

Unveiling Cultural Bias: LLMs Are Biased towards English Over Indian Stories

Akshay Rao

Received October 15, 2024

Accepted January 14, 2025

Electronic access January 31, 2025

Due to the large capacity of Large Language Models and their capacity to generate human-like text, they are increasingly being applied in the health care sector, legal industry, and education sectors. However, it faces issues like hallucination, and overconfidence. This paper tackles the issue of cultural bias in Large Language Models by contrasting its performance on stories written in English and in Indian languages. The results suggest that there is a possibility of cultural bias attributed by Large Language Models that fare better contents in the language of English, thus contributing to the marginalization of non-dominant cultures by AI-generated text. Findings are noted towards incorporating mitigating bias strategies by Large Language Models whose influence is on an increasing trajectory.

Introduction

Large language models have recently been in high demand. There are hundreds of versions they can be applied to, and lately, their application has been noticed in the most sensitive areas: from plain natural language processing up to complicated solution finding in serious spheres, like healthcare or law, and education^{1,2,3}. That is now related to their ability to analyze vast amounts of data⁴, generate text more human-like, and provide innovative solutions in various contexts.

However, despite these LLM strengths, several grave issues have been raised about LLMs. The most common among them is hallucination⁵ whereby entirely false or inaccurate information is produced by the model. There is also this tendency to exhibit overconfidence, where wrong answers are presented as facts, and there are high levels of computational resources needed for the proper functioning of the model.

Another key concern about them is that they tend to display bias, for which implications have been assessed in various contexts and are of much concern. Our paper focuses on the issue with bias in LLMs. We theorize that these models can genuinely be more accurate in representing stories from the culture of researchers who developed them than those from other cultures, for example, stories in English versus others. If that is the case, this uncontrolled bias can result in the erasure or marginalization of non-dominant cultures⁶ when training LLMs; a prospect both troubling and unsettling as such models spread.

We conducted a cross-comparison study over multiple LLMs, with a bias towards their capability of presenting the right English and Indian origin stories. We had a few stories ready, prompted the models, and started analyzing responses. From

our results, evidence can be seen that bias does exist in cross comparisons by the LLMs, but other models are still worse than yet others.

Background

Large language models like GPT-3⁷ and BERT⁸ are at present the new trend in Natural Language Processing. Such models rely on massive data and complex architectures to carry out diverse activities that include but are not limited to text generation, translation, and comprehension. Now we describe how LLMs work, then the main prominent approaches for LLMs, and finally, we will discuss some of the issues they face and common approaches to solve this.

Transformer and Attention Mechanisms

Attention mechanisms⁹, particularly in the Transformer architecture, are significantly important for training LLMs like GPT-3 and BERT. These mechanisms-Scaled Dot-Product Attention and Multi-Head Attention, to name a few-allow LLMs to weigh different words according to their importance and manage long-range dependencies in text. Transformers replace the traditional RNNs with layers that can be parallelized which improves on efficiency and accuracy of text processing and generation. This natively enables the LLMs to capture context and produce more coherent answers through a decoder-encoder structure as afforded by the architecture of Transformers¹⁰ as well as similar self-attention mechanisms.

Key Players

Thousands of influential frameworks have been developed using the Transformer framework, and hence, it is largely instrumental in governing the impact of large language models over the domain. Among them, the most famous is the Generative Pretrained Transformer, or GPT⁷ for short, which was developed by OpenAI. This model has been a landmark for LLMs and has made progress both in natural language processing and generation. Very influential also is BERT⁸, short for Bidirectional Encoder Representations from Transformers, which originates from Google. Instead, BERT masters contextual understanding from both sides and is therefore very proficient at tasks such as question answering and sentiment analysis. Then of course Google also implemented T5¹¹, Text-to-Text Transfer Transformer, where it perturbs all natural language processing problems as text generation, yet this is just another application of the versatility of the Transformer architecture. However, there are countless open-source models that have surfaced in the public eye but have not gained the same level of popularity. One of these is GPT-Neo¹², an open-source model based on work by EleutherAI, providing an open-source equivalent to the GPT-3 model that was created by OpenAI. Another one is BLOOM¹³, a multilingual open-access large language model, developed by BigScience, which seeks to make powerful language models accessible to people worldwide. However effective these models are, they do pose the problem of biases in large language models since such systems can reflect and amplify the biases present in training data¹⁴.

Biases in Large Language Models

The major problems with Large Language Models are that it involves some risks associated in themselves. For instance, this is recounted in the paper "On the Dangers of Stochastic Parrots"¹⁴ where they point out significant dangers associated with LLMs development. This includes the enormous environmental cost generated by resource-intensive training methods, thus raising carbon footprint issues. In addition, it explains how the reliance on uncurated information perpetuates current social biases, which in turn causes dangerous stereotypes to seep into such models. It further states that although so advanced LLMs generate text without understanding anything.

On similar lines, "Gender Bias in Transformers: A Comprehensive Review of Detection and Mitigation Strategies"¹⁵ elaborates on the methods transformers are used such that they help generate gender biases and thus lead to biased results. In a similar way, various approaches like WEAT and Equalized Odds for bias estimation are compared with that of bias improvement in the models. The paper also discussed issues surrounding standard metrics and mitigation strategies that allow equity in artificial intelligence systems.

However, the paper "Debiasing Pre-Trained Language Models

via Efficient Fine-Tuning"¹⁶ established a new way of debiasing LLMs without losing performance. Here, the authors detail an efficient way of fine-tuning only less than 1% of GPT-2's parameters on datasets such as WinoBias and CrowS-Pairs. The results on benchmarks about bias, StereoSet, show that bias can reside in a few percent of the parameters. Publicly available, fine-tuned, this model brings the community this compromise between bias mitigation and model performance retention.

