

QUENCH: Measuring the gap between Indic and Non-Indic Contextual General Reasoning in LLMs

Mohammad Aflah Khan*, Neemesh Yadav*, Sarah Masud, Md. Shad Akhtar

{aflah20082, neemesh20529, sarahm, shad.akhtar}@iiitd.ac.in

IIIT Delhi, India.

Abstract

The rise of large language models (LLMs) has created a need for advanced benchmarking systems beyond traditional setups. To this end, we introduce QUENCH, a novel text-based English **Quizzing Benchmark** manually curated and transcribed from YouTube quiz videos. QUENCH possesses masked entities and rationales for the LLMs to predict via generation. At the intersection of geographical context and common sense reasoning, QUENCH helps assess world knowledge and deduction capabilities of LLMs via a zero-shot, open-domain quizzing setup. We perform an extensive evaluation on 7 LLMs and 4 metrics, investigating the influence of model size, prompting style, geographical context, and gold-labeled rationale generation. The benchmarking concludes with an error analysis to which the LLMs are prone.

1 Introduction

The ubiquitous rise of large language models (LLMs) has driven the need for diverse benchmarking and evaluation systems (Chang et al., 2024; Peng et al., 2024). Examining the logical reasoning and world knowledge capabilities of the LLMs (Qiao et al., 2023; Lu et al., 2023) has been an active area of research. On one hand, world knowledge is primarily adjudged subject-wise (History, Law, STEM etc) via MMLU (Hendrycks et al., 2021), GSM-8k (Cobbe et al., 2021) or via deductive multichoice (MCQ) question banks such as JEEBench (Arora et al., 2023) or ScienceQA (Lu et al., 2022). The subject-specific questions fail to capture the multi-themed nature of real-world knowledge and reasoning, which involves relating concepts from multiple different “subject” areas. Furthermore, the MCQ setup already provides a plausible answer and restricts the answering domain to a closed-source one. Meanwhile, the commonsense understanding is broadly examined via

the likes of pragmatism (Sravanthi et al., 2024), syllogism (Wu et al., 2023), and truthfulness (Lin et al., 2022) with minimal references to real-world historical events and entities. Parallely, researchers have also been proving the efficacy of LLMs as evaluation tools, such as G-Eval (Liu et al., 2023).

Another frequently overlooked issue with benchmarks is their tendency to be “global” or “western-centric” (Santurkar et al., 2023; Durmus et al., 2024). For a language model to understand complex cultural queries, it must first be acquainted with the specific entities and concepts of those cultures. While efforts like Seth et al. (2024) have been made, to the best of our knowledge, none present an open-world, clue-based guessing game where disjoint clues rely on linguistic hints or require combining multiple concepts to arrive at the answer. To address this gap, we propose a quiz trivia-based framework to probe LLMs’ deductive reasoning skills across multiple themes and entities. Our work focuses on world knowledge and commonsense reasoning, with an emphasis on geographical context (Table 1).

Benchmark Dataset. To support our investigation, we curate QUENCH, a **Quizzing Benchmark**. It is a novel text-only dataset obtained from manually annotated YouTube quiz videos in English. The primary entity to be predicted is hidden in the form of “X” and can occur multiple times in the question along with multiple entities as well (example # 2 in Table 1). We also provide a manually curated free-text explanation/rationale for the correct answers. The non-trivial coreference resolutions across entities uniquely position QUENCH to access the deduction capabilities of the LLMs with respect to the world and commonsense knowledge.

Benchmarking and Evaluation Setup. We benchmark entity and rationale prediction for QUENCH across seven different LLMs, including both closed and open-weight models. These models vary significantly in terms of the number of pa-

*Equal Contribution

Prompt	Quiz Prompt: Consider yourself a participant in a quiz show where I am the quizmaster. I will ask you a question that can be from any general theme. You need to provide me with the correct answer. The question can have multiple variables to answer, and you need to provide me with the answer for variables X, Y, and Z. Hence, use the following format strictly in your response: ‘The answer is <X: answer>.’ You will lose points if you fail to follow the format.
	Question: <i>The primary protagonist of X, one of China’s classic epics, derived his literary name from items he carried and shared it with a rather famous mudra gesture. However, he is less well-known than Y, a secondary protagonist in the original text who was an allegory for the restless human mind. Z is the protagonist of one of the most famous manga ever, adapted to over 20 movies and spawned many memes thanks to its unique transformations. Z shares his name and characterization with Y.</i>
Outputs	Predicted entities: <X: Journey to the west>; <Y: Sun Wukong>; <Z: Son Goku>;
	Rationale: X is Journey to the West, which features the protagonist Sun Wukong, who is known as the Monkey King. He carries a staff and is associated with the mudra gesture. Y is Sun Wukong, who is the Monkey King and secondary protagonist in Journey to the West. He represents the restless human mind. Z is Son Goku from the Dragon Ball series. He shares his name and monkey-like characterization with Sun Wukong, the Monkey King from Journey to the West. Dragon Ball has spawned over 20 movies and many memes.
Prompt	Quiz Prompt: Consider yourself a participant in a quiz show where I am the quizmaster. I will ask you a question that can be from any general theme. You need to provide me with the correct answer. The question can have multiple variables to answer, and you need to provide me with the answer for variable X. Hence, use the following format strictly in your response: ‘The answer is <X: answer>.’ You will lose points if you fail to follow the format.
	Question: <i>Karondi is a remote village near Jabalpur. X of India was moved to Karondi from Nagpur after the partition. X was established during the Great Trigonometrical Survey of India in 1907 in Nagpur. The project, basically, was to demarcate the British territories in India.</i>
Outputs	Predicted entities: <X: Centre Point of India>;
	Rationale: The Centre Point of India was moved post the partition due to a change in the boundaries of the country. It was established during the Great Trigonometrical Survey of India in 1907 in Nagpur.

Table 1: An overview of the quiz evaluation setup with zero-shot prompt (without CoT in this case) for *predicting the masked entities* (X, Y, Z) in a quiz question taken verbatim from QUENCH. The prompt and question are provided to an LLM, and the expected generations are to be the missing entities in the form <X: answer>. For the CoT setup, the above “quiz prompt” is suffixed with the phrase ‘Let’s think step by step.’ Each question has the required masked entities (for instance, ‘X’ or ‘X, Y, Z’) in the text, and we simply refer to each of them via “The question can have multiple variables to answer, and you need to provide me with the answer for variable <variable>”.

rameters, knowledge cutoff dates, context lengths, and pretraining characteristics. We employ the standard metrics – BLEU, BERTScore, and ROUGE-L, as well as a GEval-based strategy to evaluate the performance under zero-shot prompting both with and without chain-of-thought (CoT) as outlined in Table 1. The Indic subset (example # 1 in Table 1) in QUENCH further allows for examining the Indian knowledge representation in the LLMs.

Observations. Based on GEvals, our analysis reveals that zero-shot entity prediction accuracy improves from 72% to 87% when upgrading from GPT-3.5 to GPT-4. As expected, GPT-4 leads across all four metrics. Meanwhile, the open-weight LLaMA-3-70B performs on par with GPT-3.5 and stands out as a strong alternative, mainly due to its lower variability across subsets (12 points compared to GPT-3.5’s 23-point difference). Consistent with existing literature, we see larger models outperforming their 7B counterparts and an overall performance decline across LLMs in the Indian context. *A significant disparity persists between the Indic and non-Indic subsets*, with GPT-4-Turbo showing a 12-point difference between the two. Gemini 1.5 Flash exhibits an even more consid-

erable gap, with a 32-point difference between the subsets. Additionally, in rationale prediction, LLMs tend to favor gold labels over their predictions when nudged for explanations. Interestingly, contrary to popular literature, we find the impact of chain-of-thought (CoT) prompting to be insignificant, reinforcing the challenging nature of QUENCH.

Contributions: Through this work¹:

- We develop a novel open-domain quiz trivia dataset, QUENCH, accompanied with rationales for each question.
- We benchmark QUENCH on seven LLMs across 4 evaluation metrics and 2 prompting setups.
- We perform extensive analyses examining the influence of model size, prompting strategy, the role of indic vs. non-indic context, and highlight the most common prediction errors.

