THE ROLE OF SYNTHETIC DATA IN MULTILINGUAL, MULTICULTURAL AI SYSTEMS: LESSONS FROM INDIC LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Developing AI systems that operate effectively across languages while remaining culturally grounded is a long-standing challenge, particularly in low-resource settings. Synthetic data provides a promising avenue, yet its effectiveness in multilingual and multicultural contexts remains underexplored. We investigate the creation and impact of synthetic, culturally contextualized datasets for Indian languages through a bottom-up generation strategy that prompts large open-source LLMs ($\geq$ 235B parameters) to ground data generation in language-specific Wikipedia content. This approach complements the dominant top-down paradigm of translating synthetic datasets from high-resource languages such as English. We introduce UPDESH, a high-quality large-scale synthetic instruction-following dataset comprising 9.5M data points across 13 Indian languages, encompassing diverse reasoning and generative tasks with an emphasis on enhancing long-context and multi-turn capabilities, in addition to improving alignment with Indian cultural contexts. A comprehensive evaluation incorporating both automated metrics and human annotation across 10k assessments indicates that the generated data is of high quality; however, human evaluation highlights specific areas for further improvement. Additionally, we perform downstream evaluations by fine-tuning models on our dataset and assessing their performance across 15 diverse multilingual datasets to evaluate the generalizability of our dataset. Our experiments demonstrate that models trained on UPDESH consistently obtain significant performance improvements on generative tasks and remain competitive on multiple-choice evaluations in NLU tasks. Further, the relative improvements of models fine-tuned with UPDESH are most pronounced in low and medium-resource languages, effectively narrowing the gap between these languages and high-resource languages. These findings provide empirical evidence that effective multilingual AI development requires multi-faceted data curation and generation strategies that include context-aware, culturally grounded methodologies.

1 INTRODUCTION

Building multilingual and multicultural AI is essential to ensure equitable access to technology in diverse linguistic and cultural communities. Without this inclusivity, AI systems risk reinforcing global power imbalances by privileging dominant languages and Western-centric worldviews. However, even frontier language models often underperform in languages other than English and in non-Western contexts. This is not only due to limited linguistic and cultural diversity in pre-training data, but also to English-centric practices in fine-tuning, evaluation, and other parts of the development pipeline. While pre-training data is often treated as abundant and readily available due to large-scale web crawling, fine-tuning and evaluation data must be deliberately constructed and curated to ensure quality and relevance. This often leads to skipping the creation of truly multilingual and culturally grounded data altogether, or relying on simplified approaches like translating existing English datasets, often overlooking linguistic nuance and cultural context.

As highlighted by Joshi et al. (2020), there is a significant data gap between high-resource and low-resource languages in the internet data, the major source of pre-training data. Their taxonomy classifies world languages into six resource classes, with Classes 5-6 encompassing over 2400 languages (93.87% of all world languages) serving 1.2 billion speakers, yet having minimal to no digital resources, creating a massive representation gap in current AI systems. This gap is even more pronounced for fine-tuning and evaluation datasets (Hu et al., 2025). Synthetic data generation has become popular to augment training sets and create benchmarks, by generating new data or by transforming existing data into desirable formats, such as converting raw text into structured instruction-style prompts. Synthetic data, created mainly in English, has shown promise in improving downstream performance in tasks such as reasoning (Goldie et al., 2025; Harsha et al., 2025), coding (Wei et al., 2024; Shao et al., 2025) and retrieval (Bonifacio et al., 2022; Dai et al., 2023; Chitale et al., 2025). In this work, we study whether synthetic data has the potential to supplement scarce multilingual, multicultural data when carefully designed.

Techniques for generating synthetic data are typically grounded in English-centric practices, often relying on assumptions about model quality and reliability that do not hold for many other languages. Assessing the quality

and utility of synthetic data is an ongoing area of research. Current methods for evaluating quality include human and automated evaluation to assess fluency, correctness, and diversity, which may not suffice for assessing the quality of multilingual, multicultural synthetic data. Downstream evaluation involves fine-tuning models using synthetic data and evaluating models on standardized benchmarks. In this paper, we introduce the first framework for multilingual and multicultural synthetic data generation, focusing on key dimensions such as choice of generation strategies, grounding in language-specific data, quality assessment, and evaluation. We apply this framework to a case study that involves generating synthetic data across 13 Indian languages and conduct comprehensive evaluations to check for quality and utility. Our detailed analysis across multiple multilingual synthetic datasets offers practical insights for developing more inclusive, linguistically diverse, and culturally attuned language technologies.