Methodology

Let us now look at four of the most popular LLMs. OpenAI's GPT-4o and GPT-3.5, Google's Gemini, and Microsoft's Copilot. These were especially selected for this work as these LLMs have advanced abilities. GPT-4o brought from OpenAI was updated on August 6th, 2024, while free access is available only to a certain number of prompts. At the same time, the older version of OpenAI was tested- namely, GPT-3.5- in order to test its capability in comparison with GPT-4. The selected models also include Google's Gemini and Microsoft's Copilot, which include the two most powerful natural language processing models. All the models were accessed during the period between August 10, 2024 and September 1, 2024. All the models were evaluated over the same period using the latest versions available at that time. This ensured uniformity in the testing environment and minimized the effects of versioning differences on the results. Open-source models, such as BLOOM, were excluded because of limited accessibility and relevance to general public use cases. Proprietary models, like GPT-4 and Gemini, were chosen because they are the most widely available tools to non-technical users, which fits the purpose of the study to examine the real-world implications.

We compiled nineteen stories, nine from Indian history and ten from the history of England, depicting the major events and their lessons learned from both. For Indian stories, we referred to a Sanskrit scholar for cultural authenticity. English stories have been collected from credible online platforms. Tables 1 and 2 describe the stories, for example "Moon Ridiculing Ganesha" and "The Hare and the Tortoise," along with the number of key points central to each story. These models were given extremely short prompts and were given only enough context to test for their inference capabilities. Each story started with the following template: "Can you tell me the story of [name of the story]?" That way, each story and each model was left open for standardized testing.

We divided the generated points into True Positives, False Positives, and False Negatives in order to assess the output of the models. Precision, Recall, and F1 Score were the metrics we used to measure how well the models could exactly and comprehensively generate items. Precision measured how many correctly generated points were valid and Recall measured the number of relevant points which were successfully generated.

The F1 Score bridged this gap between two metrics, hence giving a comprehensive performance measure. Figure 1 illustrates an actual computation of such metrics.

Results

In this section, we present the experimental results of the four major LLMs. We report the aggregate metrics for every model: precision, recall, and F1 score, focusing specifically on the Indian and the English text narratives. Then, we will highlight some outliers observed that are anomalously far from the mean scores.

We score each LLM based on the following performance metrics: precision, recall, and F1 score. The scores reflect how well the models perform in providing relevant story points. We summarize these in Table 1.

The research indicates that the Indian stories are found performing the poorest among the models, with higher number of false positives (hallucinations) and negatives (omissions). For example, the GPT-3.5 fabricated a whole new story on Arjuna and the Parrot out of ignorance to the extent of cultural knowledge. The mistakes in the English stories, on the other hand, were not as serious, revealing that the models are rather better knowledgeable of culture than the other. Of course, this performance does not focus on absolute numbers but on a relative gap dividing the performance divide of English from that of Indian stories, as that gap brings cultural biases into focus.

Outliers

The Indian story with one of the worst scores is ‘Moon ridiculing Ganesha’, it is Gemini’s score. A part of the response generated is presented in table 2. We refer to the appendix for the full generated response, as well as for the full responses to the outliers presented in this section.

The text which is in bold is false information. The model has missed the first four points along with the last point, this completely excludes the context of the moon making fun of the lord and it also excludes the condition that Ganesha set when he forgave the Moon. Hence it completely changes the story having missed these points.

The Indian story with one of the worst scores is ‘Arjuna and the Parrot’, it is GPT 3.5’s score.

The text which is in bold is false information. The model has missed almost all the points entirely, it has altogether generated a completely different story which is not true according to the original sources of the data. Hence it has a very poor score.

The English story with the worst score is ‘The Crab and the Fox’, which is ChatGPT’s score.

The text which is in bold is false information. The model has missed almost all the points entirely, it has altogether generated

a completely different story with a different moral. Hence it has a very poor score.

Discussion

Model comparisons were made to evaluate general-use biases. This paper did not consider variations in size, architecture, or accessibility since its focus was on finding the biases that are evident across general-purpose LLMs. While no formal statistical tests were conducted due to the limited dataset, this preliminary study acknowledges the need for such analyses in future research. Statistical validation, including tests like t-tests or ANOVA, will be critical in further studies to confirm the significance of observed performance gaps. We further aggregate the precision recall and F1 scores across models to determine if LLMs are biased in narrating more tales in English than in Indian languages. For this purpose, we have developed what we call a “Culture Gap Score”, which compares the storytelling capabilities of English and Indian cultures. Its intention is to be a score that shows gaps rather than absolute performance. Equal performance across cultures may be suboptimal, but large gaps are signs of cultural bias. The study focuses on the differences between precision, recall, and F1 scores for the English and Indian stories, rather than on their absolute values. For instance, false positives such as hallucinated details and false negatives such as omitted critical points were examined regarding their effects on cultural fidelity. In one case, Moon Ridiculing Ganesha was an Indian story that was grossly misrepresented, and key narrative elements were omitted, which in turn changed the story’s moral. In contrast, the thematic integrity of the English stories was preserved in general. Future research should explicitly control for story complexity and match narratives for length and detail. This will ensure that cultural biases are not confounded by differences in story difficulty. High positive culture gap score would suggest that the method has a bias towards content in English. Citing these stages, the table below (Table 4) provides differences in precision, recall and F1 of the assessed models.

The evaluation of the index "Culture Gap Score" presents several important results regarding the performance discrepancy of large language models (LLMs) over English versus Indian stories. Beginning with precision, the difference between these two types of content has a range of 0.06 (ChatGPT) to 0.26 (Gemini). This means that Gemini shows the strongest preference for relevant content in English rather than in Indian stories. In general, the models exhibit a weighted average precision difference of 0.17, which means that LLMs are less effective in classifying Indian stories as correct when compared to English stories.