2 Related Work

Benchmarking and evaluating LLM is an active and evolving area of research (Chang et al., 2024; Peng et al., 2024). Benchmarks such as MMLU (Hendrycks et al., 2021), SuperGLUE (Wang et al.,

¹Code and dataset available at <https://github.com/af1ah02/QUENCH>

2019), HELM (Lee et al., 2023), PromptBench (Zhu et al., 2023) and LMSys (Zheng et al., 2024) provide a holistic suite of tasks to access the real-world adaptability of LLMs.

Dedicated benchmarks have been proposed to access the mathematical (Lu et al., 2023), symbolic (Zhang et al., 2024a), commonsense/social (Davis, 2023; Gandhi et al., 2023), and logical (Pan et al., 2023; Giadikiaroglou et al., 2024; Sanyal et al., 2022) reasoning of LLMs. Datasets such as GSM-8k (Cobbe et al., 2021), JEEBench (Arora et al., 2023), MMLU (Hendrycks et al., 2021), MaScQA (Zaki et al., 2024), ScienceQA (Lu et al., 2022), and LogiQA (Liu et al., 2020) have been designed to evaluate LLMs’ knowledge across various subjects like mathematics, material science, and history. Most of these datasets utilize multiple-choice questions and focus on single themes per question. Concerns have been raised about the potential leakage of LLM pretraining data due to the public availability of these datasets (Xu et al., 2024). Additionally, there are datasets aimed at assessing commonsense reasoning (Zellers et al., 2019; Lourie et al., 2021), entity resolution (Sakaguchi et al., 2021), and perceptiveness (Lin et al., 2022).

Given that reasoning involves combining latent information of varying concepts (Wu et al., 2023), modality (Zhang et al., 2024b; Liu et al., 2022), assessments in terms of knowledge-graph, and neurosymbolic (Olausson et al., 2023) reasoning have also been proposed. However, these setups, too, tend to operate in an MCQ or cloze manner. Our work addresses the closed-knowledge gap by integrating world knowledge and commonsense reasoning into a quiz-based framework. This framework requires LLMs to infer masked entities within questions and generate rationales for their predictions.

Meanwhile, for datasets specific to Indic cultures, such as (Seth et al., 2024; Watts et al., 2024), there is a notable scarcity of challenging, open-ended, quiz-style benchmarks. Our work contributes to evaluating the performance of the LLMs under Indic and non-Indic setups. In the future, this can be extended to evaluate more fine-grained geographical and cultural setups.

3 QUENCH: Proposed Benchmark

QUENCH is a collection of 400 English questions from quiz competitions encompassing the 11 themes in Figure 1. Each question consists of a paragraph talking about some event related to

Subset	# Q	# E	Avg. QL	Avg. EL	Avg. RL
Indic	70	80	77.03	1.85	39.48
Non-Indic	330	379	84.82	1.96	40.52
Over All	400	459	83.46	1.94	40.33

Table 2: Dataset statistics of QUENCH enlisting the number of questions (Q) and masked entities in the questions (E). We also report the average length of questions (QL), masked entities (EL), and the annotated rationale (RL).

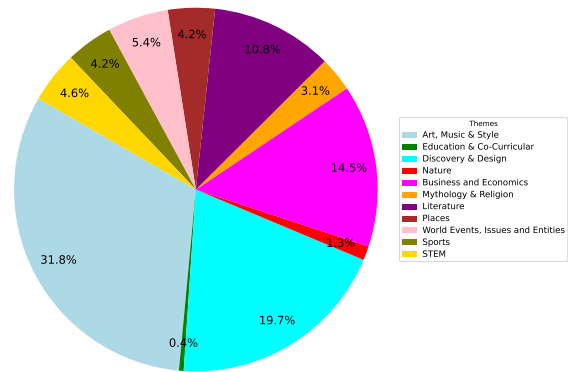


Figure 1: Themes and their distribution in QUENCH.

one of the themes. In each question, some entities are masked with ‘X.’ The aim is to connect the concepts in the question to predict ‘X.’ The questions contain adequate cues to deduce the entities. The questions in the quiz already have the entities masked, and we have not modified these. However, we manually annotate the explanations/rationale to arrive at the answer. Some sample questions, answers, and rationales are provided in Table 1.

Data Sources. Our primary source of questions is YouTube quizzing competitions videos², with a tiny portion (7%) from a website that publishes quizzing challenges³. Upon exploring these two sources, we find that the videos provide more coherent reasoning and answers, better facilitating the annotation process.

Annotation Process. The data annotation process is carried out by two male expert annotators (A1 and A2) aged 20-23. Both annotators possess previous experience with cryptic hunts, quizzes, and similar activities. The annotators spend 8-10 minutes per question listening to the questions in the video. Working with the Text-Grab OCR tool⁴, the annotators extract the question text and answers

²<https://www.youtube.com/@KumarVarunOfficial>

³<https://donquizzote.wordpress.com/>

⁴<https://learn.microsoft.com/windows/powertoys/text-extractor>

Annotation Tool

Enter the source of the question:

Enter the title of the question:

Enter the metadata of the question (if any):

Enter The Question:

Choose Themes:

<input type="checkbox"/> Architecture	<input type="checkbox"/> Art and Music	<input type="checkbox"/> Biology and Genetics
<input type="checkbox"/> Books	<input type="checkbox"/> Business and Economics	<input type="checkbox"/> Cricket
<input type="checkbox"/> Current Events	<input type="checkbox"/> Design	<input type="checkbox"/> Education
<input type="checkbox"/> Environment and Sustainability	<input type="checkbox"/> Fashion and Style	<input type="checkbox"/> Food and Cooking

(a)

<input type="checkbox"/> Sustainability	<input type="checkbox"/> Geology	<input type="checkbox"/> Health and Wellness
<input type="checkbox"/> Geography	<input type="checkbox"/> Hobbies and Interests	<input type="checkbox"/> Language and Linguistics
<input type="checkbox"/> History	<input type="checkbox"/> Literature	<input type="checkbox"/> Mathematics
<input type="checkbox"/> Law and Legality	<input type="checkbox"/> Mythology and Folklore	<input type="checkbox"/> Perfumes and Scents
<input type="checkbox"/> Movies and Entertainment	<input type="checkbox"/> Plants and Vegetation	<input type="checkbox"/> Politics and Government
<input type="checkbox"/> Pharmaceuticals & Medicine	<input type="checkbox"/> Psychology	<input type="checkbox"/> Religion and Philosophy
<input type="checkbox"/> Pop Culture	<input type="checkbox"/> Social Issues	<input type="checkbox"/> Space and Astronomy
<input type="checkbox"/> Science and Technology	<input type="checkbox"/> Travel and Tourism	<input type="checkbox"/> Wildlife and Nature
<input type="checkbox"/> Sports		

Enter comma separated variables (e.g. X, Y, Z)

X

Enter correct answer for X:

answer

Enter Rationale for X:

The rationale for why variable X refers to 'answer'.

Submit

(b)

Figure 2: Screenshots showing the upper (a) and lower (b) half of the custom annotation tool’s landing page has $6 + 2 \cdot N$ sections to fill, where N is the number of masked entities in the question. The metadata section is optional.

for the masked entities. The annotators manually rectify any issues that arise. Based on the transcribed content of the video, the annotators paraphrase the rationale into coherent, point-wise sentences that outline how the correct entity can be deduced once the rationale is read. In cases where the rationale is not discussed in the video, the annotators are free to access the internet to obtain the explanations. The annotators then populate the following fields:

- **Passage:** A passage with some entities/objects masked
- **List of Masked Entities:** The list of masked entities to be predicted. The questions are constructed to predict these entities.
- **List of Answers:** There is one answer for every mask entity.
- **List of Rationales:** The rationales behind each entity.
- **Themes:** A list of themes the passage fits in.
- **Source:** A URL to the question source.