2 FRAMEWORK

Synthetic data is utilized throughout the AI lifecycle, including pre-training, supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and evaluation (Viswanathan et al., 2025). Synthetic data generation starts with a small set of human-curated seeds, which is then provided to an LLM to generate instructions, inputs and outputs, followed by filtering for correctness and quality. An alternate approach is to retrieve the relevant datasets and use an LLM to rewrite them into the desired format. The dominant method for producing multilingual synthetic data involves translating English data with Neural Machine Translation systems or Large Language Models (LLMs) for pre-training (Doshi et al., 2024; Khan et al., 2024), Supervised Fine Tuning (SFT) (Shaham et al., 2024), and evaluation. However, translation may introduce translation artifacts (referred to as *translationese*) (Zhang & Toral, 2019; Vanmassenhove et al., 2021) and yield weaker correlations with human evaluations compared to benchmarks tailored for particular languages and contexts (Kreutzer et al., 2025; Wu et al., 2025). Recent work has explored bottom-up approaches that rely on grounding and retrieval augmentation, including instruction backtranslation (Li et al., 2024) and instruction back-and-forth translation (Nguyen et al., 2024), which generate high-quality synthetic data grounded in web corpora. These methods can potentially demonstrate superior performance compared to traditional translation-based approaches. However, there remains limited exploration of such approaches in multilingual contexts. In this work, our focus is on curating multilingual, multicultural synthetic data for SFT, although our framework can apply to other kinds of multilingual, multicultural synthetic data, including synthetic pre-training data and preference data as well. Below, we describe the key factors to consider when generating this type of data.

Base model capability: Synthetic data is usually generated by prompting large foundation models due to their fluency, ability to follow instructions, and safety guardrails. For generating multilingual, multicultural synthetic data, candidate base models should be evaluated on established multilingual and multicultural benchmarks to determine performance in target languages. If benchmarks for a specific target language are unavailable, evaluation results from related languages or overall multilingual performance can be used as indicators. Some models list supported languages and may include multilingual benchmark results in their release reports, while others do not provide these details. Additional factors to consider when selecting a base model include licensing, cost, and whether models are open-weights or restricted.

Seed data selection: Seed data is typically selected on the basis of factors such as diversity and target task coverage. Additional considerations for multilingual, multicultural seed data collection would include prioritizing tasks with coverage of cultural knowledge, norms, and values, and tasks with more relevance in specific regional and cultural contexts (for example, agriculture).

Data generation strategy: Several data generation strategies can be considered for synthetic data generation, including translation, back-translation, and retrieval-augmented synthetic data generation. Translating high-quality, curated English SFT datasets allows us to carry over critical skills the LLM needs to acquire into the multilingual training data. Back-translation using existing multilingual datasets can enhance diversity. Retrieval-augmented generation from diverse sources can potentially leverage native speaker-authored content to produce synthetic data that includes cultural knowledge and linguistic nuances.

Quality metrics: Synthetic data should be evaluated for correctness and quality before being used in downstream tasks. This may involve both automated assessment and human evaluation by native speakers. Essential quality metrics for multilingual and multicultural synthetic data include:

- *Language correctness*: Is the data in the right language, register, and dialect, or is there confusion with a similar language? (Marchisio et al., 2024)

- *Linguistic acceptability*: Does the data sound fluent and natural to a native speaker of the language? (Hada et al., 2024b;a)

- *Cultural appropriateness*: Does the data contain appropriate references to cultural artifacts and knowledge and accurately reflect values and norms?

- *Bias and safety*: Does the data amplify stereotypes, contain unsafe content, or culturally inappropriate outputs?

Downstream evaluation The primary objective of generating synthetic data is to improve model performance in real-world applications, making the choice of downstream tasks for evaluations critical. This is particularly challenging given the scarcity of multilingual and multicultural benchmarks. To address this, downstream benchmarks should cover all target languages and, wherever possible, avoid relying on translated datasets originally in English. They should also include diverse tasks, especially when synthetic data spans multiple domains. They should incorporate tests of cultural knowledge, norms, and values. In some cases, new benchmarks aligned with the target task, language, and cultural context may need to be developed to accurately assess the utility of synthetic data. Finally, care must be taken to avoid benchmark contamination in the base model, a common issue with multilingual benchmarks (Ahuja et al., 2024).

Native speaker consent, involvement and privacy: Multilingual and multicultural synthetic data creation processes should involve native speakers in seed data selection and evaluation when feasible. Informed consent is required to address cultural considerations and data sovereignty according to local regulations. Synthetic data must not contain personally identifiable information derived from seed data or online sources.