When we look at recall, the gap is larger, varying between 0.01 (ChatGPT) and 0.29 (CoPilot). Thus, CoPilot not only has a higher accuracy in identifying relevant content in English stories

Table 1: The table is structured with each model listed in the first column, followed by the precision, recall, and F1 score for each story they processed. The stories are organized into two categories: Indian and English. For each story, the model's scores are recorded in three columns representing precision, recall, and F1 score, respectively. At the end of each section (Indian and English), the average precision, recall, and F1 score for each model are calculated and displayed, including their 95% Confidence Intervals.

	Story	ChatGPT			ChatGPT (Free)			Gemini			Copilot		
		Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
English	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	1	0.87	0.93	0.6	0.5	0.54	1	0.83	0.9	1	0.87	0.93
	3	1	1	1	0.5	0.6	0.54	0.75	1	0.85	1	0.8	0.88
	4	1	1	1	0.6	0.6	0.6	0.69	0.9	0.78	1	1	1
	5	0.8	0.7	0.74	1	1	1	0.9	0.9	0.9	1	1	1
	6	1	1	1	0.7	0.7	0.7	1	1	1	1	1	1
	7	0.14	0.28	0.18	1	1	1	1	1	1	1	1	1
	8	0.8	0.88	0.83	1	1	1	0.88	1	0.94	1	1	1
	9	1	0.9	0.94	1	1	1	1	1	1	1	1	1
	10	1	0.85	0.92	0.6	0.6	0.6	1	0.9	0.95	1	1	1
	Avg.	0.87	0.84	0.85	0.9	0.8	0.8	0.92	0.95	0.93	1	0.96	0.98
95% CI	((0.68, 1.00))	((0.69, 1.00))	((0.67, 1.00))	((0.65, 0.95))	((0.65, 0.95))	((0.64, 0.95))	((0.84, 1.00))	((0.91, 1.00))	((0.88, 0.99))	((1.00, 1.00))	((0.92, 1.00))	((0.95, 1.00))	
Indian	1	0.9	1	0.94	0.6	0.5	0.54	0.3	0.4	0.34	0.8	0.6	0.68
	2	0.8	0.9	0.84	0.7	0.5	0.62	0.7	0.6	0.64	0.9	0.6	0.72
	3	1	1	1	0.07	0.1	0.09	0.7	0.7	0.7	0.9	0.8	0.84
	4	0.6	1	0.75	0.9	0.7	0.8	0.7	0.7	0.7	0.8	0.5	0.61
	5	0.9	0.8	0.84	1	0.8	0.92	0.9	0.7	0.78	0.9	0.7	0.78
	6	0.9	0.8	0.84	0.8	0.7	0.76	0.8	0.8	0.8	0.7	0.7	0.7
	7	1	1	1	0.9	0.8	0.86	1	0.8	0.88	0.8	0.6	0.68
	8	0.9	0.7	0.78	1	0.7	0.81	0.8	0.7	0.74	1	0.7	0.9
	9	0.8	1	0.8	1	1	1	0.9	0.8	0.84	0.9	1	0.94
	10	0.3	0.2	0.24	0.2	0.2	0.2	0.4	0.5	0.44	0.4	0.5	0.44
	Avg.	0.81	0.83	0.8	0.7	0.6	0.64	0.66	0.67	0.64	0.81	0.67	0.73
95% CI	((0.75, 1.00))	((0.62, 0.98))	((0.69, 0.99))	((0.56, 1.00))	((0.50, 0.90))	((0.54, 0.98))	((0.64, 0.96))	((0.70, 0.90))	((0.68, 0.92))	((0.58, 0.82))	((0.59, 0.81))	((0.60, 0.80))	

Table 2: The table contains wrong parts of the story in the response generated by the model.

Response generated
Ganesha and the Naughty Moon
Once upon a time, there was a big party in heaven. Everyone was invited, including all the gods and goddesses. Lord Ganesha, the elephant-headed god, was there too. He had a wonderful time eating delicious food and playing fun games.
When it was time to go home, Ganesha was feeling very happy and full. He was so happy that he started to laugh and his tummy jiggled. As he walked through the sky, the naughty moon saw him and started to laugh too.

compared to Indian stories but also does so at a significantly higher rate. The weighted average recall gap of 0.26 across the models indicates that they are more able to retrieve relevant English content than relevant Indian content, highlighting a potential bias in this metric. Regarding the F1 score, which is the balance of precision and recall, the difference ranges from 0.05 (ChatGPT) to 0.29 (Gemini), with the weighted mean difference standing at 0.22. This indicates that models on average generate more reliable English stories when both precision and recall are considered. Once again, Gemini has the widest gap,

suggesting that all metrics tend toward a similar bias, while the performance of ChatGPT is less extreme, suggesting the two types of stories are more equally represented in its output. The smaller performance gaps seen in ChatGPT have likely resulted from OpenAI's particular training methodology. Investigation into this would be expectative because the study did not have access to the proprietary details concerning the architecture.

Surprising observations:

One thing that has very noticeably been seen is that ChatGPT works comparatively well, showing a difference of 0.06 in precision and 0.01 in recall between the Indian and English narratives but an F1 score gap of 0.05. This means that ChatGPT can find Indian content with similar precision and recall as English content. Another interesting feature is that the gap between the F1-score is largest for Gemini and CoPilot as they both have scores that are more disparate than ChatGPT. This shows over-optimizing for precision and recall when working on English stories, at the expense of this performance on Indian stories.

Error type differences:

Another insight derived is by studying the type of error each makes-for instance, ChatGPT has a smaller F1 gap yet has a larger precision gap, meaning it is less probable to make up

things when processing the Indian narratives but is weak at inaccuracy, perhaps from a lack of domain knowledge. On the contrary, CoPilot has larger errors in all three rankings, hence it can deliver not only false information but also misinterpret or be unconscious of crucial cultural context as part of Indian stories. Such differences may indicate that models like ChatGPT address knowledge gaps by possessing a better recall of stories for India while models like CoPilot may hallucinate or make more errors due to a lack of contextual understanding.

