Benchmarking Hindi LLMs: A New Suite of Datasets and a Comparative Analysis

Anusha Kamath | BHASHA @ IJCNLP-AACL 2025

Evolution of Evaluation Datasets in English Language

The last decade

GLUE
BIG-Bench Hard
RULER
HMMT25
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
SQuAD
BIG-Bench
GPQA
AI4Math

Hindi still lacks strong foundational benchmarks for evaluating models

Translation is just not enough

✓ **Culturally rooted**

Built natively for Hindi, capturing idioms, context, and cultural references

✓ **Covers real tasks**

Closer to everyday tasks and challenges

✓ **Easy to replicate**

Well documented to reuse for other languages and integrate to existing evaluation frameworks

✓ **Human-verified**

Reviewed by diverse annotators across regions, genders, and backgrounds to reduce bias

New suite of Hindi benchmarks

IFEval-Hi	MTBench-Hi	GSM8K-Hi	ChatRAG-Hi	BFCL-Hi
848 samples	200 samples	1319 samples	5948 samples	2251 samples

IFEval

Instruction Following Evaluation benchmark

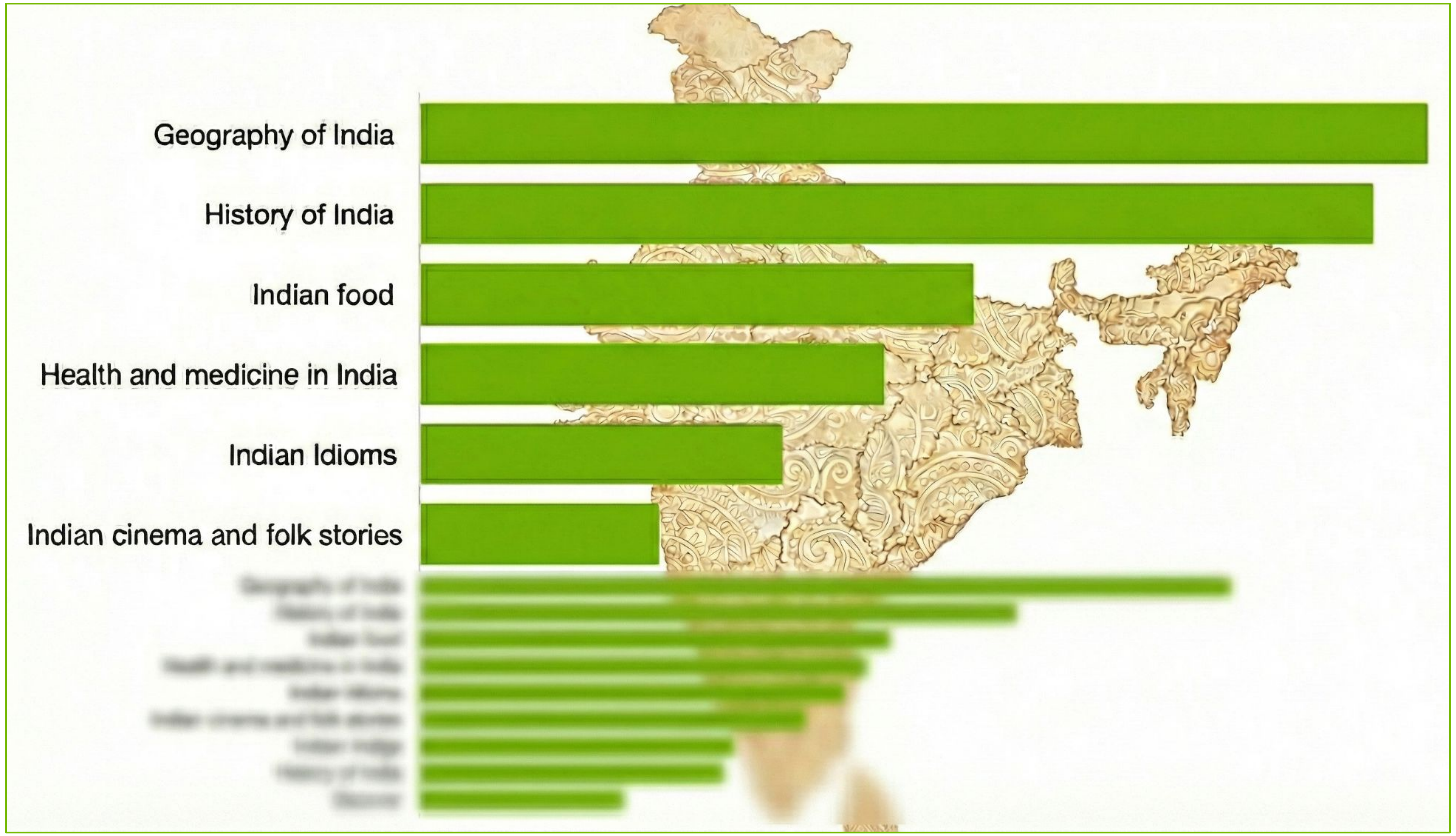
Write a short paragraph about X

- at least 500 words
- no more than 6 sentences
- include at least 1 capitalized word

भारत के प्रसिद्‌ध लेखक और नोबेल पुरस्कार विजेता रवीन्द्रनाथ ठाकुर के बारे में बताइए।

- आपका पूरा उत्तर हन्दी में होना चाहए
- कम से कम **पाँच सौ** शब्दों का लेख लिखें।
- वाक्यों की संख्या ६ से अधिक न हो।
- ~~कहीं भी कम से कम एक शब्द पूरी तरह **CAPITAL LETTERS** में लिखें।~~

