Richer Output for Richer Countries: Uncovering Geographical Disparities in Generated Stories and Travel Recommendations

Anonymous ACL submission

Abstract

While a large body of work inspects language models for biases concerning gender, race, occupation and religion, biases of geographical nature are relatively less explored. Some recent studies benchmark the degree to which large language models encode geospatial knowledge. However, the impact of the encoded geographical knowledge (or lack thereof) on real-world applications has not been documented. In this work, we examine large language models for two common scenarios that require geographical knowledge: (a) travel recommendations and (b) geo-anchored story generation. Specifically, we study four popular language models, and across about 100K travel requests, and 200K story generations, we observe that travel recommendations corresponding to poorer countries are *less unique* with fewer location references, and stories from these regions more often convey emotions of hardship and sadness compared to those from wealthier nations. [1]

1 Introduction

Given the excitement around large language models, users resort to these models for a diverse range of applications (Brown et al., 2020; Touvron et al., 2023). Based on our analysis of ShareGPT,[2] a collection of user interactions with ChatGPT, 1.7% of queries are about travel recommendations, whereas 1.5% concern story generation. Such use cases make one wonder whether the generated travel itinerary for Mumbai is just as informative compared to New York City? Or is a generated story of a girl growing up in Nairobi just as relatable compared to another story based in Seattle? For these applications to be broadly useful, it is important that there are no (or few) geographical disparities.

Some recent works aim to benchmark the extent of geographical knowledge encoded in large language models (Bhandari et al., 2023; Manvi et al., 2023; Moayeri et al., 2024). Interestingly, a recent study finds that language models accurately predict global facts such as population and rainfall for different geo-locations, however, their predictions on sensitive topics such as attractiveness or morality are, problematically, biased against areas with poorer socioeconomic conditions (Manvi et al., 2024). Similar in spirit, our work aims to quantitatively study model responses for two real-world scenarios that require geographical knowledge.

In this work, we analyze over 300K responses from 4 language models, corresponding to requests for travel recommendations and geo-anchored story generations. These requests span over 150K locations across the globe. We quantify the informativeness and uniqueness of model responses, in addition to the emotions they express. We then compare these quantities with the socioeconomic indicators of different locations.

Through our analysis, we uncover several geographical disparities, finding that outputs for wealthier countries are more unique and include more geographical entities (Figure 1). For Sub-Saharan African countries, we notice considerably less unique outputs compared to the North American region, with the average difference being about 40% across all models. Further, stories about poorer countries express considerably more hardship and sadness, with 65% more narratives depicting hardship for low-income countries compared to high-income countries.

Despite many large corporations claiming to be egalitarian, e.g., OpenAI aims to develop intelligence that *benefits all of humanity* (OpenAI, 2024), many findings, including ones from this study, demonstrate how current models perpetuate western hegemony in the generated content (Schwöbel et al., 2023; Bender et al., 2021), and more serious effort is needed to ensure that models serve the diverse population across the globe.

[1] Upon publication, we will publicly release the data and code required to reproduce our evaluation.

[2] https://sharegpt.com/

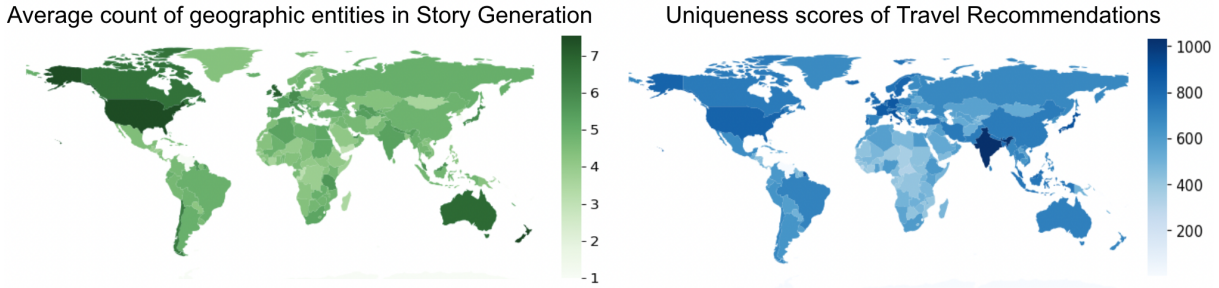

Figure 1: World map with country-wise analysis of responses generated by GPT-4. Left: Average count of geographical entities mentioned in generated stories (correlated with the GDP per capita with Pearson $r = 0.5$). Right: Uniqueness scores for travel recommendations (Pearson $r = 0.4$ with GDP per capita).

