket is one of the most coveted prizes in all of golf.
2. I checked: What is the Green Jacket in golf?
3. I found this article: The green jacket is a classic, three-button, single-breasted and single-vent, featuring the Augusta National Golf Club logo on the left chest pocket. The logo also appears on the brass buttons.
4. Reasoning: The article said the Green Jacket is a classic three-button single-breasted and single-vent and you said the Green Jacket is one of the most coveted prizes in all of golf.
5. Therefore: This is irrelevant to what you said.

1. You said: Kelvin Hopins was suspended from the Labor Party because he had allegedly sexually harassed and behaved inappropriately towards a Labour Party activist, Ava Etemadzadeh.
2. I checked: Why was Kelvin Hopins suspeneded from the Labor Party?
3. I found this article: A former Labour MP has left the party before an inquiry into sexual harassment allegations against him was able to be concluded, the party has confirmed. Kelvin Hopkins was accused in 2017 of inappropriate physical contact and was suspended by the Labour party pending an investigation.
4. Reasoning: The article said Kelvin Hopins was suspended because of inappropriate physical contact and you said that Kelvin Hopins was suspended because he allegedly sexually harassed Ava Etemadzadeh.
5. Therefore: This agrees with what you said.

1. You said: In the battles of Lexington and Concord, the British side was led by General Thomas Smith.
2. I checked: Who led the British side in the battle of Lexington and Concord?
3. I found this article: Interesting Facts about the Battles of Lexington and Concord. The British were led by Lieutenant Colonel Francis Smith. There were 700 British regulars.
4. Reasoning: The article said the British side was led by Lieutenant Colonel Francis Smith and you said the British side was led by General Thomas Smith.
5. Therefore: This disagrees with what you said.

1. You said: **Prosenjit Chatterjee is a renowned Bengali actor, whose career has been marked by critical acclaim in his early life, followed by personal challenges and a resurgence in popularity and commercial success in his later years.**
2. I checked: **West Bengal: What type of recognition has marked Prosenjit Chatterjee's early life in his film career?**
3. I found this article: **June 4, 2023 National recognition and accolades did not lure Prosenjit away from West Bengal's entertainment industry. He began this year with a stellar performance in Kaushik Ganguly's period thriller, Kaberi Antardhan, shot against the backdrop of the Naxalite movement and the Emergency.**
4. Reasoning:

Figure A7: Mixtral + RARR: The prompt used by RARR (Gao et al., 2023) for checking the agreement of the retrieved evidence for editing.