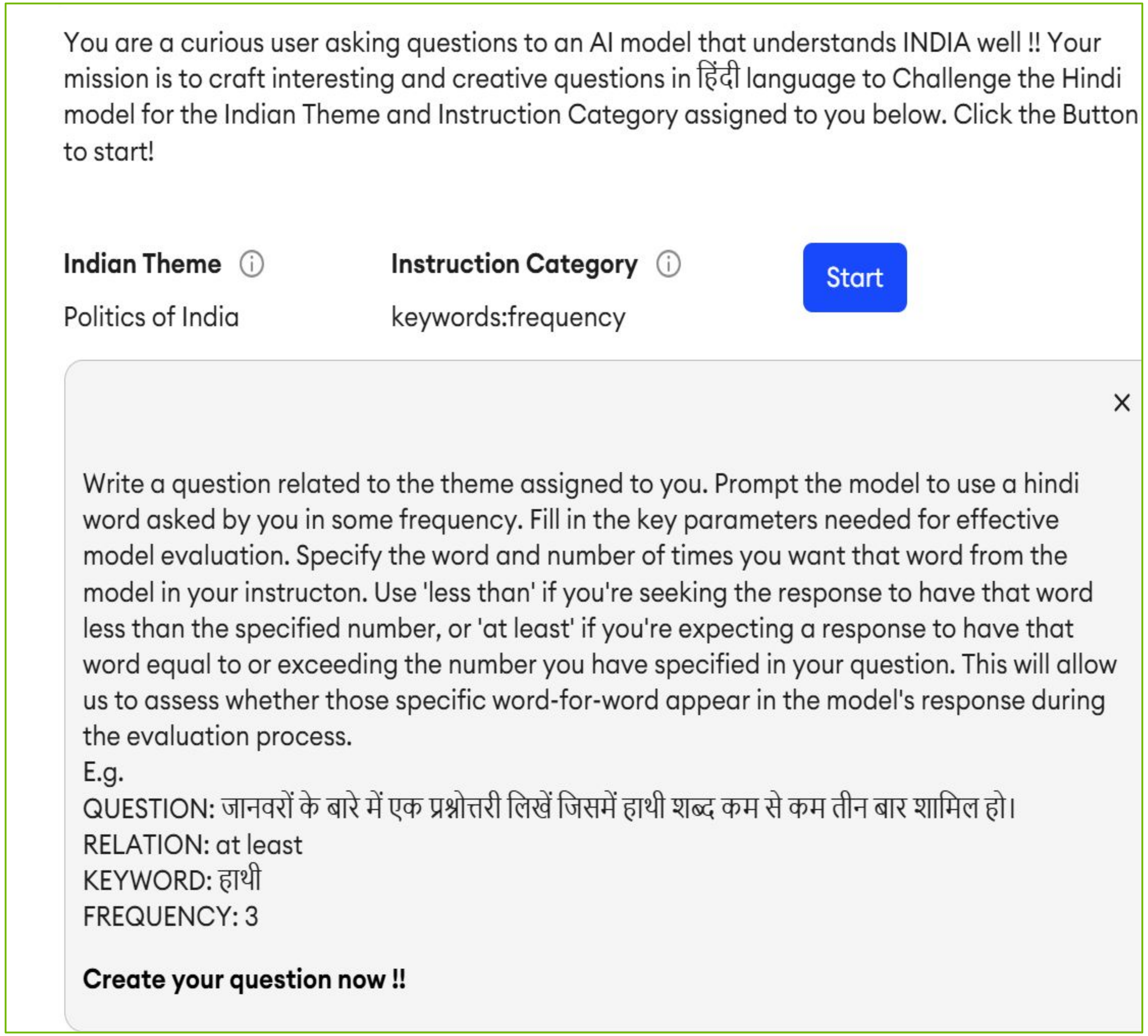
IFEval-Hindi



Distribution of samples by Indian cultural themes in the IFEval-Hi dataset.

- Native, India-centric prompts: Created from Indian Wikipedia sources
- 22 core instruction types: Filtered on language, removed irrelevant ones (e.g., case conversion, punctuation change).
- Three complexity levels: Simple, moderate, and complex - to test models comprehensively.

IFEval-Hindi Annotation



You are a curious user asking questions to an AI model that understands INDIA well !! Your mission is to craft interesting and creative questions in हिंदी language to Challenge the Hindi model for the Indian Theme and Instruction Category assigned to you below. Click the Button to start!

Indian Theme ⓘ

Politics of India

Instruction Category ⓘ

keywords:frequency

Write a question related to the theme assigned to you. Prompt the model to use a hindi word asked by you in some frequency. Fill in the key parameters needed for effective model evaluation. Specify the word and number of times you want that word from the model in your instructon. Use 'less than' if you're seeking the response to have that word less than the specified number, or 'at least' if you're expecting a response to have that word equal to or exceeding the number you have specified in your question. This will allow us to assess whether those specific word-for-word appear in the model's response during the evaluation process.

E.g.

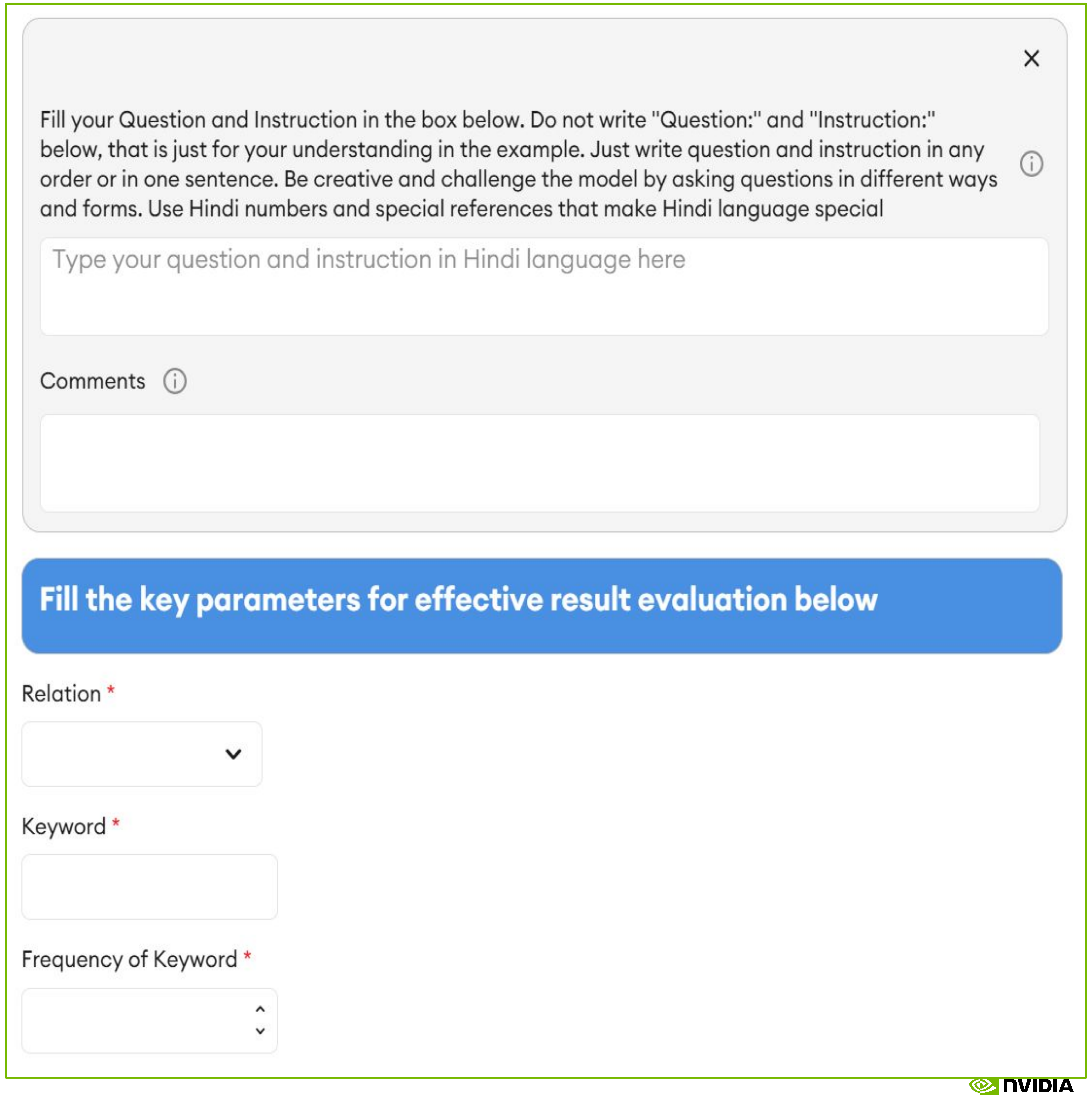
QUESTION: जानवरों के बारे में एक प्रश्नोत्तरी लिखें जिसमें हाथी शब्द कम से कम तीन बार शामिल हो।

RELATION: at least

KEYWORD: हाथी

FREQUENCY: 3

Create your question now !!