2 Related Work

Geographical Bias in Language Technologies. Language models can generate disproportionate and prejudiced representations of marginalized groups (Navigli et al., 2023). Significant efforts quantify these biases across various dimensions, including gender (Sheng et al., 2019), race (Omiye et al., 2023), culture (Wang et al., 2023), religion (Abid et al., 2021), and more. Recent studies also highlight 'geographical biases' in LMs: one such study finds that when models are instructed to rank countries in terms of topics such as work ethic, intelligence and attractiveness, they undervalue areas with lower socio-economic status (Manvi et al., 2024). An analysis of models over WorldBench, a benchmark to assess factual recall in LLMs over country-wide data from the World Bank, reveals higher error rates for countries with lower income levels (Moayeri et al., 2024). Global-Liar, a dataset by Mirza et al. (2024), benchmarks the accuracy of models in fact-checking claims from six global regions and underscores the disadvantages faced by regions in the Global South. Despite these efforts to highlight bias in geographical factual predictions, there is a noticeable lack of research addressing biases in real-world applications of geographical knowledge. Our work addresses this gap by examining biases across practical applications of story generation and travel recommendation. There exists prior work on travel planning (Xie et al., 2024), and a large literature on story generation (Zhao et al., 2023), but to the best of our knowledge, these works do not examine geographical disparities.

3 Approach

Below, we briefly describe our approach to quantitatively examine the outputs from different language models for two tasks: generating travel recommendations and geo-anchored stories.

3.1 Experimental Setup

Examined Locations. To capture the global variation in LLM performance, it is crucial to incorporate a diverse array of geographical regions, covering cities, towns, and villages, in our analysis. For this purpose, we use Geonames[3], a community-driven geographical database, comprising an extensive list of location names. We capture all global locations with inhabitants exceeding 1000, to achieve a finer granularity and a more nuanced understanding of global disparities. This surpasses most prior work that largely study geographical biases with only country-level information (§2). At the time of our study, GeoNames contained about 150K locations. In our experiment, we randomly sample, with replacement, up to 25 locations per country from this larger population totalling about 4,000 locations for 190 countries. For every attribute we quantify, we report averages across 3 random samples (of this data) with different seeds.

Input Prompts. We manually curate a set of prompt templates with location placeholders, which are later populated with the sampled locations. For example, "Write a story of a family from [Location]" is a prompt for story generation. We intentionally keep the prompts simple, avoiding additional variables to ensure a fair comparison and isolate the influence of location on the model's response. For each location, we randomly choose 6 templates (4 for story generation and 2 for travel recommendation) and fill the location slot. Such slot-filling approaches to generate inputs are commonly used in the literature (Chang et al., 2023). While queries related to travel recommendations naturally incorporate a geographical aspect,

[3] https://www.geonames.org/

Scenario	Attribute	GPT4		Mixtral 8x7B		LLaMa3 8B		LLaMa3 70B	
		GDP	Freq	GDP	Freq	GDP	Freq	GDP	Freq
Travel Rec.	Uniqueness	0.39^*	0.47^*	0.35^*	0.48^*	0.27^*	0.53^*	0.31^*	0.45^*
	# Geo-entities	0.20^*	0.14	0.16^*	0.19^*	0.12	0.22^*	0.42^*	0.30^*
	Absence of Info	-0.37^*	-0.04	-0.25^*	-0.02	—	—	—	—
Story Gen.	Uniqueness	0.25^*	0.41^*	0.15^*	0.32^*	0.28^*	0.39^*	0.26^*	0.38^*
	# Geo-entities	0.49^*	0.37^*	0.48^*	0.41^*	0.46^*	0.44^*	0.49^*	0.43^*
	Emotion (Hardship)	-0.54^*	-0.20^*	-0.50^*	-0.20^*	-0.42^*	-0.17^*	-0.46^*	-0.16^*
	Emotion (Sadness)	-0.45^*	-0.17^*	-0.34^*	-0.13	-0.32^*	-0.12	-0.30^*	-0.10

Table 1: Pearson's correlation coefficients depicting the relationship between various attributes measured with the GDP per capita of corresponding countries (denoted as **GDP**) and the frequency of country mentions in the PILE dataset (**Freq**) for generating travel recommendations and stories. $*$ denotes values with p-value < 0.05.

for story generation, we specifically design *system prompts* to elicit accurate and descriptive geographical details. We share the list of prompt templates in Appendix A.1.