Outliers and Error trends:

One worthy of mention outlier that truly deserves a closer look is CoPilot, which has the highest recall difference of 0.29, which suggests that it has a better recall on English stories. As this model appears to favor recall-centered strategies with the trade-off that provides higher recalls in comparison to precision, it may give off hints of hallucinations in processing Indian stories. It's interesting that ChatGPT shows a minimum gap in precision compared to other alternatives and exemplifies neutrality while classifying relevant information in not only the English, but the Indian narratives as well.

Conclusion

Our research work started with the question, "Are Large Language Models biased in their understanding and reply to English cultural history?". From the literature available on this topic, we had hypothesized that there was a bias in these models. Our answer, as it turns out at the end of our research, is affirmative to this question but not uniformly across all the models considered.

Culture bias in LLMs can have deep real-world implications, especially in the space of education, public policy, and media. A biased representation in educational tools can result in a marginalization of non-dominant cultures and create less homogenous views that fail to see and appreciate cultural difference. For example, suppose the generation of educational content in LLMs prioritizes the narratives of dominant cultures; in that case, it may erase or misrepresent significant cultural traditions and thereby alienate students of marginalized backgrounds. The same can be said of public policy, where biased outputs may further systemic inequalities through slanted insights or recommendations because the model understands one culture and not another. Such policies may at times unintentionally favor the norms of dominant cultures while ignoring the needs of diverse communities. Then again, in media and communication, this kind of bias will go on to strengthen stereotypes, cause misinformation, and promote cultural homogenization. Curbing such biases is not merely a technical problem, but a societal necessity because equal AI

applications would be important in making sure that technology reflects the various realities of its global users.

Addressing cultural bias at the training level is a much more complicated task that requires far more extensive research than that of this study. However, prompt engineering was proposed as an immediate mitigation strategy. As such, the subsequent work on these models should concentrate more on nullifying cultural biases surfacing through these models, especially in connection with the English cultural heritage. Our paper itself is nothing but proof for the bias in such models, variations in biasness perceived between models can suggest that a single solution may not yield. Improving the prompts used should serve as the starting point for a wholesome approach to deal with the discrepancies arising from here. When such prompts are presented clearly and in an unambiguous manner especially when there are those stories or personalities across cultures, it helps models differentiate clearly and thus receive more accurate results. It would also help prevent possible misinterpretation, hence the portrayal of cultural characteristics being more faithful. An important area for further improvement is the information sources on which these models rely. While internet resources are voluminous, they often include error or slanted opinion and may amplify the cultural biases that already exist. Future research will require inclusion of validation and source appraisal steps led by experts.

More importantly, the dividing line between spoken and written traditions should be better accounted for because oral histories carry cultural meanings and nuances that written forms ignore. More inclusion of oral histories and consideration of a wider range of communication would allow models to more accurately reflect the diversity of cultural contexts.

Methods

The capabilities of the two AI models: OpenAI's GPT-4 and GPT-3.5 and those developed by Google, known as Gemini, and Microsoft Copilot are investigated in this research. The model GPT-4¹⁷ has been designed by OpenAI to suit all its users-different limitations have been put in place between paid and free users who have unlimited access to it. The prior model before the release of GPT-4 was that of GPT-3.5¹⁸, the most recent update the model received was on February 16, 2024. In the same line of thinking, Google's Gemini and Microsoft's Copilot are two other heavyweights in the AI industry with many features as well. Each of these was tested and rated as being great in the natural language processing ecosystem and for its frequent usage and now can be accessed in free as well as paid versions. But in this paper, there is no open-source model employed as testing and availability were limited. All models were accessed within the period of August 10th, and September 1st, 2024, therefore their assessment took place on the most updated versions of these models at that time.

Story Collection

Stories were deliberately chosen for their simplicity, often being narratives told to children. This ensured comparability across cultures and minimized the influence of complexity as a confounding factor.

The English stories were obtained from reputable online sources, which would most probably coincide with the training data of LLMs. This inherent overlap may have contributed to the bias as these stories are more familiar to the models. Attempts were made to verify this overlap by trying to find the same stories or similar ones online. Indian stories are available online, but also include many errors, omissions or nuances which are completely missing in the culturally authentic forms. This further reflects how systemic advantages English stories might have during model training, by their availability and uniformity of presentation in online datasets.

The Indian stories were selected and validated by a Sanskrit expert for cultural and historical authenticity, while the English stories were reviewed and cross-validated by another reviewer. The emphasis of the review was more on the message of the story rather than on elaborate cultural or historical details. This way, both sets of stories were parallel in their thematic and moral intent, thus comparable for the study.

In total, twenty stories were compiled—ten of which belonged to the Indian history domain and ten of which had roots in English history. The Indian stories were obtained from a Sanskrit expert for cultural as well as historical authenticity, while the English stories were taken from reliable online sources and were reviewed by a native English citizen. Each story tried to keep the background part as brief as possible with only one or two points relevant to the context being created.

Focus was then directed to the critical events in every story and the driving moral or moral behind it. Tables 5 and 6 present the stories used in the study, indicating the number of key points gleaned to be important in stating the story. To view the full exposition of these key points, the reader is referred to the appendix section where full exposition can be found.

Prompt Design

In the design of the questions for this study, our goal was to provide as little context as possible.

This methodology intentionally avoided the provision of rich context in order to obtain responses free from bias. Differences in phrasing were not explored, as the aim of the study was to emphasize biases in general model responses rather than to optimize for cultural representation. Therefore, each prompt was created using a new context so that the models can reset between different stories, ensuring they are not swayed by possibly previous outputs. The question asked was: "Could you repeat the story of [.]."