Fill your Question and Instruction in the box below. Do not write "Question:" and "Instruction:" below, that is just for your understanding in the example. Just write question and instruction in any order or in one sentence. Be creative and challenge the model by asking questions in different ways and forms. Use Hindi numbers and special references that make Hindi language special ⓘ

Type your question and instruction in Hindi language here

Comments ⓘ

Fill the key parameters for effective result evaluation below

Relation *

Keyword *

Frequency of Keyword *

MTBench

Multi-Turn Evaluation Benchmark

Question:

Pretend yourself to be **Elon Musk** in all the following conversations.

Why do we need to go to Mars?

Follow up:
How do you like dancing? Can you teach me?

Question:

मान लीजिए, आप **स्वामी विवेकानंद** हैं,

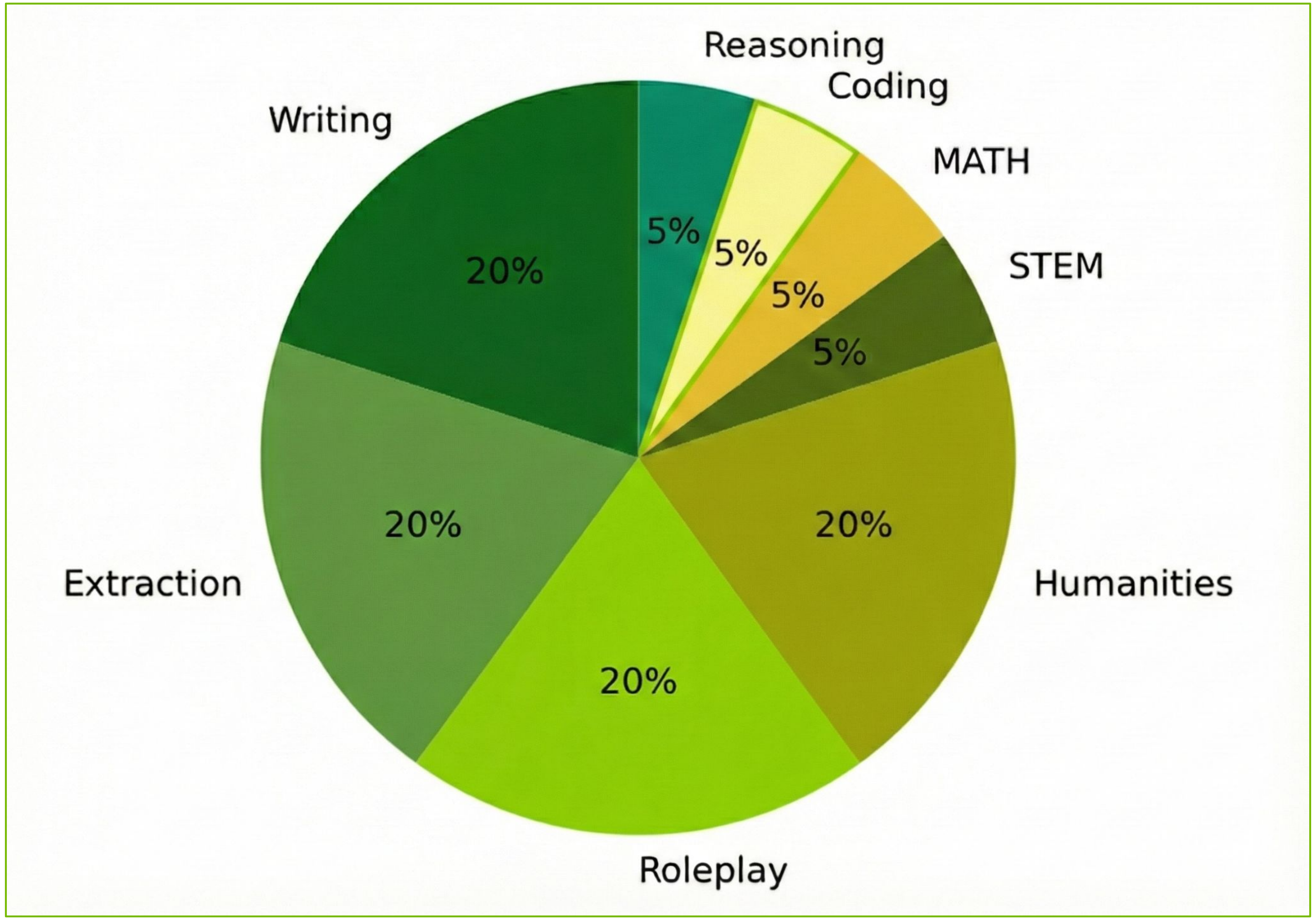
और अब से सभी वार्तालापों में विवेकानंद जी की तरह बोलें।

हमें अपने देश के विकास में योगदान क्यों देना चाहिए?

Follow up:

वैसे, आपको वीडियो गेम खेलना कैसा लगता है?

MTBench-Hindi



Category distribution in MT-Bench-Hi, adapted with Indian cultural themes to increase focus on culturally relevant instructions.

- Expanded only Cultural categories (Writing, Roleplay, Humanities, Extraction): Created natively by specialists.

- Added diverse Indian themes for richer context.

MTBench-Hindi Annotation

You are a curious user, conversing with an AI model that understands India so well!! Your mission is to craft interesting and creative questions in हिंदी language to challenge the model in TWO steps. You will be given detailed instructions at each stage. Observe the example below for example.

Reference question asked by the user to the model

Pretend yourself to be Elon Musk in all the following conversations. Speak like Elon Musk as much as possible. Why do we need to go to Mars?

Reference response from the model

Well, there are several reasons why going to Mars is crucial for humanity. One of the primary reasons is to ensure the survival of our species. By becoming a multiplanetary species, we can safeguard ourselves from potential extinction

What we observed from this response ....

The assistant's response is accurate, relevant, and detailed. It captures the essence of Elon Musk's vision for Mars colonization, emphasizing the importance of becoming a multiplanetary species, the potential for scientific

★★★★★★★★★☆

Your turn now !