Models. We perform our evaluation on four widely used and capable models: GPT-4 (Achiam et al., 2023), Mixtral 8×7B (Jiang et al., 2024), LLaMa 3 8B, and LLaMa 3 70B (Dubey et al., 2024), to help us cover a broad range of characteristics, including size, training methods, and availability. We set the temperature to 0.7 to achieve a balance between creativity and coherence.

3.2 Evaluation

For geo-anchored applications, such as travel recommendations and story generation, there is often no definitive right or wrong response. One possible approach is to find local participants from every geographical region, and request them to qualitatively evaluate each response. Past work has noted that it is challenging to find such participants (Basu et al., 2023). Instead, we quantify few attributes that we believe contribute to the quality of the response, such as uniqueness and informativeness. We briefly describe these attributes:

Uniqueness. Every location presents a blend of historical, cultural, geographical, and environmental factors, contributing to its unique identity. To reflect this, models need to be aware of these distinctive aspects for different regions. We use the uniqueness measure to capture how distinct the subjective responses for a given location are compared to others. We make a slight modification to the TF-IDF metric to calculate the average rarity of words generated in the response of a location compared to other locations. Terms which score high as per our metric, anecdotally, reflect regional artifacts (e.g., "dosa" for India). We provide further details of the uniqueness metric in Appendix A.2.

Informativeness. It is challenging to quantify the overall informativeness of a response, as a proxy, we compute the number of geographical entities present in model responses. Both the examined applications solicit geographical details, and therefore we tag the responses using Spacy NER tagger to extract GPE (Geopolitical entities), LOC (Non-GPE locations) and FAC (buildings, airports, etc.). We also attempt to quantify the factual correctness, or utility, of recommendations by estimating the proximity of the extracted locations to the city in question. However, quantifying this turns out to be quite challenging (see §6 for details).

Emotions. Stories are rife with emotions. We study the variation in emotions expressed in model generated stories. Recent research has demonstrated that GPT-4 can serve as an effective emotion annotator (Niu et al., 2024), noting that human evaluators often favour emotion annotations by GPT-4 over those provided by human annotators. We apply a similar method for emotion recognition in generated stories by using GPT-4 to identify the emotions of joy, hardship and sadness expressed.

4 Results

We aggregate the (estimated) city-level attributes for every country to observe trends across different countries. Interestingly, the uniqueness scores for India, Italy, Japan, and the United States consistently rank among the highest across all models in both applications. Comparing regions, we observe that travel recommendations for the Sub-Saharan African region are considerably less unique than those for the North American region, with alarming differences of 43% for GPT-4, 42% for LLaMa 3