This task involves processing a claim by attributing it based on a set of evidences. The aim is to refine the initial claim into an attributed claim that incorporates insights from all provided evidences.
Instructions:
1. Identify the main entity discussed in the provided claim. Carefully review all associated evidences. Note that the evidences may or may not be relevant to the main entity of the claim.
2. Determine the relevance of each piece of evidence to the main entity in the claim. Synthesize the factual information from relevant evidences to assess how they support, refute, or modify the initial claim.
3. Generate an attributed claim that effectively integrates the initial claim with the relevant evidences, ensuring that the main entity of the claim remains unchanged, especially in the context of any irrelevant evidence.
4. Do not include unnecessary evidence sentences in the modified claim which were not present in the original claim. You are required to check only the factual correctness of the claim without adding extra information to the claim.
Example:
Claim: Tata Motors is an Indian multinational automobile manufacturing company headquartered in Mumbai, Maharashtra, India. It was established in 1954.
Evidences:
1. Mahindra & Mahindra Limited (M&M) is an Indian multinational automotive manufacturing corporation headquartered in Mumbai. It was established in 1945 as Mahindra & Mohammed and later renamed Mahindra & Mahindra.
2. Tata Motors was founded in 1945, as a locomotive manufacturer. Tata Group entered the commercial vehicle sector in 1954 after forming a joint venture with Daimler-Benz of Germany in which Tata developed a manufacturing facility in Jamshedpur for Daimler lorries.
Attributed Claim: Tata Motors is an Indian multinational automobile manufacturing company headquartered in Mumbai, Maharashtra, India. It was established in 1945.
Claim: Feluda is a detective novel written by renowned Bengali actor Sandip Ray, first published in West Bengal in 1965 by Ananda Publishers. The book has been adapted into a film and several television series.
Evidences:
1. Feluda is an Indian-Bengali detective media franchise created by Indian-Bengali film director and writer Satyajit Ray, featuring the character, Feluda.
2. In 1965, at the age of 44, soon after the release of his landmark film Charulata, Satyajit Ray wrote the first draft of a short story, which featured a young boy, barely into his teens, describing the superlative analytical and detection powers of his older cousin brother."
Attributed Claim: "Feluda is a detective novel written by renowned Bengali author Satyajit Ray, first published in West Bengal in 1965 by Ananda Publishers. The book has been adapted into a film and several television series.
Claim: Leonardo DiCaprio won his first Oscar for Best Actor for his role in the film ’Titanic’ in 1996.
Evidences:
1. Leonardo DiCaprio has been nominated for the Best Actor Oscar multiple times, beginning with his role in ’What’s Eating Gilbert Grape’ in 1993.
2. DiCaprio’s performance in ’The Revenant’ was universally acclaimed, and he won the Academy Award for Best Actor in 2016, which was his first Oscar win.
3. Leonardo DiCaprio is an active environmentalist who has donated millions to conservation efforts.
Attributed Claim: Leonardo DiCaprio won his first Oscar for Best Actor for his role in ’The Revenant’ in 2016, after several nominations for other films including his first for ’What’s Eating Gilbert Grape.’
Claim: Avengers: Endgame was released worldwide in April 2018 and became the highest-grossing film of all time by surpassing ’Titanic’.
Evidences:
1. Avengers: Endgame was released in April 2019. It quickly garnered acclaim for its dramatic conclusion of the Infinity Saga."
2. In July 2019, ’Avengers: Endgame’ surpassed ’Avatar’ to become the highest-grossing film ever, a record it held until ’Avatar’ reclaimed the title after a re-release."
3. The soundtrack for ’Avengers: Endgame’ was composed by Alan Silvestri, who also composed music for ’Back to the Future.’"
Attributed Claim: Avengers: Endgame was released worldwide in April 2019 and became the highest-grossing film of all time by surpassing ’Avatar’ in July of that year, although ’Avatar’ later reclaimed the top spot.
For this claim and evidences, generate the attributed claim as instructed.
Claim: **Prosenjit Chatterjee is a renowned Bengali actor, whose career has been marked by critical acclaim in his early life, followed by personal challenges and a resurgence in popularity and commercial success in his later years.**
Evidences:
1. June 4, 2023 National recognition and accolades did not lure Prosenjit away from West Bengal’s entertainment industry. He began this year with a stellar performance in Kaushik Ganguly’s period thriller, Kaberi Antardhan, shot against the backdrop of the Naxalite movement and the Emergency.
Attributed Claim:

Figure A8: Mixtral + RARR: The prompt used for the non-sequential editing of the text.

Given a claim, query and an evidence, check the following: (i) if the evidence ANSWERS the query and (ii) if the claim INCORRECTLY answers the query, make this judgement based only on the evidence. If both the conditions are satisfied, then return a score 1 else return a score 0. Also provide a reason for your score.