Custom Annotation Tool. The annotation process is carried out online without the need to scrape any videos. Once the transcription is obtained, the annotators compile the above information for each sample. We use a custom annotation tool with the help of Streamlit⁵. A screenshot of the annotation process is highlighted in Figure 2. The themes are multi-choice and subjective to the annotator’s reasoning if not pre-defined. The variables section is constrained by the number of comma-separated

⁵<https://streamlit.io/>

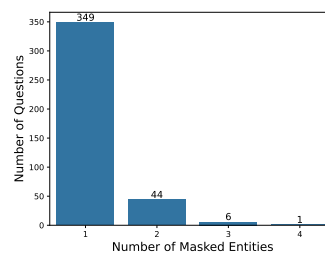


Figure 3: The number of masked entities in QUENCH.

variables inserted by the annotator. We also store the question’s source so that anyone can proof-check these annotations if needed.

Curated Dataset. The annotation process spanning 4 months is completed via a custom annotation tool as described below. We observe that a subset of questions exclusively pertains to Indian entities. Thus, we further tag each question as *Indic* or *Non-Indic*. The dataset statistics are outlined in Table 2. As a result of the annotations, the distribution of the number of masked entities across different questions is illustrated in Figure 3.

Inter-annotator Agreement. Since the rationale annotations are performed in a free-text manner, we carry out an inter-annotator evaluation to adjudge the quality of the rationale text. We randomly provide one annotator with 10 samples compiled by the other. Both rank the free-text rationale on a 5-point Likert scale, with five being the highest quality of annotation. Based on this assessment, we obtain an inter-annotator agreement of 4.9 from A1 to A2 and 4.85 from A2 to A1.

Unique Characteristics of QUENCH. *Firstly, a critical aspect of our dataset is its objective yet open-domain nature.* While the answers to the quiz question are objective (one word/one phrase), the queries do not have fixed gold labels (for example, ‘Barack Hussein Obama,’ ‘Barack Obama,’ and ‘Obama’ are all correct answers to the question ‘X was a civil rights attorney turned 44th President of USA’). Further, unlike the popular quiz show (Who Wants to Be Millionaire), we do not provide multiple-choice (MCQ) answers, which increases the difficulty of predicting the entities as the range of possibly correct entities is unrestricted. Simply removing the options in MCQ questions to produce a new dataset is not sufficient, as many of those questions depend on or refer to the options, such as, “Which of the following is closest to X?”. *Secondly, the multi-hop reasoning and multiple co-reference resolution setup in QUENCH spans inter-sectional themes within a question.* It provides a challenging environment to assess the world knowledge and entity recall capabilities of the LLMs. *Thirdly, it is also noteworthy that a subset of the quiz questions pertains to India-specific entities, allowing us to judge the indic-specific knowledge of the LLMs.* The questions mention sufficient Indian-specific concepts to nudge the deduction toward an India-specific answer without an explicit hint. *Lastly, the dataset allows for a multifold assessment of LLMs covering both entity recall as well as rationale-building capabilities.* Overall, based on the above characteristics, it is evident QUENCH provides a novel and challenging benchmark to access both world knowledge and recall capabilities of the LLM as well as establish the efficacy of the systems under open-domain setups beyond standardized NLP tasks. To ensure the quiz questions curated in QUENCH are not already present in LLM pretraining datasets, we also conduct a data contamination check (Appendix A).

4 Benchmarking Setup

This section outlines the models we employ for benchmarking QUENCH along with the prompting and evaluation setups.

Benchmarked LLMs. We experiment with various open-weight and closed-sourced instruction-tuned LLMs. Non-instruct LLMs are excluded from our assessment as they generate incoherent outputs. Similarly, formatting issues were registered from the Pythia family (Biderman et al., 2023)

Model	CW	Knowledge Cutoff
GPT-4-Turbo* (gpt-4-turbo-2024-04-09)	128K	Dec’23
GPT-3.5-Turbo* (gpt-3.5-turbo-0125)	16k	Sep’21
Gemini-1.5-Flash*	1M	Nov’23
Gemma-1.1-7B-Instruct†	8K	Unknown
Mixtral-8x7B-Instruct-v0.1†	32k	Unknown
Meta-Llama-3-8B-Instruct†	8k	Mar’23
Meta-Llama-3-70B-Instruct†	8k	Dec’23

Table 3: Details of LLMs employed in this study. CW captures the context window based on token length. Star(*) refers to closed-sourced LLMs and the dagger (†) signifies open-weight LLMs.

models. Our model shortlisting was performed on 2xH100 and 2xA100 machines. The LLMs eventually shortlisted for benchmarking QUENCH are furnished in Table 3. Amongst the closed-sourced models, we use the ones supported via APIs. Here, we employ **GPT-4-Turbo**, **GPT-3.5-Turbo**, and **Gemini-1.5-Flash**. For open-weighted models, we run inference via the free tier provided by Groq⁶. This setup allows fast inference at minimal infrastructural cost. Here we employ, **Meta-Llama-3-70B-Instruct**, **Meta-Llama-3-8B-Instruct**, **Mixtral-8x7B-Instruct-v0.1** and **Gemma-1.1-7B-Instruct**.

Prompting Predictions from LLMs. Given a quiz question and a list of missing entities, we prompt LLMs to predict the masked entities in the question. In the next stage of the pipeline, we use the predicted entities to prompt the model to provide a free text reason/rationale for why the expected entity correctly fits the context in the question. To separately examine the role of predicted entities in nudging the rationale, we also repeat the rationale prediction task with the gold-labeled entity to generate rationales. Here, we explore zero-shot prompting both with and without chain-of-thought (CoT) prompting (Wei et al., 2022). For CoT prompting, we add the phrase “Let’s think step by step” similar to Arora et al. (2023) to the prompts outlined in Appendix B.

Evaluation Metrics. We employ a suit of natural language generation (NLG) based metrics for evaluation. Among the standard ones, we report BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang* et al., 2020) for semantic similarity. The standard automated metrics, however, will penalize the open-ended responses with variations like ‘U.S.A.’ versus ‘United States’ or ‘United States of America’. For fairer and more comprehensive analysis, we also explore LLM-

⁶<https://console.groq.com/>

Model	Category	Without Chain-of-Thought				With Chain-of-Thought			
		Answer		Rationales with Predicted Answer		Answer		Rationales with Predicted Answer	
		BS	GEval	BS	GEval	BS	GEval	BS	GEval
Gemini 1.5 Flash	Indic	91.7	40	86.8	56.2	91.3	38	86.7	56.4
	Non-Indic	93.6	71	86.9	76	93.7	70	87.1	75.4
	All	93.3	66	86.9	72.6	93.3	64	87.1	72.2
	$\pm \Delta(N - I)$	+1.9	+31.0	+0.1	+19.8	+2.4	+32.0	+0.4	+19.0
	<hr/>								
Gemma-1.1-7B-it	Indic	89.2	14	86	38.6	87.5	20	86.2	43.6
	Non-Indic	90.8	41	85.9	58.2	89.3	43	86.1	59.6
	All	90.6	36	86	54.8	89	39	86.1	56.8
	$\pm \Delta(N - I)$	+1.6	+27.0	-0.1	+19.6	+1.8	+23.0	-0.1	+16.0
	<hr/>								
GPT-3.5-Turbo	Indic	95	53	87.4	67.2	94.6	54	87.6	66.2
	Non-Indic	95.9	76	87.6	83	96	77	87.6	82.6
	All	95.8	72	87.5	80.2	95.8	73	87.6	79.8
	$\pm \Delta(N - I)$	+0.9	+23.0	+0.2	+15.8	+1.4	+23.0	0.0	+16.4
	<hr/>								
GPT-4-Turbo	Indic	96.8	78	87.5	86.8	97	77	87.5	82.8
	Non-Indic	97.4	88	87.3	91	97.5	89	87.3	91
	All	97.3	86	87.3	90.4	97.4	87	87.3	89.6
	$\pm \Delta(N - I)$	+0.6	+10.0	-0.2	+4.2	+0.5	+12.0	-0.2	+8.2
	<hr/>								
Meta-Llama-3-8B-Instruct	Indic	94	29	86.3	48.8	93.3	25	86.5	46.6
	Non-Indic	95.3	46	86.4	60.8	95.4	47	86.4	62.6
	All	95.1	43	86.3	58.6	95	43	86.4	59.8
	$\pm \Delta(N - I)$	+1.3	+17.0	+0.1	+12.0	+2.1	+22.0	-0.1	+16.0
	<hr/>								
Meta-Llama-3-70B-Instruct	Indic	95.4	58	87	70.8	95.7	62	87	71.8
	Non-Indic	96.7	73	87.3	82	96.7	74	87.3	82.4
	All	96.5	70	87.3	80	96.5	72	87.2	80.6
	$\pm \Delta(N - I)$	+1.3	+15.0	+0.3	+11.2	+1.0	+12.0	+0.3	+10.6
	<hr/>								
Mixtral-8x7B-Instruct-v0.1	Indic	86.1	43	87	61.4	86	42	86.9	59.2
	Non-Indic	88	68	87	79.4	88.4	71	87.1	78.4
	All	87.7	64	87	76.2	88	66	87	75
	$\pm \Delta(N - I)$	+1.9	+25.0	0.0	+18.0	+2.4	+29.0	+0.2	+19.2
	<hr/>								