Craft a similar question in Indian context to test the model. Make sure you are writing in हिंदी language.

Question to the model

खुद को रत्न टाटा के रूप में प्रस्तुत करें और अगले संवाद में उनके जैसा बोलने की कोशिश करें। भारत की सौर ऊर्जा में निवेश क्यों करना चाहिए, और क्या यह देश के भविष्य के लिए महत्वपूर्ण हैं।

Follow-up question

भारत मे सौर ऊर्जा के क्षेत्र में निजी क्षेत्र और सरकार को मिलकर किस प्रकार की नीतियाँ अपननी चाहिए, ताकि यह क्षेत्र और भी तेजी से विकसित हो सके?

GSM8K

GSM8K En Instance

Tim decides to light off some fireworks for the fourth of July. He buys a package of fireworks worth $400 and another pack worth twice that much. He gets a 20% discount on them. He also buys a finale firework that costs $150. How much did he spend in total?

Translated Instance

टिम ने चौथी जुलाई के लिए कुछ आतिशबाजी जलाने का फैसला किया। वह $400 के पैकेज और उतनी ही कीमत के दो पैक खरीदता है। उसे उन पर 20% की छूट मिलती है। वह एक फिनाले आतिशबाजी भी खरीदता है जिसकी कीमत $150 है। उसने कुल कितना खर्च किया?

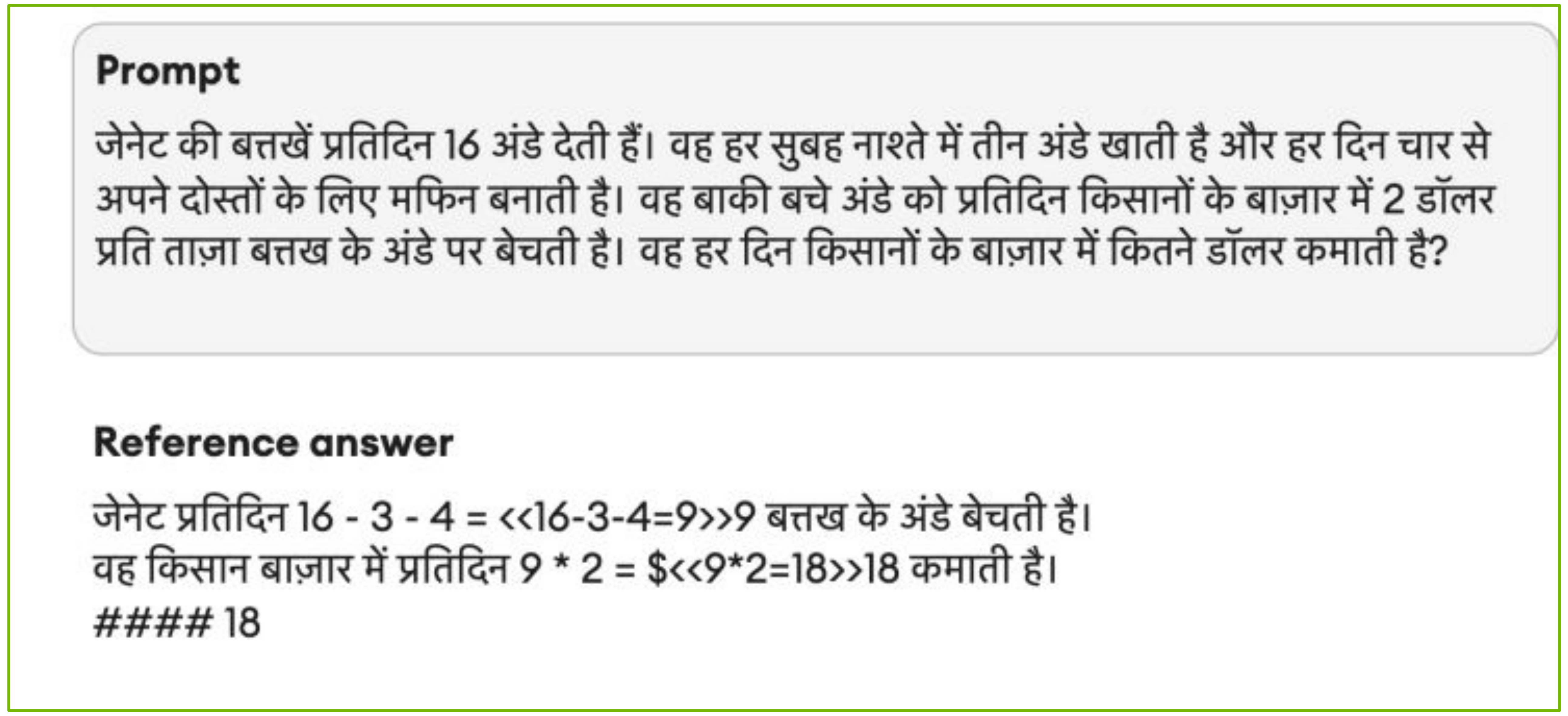
Tim decides to light some fireworks for the Fourth of July. He buys a $400 package and two packs of the same price. He gets a 20% discount on them. He also buys a finale firework that costs $150. How much did he spend in total?

Revised Instance

टिम ने चौथी जुलाई के लिए कुछ आतिशबाजी जलाने का फैसला किया। वह 400 डॉलर मूल्य का एक पैकेट पटाखों का खरीदता है तथा उससे दुगुना मूल्य का एक और पैकेट खरीदता है। उसे उन पर 20% की छूट मिलती है। वह एक फिनाले आतिशबाजी भी खरीदता है जिसकी कीमत $150 है। उसने कुल कितना खर्च किया?