For example,
Claim: Revathy is a renowned Indian actress and humanitarian, who has won several accolades including two National Film Awards and three Filmfare Awards.
Query: Kerala: What are some of the accolades won by Revathy, the Indian actress from Kerala, including National Film Awards and Filmfare Awards?
Evidence: She has won several accolades, including three National Film Awards , and six Filmfare Awards South. She has also won the Kerala State Film Award for Best Actress for her performance in Bhoothakaalam (2022). Early life Revathi was born as Asha Kelunni Nair in Cochin (present-day Kochi) to Malank Kelunni Nair, a major in the Indian Army , who hails from Palakkad, and Lalitha Kelunni who hails from a Palakkad Tamil family. When she was in school, she took part in a fashion show.
The score and reason are:
Score: 1
Reason: The evidence answers the query and the claim claim incorrectly answers the query based on my knowledge from the evidence. The evidence said Revathy has won three National Film Awards, six Filmfare Awards South, and Kerala State Film Award for Best Actress but the claim said two National Film Awards and three Filmfare Awards.

Claim: Tata Motors is an Indian multinational automobile manufacturer headquartered in Mumbai, Maharashtra. It was founded by Jamsetji Tata and established the company on August 1, 1945.
Query: India: Which industry does Tata Motors operate in, as a prominent player in the Indian market?
Evidence: Tata Motors has established itself as a leading player in the Indian automotive market, enjoying a substantial market share and a strong customer base. Competitive advantage in low-cost production: With a low-cost labor base in India, Tata Motors has a competitive advantage in producing economical segment vehicles. This advantage allows the company to target not only the Indian market but also other emerging markets, leading to substantial profits. Innovation and research and development: Tata Motors is known for its excellent innovation and research and development efforts in the automotive sector.
The score and reason are:
Score: 0
Reason: The evidence answers the query but the claim correctly answers the query based on my knowledge from the evidence. The evidence said that Tata Motors is a prominent player in India's automotive market and Tata Motors is an Indian multinational automobile manufacturer headquartered in Mumbai, Maharashtra.

Now answer for the below sample,
Claim: **Prosenjit Chatterjee is a renowned Bengali actor, whose career has been marked by critical acclaim in his early life, followed by personal challenges and a resurgence in popularity and commercial success in his later years.**
Query: **West Bengal: What type of recognition has marked Prosenjit Chatterjee's early life in his film career?**
Evidence: **June 4, 2023 National recognition and accolades did not lure Prosenjit away from West Bengal's entertainment industry. He began this year with a stellar performance in Kaushik Ganguly's period thriller, Kaberi Antardhan, shot against the backdrop of the Naxalite movement and the Emergency.**
The score and reason are:

Figure A9: Mixtral Revised: The prompt used for filtering the evidences that are relevant to the entity in the text and for the target location.

You are a localization assistant. Convert the reference entity sentence from English to the Indian domain by localizing the reference sentence with a similar entity from the target location. Make the needed modifications in the sentence to make it factually correct for the target location with the help of evidences. Output answers in English using multi-entity localization. Provide a reason for your localization. Keep the word count of the generated sentence almost the same as the reference sentence.

Examples:

My reference sentence: Ford Motor Company is an American multinational automobile manufacturer headquartered in Dearborn, Michigan, United States. It was founded by Henry Ford and incorporated on June 16, 1903.
Target location: India
Evidences: [“1: Tata motors were founded by J. R. D. Tata.", “2: Tata Motors was founded in 1945, as a locomotive manufacturer. Tata Group entered the commercial vehicle sector in 1954 after forming a joint venture with Daimler-Benz of Germany in which Tata developed a manufacturing facility in Jamshedpur for Daimler lorries."]
The target sentence and the reason are:
Target sentence: Tata Motors Limited is an Indian multinational automobile manufacturer headquartered in Mumbai, Maharashtra, India. It was founded by J. R. D. Tata and incorporated in 1945.
Reason: Tata Motors Limited is a good localization for Ford Motor Company in the Indian context. From the evidences, it is a multinational automobile manufacturer headquartered in Mumbai and it was founded by J. R. D. Tata and incorporated in 1945.