Table 4: LLM performances on QUENCH with and without Chain-of-Thought prompting. Here, $\Delta(N - I)$ is the difference in performance between the Non-Indic and Indic subset. BS: BERTScore; GEval: Jury Evaluation.

driven evaluations. Recent works have shown that LLMs favor their own outputs (Panickssery et al., 2024) during LLM-based evaluations. To address this, Verga et al. (2024) proposed using multiple LLMs as a jury. We, thus, employ the most robust models (GPT-4-Turbo, Mixtral-8x7B-Instruct-v0.1, and Meta-Llama-3-70B-Instruct) as our judges. Each judge individually scores the responses of every other benchmarked LLM, resulting in 21 judge-model combinations. In the case of entity prediction, we employ a binary scale to determine whether or not the entity is correct. For rationale prediction, we use a more granular 5-point Likert scale to capture semantics and handle nuances in lengthy texts. Since each question may contain multiple masked entities, we treat each entity as a distinct prediction. We aggregate the results across all entities for each question ($\sum_{i=1}^{num_q} \sum_{j=1}^{num_ent(q_i)} score(q_i, ent_j)$) where $num_ent(q_i)$ counts how many masked entities are present in the question q_i and finally scale the aggregated results to 100. The rationales are assessed under both predicted and gold entities. The evaluation prompts are outlined in Appendix C.

5 Results and Discussion

Comprehensive results for all metrics are provided in Appendix D, with a shorter aggregate provided

in Table 4 for reference here. In each table, we refer to $\Delta(N - I)$ as the difference in performance between the Non-Indic and Indic subsets.

5.1 Entity Prediction

In terms of standard metrics, we observe that GPT-4-Turbo is the best-performing model on QUENCH, with Meta-Llama-3-70B-Instruct performing comparably well as the second-best model in terms of overall BERTScore. In Table 4, GPT-4-Turbo produces the highest BERTScore of 97.3 (97.4) when prompted without (with) CoT. Meta-Llama-3-70B-Instruct lags only by ≈ 1 point with scores of 96.5 (96.5) when prompted without (with) CoT. Similarly, in terms of LLM-Juries, we observe (Figure 4 (a)) that GPT-4-Turbo outperforms other models with a score of 86 (87) when prompted without (with) CoT. Here again, the second-best models trail by ≈ 14 points, with GPT-3.5-Turbo and Meta-Llama-3-70B-Instruct scoring 72 (73) and 70 (72) respectively when prompted without (with) CoT.

5.2 Rationale Prediction

Based on the entity predicted in the previous step, the LLMs are then prompted to explain how the predicted entities can be deduced/thought through. Interestingly, from Table 4, we observe GPT-3.5-Turbo performs slightly better than the rest of the models, beating GPT-4-Turbo and Meta-Llama-3-

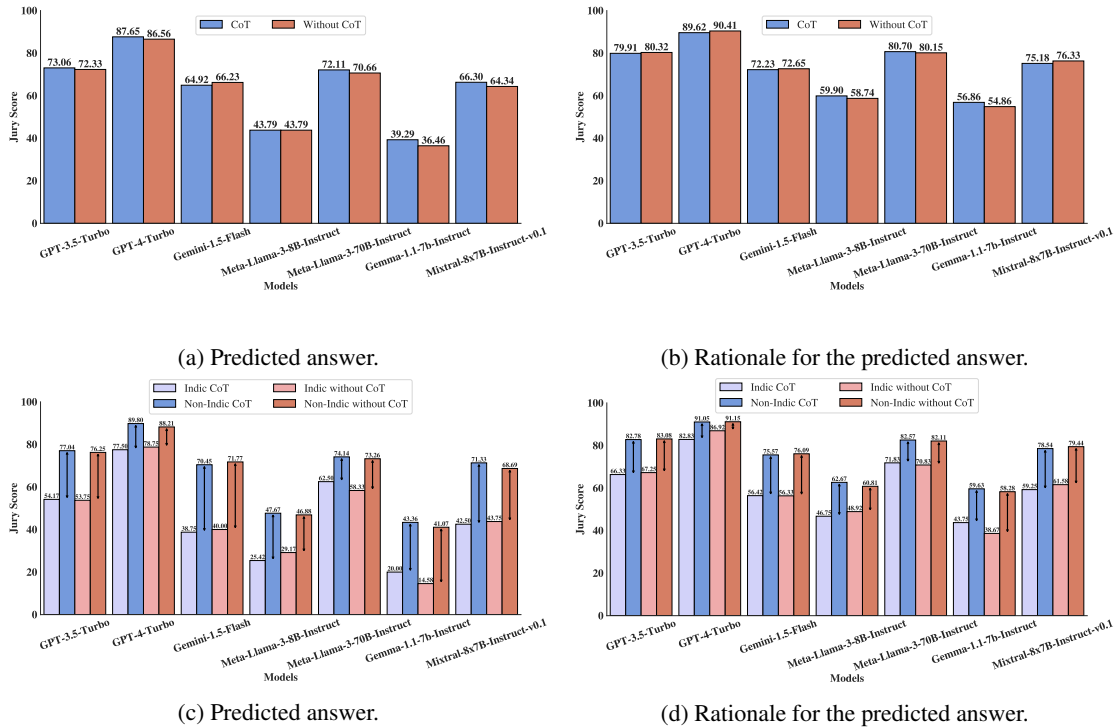


Figure 4: Figures (a) and (b) display our aggregated results comparing scenarios with and without Chain-of-Thought (CoT) prompting. Figures (c) and (d) present a comparison between Indic and Non-Indic languages. The results are evaluated across three types of metrics: (i) the correctness of the predicted answer, and (ii) the correctness of the rationale for the prediction. All metrics are scaled from 0 to 100.

70B-Instruct by $\approx 0.2\%$ (0.3%) and 0.2% (0.4%) BERTScores respectively, in terms of without (with) CoT. On the contrary, in terms of LLM-metrics (Figure 4 (b)), we observe that GPT-4-Turbo outshines other models in terms of producing coherent and correct rationals with scores as high as 90.41 (89.62) for rationales generated without (with) CoT. The meta-best systems trail by ≈ 10 points, with GPT-3.5-Turbo and Meta-Llama-3-70B-Instruct scoring 80.32 (79.91) and 80.15 (80.7), respectively for without (with) CoT.

Note on standard metrics. BERTScore is relatively more helpful than n-gram metrics like BLEU and ROUGE. *Specifically, for rationale generation, syntactic metrics fail to convey meaningful information. This is because the generated rationales can vary significantly in structure, format, and length while conveying the same reasoning.*

5.3 Factors Influencing the Performance of Benchmarked LLMs

We comment on the performance w.r.t standard and LLM-based metrics in terms of the training parameters, prompting strategy, the non-western context in the question, as well as the role of predicted/human entities in generating rationales.