The uniformity in the design of the prompts allowed us to test the performance capabilities of each model given the constrained input and to assess their ability in understanding and communicating the core elements of a story.

Evaluation Metrics

The researcher used a preset prompt and a list of key points for reference to assess model outputs. Responses were saved to ensure reproducibility and transparency, and these can be made available upon request. The study sought to minimize evaluation bias by using a reference list of key points for comparison. For instance, in the story Hanuman in Ravana's Court, the model output was compared against points like "Hanuman setting Lanka ablaze" to assess completeness and accuracy. Future work should involve multiple annotators to ensure inter-rater reliability. To verify the efficiency of our LLMs, we assigned the points produced into three different categories: True Positives, False Positives, and False Negatives. True Positives referred to points indicating the right information in the story, synonymous with the plot and message of the story. Points that were wrong or had nothing to do with the story were False Positives, also known as Hallucinations. Finally, False Negatives, also known as Omissions, represent important points of the original story that the model failed to generate. These classes allowed us to test strengths and weaknesses in the quality of their outputs. In the story Moon Ridiculing Ganesha, the omission of key details such as the Moon's mocking of Ganesha and the conditional forgiveness fundamentally altered the story's moral and narrative arc. Similarly, in Arjuna and the Parrot, the model fabricated(hallucinated) a completely new narrative involving a parrot named Chaitanya, replacing the original context and significantly deviating from the story's intended focus on concentration. Precision is said to be the number of all the true positives present in all the positive predictions. It gives the number of points that the model has generated and are correct. It can be given as:

$$\text{Prec} = \frac{TP}{TP + FP}$$

For instance, if a model generates 8 correct points and returns 2 unrelated ones, its precision would be 80. Another significant measure is recall which pertains to the ability of the model to recognize all relevant aspects of the problem at hand. This aspect can be expressed mathematically as the number of True Positives divided by the sum of True Positives and False Negatives. This helps in finding out the amount of information that the model fails to capture to be considered significant. It is given as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

In such a case, if a model produces 8 accurate points and loses 2 essential ones, in this situation, the recall will also

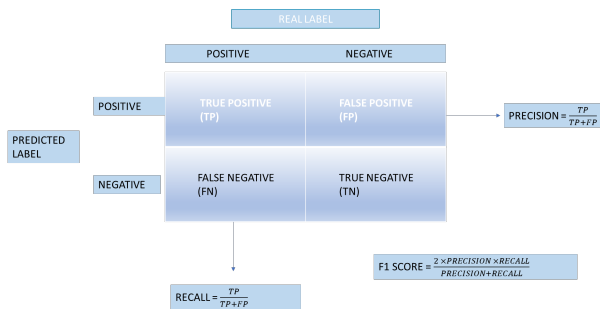


Fig. 1 These are the formulas which have been used along with the confusion matrix

be equal to 80%. F1 score is the last metric computed that gives out a report of the performance of the model by taking into consideration the two, namely precision and recall. The statement can mathematically be put as:

$$F_1 = \frac{2 \times \text{Prec} \times \text{Recall}}{\text{Prec} + \text{Recall}}$$

This means that if the model in question could maintain a satisfactory F1 score, it has a good trade-off between precision and recall. These metrics are further summarized in figure 1.

Statistical Testing

For all metrics we report 95% confidence intervals using the t-distribution. In addition to this, for the reported cultural gap values we performed a one-tailed Welch’s t-test to determine the significance of the reported change. For this, we used the standard t-distribution and we used the Welch–Satterthwaite method to compute degrees of freedom. Significance level was set at 0.05.

Table 3: The table contains wrong parts of the story in the response generated by the model.

Response generated
<p>Sure! The story of "Arjuna and the Parrot" is a popular tale from Indian folklore that showcases themes of wisdom, friendship, and the bond between humans and animals.</p> <p>### The Story of Arjuna and the Parrot</p> <p>Once upon a time, in a lush forest, there lived a wise young prince named Arjuna. He was known for his valor and keen intellect. One day, while exploring the woods, he stumbled upon a beautiful parrot trapped in a hunter’s net. The parrot, with its vibrant green feathers and intelligent eyes, pleaded for help.</p> <p>Moved by compassion, Arjuna carefully freed the bird. The parrot, grateful for its newfound freedom, introduced itself as Chaitanya and promised to be Arjuna’s friend. They quickly formed a deep bond, sharing stories and adventures in the forest.</p> <p>One day, while wandering through the woods, they came across a magnificent lake. Arjuna noticed that the water was shimmering strangely and decided to investigate. The parrot warned him, saying, “That water is enchanted. It can reveal your deepest desires, but be cautious.”</p> <p>Curiosity got the better of Arjuna, and he drank from the lake. Suddenly, visions of power and glory flooded his mind. He saw himself as a great warrior, conquering kingdoms and earning fame. However, as he reveled in these fantasies, the parrot spoke up again, reminding him, “True greatness lies in wisdom, not just power.”</p> <p>Realizing the parrot’s wisdom, Arjuna stepped back from the lake. He understood that while ambition was important, it should not come at the cost of his values. The friendship with Chaitanya had taught him the importance of humility and wisdom.</p> <p>From that day on, Arjuna focused on becoming a just and wise leader, valuing friendship and integrity over mere conquest. He and the parrot continued to explore the forest together, their bond growing stronger as they faced challenges and learned from one another.</p> <p>In time, Arjuna became a revered king, known not just for his strength but for his wisdom. He often credited his friendship with Chaitanya for guiding him on the right path.</p> <p>### Moral of the Story</p> <p>The tale highlights the importance of true friendship and the value of wisdom over ambition. It teaches us that the relationships we nurture can shape our character and lead us toward the right choices in life.</p>

Table 3: The table contains wrong parts of the story in the response generated by the model.