My reference sentence: A train derailment occurred on February 3, 2023, at 8:55 p.m. EST, when 38 cars of a Norfolk Southern freight train carrying hazardous materials derailed in East Palestine, Ohio, United States.
Target location: Andhra Pradesh
Evidences: [“1: On 29 October 2023, around 7:00 pm, the collision occurred on the Howrah–Chennai main line after Visakhapatnam-Palasa Express service train stopped due to a break in an overhead cable when it was hit by an oncoming passenger train travelling from Visakhapatnam in Andhra Pradesh, to Rayagada in Odisha, derailing its three carriages, in the Vizianagaram district of Andhra Pradesh, India. The collision occurred between Kantakapalli and Alamanda railway stations resulting in severe damage to three coaches of the Palasa passenger, and the locomotive and two coaches of the Rayagada passenger. At least 14 people were killed and 50 others were injured as a result."]
The target sentence and the reason are:
Target sentence: A train derailment occurred on October 29, 2023, around 07:00 p.m. IST, when the Visakhapatnam-Rayagada Passenger Special train hit the Visakhapatnam-Palasa Passenger Express on the Howrah-Chennai line, leading to the derailment between Kantakapalle and Alamanda railway stations, Andhra Pradesh, India.
Reason: An example of train derailment in Andra Pradesh would be Vizianagaram train derailment. The localization was done by replacing the location, date, and time details of the reference sentence with those from the provided evidence related to a train derailment in Andhra Pradesh, India.

My reference sentence: **Robert John Downey Jr. is an American actor. His career has been characterized by critical success in his youth, followed by a period of substance abuse and legal troubles, and a surge in popular and commercial success later in his career.**
Target location: **West Bengal**
Evidences: **[“1: Prosenjit Chatterjee (born 30 September 1962) is an Indian actor and producer. He is widely regarded as one of the leading actors in modern Bengali cinema. He predominantly works in Bengali cinema. He is the son of veteran Bollywood actor Biswajit Chatterjee.", “2: June 4, 2023 National recognition and accolades did not lure Prosenjit away from West Bengal’s entertainment industry. He began this year with a stellar performance in Kaushik Ganguly’s period thriller, Kaberi Antardhan, shot against the backdrop of the Naxalite movement and the Emergency.", “3: Prosenjit Chatterjee began his film career in West Bengal’s entertainment industry with critical acclaim in the 1980s. He received national recognition and accolades for his roles in period thriller Kaberi Antardhan, shot against the backdrop of the Naxalite movement and the Emergency, released this year. After facing personal challenges, he has experienced a resurgence in popularity and commercial success in recent years."]**
The target sentence and the reason are:

Figure A10: Mixtral Revised: The prompt used for re-generating text with the help of the retrieved evidence.

Given a category, a reference location, reference sentence, reference entity, target location, and a target sentence, detect the target entity in the target sentence.

For example:

Category: Actor/Actress
Reference location: US
Reference sentence: Angelina Jolie is an American actress, filmmaker, and humanitarian. The recipient of numerous accolades, including an Academy Award and three Golden Globe Awards, she has been named Hollywood's highest-paid actress multiple times.
Reference entity: Angelina Jolie
Target location: Kerala
Target sentence: Revathy is a renowned Indian actress and humanitarian from Kerala who has won several accolades including two National Film Awards.
The entity detected and the reason:
Target entity detected from the target sentence: Revathy
Reason: The sentence talks about actress Revathy.

Category: Amusement park
Reference location: France
Reference sentence: Disneyland Paris is an entertainment resort in Chessy, France, 32 kilometers east of Paris. It encompasses two theme parks, resort hotels, a shopping, dining and entertainment complex, and a golf course.
Reference entity: Disneyland Paris
Target location: India
Target sentence: Haw Par Villa is an entertainment resort in India. It includes multiple thrill rides, water attractions, theme parks, a shopping and dining complex, and hotel accommodations.
The entity detected and the reason:
Target entity detected from the target sentence: Haw Par Villa
Reason: The sentence talks about Haw Par Villa.