3 RELATED WORK

English Instruction Fine-tuning Datasets: Instruction Fine-tuning (IFT) (Ouyang et al., 2022) adapts pre-trained language models to follow human instructions through supervised learning on instruction-output pairs. English IFT datasets include the FLAN collection(Wei et al., 2022), which scaled instruction tuning to over 1,800 tasks and chain-of-thought prompting, and Self-Instruct (Wang et al., 2023), which demonstrated effectiveness of LLM-generated synthetic data, leading to Stanford Alpaca (Taori et al., 2023) and the enhanced Alpaca-GPT4 (Peng et al., 2023). Subsequent work expanded through conversational approaches like Vicuna (Chiang et al., 2023) using ShareGPT conversations, WizardLM's Xu et al. (2024) Evol-Instruct methodology, and LIMA's (Zhou et al., 2023a) demonstration that 1K curated prompts suffice for effective tuning. The ORCA series (Mukherjee et al., 2023; Mitra et al., 2023) introduced explanation tuning and Prompt Erasure techniques, culminating in ORCAAGENT-INSTRUCT (Mitra et al., 2024a) with 25.8M synthetic instruction-response pairs.

Multilingual Instruction Fine-tuning Datasets: Multilingual instruction-following capabilities have been developed through three primary methodologies. Translation-based methods create multilingual datasets by machine translation of existing English instruction data. BACTRIAN-X (Li et al., 2023) translates Alpaca (Taori et al., 2023) and Dolly (Conover et al., 2023) to produce 3.4M pairs across 52 languages, mAlpaca (Chen et al., 2024) demonstrates cost-effective multilingual tuning through Alpaca translation, and LIMA-X achieves 9.9% cross-lingual improvements using ∼1K instructions. Template-based approaches like M2Lingual leverage Evol-guided taxonomy (Xu et al., 2024) for synthetic generation across 70 languages. Hybrid strategies combine multiple techniques: AYA-COLLECTION (Singh et al., 2024b) integrates crowdsourcing across 65 languages with templating and translation, repurposing xP3 (Muennighoff et al., 2023), Flan (Longpre et al., 2023), and Dolly datasets while employing NLLB 3.3B (Team et al., 2022) for n-way translation. Similarly, INDICALIGN aggregates 74.7M prompt-response pairs across 20 Indian languages through aggregation of existing datasets, translations via the INDICTRANS2 (Gala et al., 2023) model, synthetic conversation generation from India-centric Wikipedia, and crowdsourcing.

Data Generation Strategies: Most prior efforts rely on **output distillation** from advanced LLMs like GPT-4, where models learn from teacher-generated responses. However, recent approaches have explored alternative strategies to address distillation limitations. **Instruction Backtranslation** (Li et al., 2024) starts with a model fine-tuned on seed data to generate instruction prompts for web documents, followed by self-curation to select high-quality examples. This leverages web content diversity while avoiding over-reliance on teacher models through iterative self-augmentation and self-curation. **Instruction Back-and-Forth Translation** (Nguyen et al., 2024) extends backtranslation by generating synthetic instructions and then rewriting responses using an LLM to improve quality based on initial documents. This combines web information diversity with high-quality response generation, outperforming distillation and yielding better win rates than ShareGPT and Alpaca-GPT4 (Peng et al., 2023).

Limitations of prior multilingual IFT Datasets: Existing multilingual instruction datasets suffer from critical limitations that hinder effective cross-lingual alignment. Translation-based approaches like Bactrian-X, mAlpaca, and LIMA-X focus primarily on basic instruction-following while excluding advanced reasoning capabilities and lack demographic or cultural representation for local alignment. These datasets suffer from translation inaccuracies due to context-length limitations of traditional sentence-level translation methods (NLLB 3.3B and INDICTRANS2) which fail to capture linguistic nuances and introduce errors that can potentially creep in during training. Despite impressive language coverage, AYA-COLLECTION provides limited culturally-specific content, while INDICALIGN sources a large proportion of data from WordNet (Miller, 1994) with a focus on question-answering offering limited task diversity. Most existing datasets are short-context and single-turn, failing to ad-

dress real-world usage patterns including long-context or multi-turn dialogue capabilities. This can significantly constrain model performance in practical applications that demand sustained, contextually-aware interactions.

4 DATA GENERATION

The UPDESH dataset comprises two complementary subsets addressing distinct aspects of multilingual instruction following: reasoning data and open-domain generative data. This design recognizes that reasoning capabilities are largely language- and culture-agnostic, making translation-based approaches suitable for tasks like mathematical problem-solving and logical inference (Shaham et al., 2024). Existing high-quality reasoning datasets such as ORCAAGENT-INSTRUCT and ORCAMATH, thus are valuable resources for multilingual adaptation. However, generative capabilities requiring cultural awareness, linguistic naturalness, and factual grounding in local contexts cannot be adequately addressed through translation due to inherent Western-centric biases and lack of cultural specificity in existing datasets (Yao et al., 2024). Therefore, our generative subset employs a grounded approach that ensures factuality through Wikipedia-based content, maintaining linguistic naturalness through native language generation, and preserves cultural adherence through systematic curation of India-specific cultural artifacts.