Translate and verify of GSM8K sample

GSM8K-Hindi annotation

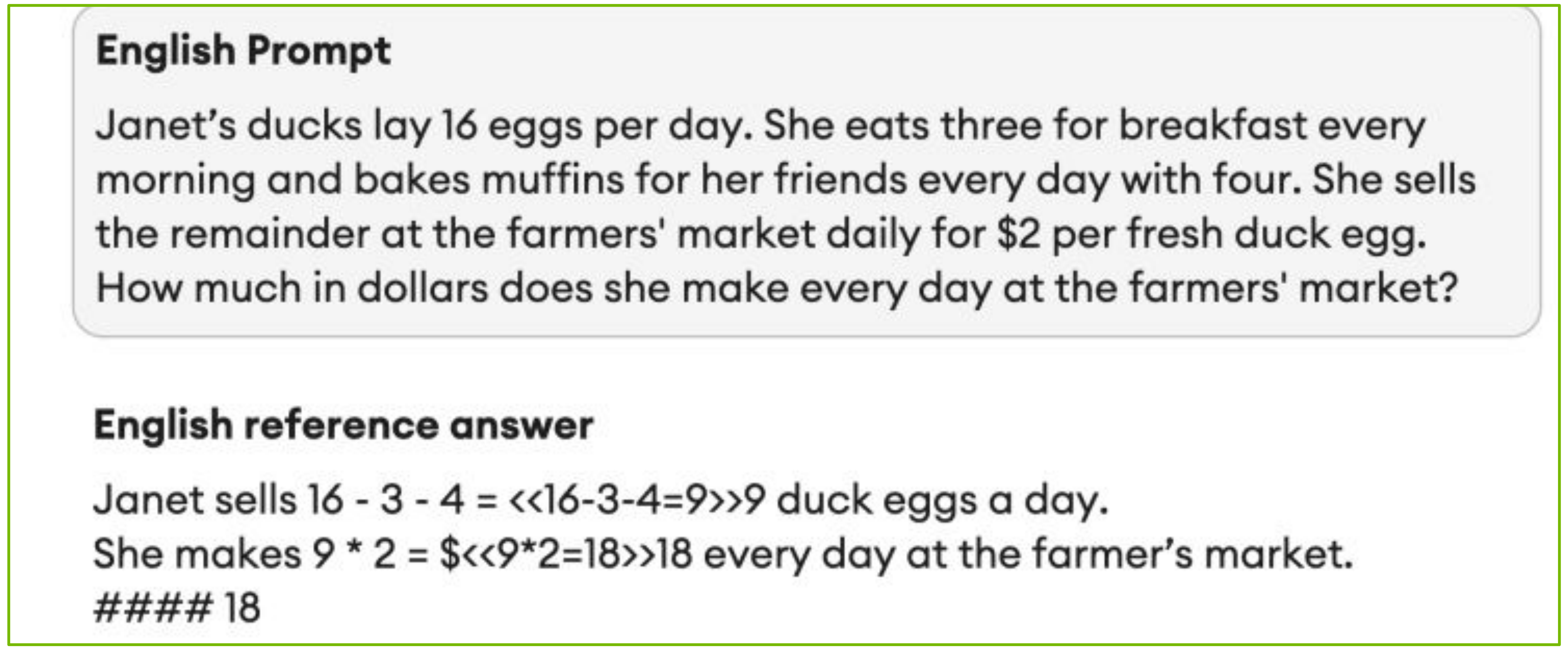


Prompt

जेनेट की बत्तखें प्रतिदिन 16 अंडे देती हैं। वह हर सुबह नाश्ते में तीन अंडे खाती है और हर दिन चार से अपने दोस्तों के लिए मफिन बनाती है। वह बाकी बचे अंडे को प्रतिदिन किसानों के बाज़ार में 2 डॉलर प्रति ताज़ा बत्तख के अंडे पर बेचती है। वह हर दिन किसानों के बाज़ार में कितने डॉलर कमाती है?

Reference answer

जेनेट प्रतिदिन 16 - 3 - 4 = <<16-3-4=9>>9 बत्तख के अंडे बेचती है।
वह किसान बाज़ार में प्रतिदिन 9 * 2 = $<<9*2=18>>18 कमाती है।
18

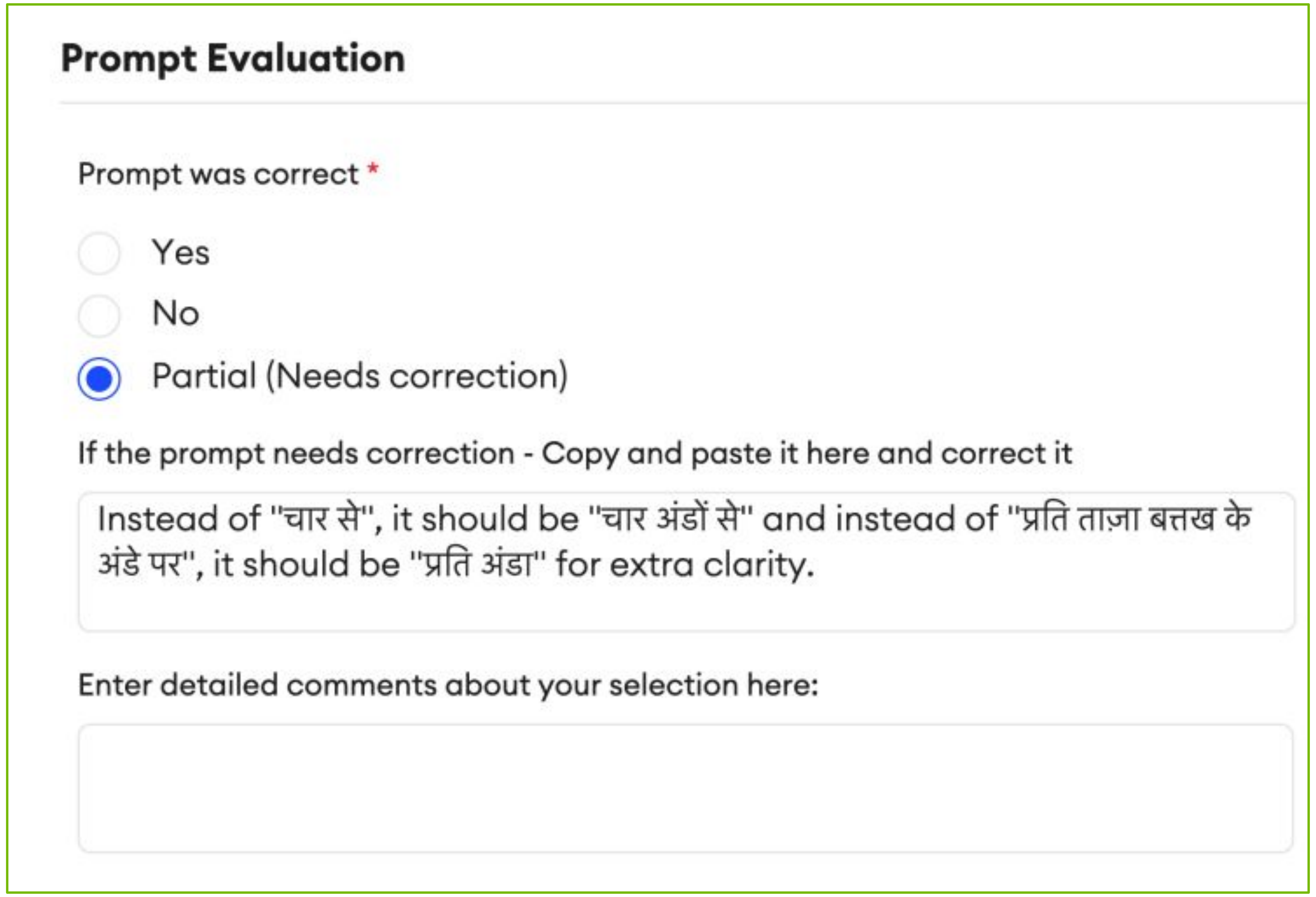


English Prompt

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

English reference answer

Janet sells 16 - 3 - 4 = <<16-3-4=9>>9 duck eggs a day.
She makes 9 * 2 = $<<9*2=18>>18 every day at the farmer’s market.
18



Prompt Evaluation

Prompt was correct *

○ Yes
○ No
● Partial (Needs correction)

If the prompt needs correction - Copy and paste it here and correct it

Instead of "चार से", it should be "चार अंडों से" and instead of "प्रति ताज़ा बत्तख के अंडे पर", it should be "प्रति अंडा" for extra clarity.