Number of Parameters. Across both standard and LLM-based metrics, there is a clear distinction in terms of the number of model parameters. Within the same family, Meta-Llama-3-70B-Instruct outperforms Meta-Llama-3-8B-Instruct on both entity and rationale predictions. However, this performance drop is not consistent even among models with a similar number of parameters. The variation among models with similar parameter sizes reiterates that not only just the number of parameters but also the pretraining strategy plays a role in the downstream reasoning ability of the LLMs.

Impact of Indic Subset. All the benchmarking LLMs perform poorly at questions with an Indic context for both entity and rationale generation tasks. It is vital to reiterate that the only difference between indic and non-indic questions is the *context*, as the dataset and predictions are curated in English. The performance variation is significantly noticeable across all metrics, as evident from Table 4. In terms of GEval, the difference in performance is $\approx 21\%$ on average for predicting the correct entity (Figure 4 (c)) and $\approx 14.7\%$ for predicting the rationale (Figure 4 (d)) for the expected answer for non-Indic vs Indic entities. We hypothesize this

is because the pretraining datasets for LLMs have a predominantly North American context (Zhou et al., 2022). *It also means that cultural cues are critical while answering open-domain questions (Lee et al., 2024).*

Rationale Predictions with Gold Labels. In the case of generating a rationale for a given question, we alternatively provided the list of correct masked entities instead of the predicted ones. As expected, inserting gold labels in the prompts dramatically improves the quality of the rationale generated by the model. This difference is highest for Gemma-1.1-7b at $\approx 32\%$ (Tables 7 and 8). Similar behavior is observed even within Indic and non-Indic subsets. *It shows that LLMs are capable of reasoning the path between the correct answer and question much better than generating the rationale behind their own predicted answers (Huang et al., 2024).*

CoT Prompting. Contradictory to expected behavior, we observe via Figure 4 (a), (b) that the influence of CoT is inconsistent on QUENCH. It may be a result of the challenging nature of the quiz-based task. LLMs need improvements in reasoning to make connections between multiple entities in the real world. *In line with prior studies, we, too, observe that CoT optimization requires LLMs with parameters $> 7B$ (Chowdhery et al., 2023).*

6 Human Benchmarking

To highlight the competitive nature of our benchmark, we also perform a human benchmarking.

Setup. Given the resource-intensive nature of generating explanations, we randomly sample 20 questions for this assessment. We sample equal numbers from both subsets and across all themes. In total, we have 10 Non-Indic and 13 Indic questions, with some questions having more than one mask to predict. The participants answered both the missing entities and the rationale for the entities. The questions are distributed via a Google form (Appendix E). The participants are given the option to respond with "NA" if they do not have an answer for an entity or rationale. This setup allows us to analyze refusal rates effectively. We recruit 18 participants for benchmarking. The participants consist of college students pursuing a range of degrees, from Bachelor’s to Master’s and PhD programs.

Observations. Based on the number of unanswered questions, Figure 5 (a) demonstrates the

Error Types	Counts	
	w/o CoT	w/ CoT
Unrelated to theme	5	0
Unrelated but same theme	10	0
Similar entities in same theme	45	1
Wrong entity predicted but correct in rationale	5	9
Correct Answer	35	11

Table 5: Error analysis to classify each sample into one of the five error types (a–e, respectively).

difficulty humans face with questions from QUENCH. Moreover, both the task of predicting the correct answer and providing a rationale for it prove challenging, with the highest scores reaching only 30%. Additionally, Figure 5 (b-c) reveals that participants generally find Indic questions more difficult to answer. Interestingly, there are more instances where Indic rationales receive higher scores, which might suggest that writing rationales for Indic questions is easier once an answer is known. However, due to the limited number of participants, we refrain from making any broad conclusions.

7 Error Analysis

Even the best-performing LLM (GPT-4-Turbo in our case) is prone to generative errors. In this section, we focus on the incorrect predictions arising from entity recognition and reasoning.

Incorrect Entity Recognition. The model struggles with identifying the correct entities, even when using Chain-of-Thought (COT) reasoning. It becomes evident when, in 90.19% of cases, the model receives the same jury score for both CoT and non-CoT generations.

When prompted to identify entities within a specific domain, the model tends to favor well-known figures over the correct but less famous ones. For example, when asked about an Indian singer, it incorrectly identifies a widely recognized singer (e.g., “Sonu Nigam”) instead of the correct and lesser well-known individual (e.g., “Lucky Ali”). Consequently, the model has a tendency to link entities to more common or prominent organizations within a field rather than accurately identifying the rare entity. For instance, in a scenario where the correct answer is “Amrutanjan” (a patent medicine business), the model incorrectly identifies “All India Radio,” possibly due to the mention of a journalist in the question. The model can provide misinformed entities, revealing significant gaps in its knowledge and ability to differentiate between similarly categorized entities. For example, it in-

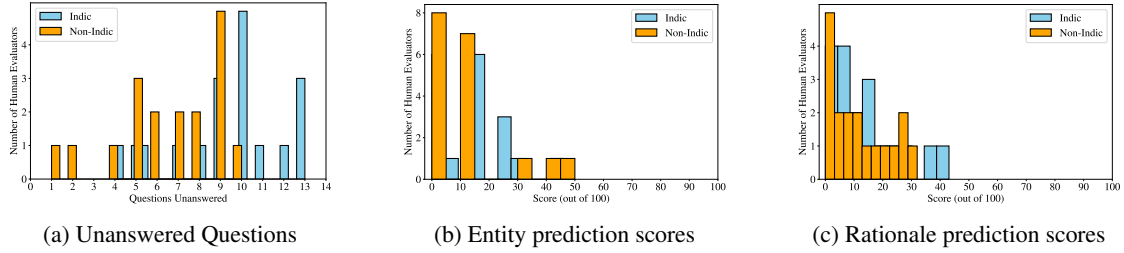


Figure 5: Analysis of human benchmarking split across the indic and non-indic subsets capturing the distribution.

	Correct Rationale for incorrect prediction?	
	Yes	No
Counts (w/o CoT)	17	83
Counts (w/ CoT)	8	2

Table 6: Assessing when LLMs produce a correct rationale despite making an incorrect entity prediction.

correctly identifies “Yogi Adityanath” (the Chief Minister of Uttar Pradesh) as the Chief Minister of Odisha instead of “Naveen Patnaik.”

Despite the shortcomings in specific entity recognition, the model shows a surprising proficiency in identifying high-level relationships. Suppose the answer involves identifying a cricketer. In that case, the model generally generates a name within the correct category (i.e., a cricketer), even if it does not pinpoint the exact individual. It suggests that while the model has a reasonable grasp of broad categories, roles, and domain leaders within that role, it may struggle with precise identification within those categories. Some other examples are predicting “Waheeda Rehman” in place of “Bhanu Athaiya,” who are both famous Indian actresses.

Incorrect Rationale Generation: LLMs can also struggle to generate specific rationales. We conduct a 2-dimensional analysis of GPT-4 over 100 randomly sampled predictions and record:

1. The type/category of error in the generations outlined in (Table 5).
2. If the model is able to generate the correct rationale even if the entity predicted is incorrect as outlined in (Table 6).

Table 5 shows the statistics for the categorical error analysis. We highlight that the CoT sample count does not total 100 because we only mark samples where the reasoning differed significantly between the two experiments (without and with CoT). We see that most of the errors are of type c, where the theme is correct, but the LLM gets confused between two very similar entities from the same domain (possibly due to differences in

entity popularity/richness of embeddings or their closeness in terms of semantic overlap).

Through manual assessment, we also notice that CoT may not be effective in the entity prediction part. When employed for rationale generation, it can rectify the mistakes in initial entity predictions. Table 6 shows that for $\approx 20\%$ of the subset of samples, the model is able to predict the rationale correctly without needing first to predict the correct entity, and for 80% of the anomalous cases in CoT, we see that the rationales are correct. These observations show that LLMs are incredibly sensitive to the way prompts are framed, and predicting entities in a subjective fashion is a relatively more complex challenge than in an MCQ setting.

