

Kathakar

Kathakar is a project that leverages fine-tuning to enable a Large Language Model (LLM) to generate short stories effectively in Indic languages. It is powered by Sarvam-1.

Kathakar has been fine-tuned using a custom dataset of 3,000 Indian folklore and folktales sourced from various books that have been open-sourced through Project Gutenberg. Due to the limited availability of stories in specific Indic languages such as Marathi and Hindi, the initial 3,000 Indian stories were translated into multiple target languages to enhance the fine-tuning process. This model currently supports story generation in:

1. Hindi (hi)
2. Bengali (bn)
3. Gujarati (gu)
4. Kannada (kn)
5. Malayalam (ml)
6. Marathi (mr)
7. Oriya (or)
8. Punjabi (pa)
9. Tamil (ta)
10. Telugu (te)
11. English (en)

Note - Through extensive experimentation, I observed that the model generates stories in regional languages (such as Hindi or Marathi) only when the input prompt is provided in that respective language. If the prompt is in English, the output often defaults to English, even when a regional language is selected. To address this, I integrated [IndicTrans2](#), which first translates the user's prompt from English to the selected regional language before passing it to the story generation model. This ensures that the model receives a prompt in the correct language, leading to more accurate and fluent story generation in the desired regional language.

To further augment the training data, additional datasets were incorporated from Kaggle, including:

- [1002 Short Stories from Project Gutenberg](#)
- [Reddit Short Stories](#)

Fine-tuned model & Dataset

The Finetuned model as well as the dataset used for fine-tuning is available through this [link](#).