Enter detailed comments about your selection here:

ChatRAG

QA over retrieved documents

Context 1:

Once upon a time, in a barn near a farm house, there lived a little white kitten named Cotton. But Cotton wasn't alone in her little home above the barn, oh no. She shared her hay bed with her mommy and 5 other sisters. But she was the only white one in the bunch. The rest of her sisters were all orange with beautiful white tiger stripes like Cotton's mommy. Being different made Cotton quite sad. She often wished she looked like the rest of her family. So one day, when Cotton found a can of the old farmer's orange paint, she used it to paint herself like them. When her mommy and sisters found her they started laughing.

Context 2: .... Context n: ....

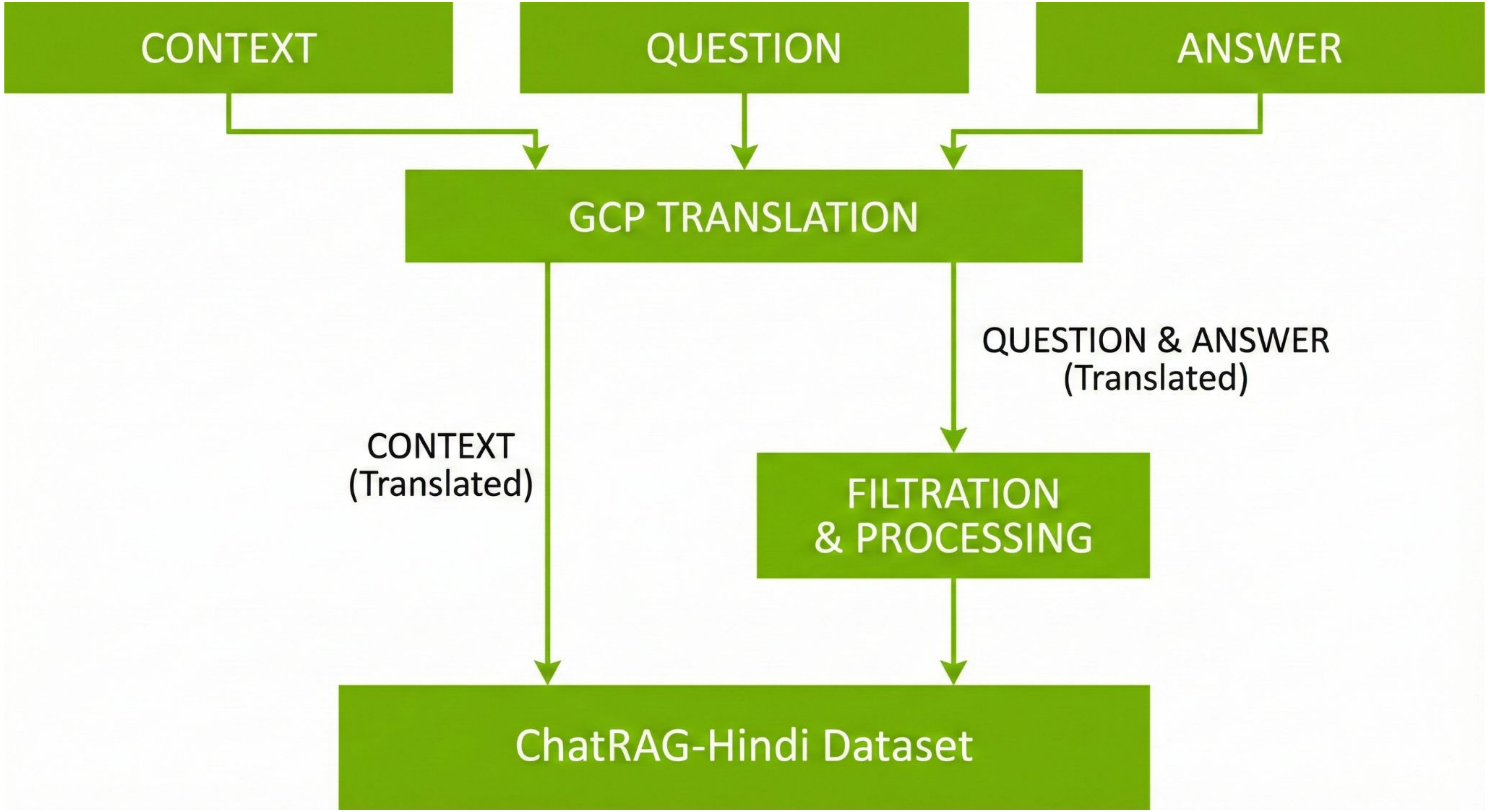
Question:

What color was Cotton?

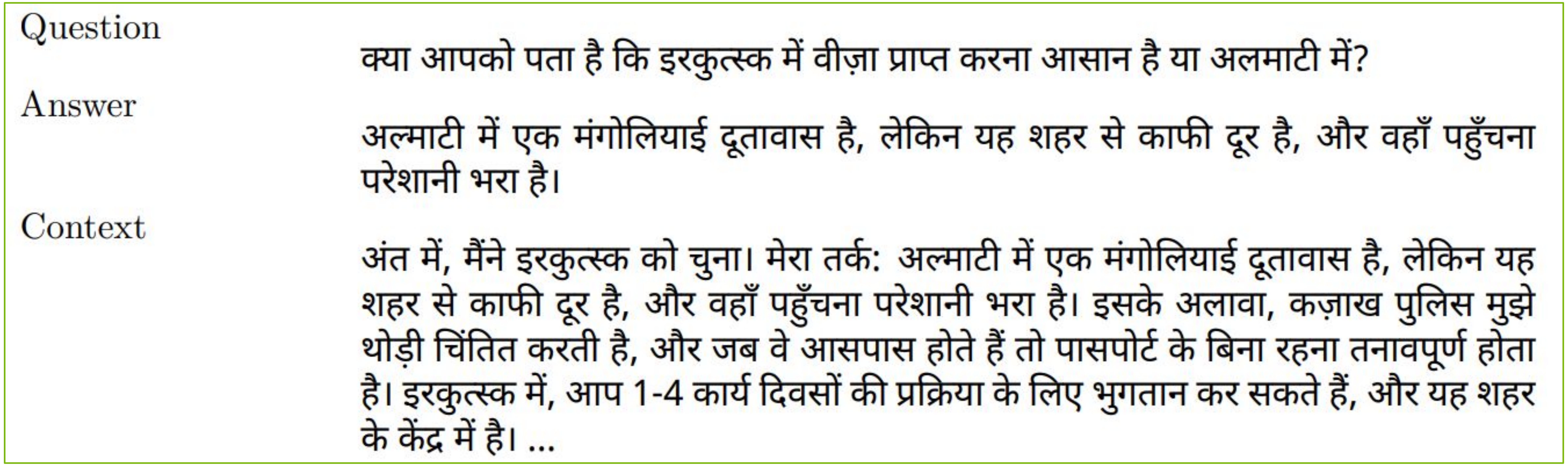
Answer:

White, white, colour WHITE…

ChatRAG-Hindi



Differential Translation Pipeline for ChatRAG-Hi creation



Question	क्या आपको पता है कि इरकुत्स्क में वीज़ा प्राप्त करना आसान है या अलमाटी में?
Answer	अल्माटी में एक मंगोलियाई दूतावास है, लेकिन यह शहर से काफी दूर है, और वहाँ पहुँचना परेशानी भरा है।
Context	अंत में, मैंने इरकुत्स्क को चुना। मेरा तर्क: अल्माटी में एक मंगोलियाई दूतावास है, लेकिन यह शहर से काफी दूर है, और वहाँ पहुँचना परेशानी भरा है। इसके अलावा, कज़ाख पुलिस मुझे थोड़ी चिंतित करती है, और जब वे आसपास होते हैं तो पासपोर्ट के बिना रहना तनावपूर्ण होता है। इरकुत्स्क में, आप 1-4 कार्य दिवसों की प्रक्रिया के लिए भुगतान कर सकते हैं, और यह शहर के केंद्र में है। ...

ChatRAG-Hi sample

BFCL

Berkeley Function Calling Leaderboard

Question:

"Book a one-way economy flight from San Francisco to New York next Friday after 5 PM for two adults, window seats if available."

Functions:

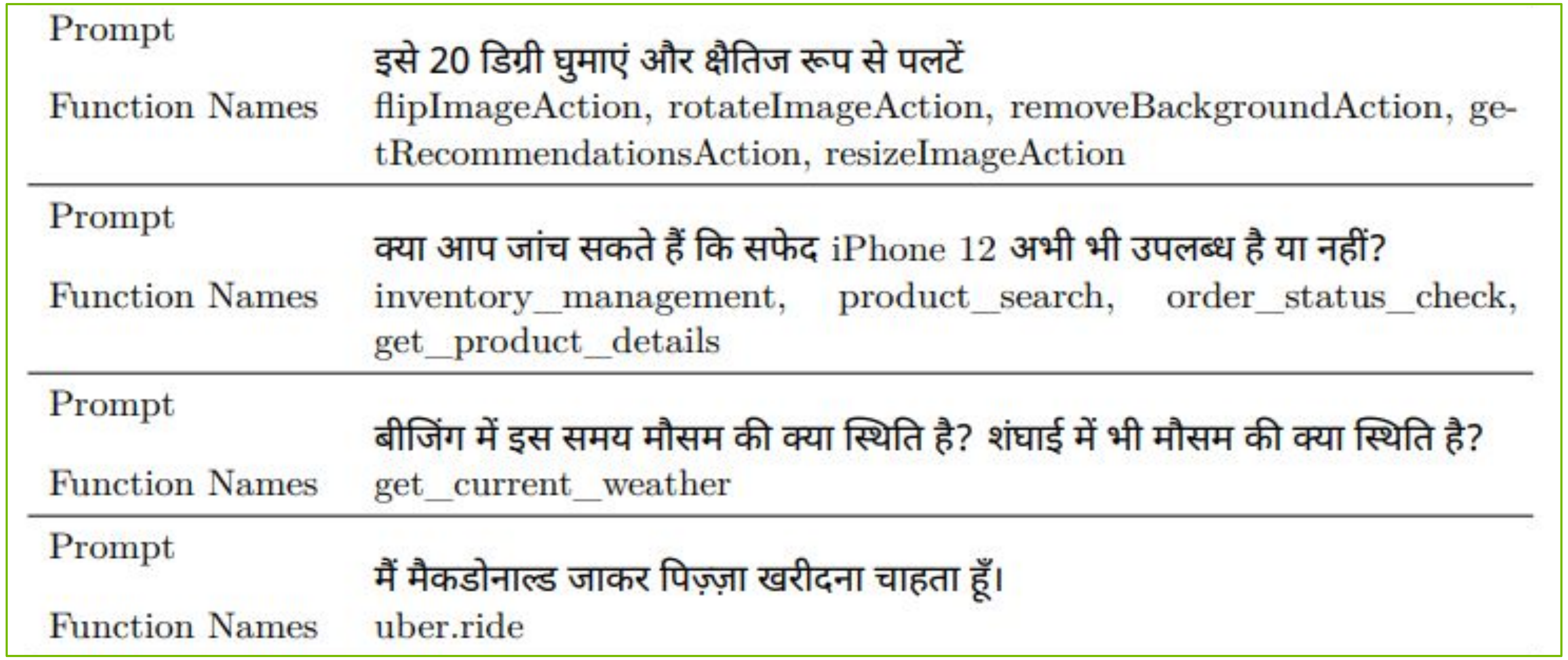
search_flights*(origin, destination, date, ...)*

book_flight*(flight_id, passengers, ...)*

Expected behavior:

- First call search_flights with *origin="SFO", destination="JFK", date=<next Friday>*
- Then choose a suitable flight from the result and call book_flight with the chosen *flight_id, passengers=2*

BFCL-Hindi



Prompt	इसे 20 डिग्री घुमाएं और क्षैतिज रूप से पलटें
Function Names	flipImageAction, rotateImageAction, removeBackgroundAction, getRecommendationsAction, resizeImageAction
Prompt	क्या आप जांच सकते हैं कि सफेद iPhone 12 अभी भी उपलब्ध है या नहीं?
Function Names	inventory_management, product_search, order_status_check, get_product_details
Prompt	बीजिंग में इस समय मौसम की क्या स्थिति है? शंघाई में भी मौसम की क्या स्थिति है?
Function Names	get_current_weather
Prompt	मैं मैकडोनाल्ड जाकर पिज़्ज़ा खरीदना चाहता हूँ।
Function Names	uber.ride