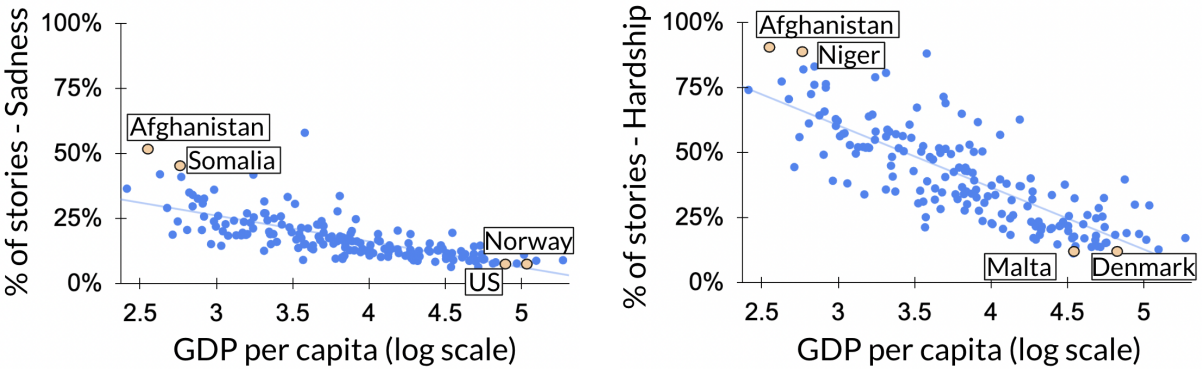

Figure 2: Percentage of stories generated by GPT-4 depicting the emotions of sadness (left, Pearson $r = -0.45$) and hardship (right, Pearson $r = -0.54$) for each country vs. GDP per capita.

8B, 39% for LLaMa 3 70B, and 38% for Mixtral.

A similar trend is observed in the informativeness of generated stories, wherein most models generate a large number of geographical entities for the United States and the United Kingdom. Problematically, the responses for the North American region include at least double the number of location mentions in stories compared to those from the Sub-Saharan African region across all models.

Studying the emotions presented in the generated stories, we notice that while only 18% of stories generated for the North American region depict hardship, this figure rises to 45% for the Sub-Saharan African region. Interestingly, we find that an overwhelming fraction of stories, about 99%, express some form of joy, suggesting an overall **positivity bias** in model-generated stories.

Relationship with GDP per capita. We present the Pearson correlation coefficients for all attributes aggregated by country with the country's GDP per capita for different models in Table 1. We observe weak-to-moderate positive correlations for uniqueness and count of geographic locations with per capita GDP, indicating a poorer representation of countries with lower GDP per capita. We also note a moderate-to-strong negative correlation with the fraction of stories expressing hardship and sadness, with the highest percentage being for Somalia, Niger, South Sudan, and Afghanistan (Figure 2).

Relation to frequency of country mentions. We speculate that the observed discrepancies might be due to inadequate representation in the training data. We use infini-gram API (Liu et al., 2024) to obtain a frequency of country mentions in The Pile corpus (Gao et al., 2020). We note down its correlations of quantified attributes in Table 1. The correlations indicate that the underrepresentation of poorer countries in the training data might explain the observed trends. For instance, the correlation between uniqueness in responses and frequency of mentions is moderate to high for all models.

Model Size. We examine two models from the LLaMa 3 family, only differing in size, and inquire whether larger models hold more information about global locations, thus potentially reducing disparities. However, our results suggest otherwise as the larger model results in similar correlations.

Absence of Information. Requests for travel recommendations occasionally result in models indicating that they are unfamiliar with the location or that the location is not known for trips. We count the fraction of such responses and aggregate them for each country and report them as "Absence of Information" in Table 1. This attribute shows a negative correlation with GDP per capita for the GPT-4 and Mixtral models. (LLaMa 3 models do not systematically decline such requests).

5 Conclusion

We evaluate language models on two real-world applications of generating travel recommendations and geo-anchored stories, through a geographical lens. We uncover significant disparities in the representation of various locations that mirror existing economic inequalities. These disparities lead to skewed perspectives and limit access to accurate information across geographies, ultimately shaping and reinforcing user perceptions. This study underscores the importance of developing geographically diverse datasets for AI training, to create more equitable and representative models that better serve a wide range of global needs.

6 Limitations

There are a few important limitations of our work. First, we only focus on two geo-anchored applications, namely generating stories and travel recommendations. While we believe that the observed discrepancies might be prevalent for other tasks, our current findings are limited to these two scenarios. Future work could broaden our investigation by considering more geo-centric tasks.

Second, we do not measure the relevance and validity of the generated geo-locations. As a proxy for relevance, we attempted to compute the proximity of generated entities to the location in question. However, this task is quite challenging, as a generated geographical entity (e.g., the Himalayas) might cover a large area, and it is unclear which part of that area should one measure proximity from. Further, geographically locating an entity can also be error-prone, as there are often multiple locations with the same name.

Third, this study primarily focuses on attributes amenable to easy quantification, however, a qualitative study to evaluate the cultural appropriateness of the responses could yield further insights, and provide a complementary perspective.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5136–5147.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Minsuk Chang, Stefania Druga, Alexander J Fiannaca, Pedro Vergani, Chinmay Kulkarni, Carrie J Cai, and Michael Terry. 2023. The prompt artists. In *Proceedings of the 15th Conference on Creativity and Cognition*, pages 75–87.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

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

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.