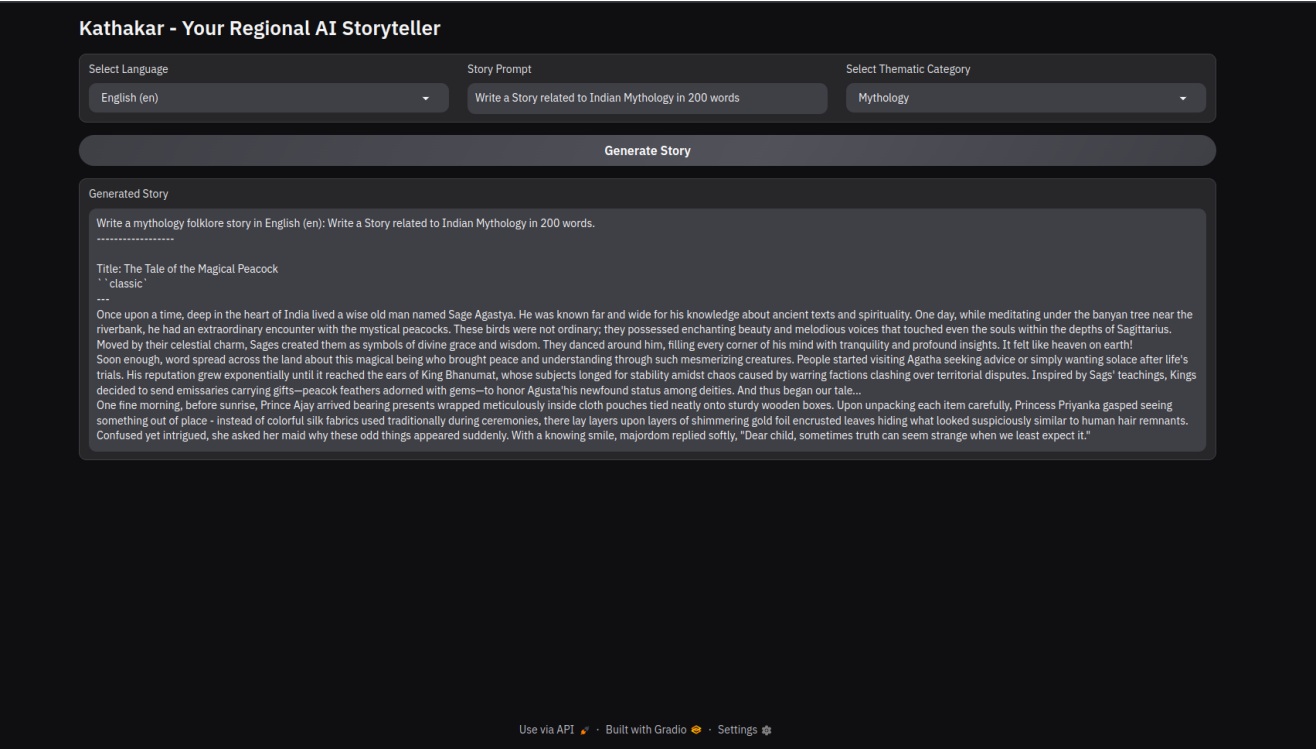
Tech-stack

The project is built with the help of the following libraries

1. Python Programming language
2. Gradio (For UI)
3. Huggingface suite (for finetuning)

Screenshots

Example of a story in English Language



Example of a story in Marathi Language

Kathakar - Your Regional AI Storyteller (IndicTrans2)

Select Language

Marathi (mr)

Story Prompt (in English)

Write a story in about 200 words detailing the Regional History of the people in Maharashtra

Select Thematic Category

Regional History

Generate Story

Generated Story

What is the theme of your regional history?

शीर्षक: गिरीचैलीची कथा

विषय: निसर्ग आणि मैत्री

वर्णन:

गिरिचैली ही एक माळी होती, जी तिच्या बागकामाच्या कौशल्यासाठी प्रसिद्ध होती. तिला फळे आणि फुले वाढवण्याची कला अवगत होती आणि ती त्यांच्या सौंदर्याने सर्वत्र लोकांना आकर्षित करायची. मात्र, तिला सर्वात जास्त आवडणारी गोष्ट म्हणजे गवत. तिने प्रत्येक गवताच्या कंदिलाची कापणी करण्यात आणि त्याच्या सौम्यतेत आनंद घेण्यात अनेक तास घालवले.

इतर लोकांप्रमाणे, गिरिचैलीला गवताळ जमिनीचे सौंदर्य समजले नाही. त्याऐवजी, त्याने त्याला एक अडथळा म्हणून पाहिले. एक दिवस, इतरांनी गवतातून चालत जाण्याबद्दल तक्रार केली तेव्हा, तिने त्यांच्याकडे दुर्लक्ष केले आणि तिचा बागकाम सुरू ठेवला. पण, शेवटी, तिची बाग खूप ओली आणि दुर्गंधीयुक्त झाली. इतर लोक त्यांच्या चिंतांकडे दुर्लक्ष करू शकले नाहीत आणि त्यांनी त्यांना जाण्यास सांगितले. ते निघून गेल्यावर, त्यांना जाणीव झाली की त्यांनी त्यांच्या बागेची किती वाईट अवस्था केली आहे. गिरिचैल्याने लगेचच गवती पुन्हा उगवली आणि त्याचे नूतनीकरण केले. तेव्हापासून, त्या गवतातील वाऱ्याच्या झोतांचा आनंद लुटायला शिकल्या. </s>

Literature

The following papers were reviewed and utilized in the development of this project, categorized for clarity:

Fine-Tuning Techniques:

- [LoRA: Low-Rank Adaptation of Large Language Models](#) - Focused on efficient adaptation of LLMs.
- [QLoRA: Efficient Finetuning of Quantized LLMs](#) - Explores efficient fine-tuning of quantized LLMs.

Large Language Models & Indic Languages:

- [Sarvam 1: The first Indian language LLM](#) - Introduces the base LLM used in this project.
- [IndicGenBench: A Multilingual Benchmark to Evaluate Generation Capabilities of LLMs on Indic Languages](#) - Provides a benchmark for evaluating LLM performance in Indic languages.