8 Conclusion

We devise a novel benchmark, QUENCH of about 400 quizzing questions in English from a diverse set of themes. QUENCH helps evaluate the deductive reasoning capabilities of LLMs. Interestingly, by accessing CoT and without CoT prompting techniques, we recorded the non-consequent differences between the two setups for our dataset. We also find that LLMs are much better at answering questions that have a general/non-indic context. Overall, we observe QUENCH to be a challenging benchmarking necessitating future research in the area of open-domain deductive reasoning. In the future, we would like to extend our assessment under multilingual settings and benchmark other reasoning techniques such as tree-of-thoughts, question decomposition, and self-consistent CoT.

Limitations

Despite our best efforts, we could not evaluate a broader range of models and prompting techniques due to resource constraints. This study examines the behavior of entity and rationale predictions in a 2-step fashion, and the real impact of CoT may come into play if both tasks are prompted in a single prompt with the liberty first to rationalize

and then predict the entities. However, parsing such responses will be complex as a strict format may not be followed, and it may require increased human efforts for evaluation. While a jury can lead us to a better assessment, each jury-LLM incurs a cost in terms of hardware resources and API.

Ethical Considerations

Human annotators play a crucial role in the development of our dataset. We ensure that all annotators are fairly compensated for their work and provided with clear instructions to minimize subjectivity and bias in their annotations. Further, the annotators were offered sufficient time to annotate so as not to burden them. Annotators were also given the option to decline participation without any repercussions. We maintained a respectful and supportive work environment throughout the annotation process. Secondly, we source all our data from public platforms and test both closed-source and open-weight models, which allow for a fair benchmarking of LLMs. Lastly, we also make minimal use of LLMs for re-writing and grammatical corrections and, in some cases, Copilot for code completion during experiments.

Acknowledgments

The authors acknowledge the support of Infosys foundation through Center for AI (CAI) at IIT Delhi.

References

- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. *arXiv preprint arXiv:2201.06642*.
- Daman Arora, Himanshu Singh, and Mausam. 2023. [Have LLMs advanced enough? a challenging problem solving benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. [A survey on evaluation of large language models](#). *ACM Trans. Intell. Syst. Technol.*, 15(3).
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pili, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. [Palm: Scaling language modeling with pathways](#). *Journal of Machine Learning Research*, 24(240):1–113.
- Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Together Computer. 2023. [Redpajama: an open dataset for training large language models](#).
- Ernest Davis. 2023. [Benchmarks for automated commonsense reasoning: A survey](#). *ACM Comput. Surv.*, 56(4).
- Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. [Towards measuring the representation of subjective global opinions in language models](#). *Preprint*, arXiv:2306.16388.
- Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer

- Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. [What’s in my big data?](#) *Preprint*, arXiv:2310.20707.
- Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. [Understanding social reasoning in language models with language models.](#) *Preprint*, arXiv:2306.15448.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The pile: An 800gb dataset of diverse text for language modeling.](#) *Preprint*, arXiv:2101.00027.
- Panagiotis Giadikiaroglou, Maria Lymperaioi, Giorgos Filandrianos, and Giorgos Stamou. 2024. [Puzzle solving using reasoning of large language models: A survey.](#) *Preprint*, arXiv:2402.11291.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding.](#) In *International Conference on Learning Representations*.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. [Large language models cannot self-correct reasoning yet.](#) In *The Twelfth International Conference on Learning Representations*.
- Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. [Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis.](#) *Preprint*, arXiv:2308.16705.
- Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy S Liang. 2023. [Holistic evaluation of text-to-image models.](#) In *Advances in Neural Information Processing Systems*, volume 36, pages 69981–70011. Curran Associates, Inc.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries.](#) In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- 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*.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. [Logiqa: A challenge dataset for machine reading comprehension with logical reasoning.](#) In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. [Things not written in text: Exploring spatial commonsense from visual signals.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376, Dublin, Ireland. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. [Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark.](#) *AAAI*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. [Learn to explain: Multimodal reasoning via thought chains for science question answering.](#) In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. [A survey of deep learning for mathematical reasoning.](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.
- Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. [LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.
- Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. [Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [Llm evaluators recognize and favor their own generations.](#) *Preprint*, arXiv:2404.13076.

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.
- Ji-Lun Peng, Sijia Cheng, Egil Diau, Yung-Yu Shih, Po-Heng Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. [A survey of useful llm evaluation](#). *Preprint*, arXiv:2406.00936.
- Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. [Reasoning with language model prompting: A survey](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *arXiv e-prints*.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. [Winogrande: an adversarial winograd schema challenge at scale](#). *Commun. ACM*, 64(9):99–106.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. [FaiRR: Faithful and robust deductive reasoning over natural language](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1075–1093, Dublin, Ireland. Association for Computational Linguistics.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.
- Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. [DOSAs: A dataset of social artifacts from different Indian geographical subcultures](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. [Dolma: an open corpus of three trillion tokens for language model pretraining research](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.
- Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. [Pub: A pragmatics understanding benchmark for assessing llms’ pragmatics capabilities](#). *Preprint*, arXiv:2401.07078.
- Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. [Replacing judges with juries: Evaluating llm generations with a panel of diverse models](#). *Preprint*, arXiv:2404.18796.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. [Pariksha : A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data](#). *Preprint*, arXiv:2406.15053.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Yongkang Wu, Meng Han, Yutao Zhu, Lei Li, Xinyu Zhang, Ruofei Lai, Xiaoguang Li, Yuanhang Ren, Zhicheng Dou, and Zhao Cao. 2023. [Hence, socrates is mortal: A benchmark for natural language syllogistic reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2347–2367, Toronto, Canada. Association for Computational Linguistics.
- Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. [Benchmarking benchmark leakage in large language models](#). *arXiv preprint arXiv:2404.18824*.
- Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. 2024. [Mascqa: investigating materials science knowledge of large language models](#). *Digital Discovery*, 3(2):313–327.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of*

the 57th Annual Meeting of the Association for Computational Linguistics, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024a. [Llm as a mastermind: A survey of strategic reasoning with large language models](#). *Preprint*, arXiv:2404.01230.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024b. [Multi-modal chain-of-thought reasoning in language models](#). *Transactions on Machine Learning Research*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. [LMSYS-chat-1m: A large-scale real-world LLM conversation dataset](#). In *The Twelfth International Conference on Learning Representations*.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022. [Richer countries and richer representations](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2074–2085, Dublin, Ireland. Association for Computational Linguistics.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. [Promptbench: Towards evaluating the robustness of large language models on adversarial prompts](#). *arXiv preprint arXiv:2306.04528*.

A Data Leakage Analysis

Due to the potential for benchmark leakage into pretraining corpora, we investigate several popular data sources to ensure our benchmark remains unaffected.

Phase 1: Checking for presence of data sources in pretraining corpora

Specifically, we examine [youtube.com](https://www.youtube.com) (YT) and donquizote.wordpress.com (DQ), which are primary sources using Elazar et al. (2024), to determine if data from these sources has been used in pretraining. We check for contamination in the C4 (Raffel et al., 2019), mC4-en (Chung et al., 2023), OSCAR (Abadji et al., 2022), RedPajama (Computer, 2023), LAION-2B-en (Schuhmann et al., 2022) and Dolma (Soldaini et al., 2024) datasets.

We observe a negligible presence of both data sources across all public pretraining corpora, with a maximum of around 22K tokens, and most instances below 1K tokens. Notably, older datasets like OSCAR and C4 show a slightly higher presence of DQ, with 23K tokens (0.0000046% of total tokens) and 9K tokens (0.00000032% of total tokens), respectively. In contrast, newer datasets such as Dolma contain only 40 tokens (0.00000091% of total tokens). Other datasets show no trace of the site.

Similarly, YT follows the same pattern. WIMBD’s prefix search also picks up other URLs like youtube.com/activate.org, which share the "youtube.com" prefix, but their presence is minimal (<0.00000001%). The highest YT presence is found in mC4-en with 22.3K tokens (0.00000081% of total tokens), indicating that both sources have an insignificant presence in the pretraining corpora.