Response generated
Certainly! "The Crab and the Fox" is a fable that teaches a lesson about cunning and overconfidence. Here's a version of the story:
Once upon a time, a crab lived by the sea. The crab often ventured out of the water to explore the land. One day, while strolling along the shore, the crab met a sly fox.
The fox, seeing the crab out of its natural element, thought it could easily make a meal out of the crab. With a cunning plan in mind, the fox approached the crab and said, "Dear crab, I've heard so much about your delicious taste. Why don't you join me for a walk in the forest? There, we can enjoy each other's company."
The crab, aware of the fox's reputation, was wary. However, it thought it could outsmart the fox by being cautious. "Thank you, dear fox," replied the crab, "but I'm quite happy here by the sea. It's where I belong, after all."
The fox, determined to have the crab for lunch, continued to persuade it. "But the forest is so beautiful and full of wonders," the fox insisted. "You should see it at least once."
The crab, realizing the fox wouldn't give up easily, decided to use its own wit. "Alright, fox, let's go for a walk," it said. "But first, allow me to find a suitable shell for protection."
The crab slowly moved toward the water, pretending to search for a shell. As soon as it reached the edge of the sea, it quickly scurried into the water and disappeared beneath the waves, safe from the fox's clutches.
The fox, outwitted and disappointed, realized it had underestimated the crab's intelligence and returned to the forest, empty-handed.
The moral of the story is that overconfidence can lead to one's downfall, and sometimes, the seemingly weaker or smaller can outsmart the strong and cunning.

Table 4: This table shows the differences in precision, recall, and F1 scores between English and Indian stories across the models, as well as their confidence intervals. *: P<.05, **:P<0.01, and ***:P<001.

	Precision	Recall	F1
ChatGPT	0.06 ([-0.15, 0.28])	0.01 ([-0.20, 0.22])	0.05 ([-0.16, 0.26])
FreeGPT	0.08 ([-0.17, 0.33])	0.20 ([-0.02, 0.42]) *	0.14 ([-0.10, 0.37])
Gemini	0.20 ([0.05, 0.36]) *	0.28 ([0.19, 0.38]) ***	0.25 ([0.13, 0.36]) ***
CoPilot	0.19 ([0.09, 0.29]) **	0.30 ([0.19, 0.40]) ***	0.25 ([0.16, 0.35]) ***
Avg.	0.17 ([0.09, 0.24]) ***	0.26 ([0.19, 0.32]) ***	0.22 ([0.15, 0.29]) ***

Table 5: These are the names of Indian stories along with the number of key points in each story

Moon ridiculing Ganesha	9 points
Hanuman in Ravana's court	10 points
Arjuna and the parrot	7 points
Hanuman crossing the ocean	7 points
Holika Dahan	10 points
Kaliya Mardhana	8 points
Krishna and Sudhama	10 points
Krishna, Arjuna's charioteer	6 points
Ajamila	12 points
Sant Tukaram's forgiveness	7 points

Table 6: These are the names of English stories along with the number of key points in each story

Golden eggs	5 points
Two bulls and the frogs	6 points
Shepherd	8 points
Hare and the tortoise	10 points
The dog and the Oyster	7 points
The man and the Lion	5 points
The crab and the Fox	4 points
The cat and the Birds	7 points
An old Lion and the Fox	4 points
The sick stag	5 points

References

- 1 T. Dave, S. A. Athaluri and S. Singh, *Frontiers in Artificial Intelligence*, 2023, **6**, 1169595.
- 2 H. Surden, *Fordham Law Review*, 2023, **92**, 1941.
- 3 S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlikad and V. Stankovski, *Internet of Things and Cyber-Physical Systems*, 2024, **4**, 19–23.
- 4 J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong and L. S. Chao, *arXiv preprint arXiv:2310.14724*, 2023.
- 5 L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin and T. Liu, *arXiv preprint arXiv:2311.05232*, 2023.
- 6 I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang and N. K. Ahmed, *Computational Linguistics*, 2024, 1–79.
- 7 T. B. Brown, *arXiv preprint arXiv:2005.14165*, 2020.
- 8 J. Devlin, *arXiv preprint arXiv:1810.04805*, 2018.
- 9 A. Vaswani, *Advances in Neural Information Processing Systems*, 2017.
- 10 S. Tipirneni, M. Zhu and C. K. Reddy, *ACM Transactions on Knowledge Discovery from Data*, 2024, **18**, 1–20.
- 11 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, *Journal of Machine Learning Research*, 2020, **21**, 1–67.
- 12 S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonnell, J. Phang and M. Pieler, *arXiv preprint arXiv:2204.06745*, 2022.
- 13 T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé and J. Tow, 2023.
- 14 E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.
- 15 P. Nemani, Y. D. Joel, P. Vijay and F. F. Liza, *Natural Language Processing Journal*, 2023, 100047.
- 16 M. Gira, R. Zhang and K. Lee, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 2022, pp. 59–69.
- 17 J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat and R. Avila, *arXiv preprint arXiv:2303.08774*, 2023.
- 18 J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen and J. Zhou, *arXiv preprint arXiv:2303.10420*, 2023.

Appendix

Below is a list of the collected stories, along with the main criteria for each, presented in bullet points.

Moon ridiculing Ganesha:

- Lord Ganesha is very fond of Modaks and sweets
- Once it so happened that Lord Ganesha ate a sumptuous dinner, with lots of sweets.