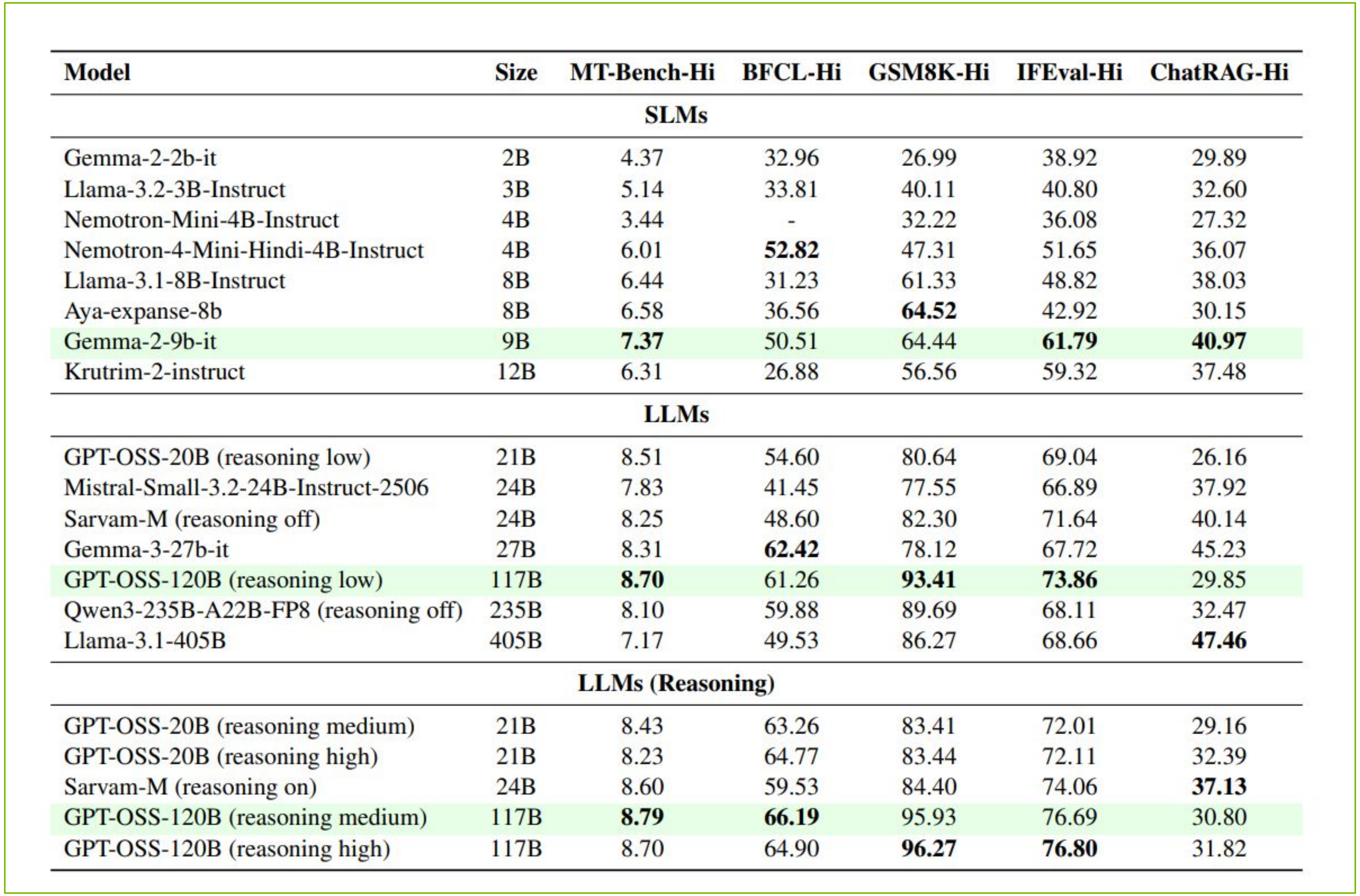
BFCL-V2-Hi sample

- Conversational history: Translated to Hindi using GCP
- Function calls & parameters: Retained in original English
- Tests model's ability to understand Hindi queries and map them to English-defined tools

Evaluation:

- Ground truth unchanged for simple, multiple, and parallel categories
- BFCL Abstract Syntax Tree (AST) methodology for thorough analysis

Results



Model	Size	MT-Bench-Hi	BFCL-Hi	GSM8K-Hi	IFEval-Hi	ChatRAG-Hi
			SLMs			
Gemma-2-2b-it	2B	4.37	32.96	26.99	38.92	29.89
Llama-3.2-3B-Instruct	3B	5.14	33.81	40.11	40.80	32.60
Nemotron-Mini-4B-Instruct	4B	3.44	-	32.22	36.08	27.32
Nemotron-4-Mini-Hindi-4B-Instruct	4B	6.01	**52.82**	47.31	51.65	36.07
Llama-3.1-8B-Instruct	8B	6.44	31.23	61.33	48.82	38.03
Aya-expanse-8b	8B	6.58	36.56	**64.52**	42.92	30.15
Gemma-2-9b-it	9B	**7.37**	50.51	64.44	**61.79**	**40.97**
Krutrim-2-instruct	12B	6.31	26.88	56.56	59.32	37.48
			LLMs			
GPT-OSS-20B (reasoning low)	21B	8.51	54.60	80.64	69.04	26.16
Mistral-Small-3.2-24B-Instruct-2506	24B	7.83	41.45	77.55	66.89	37.92
Sarvam-M (reasoning off)	24B	8.25	48.60	82.30	71.64	40.14
Gemma-3-27b-it	27B	8.31	**62.42**	78.12	67.72	45.23
GPT-OSS-120B (reasoning low)	117B	**8.70**	61.26	**93.41**	**73.86**	29.85
Qwen3-235B-A22B-FP8 (reasoning off)	235B	8.10	59.88	89.69	68.11	32.47
Llama-3.1-405B	405B	7.17	49.53	86.27	68.66	**47.46**
			LLMs (Reasoning)			
GPT-OSS-20B (reasoning medium)	21B	8.43	63.26	83.41	72.01	29.16
GPT-OSS-20B (reasoning high)	21B	8.23	64.77	83.44	72.11	32.39
Sarvam-M (reasoning on)	24B	8.60	59.53	84.40	74.06	**37.13**
GPT-OSS-120B (reasoning medium)	117B	**8.79**	**66.19**	95.93	76.69	30.80
GPT-OSS-120B (reasoning high)	117B	8.70	64.90	**96.27**	**76.80**	31.82

SLMs (<20B params):

- Gemma-2-9B-it: Best overall on MT-Bench-Hi, IFEval-Hi, ChatRAG-Hi.
- Aya-Expanse-8B: Top on GSM8K-Hi.
- Nemotron-4B-Hindi: Leads BFCL-Hi (tool-calling).

LLMs (>20B params):

- GPT-OSS-120B: Best on MT-Bench-Hi, GSM8K-Hi, IFEval-Hi.
- Gemma-3-27B-it: Highest on BFCL-Hi.
- Llama-3.1-405B: Excels on ChatRAG-Hi.

Key Insights

- No single model dominates all tasks.
- Reasoning modes significantly boost performance on complex tasks.
- Human in the loop is unavoidable at this stage
- The work serves as a blueprint for broader Indic language

References

Dataset:

- https://huggingface.co/datasets/nvidia/IFEval-Hi
- https://huggingface.co/datasets/nvidia/MT-Bench-Hi
- https://huggingface.co/datasets/nvidia/GSM8K-Hi
- https://huggingface.co/datasets/nvidia/ChatRAG-Hi
- https://huggingface.co/datasets/nvidia/BFCL-Hi

Q&A