Phase 2: Checking for exact match count in pretraining corpora

We utilize the Infinigram API (Liu et al., 2024) to search for exact question matches within the following pretraining corpora: Dolma (Soldaini et al., 2024), RedPajama (Computer, 2023), Pile (Gao et al., 2020), and C4 (Raffel et al., 2019). Our analysis shows zero contamination across all corpora, confirming that our dataset has not been leaked and any model performance on it is unrelated to memorization.

B Prediction Prompts

- **For predicting entities:** *Consider yourself a participant in a quiz show where I am the quizmaster. I will ask you a question that can be from any general theme. You need to provide me with the correct answer. The question can have multiple variables to answer, and you need to provide me with the answer for variable {}. Hence, use the following format strictly in your response: ‘The answer is <X answer>.’ You lose points if you fail to follow the format.*
- **For generating rationale using prediction:** *Consider yourself a participant in a quiz show where I am the quizmaster. I will ask you a question that can be from any general theme. The prediction for variable {} is {}. Provide me with the rationale followed for your answer. Use the following format in your response: ‘The rationale is <rationale>’. You lose points if you fail to follow the format."*
- **For generating rationale using gold labels:** *Consider yourself a participant in a quiz show where I am the quizmaster. I will ask you a question that can be from any general theme. You need to provide me with the correct answer. The question can have multiple variables to answer, and you need to provide me with the answer for variable {}. Hence, use the following format strictly in your response: ‘The answer is <X answer>.’ You lose points if you fail to follow the format.*

Note that the phrase “You lose points if you fail to follow the format” is only added to induce a quizzing setup. Since ours is a zero-shot setup, there is no explicit punishment/loss sent as feedback to the model if the answer is not correctly predicted. Further, the models are evaluated independently and not competing against each other.

C Evaluation Prompts

- **Entity Evaluation:** *You are the host of a quiz show where you ask complex and tricky questions to the contestants. Now, once you ask one such question, the contestant gives an answer that might not be the exact answer but is still correct. For instance, the answer provided to you might be ‘U.S.A’ while the actual answer is ‘United States of America’ or ‘United States’ or ‘America’, etc. Use your wise judgment to decide, based on the ques-*

tion given, whether the answer is correct or not. **YOU ARE THE JUDGE AND YOUR WORD IS FINAL.** Be fair and just in your judgment. Always respond with 'correct' or 'incorrect' based on the answer provided by the contestant. You will be provided the question, the true answer, and the answer provided by the contestant, and you need to decide whether the answer is correct or not. ## Question

<question>

True Answer

<true_answer>

Answer Given by contestant

<answer_given_by_contestant>

Your Judgement (correct/incorrect)

- **Rationale Evaluation:**

*You are the host of a quiz show where you ask complex and tricky questions to the contestants. Now, once you ask one such question, the contestant gives an answer as well as the rationale behind that answer. You need to decide whether the rationale provided is correct or not by comparing it with the true rationale. Use your wise judgment to decide based on the question given whether the rationale is correct or not. **YOU ARE THE JUDGE AND YOUR WORD IS FINAL.** Be fair and just in your judgment. Provide a score between 1 to 5 based on the rationale provided by the contestant. 1 being the least and 5 being the highest score. ## Question*

<question>

True Rationale

<true_answer>

Rationale Given by contestant

<rationale_given_by_contestant>

Your Judgement (Score between 1 to 5 do not provide any other score or text)

D Evaluation Results

Tables 8 and 7 capture the results with and without CoT prompting on varying subsets of the dataset and record the evaluation on all metrics BLEU, ROUGE, BERTScore and LLM-metrics.

BLEU and ROUGE. From Tables 7 and 8, with a BLEU of 65.7 (66.2), GPT-4-Turbo leads among all models when prompted without (with) CoT. In comparison, Meta-Llama-3-70B-Instruct trails by ≈ 4 points with a BLEU of 61.2 (61.5) when prompted without (with) CoT. Meanwhile, in terms of ROUGE-L (Tables 7 and 8), we ob-

serve similar patterns. GPT-4-Turbo leads among all models when prompted without (with) CoT with a score of 89.8 (89.9), and Meta-Llama-3-70B-Instruct trails behind with 83.3 (83.6) when prompted without (with) CoT. On the other hand, GPT-3.5-Turbo leads with 22.5 (22.7) BLEU and 29.1 (29.7) ROUGE-L when prompted without (with) CoT. Meta-Llama-3-70B-Instruct trails by 1 point with a BLEU of 21.4 (21.3) and a ROUGE-L of 28.7 (28.7) when prompted without (with) CoT.

Tradeoff between Closed Source v/s Open Weight Models. Across all metrics and experiments, GPT-4-Turbo consistently outperformed other models. While its performance is closely followed by both GPT-3.5-Turbo and the open-weighted Meta-Llama-3-70B-Instruct, the tradeoff between using open and closed sources is apparent. Speaking broadly about the LLM-evaluated metrics, if the only parameter to optimize is performance (in lieu of cost or open access), then GPT-4-Turbo should be the preferred model for world knowledge-related tasks. Meanwhile, Meta-Llama-3-70B-Instruct can efficiently serve as a substitute for the closed-sourced API models. It registers a slight drop in performance against GPT-4-Turbo and clearly outperforms Gemini-1.5-Flash, while performance is very comparable to GPT-3.5-Turbo. However, the 70B model comes with resource constraints. In terms of the LLM-evaluated metric, Mixtral-8x7B-Instruct-v0.1 is an excellent alternative to Meta-Llama-3-70B-Instruct in case of limited hardware, albeit with a significant performance drop.

E Human Benchmarking


Figure 6 provides an overview of the Google form employed for human benchmarking of QUENCH.

QUENCH Human Evaluation

In this task, you'll be presented with passages where certain details are missing and replaced by variables such as X and Y. Your objective is to carefully analyze the provided context and guess the missing entity while also providing rationales for why you think that answer fits the question. If you do not have any idea write "NA" in both blanks.

Variables can represent entities, concepts etc. and you need to think of a reasonable answer (we do not expect an exact match with our answer and hence will not penalize for the same)

Please do not search answers on the internet. This form is to evaluate the difficulty of our benchmark.



* Indicates required question

Email *

Your email

(a)

Q3. X co-founded Y in 2007. In 2018, X signed a non- compete clause with Walmart and exited Y. X then founded Navi in December 2018 along with IIT-Delhi batchmate Ankit Agarwal. Navi operates in the space of digital Loans, home loans, mutual funds, health insurance and micro-loans. On 12 March 2022 it filed a draft for an INR 3350 crore IPO. *

ID X, Y with the rationale you used to arrive at the answer.

Your answer

Rationale for X in Q3 *

Your answer

Rationale for Y in Q3 *

Your answer

(b)

Figure 6: Screenshots showing the instructions (a) and one of the questions with multiple rationales (b) for the human evaluation of QUENCH.

