

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. English
2. Marathi
3. Hindi

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 IndicTransToolkit (GitHub Repository), 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](#)

Dataset

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

Kathakar - Your Regional AI Storyteller

Select Language

English (en)

Story Prompt

Write a Story related to Indian Mythology in 200 words

Select Thematic Category

Mythology

Generate Story

Generated Story

Write a mythology folklore story in English (en): Write a Story related to Indian Mythology in 200 words.

Title: The Tale of the Magical Peacock

"" classic

Once upon a time, deep in the heart of India lived a wise old man named Sage Agastya. He was known far and wide for his knowledge about ancient texts and spirituality. One day, while meditating under the banyan tree near the riverbank, he had an extraordinary encounter with the mystical peacocks. These birds were not ordinary; they possessed enchanting beauty and melodious voices that touched even the souls within the depths of Sagittarius. Moved by their celestial charm, Sages created them as symbols of divine grace and wisdom. They danced around him, filling every corner of his mind with tranquility and profound insights. It felt like heaven on earth! Soon enough, word spread across the land about this magical being who brought peace and understanding through such mesmerizing creatures. People started visiting Agatha seeking advice or simply wanting solace after life's trials. His reputation grew exponentially until it reached the ears of King Bhanumat, whose subjects longed for stability amidst chaos caused by warring factions clashing over territorial disputes. Inspired by Sage's teachings, Kings decided to send emissaries carrying gifts—peacock feathers adorned with gems—to honor Agusta's newfound status among deities. And thus began our tale...

One fine morning, before sunrise, Prince Ajay arrived bearing presents wrapped meticulously inside cloth pouches tied neatly onto sturdy wooden boxes. Upon unpacking each item carefully, Princess Priyanka gasped seeing something out of place - instead of colorful silk fabrics used traditionally during ceremonies, there lay layers upon layers of shimmering gold foil encrusted leaves hiding what looked suspiciously similar to human hair remnants. Confused yet intrigued, she asked her maid why these odd things appeared suddenly. With a knowing smile, majordom replied softly, "Dear child, sometimes truth can seem strange when we least expect it."

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.