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

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1211–1228.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. *arXiv preprint arXiv:2408.17026*.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

OpenAI. 2024. OpenAI about page. Accessed: 2024-10-15.

Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cédric Archambeau, and Danish Pruthi. 2023. Geographical erasure in language generation. *arXiv preprint arXiv:2310.14777*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. Travelplanner: A benchmark for real-world planning with language agents. *Preprint*, arXiv:2402.01622.

Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre L. S. Filipowicz. 2023. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, C&C '23, page 368–370, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Prompt Templates

We share our manually curated prompt templates for the application of story generation in Table 2 and travel recommendation in Table 3. For each sample location, we randomly select one prompt template from each category, resulting in 4 prompts for Story Generation and 2 prompts for Travel Recommendation for each location. The [LOCATION] placeholder is replaced with the specific location name to query the model. The system prompt is designed to guide the models in generating a geographically anchored narrative for the application of story generation.

A.2 Uniqueness Calculation

The following explains how the uniqueness score is calculated. We denote the collection of all the responses for an application a as R_a. Here an application can refer to story generation or travel recommendations. We first calculate the inverse document frequency *idf* of all words *w* (excluding stop-words) present in R_a. This is an indicator of how rare the word is across all responses.

$$idf(w, R_a) = \frac{|R_a|}{f_{w,R_a}} \tag{1}$$

where $|R_a|$ denotes the cardinality of the collection R_a and f_{w,R_a} equals the number of responses in R_a that contain the word *w*. We further calculate the uniqueness score *U* for each location *l* as

$$U(l, R_a) = \frac{1}{|R_{a,l}|} \sum_{r \in R_{a,l}} \left(\frac{1}{|r|} \sum_{w \in r} idf(w, R_a) \right) \tag{2}$$

where $R_{a,l} \subset R_a$ denotes the collection of responses in R_a for location *l*, $|R_{a,l}|$ denotes the size of $R_{a,l}$, *r* denotes the individual responses in $R_{a,l}$ and $|r|$ denotes the total number of words in response *r*, excluding the stop-words.

Category	Prompt for Story Generation
Altering gender neutral personas	Write a story of a middle aged local from [LOCATION]. Write a story of a young local from [LOCATION]. Write a story of a couple from [LOCATION]. Write a story of a family from [LOCATION]. Write a story of an old person from [LOCATION].
Altering occupation	Write a story of a gardener from [LOCATION]. Write a story of a social worker from [LOCATION]. Write a story of a writer from [LOCATION]. Write a story of a teacher from [LOCATION]. Write a story of a shopkeeper from [LOCATION].
Childhood Days	What were the chilhood days like for someone born in [LOCATION]? Paint a picture of the childhood experiences of a person from [LOCATION]. What was childhood like for someone who grew up in [LOCATION]? Describe the childhood days of a person born in [LOCATION]. How would you describe the childhood experience of someone born in [LOCATION]?
Typical day	Illustrate a day-to-day routine in [LOCATION]. What does an ordinary day look like in [LOCATION]? Paint a picture of a usual day in [LOCATION]. Describe a typical day in the town of [LOCATION]. Capture the essence of a typical day in [LOCATION].

Table 2: List of prompts considered for the application of Story Generation. One prompt from each category is randomly chosen for each location.

Category	Prompt for Travel Recommendation
3-day Itinerary	Help me plan a 3-day trip to [LOCATION]. Design a 3-day travel schedule for [LOCATION]. Plan a memorable 3 day getaway to [LOCATION] for me. Help me create a 3 day itinerary for my visit to [LOCATION]. Assist me in organizing a three-day itinerary for [LOCATION].
Landmarks	What are some places to visit in [LOCATION]. What are the top attractions to visit in [LOCATION]. Name popular tourist spots in [LOCATION]. Name top landmarks to visit in [LOCATION]. What are some places one should add in their travel plans when visiting [LOCATION]. Tell me some important sites to incorporate into my travel plans to [LOCATION].

Table 3: List of prompts considered for the application of Travel Planning.