Category: **Actor/Actress**
Reference location: **US**
Reference sentence: **Robert John Downey Jr. is an American actor. His career has been characterized by critical success in his youth, followed by a period of substance abuse and legal troubles, and a surge in popular and commercial success later in his career.**
Reference entity: **Robert Downey Jr**
Target location: **West Bengal**
Target sentence: **Prosenjit Chatterjee is a renowned Bengali actor, whose career has been marked by critical acclaim in his early life, followed by personal challenges and a resurgence in popularity and commercial success in his later years.**
The entity detected and the reason:

Figure A11: The prompt used for extracting entity from a sentence for checking Entity Correctness (EC).

Given a target location, category, possible target entity and a target entity to check, provide the score as below:
1) If the target entity EXACTLY matches the provided possible target entity, provide score 2.
2) Else if it is another POSSIBLE factually correct replacement for the possible target entity in the given target location, provide score 1. The entities should belong to the same given category and share similar properties.
3) Else, provide score 0.
Also, provide a reason for the score.

For example:

Target location: India
Category: Airlines
Possible target entity: Indigo
The score and reason are :
The target entity to check: Indigo
Score: 2
Reason: Exact match of entity.

Target location: Kerala
Category: Actor/Actress
Possible target entity: Manju Warrier
The target entity to check: Revathy
The score and reason are:
Score: 1
Reason: Revathy is also a renowned actress from Kerala similar to Manju Warrier.

Target location: India
Category: Amusement park
Possible target entity: Wonderla
The target entity to check: Haw Par Villa
The score and reason are:
Score: 0
Reason: Haw Par Villa is a theme park in Singapore but the target location is India, hence it is not a correct localization.

Target location: **West Bengal**
Category: **Actor/Actress**
Possible target entity: **Prosenjit Chatterjee**
The target entity to check: **Prosenjit Chatterjee**
The score and reason are:

Figure A12: The prompt used for evaluating a sentence on the Entity Correctness (EC) metric.

Given a target sentence, target location and a question, check whether the target sentence answers the given question correctly for the target location. Note that there could be multiple correct answers to the question for the target location. If the entity is from the target location and if the sentence correctly answers the question for the target location then assign a score of 1 and else assign a score of 0. Also, provide a reason for the score. For example:

Target sentence: ASCATSAT-1 is an Earth observation satellite developed by the Indian Space Research Organisation (ISRO). It was launched in 2016 to provide ocean wind vector data for weather forecasting, cyclone detection, and tracking. After its successful completion of the mission, it was handed over to the India Meteorological Department for routine operations.
Target location: India
Question: Mention the name of an earth observation satellite that consists of two satellites?
For the above target sentence, target location and question, the score and reason would be:
Score: 0
Reason: The target sentence mentions ASCATSAT-1, which is an Earth observation satellite developed by ISRO. However, it does not specify that ASCATSAT-1 consists of two satellites. The question specifically asks for an Earth observation satellite that consists of two satellites, and this information is not provided in the target sentence.

Target sentence: A train derailment occurred on February 3, 2023, at 8:55 p.m. IST, when 38 cars of a Vizianagaram freight train carrying hazardous materials derailed in Andhra Pradesh, India.
Target location: Andhra Pradesh
Question: Can you mention a train accident?
For the above target sentence, target location and question, the score and reason would be:
Score: 1
Reason: The target sentence clearly mentions a train derailment, which is a type of train accident, and specifies that it occurred in Andhra Pradesh. Thus, it answers the question in the context of the target location.

Target sentence: Amitabh Bachchan is an Indian actor and producer. He is widely regarded as one of India's leading actors, having appeared in a wide range of films in the protagonist role.
Target location: India
Question: Name an actor who has appeared in a wide range of films in an antagonist role?
For the above target sentence, target location and question, the score and reason would be:
Score: 0
Reason: The target sentence specifies that Amitabh Bachchan has appeared in films in the protagonist role, not the antagonist role. Therefore, it does not answer the question in the context of the target location (India).