- After the meal, He set out for a breath of fresh air, on His beloved Mooshak.
- During his outing, Lord Ganesha fell off His Mooshak, and Lo! Out came all the
- Modaks and sweets that he had eaten!
- Chandra (Moon), who was watching this, laughed heartily and made fun of Lord Ganesha.
- Lord Ganesha got angry and cursed the Moon to lose His radiance.
- Accordingly, the Moon became dull and had no radiance.
- Chandra, realising His folly, begged for forgiveness from Lord Ganesha.
- The Lord forgave Chandra, but altered His words - anyone who sees Chandra on Ganesh Chaturthi night shall be falsely accused of a wrongdoing.

Hanuman in Ravana's court:

- Ravana's son Indrajit bound Hanuman with the power of Brahmastra.
- Hanuman took the opportunity to view Ravana's court, and have a meeting with the demon king.
- During His conversation with Ravana, Hanuman coiled His tail such that He sat at a height higher than that of Ravana.
- Hanuman the messenger warned Ravana and asked him to return mother Sita to Lord Rama.
- Ravana refused and ordered Hanuman to be killed.
- Vibhishana, the righteous brother of Ravana explained that a messenger should not be killed.
- Upon hearing this, Ravana ordered that Hanuman's tail be set ablaze.
- Once Hanuman's tail was set ablaze, He set fire to the whole of Lanka except
- Vibhishana's house.
- He then went to the ocean and cooled off His tail.
- His mission being more than accomplished, Hanuman flew back to His Lord.

Arjuna and the parrot:

- Dronacharya, Arjuna's teacher, decides to test the focus of each of his students.
- He places a parrot on the tree nearby and instructs all his students to assemble.
- Acharya then calls each of his students one by one.
- As each student gets set to hit the parrot with his arrow, the teacher asks what the student can view.
- Each student provides his answer based on his focus and view.
- Finally when it was Arjuna's turn, he said he could see only the eye of the parrot.
- Dronacharya was extremely impressed with the level of concentration of his favourite disciple.

Hanuman crossing the ocean:

- Hanuman, the mighty warrior decides to cross over the ocean.
- His purpose is to land in Lanka, the kingdom of Ravana and search for Mother Seeta.
- As he leaps across the ocean, he meets Mainaka, the mountain.
- Mainaka requests Hanuman to rest on Him for a while, but Hanuman politely refuses since He has an errand to complete!
- On His way, he also meets Surasaa, NaagaMaata. Recognizing Her, he alters

- His size and wins.
- Next comes a demoness who captures shadows. She tries her tricks with
- Hanuman, but He fights and wins over her, too.
- Hanuman finally crosses over the ocean and reaches Lanka, the glittering city of Ravana.

Holika Dahan:

- Prahlada, son of Hirankashipu, the king of demons was born in Sage Narada's hermitage.
- After failed attempts to convince Prahlada that he is very powerful, Hiranyakashipu decided to kill Prahlada.
- The demon king's sister had a received a boon.
- The boon given to Holika was that she cannot be set ablaze even during fire.
- Using this boon, Hirankashipu makes Holika and Prahlada sit in a pyre, and sets it ablaze.
- As the pyre burnt, Holika was set on fire, but not Prahlada.
- This surprised the onlookers.
- The reason for this occurrence was that Vishnu was always present to protect
- Prahlada, his devotee.
- Also, the boon given to Holika was to protect, and not to hurt.
- Even today, Holika Dahan is conducted as a social and religious event to mark the end of all evil, inside and outside of us.

Kaliya Mardhana:

- Krishna and His friends were playing in the Yamuna river along with their cows.
- Suddenly the waters of the Yamuna started turning into darker shade.
- As they warned each other and were all quickly getting out, some boys and cows fainted and lay still.
- Soon the multi headed serpent Kaliya was seen in the river.
- The cowboys called out to Krishna for help.
- Krishna dived into the waters and danced on the hood of the dangerous snake.
- This went on till Kaliya was tired and all his poison was spewed out.
- Krishna defeated Kaliya and told him never to return to the waters of Yamuna river.

Krishna and Sudhama:

- Sudhama was Lord Krishna's classmate and friend in the Gurukulam of Sandeepani Maharshi.
- Once learning at Gurukulam was completed, Krishna and Sudhama parted ways.
- Lord Krishna went on to become the King, while Sudhama lived as a poor Brahmin struggling to make ends meet.
- As Sudhama and his family struggled through poverty-stricken days, his wife suggested that he should meet his dear comrade Krishna and request help.
- Sudhama sets out to meet Krishna.
- After a long journey, he reaches the Lord's palace.
- Sudhama is treated with utmost affection and warmth by Lord Krishna.
- Sudhama gives Krishna a handful of puffed rice that his wife would have sent with him.
- Krishna happily eats His favorite puffed rice and distributes it too.

- Though Sudhama does not ask for help, by the time he goes back home, his family will be blessed with abundant wealth and prosperity.

Krishna, Arjuna's charioteer:

- There prevailed an interesting custom in ancient India. After a war, the charioteer of the victorious leader had to bow down to his master, and then, the master, being pleased with his services would gift him something precious.
- Lord Krishna, who had played the role of his charioteer had to follow this custom.
- Arjuna continued to wait for the Lord, his best friend and guide. But to his surprise, Krishna asked Arjuna to get off the chariot before him!
- Surprised and confused, Arjuna still chose to obey the omniscient Lord.
- And lo! The beautiful chariot that till then stood shimmering was set ablaze!
- Arjuna realized that the Lord had indeed saved him from this grave danger by being the last one to vacate the divine chariot.

Ajamila:

- Ajamila was a very devout and righteous man.
- He lead a disciplined life of austerities.
- But he fell into the company of a woman and started leading a life full of mundane pleasures.
- He also had children with this woman, one of whom was named Narayana.
- Narayana is one of the many beautiful names of Lord Vishnu.
- As Ajamila aged and grew weak, his end seemed near.
- Yama Dutas (the messengers of the Lord of death) arrived to take him.
- But he called out to his dear son, Narayana, which fortunately also was the Lord's name.
- So Vishnu Dutas (messengers of Lord Vishnu) also came.
- Vishnu Dutas did not allow Ajamila to be taken away by the messengers of Yama as he had chanted the Lord's name.
- Realizing the power of Lord's name, and re-enforcing upon himself the righteous ways of life, Ajamila used the rest of his life span to contemplate on God.
- Finally, when his end arrived, Ajamila attained Narayana Loka, the place where Lord Vishnu's devotees are present.