Model	Category	Answer				Rationales w/ Predicted Answer				Rationales w/ Gold Answer			
		BLEU	Rouge	BS	GEval	BLEU	Rouge	BS	GEval	BLEU	Rouge	BS	GEval
Gemini 1.5 Flash	Indic	31.3	72.8	91.7	40	6.6	29.2	86.8	56.2	19.7	30.6	87.6	87.6
	Non-Indic	36.4	81.9	93.6	71	18.4	28	86.9	76	17.5	27.9	87.2	93.8
	All	35.5	80.2	93.3	66	13.7	28.2	86.9	72.6	17.8	28.4	87.3	92.6
	$\pm\Delta(N - I)$	+5.1	+9.1	+1.9	+31.0	+11.8	-1.2	+0.1	+19.8	-2.2	-2.7	-0.4	+6.2
Gemma-1.1-7B-it	Indic	13.2	45.7	89.2	14	19.4	28.5	86	38.6	22	30.3	87	85.2
	Non-Indic	15.3	52.3	90.8	41	17.2	26	85.9	58.2	19.4	28.2	86.9	88.2
	All	14.9	51.1	90.6	36	17.6	26.4	86	54.8	19.9	28.5	86.9	87.6
	$\pm\Delta(N - I)$	+2.1	+6.6	+1.6	+27.0	-2.2	-2.5	-0.1	+19.6	-2.6	-2.1	-0.1	+3.0
GPT-3.5-Turbo	Indic	54.8	76.2	95	53	23.2	29.3	87.4	67.2	27.8	34.7	88.7	94
	Non-Indic	52.8	81	95.9	76	22.3	29.1	87.6	83	23.9	31.4	88.2	94
	All	53.1	80.1	95.8	72	22.5	29.1	87.5	80.2	24.6	32	88.3	94
	$\pm\Delta(N - I)$	-2.0	+4.8	+0.9	+23.0	-0.9	-0.2	+0.2	+15.8	-3.9	-3.3	-0.5	0.0
GPT-4-Turbo	Indic	66.4	87.5	96.8	78	18.8	27.6	87.5	86.8	19	28.2	88.1	97
	Non-Indic	65.5	90.4	97.4	88	17.1	25.6	87.3	91	16.8	27	87.6	97.6
	All	65.7	89.8	97.3	86	17.4	26	87.3	90.4	17.1	27.2	87.7	97.6
	$\pm\Delta(N - I)$	-0.9	+2.9	+0.6	+10.0	-1.7	-2.0	-0.2	+4.2	-2.2	-1.2	-0.5	+0.6
Meta-Llama-3-8B-Instruct	Indic	53.4	71.6	94	29	18.7	26.8	86.3	48.8	20.2	27.5	87.2	84.8
	Non-Indic	54.2	75.1	95.3	46	18	25.2	86.4	60.8	19.4	27.4	87	84.6
	All	54.1	74.4	95.1	43	18.1	25.5	86.3	58.6	19.5	27.4	87.1	84.6
	$\pm\Delta(N - I)$	+0.8	+3.5	+1.3	+17.0	-0.7	-1.6	+0.1	+12.0	-0.8	-0.1	-0.2	-0.2
Meta-Llama-3-70B-Instruct	Indic	59.1	79.4	95.4	58	21.7	28.5	87	70.8	22.5	32.1	88.1	92.2
	Non-Indic	61.7	84.2	96.7	73	21.3	28.7	87.3	82	21	29.2	87.6	94.2
	All	61.2	83.3	96.5	70	21.4	28.7	87.3	80	21.2	29.7	87.7	93.8
	$\pm\Delta(N - I)$	+2.6	+4.8	+1.3	+15.0	-0.4	+0.2	+0.3	+11.2	-1.5	-2.9	-0.5	+2.0
Mixtral-8x7B-Instruct-v0.1	Indic	4	21.5	86.1	43	16.8	26.3	87	61.4	16	25.6	87.2	93.4
	Non-Indic	4.9	29.8	88	68	16.1	25.7	87	79.4	15.7	25.9	87.3	95.6
	All	4.7	28.5	87.7	64	16.3	25.8	87	76.2	15.7	25.9	87.3	95.2
	$\pm\Delta(N - I)$	+0.9	+8.3	+1.9	+25.0	-0.7	-0.6	0.0	+18.0	-0.3	+0.3	+0.1	+2.2

Table 7: This table shows our complete evaluation results without using the Chain-of-Thought prompting technique. Here, $\Delta(N - I)$ is the difference in performance between the Non-Indic and Indic subset. BS: BERTScore; GEval: Jury Evaluation.

Model	Category	Answer				Rationales w/ Predicted Answer				Rationales w/ Gold Answer			
		BLEU	Rouge	BS	GEval	BLEU	Rouge	BS	GEval	BLEU	Rouge	BS	GEval
Gemini 1.5 Flash	Indic	1.3	72.6	91.3	38	4.7	28.2	86.7	56.4	18.9	29.2	87.4	92.4
	Non-Indic	34.8	81.7	93.7	70	18.5	28.3	87.1	75.4	17.7	27.9	87.2	93.6
	All	6.9	80.1	93.3	64	12.1	28.3	87.1	72.2	17.9	28.1	87.2	93.4
	$\pm\Delta(N - I)$	+33.5	+9.1	+2.4	+32.0	+13.8	+0.1	+0.4	+19.0	-1.2	-1.3	-0.2	+1.2
Gemma-1.1-7B-it	Indic	8.4	31.9	87.5	20	20.2	30.5	86.2	43.6	19.8	29.3	86.7	87
	Non-Indic	10.1	39.9	89.3	43	17.2	26.5	86.1	59.6	18	27.4	86.6	88.4
	All	9.8	38.6	89	39	17.8	27.2	86.1	56.8	18.3	27.7	86.6	88.2
	$\pm\Delta(N - I)$	+1.7	+8.0	+1.8	+23.0	-3.0	-4.0	-0.1	+16.0	-1.8	-1.9	-0.1	+1.4
GPT-3.5-Turbo	Indic	44.8	74	94.6	54	24.7	30.2	87.6	66.2	28	34.7	88.7	94.8
	Non-Indic	53.7	82.2	96	77	22.3	29.6	87.6	82.6	23.1	31.4	88	95.6
	All	52	80.7	95.8	73	22.7	29.7	87.6	79.8	23.9	32	88.2	95.4
	$\pm\Delta(N - I)$	+8.9	+8.2	+1.4	+23.0	-2.4	-0.6	0.0	+16.4	-4.9	-3.3	-0.7	+0.8
GPT-4-Turbo	Indic	67.1	87.3	97	77	18.5	27.2	87.5	82.8	18.9	29.6	88.2	97.6
	Non-Indic	66.6	90.4	97.5	89	16.6	26	87.3	91	16.6	26.5	87.5	97.6
	All	66.6	89.9	97.4	87	16.9	26.2	87.3	89.6	17	27	87.6	97.6
	$\pm\Delta(N - I)$	-0.5	+3.1	+0.5	+12.0	-1.9	-1.2	-0.2	+8.2	-2.3	-3.1	-0.7	0.0
Meta-Llama-3-8B-Instruct	Indic	48.6	67.4	93.3	25	19.7	27.7	86.5	46.6	20.7	29	87.2	77
	Non-Indic	55.3	75.7	95.4	47	17.6	25.1	86.4	62.6	19.8	27.9	87.1	82.8
	All	54.1	74.2	95	43	17.9	25.6	86.4	59.8	20	28.1	87.1	81.8
	$\pm\Delta(N - I)$	+6.7	+8.3	+2.1	+22.0	-2.1	-2.6	-0.1	+16.0	-0.9	-1.1	-0.1	+5.8
Meta-Llama-3-70B-Instruct	Indic	59.5	81	95.7	62	21.6	28.3	87	71.8	24.9	32.2	88.2	95.6
	Non-Indic	62	84.1	96.7	74	21.2	28.8	87.3	82.4	21.5	30	87.8	94.2
	All	61.5	83.6	96.5	72	21.3	28.7	87.2	80.6	22.1	30.4	87.9	94.4
	$\pm\Delta(N - I)$	+2.5	+3.1	+1.0	+12.0	-0.4	+0.5	+0.3	+10.6	-3.4	-2.2	-0.4	-1.4
Mixtral-8x7B-Instruct-v0.1	Indic	3.8	21.4	86	42	16.8	27.1	86.9	59.2	16.8	26.2	87.6	94
	Non-Indic	4.9	32.4	88.4	71	16.1	25.8	87.1	78.4	15.7	26.1	87.4	95
	All	4.7	30.5	88	66	16	26.1	87	75	15.7	26.2	87.4	94.8
	$\pm\Delta(N - I)$	+1.1	+11.0	+2.4	+29.0	-0.7	-1.3	+0.2	+19.2	-1.1	-0.1	-0.2	+1.0

Table 8: This table shows our complete evaluation results using the Chain-of-Thought prompting technique. Here, $\Delta(N - I)$ is the difference in performance between the Non-Indic and Indic subset. BS: BERTScore; GEval: Jury Evaluation.