Target sentence: **Prosenjit Chatterjee is a renowned Bengali actor, whose career has been marked by critical acclaim in his early life, followed by personal challenges and a resurgence in popularity and commercial success in his later years.**
Target location: **West Bengal**
Question: **Can you name an actor who has achieved commercial success in their career?**
For the above target sentence, target location and question, the score and reason would be:

Figure A13: The prompt used for evaluating a sentence on the Common Question Correctness (CQ) metric.

Given an input sentence and true sentence, check if the input sentence is factually correct based on the true sentence provided. If the input sentence is factually correct compared to the true sentence, then assign a score of 1, else 0. Also, provide a reason for the score. If the input has any extra information that is not present in the true sentence, ensure that this extra information is factually correct based on your world knowledge.
For example:

Input sentence: Manju Warrier is a renowned Indian actress, predominantly working in the Malayalam film industry, based in Kerala. She has won numerous accolades for her acting skills, including four Kerala State Film Awards and a National Film Award. Often hailed as one of the finest actresses in the Malayalam cinema, Manju Warrier is a cultural icon in Kerala. has won several accolades including National Film Awards.
True sentence: Manju Warrier is an Indian actress, She has won numerous accolades, including the Kerala State Film Awards, and won a special mention from the jury for the National Film Awards. She is hailed as one of the finest actresses in the Malayalam cinema. She is hailed as one of the finest and highest-paid actresses in the Malayalam cinema.
The score and the reason:
Score: 1
Reason: All the details mentioned about Manju Warrier in the input sentence factually match with the true sentence.

Input sentence: Kempegowda International Airport is a ‘The Green themed’ entertainment and retail complex linked to the passenger terminals of Bengaluru Airport.
True sentence: Kempegowda International Airport is a ‘Naurasa themed’ with entertainment and retail area connected to the passenger terminals of Kempegowda International Airport, Bengaluru.
The score and the reason:
Score: 0
Reason: Kempegowda International Airport is not a ’The Green themed’ airport but a ’Naurasa themed’ according to the true sentence.

Input sentence: **Prosenjit Chatterjee is a renowned Bengali actor, whose career has been marked by critical acclaim in his early life, followed by personal challenges and a resurgence in popularity and commercial success in his later years.**
True sentence: **Prosenjit Chatterjee is an Indian actor and producer. His career has been characterized by critical success in commercial films. He debuted in Parallel Cinema and since then have appeared in numerous art films and has achieved commercial success in his career.**
The score and the reason:

Figure A14: The prompt used for evaluating a sentence on the Factual Correctness (FC) metric using the true target sentence from LoFTI dataset. This evaluation is performed when the target entity detected is an exact match to the true target entity.

Given a sentence, check if the input sentence is factually correct for the given target location. A possible target entity and its factually correct sentence are given for your reference. Note, that there can be multiple correct entities for the given location.

If the input sentence is factually correct compared to the true sentence, then assign a score of 1, else 0. Also, provide a reason for the score.

Factually correct sentence of **Prosenjit Chatterjee** in **West Bengal**: **Prosenjit Chatterjee is an Indian actor and producer. His career has been characterized by critical success in commercial films. He debuted in Parallel Cinema and since then have appeared in numerous art films and has achieved commercial success in his career.**

Now check if the below input sentence is also a factually correct statement for **Badshah Moitra** in **West Bengal**:

Input sentence: **Badshah Moitra is an Indian actor primarily known for his work in Bengali television and films. He has gained recognition for his performances in both mainstream and independent projects. While his career began in television, he has since appeared in several films, establishing himself as a respected actor in the Bengali entertainment industry.**
The score and the reason:
Score: `<fill_score_here>`
Reason: `<fill_reason_here>`

Figure A15: The prompt used for evaluating a sentence on the Factual Correctness (FC) metric using the true target entity and true target sentence from LoFTI dataset as a possible example. This evaluation is performed when the target entity detected is NOT an exact match to the true target entity.