Sant Tukaram's forgiveness:

- Amongst the seventeenth century poets of Bharat, Sage Tukaram was very popular. He was a Marathi poet, steeped in devotion to Lord Vithoba, a form of Sri Krishna.
- There existed in ancient Bharat, a unique tradition known as Bhiksha. Sages, Yogis and others on the spiritual path would go from house to house begging for alms.
- As per this tradition, Sage Tukaram, too, would go begging for Bhiksha. A devout couple, whom Sant Tukaram daily visited, offered him alms with utmost affection, reverence and devotion.
- One morning when the couple had indulged in an argument, the lady threw out a piece of cloth when Saint Tukaram came begging for alms!
- Regretful of what she had done by evening, the couple immediately set out to look for the sage and beg his forgiveness.
- The piece of cloth that she had thrown was used by the saint to light numerous oil lamps in the temple.

Golden eggs:

- Once upon a time, a countryman possessed a goose who laid golden eggs every single day

- The countryman used to sell these eggs in the market. He soon began to become rich in this way
- The countryman grew impatient as the goose used to give him only one golden egg per day and he wished to grow his wealth faster
- Then one day, he decided to cut the goose open thinking it would give him all the golden eggs at once
- But when the deed was done, he did not find a single golden egg and now his source of golden eggs was gone

Two bulls and the frogs:

- Once upon a time, two bulls were fighting fiercely in a field.
- In the same field, there was a marsh in which a few frogs lived
- As one of the old frogs watched the battle, he trembled
- A young frog asked him what he was afraid of
- The old frog replied that once one of the bulls is defeated, he will be forced away from where he is in the field now up to the marsh here and we will be crushed
- As the old frog had said, the beaten bull was driven to the marsh, where his great hoofs crushed the frogs to death

Shepherd:

- Once a shepherd who was grazing his sheep wanted to amuse himself.
- He thought of playing a trick on the villagers.
- When the sheep were grazing one day, he cried 39; Wolf! Wolf!39;.
- All the villagers rushed to help.
- But the shepherd laughed aloud.
- He repeated the same act couple of times.
- One evening, the wolf actually attacked the sheep, and the shepherd again cried for help.
- But none of the villagers came to help, assuming that it is a prank that the boy is playing.

Hare and the tortoise:

- Once upon a time, a hare and a tortoise lived in a forest
- The hare was very proud of its fast speed
- It made fun of the tortoise for its slow speed
- The tortoise challenged the hare to have a race with him
- The hare accepted the challenge
- The race started and crow was the referee
- The hare ran very fast while the tortoise was left behind
- In the middle of the race, the hare stopped to take rest under a tree
- However the hare fell asleep and the tortoise passed him and reached the finish of the race
- The hare woke up and ran as fast as he could, but the tortoise had already won the race

The dog and the Oyster:

- There was once a dog who was very fond of eggs.
- He used to visit the hen very often and swallowed eggs.
- He was still very greedy and used to swallow anything that looked like eggs.
- One day he wandered to the seashore, where he spotted an oyster.
- In the next moment the Dog ate the oyster.
- It resulted in a lot of pain for the Dog.
- Then he painfully realized that not all round things were eggs.

The man and the Lion:

- A Lion and a Man chanced to travel in company through the forest.
- They soon began to quarrel, for each of them boasted that he and his kind were far superior to the other both in strength and mind.
- Now they reached a clearing in the forest and there stood a statue. It was a representation of Heracles in the act of tearing the jaws of the Nemean Lion.
- "See," said the man, "that's how strong we are! The King of Beasts is like wax in our hands!"
- "Ho!" laughed the Lion, "a Man made that statue. It would have been quite a different scene had a Lion made it!"

The crab and the Fox:

- There was once a crab who got disgusted in the sands he lived.
- He decided to take a stroll to the meadow.
- He imagined that he would find better fare than briny water and sand mites.
- When he crawled to the meadow, a hungry fox saw him and ate him in the same instant.

The cat and the Birds:

- Once upon a time there was a cat who did not get enough to eat.
- As a result, he grew very thin.
- One day, he got to know that the birds in the neighborhood were ailing and needed a doctor.
- So, he put on a pair of spectacles and with a leather box in his hand, knocked at the door of the Bird's home.
- The Birds peeped out and the Cat asked how they were.
- He said he would be very happy to give them some medicine.
- The Birds answered that they were very well and would only get better if the Cat kept away.

An old Lion and the Fox:

- An old Lion, whose teeth and claws were so worn that it was not so easy for him to get food as in his younger days, pretended that he was sick.
- He took care to let all his neighbors know about it, and then lay down in his cave to wait for visitors. And when they came to offer him their sympathy, he ate them up one by one.
- The Fox came too, but he was very cautious about it. Standing at a safe distance from the cave, he inquired politely after the Lion's health.
- Master Fox very wisely stayed outside, thanking the Lion very kindly for the invitation. "I should be glad to do as you ask," he added, "but I have noticed that there are many footprints leading into your cave and none coming out. Pray tell me, how do your visitors find their way out again?"

The sick stag:

- There was once a stag who had become ill.
- He had just enough strength to gather some food and find a suitable place to rest until he recovers.
- All the animals soon heard about the Stag's illness and came to visit.
- They were all hungry and helped themselves to the food the Stag had gathered for himself.
- As a result, the Stag starved to death.