Reasoning Data: Inspired by prior work (Ahuja et al., 2025; Khan et al., 2024), we translate eight subsets of the ORCAAGENT-INSTRUCT (Mitra et al., 2024a) and ORCAMATH (Mitra et al., 2024b) datasets into 13 Indic languages. Specifically, we consider seven reasoning-related subsets from ORCAAGENT-INSTRUCT[1] along with the *Math* subset from ORCAMATH[2] (Table 1). Both datasets have been attributed to induce significant chain-of-thought and reasoning capabilities in models during instruction-tuning without the need for specific preference optimization. We employ LLAMA-3.1-405B-INSTRUCT for selective translation given its strong coverage in Indian languages and instruction-following capabilities that enable adaptation to various conversational styles (Sankar et al., 2025). Decoding is performed using nucleus sampling with $top\_p = 0.95$ and temperature set to 1.0. After generation, all outputs undergo a series of quality control filters, described in detail below.

Table 1: Reasoning task categories and their descriptions

Task Type	Description
ANALYTICAL REASONING	MCQ-style questions requiring step-by-step logical inference
MULTIPLE-CHOICE QUESTIONS	General-purpose problems across diverse knowledge domains
FERMI (GUESSTIMATION)	Open-ended estimation problems using logical assumptions
FEW-SHOT CHAIN-OF-THOUGHT	Tasks with 4-5 in-context examples for learning
BRAIN TEASERS	Puzzles stimulating lateral thinking and creativity
TEXT CLASSIFICATION	Categorization tasks for predefined labels
MATH	Grade-school arithmetic, algebra, and geometry word problems

Open-Domain Generative Data: Synthesizing generative data presents greater challenges than translating existing content, as it introduces increased risks of hallucinations, factual inaccuracies, and demographic misalignment. We conducted a comparative analysis between LLAMA-3.1-405B-INSTRUCT and QWEN3-235B-A22B models across both reasoning and non-reasoning paradigms to evaluate qualitative differences in generated outputs. Manual assessment revealed that the QWEN3-235B-A22B model demonstrated superior proficiency in generative tasks and excelled in complex instruction following owing to its strong reasoning traces. Therefore, we used the QWEN3-235B-A22B for our generative data curation process. Drawing inspiration from instruction backtranslation techniques (Li et al., 2024), our methodology begins with unlabelled content to construct questions, followed by LLM-generated answers. To ensure generation diversity and contextual grounding that maintains factual accuracy and demographic relevance, we leverage Wikipedia pages in respective target languages as our knowledge base. Table 2 summarizes the eight generative task categories. As demonstrated, some tasks required two phases, which indicates two LLM inferences in order to generate the synthetic data instance.

Further, to ensure cultural representation in synthetic data generation, we systematically curated culturally relevant content from Wikipedia using the MediaWiki API. Following (Yao et al., 2024)'s cultural taxonomy framework, we traversed Wikipedia category structures starting from `Category:Culture of India` and `Category:Culture of India by state or union territory`, exploring 2-3 levels deep. This yielded two complementary datasets: a general collection (54 categories, 19,143 artifacts) covering pan-Indian elements, and a state-wise collection (28 states + 8 union territories, 37,227 artifacts) with region-specific content spanning festivals, cuisine, traditional arts, architecture, and religious practices. We sampled 26.8K cultural artifacts from this collection to create multi-hop question-answer pairs for synthetic data generation.

Data Filtering After generating approximately 7.5 million data points across 13 languages for both the Reasoning and Open-Domain subsets, manual validation was not feasible, therefore, following the approach of Shen et al.

---

[1]https://huggingface.co/datasets/microsoft/orca-agentinstruct-1M-v1
[2]https://huggingface.co/datasets/microsoft/orca-math-word-problems-200k

Table 2: Generative task categories with synthesis methods, phases, and model configuration

Task Type	Synthesis Method	Phases	Qwen3-Mode
LOGICAL REASONING	Generate implicit inferences from text passages	(1) Direct inference generation	Reasoning
MULTI-HOP QA	Create questions requiring information synthesis across text segments	(1) Question generation (2) Answer generation	Reasoning
CREATIVE WRITING	Transform factual content into engaging narratives	(1) Generate creative piece (2) Generate eliciting prompt	Reasoning
MULTI-TURN DIALOGUE	Agentic workflows with 3-5 turn conversations between personas	(1) Generate dialog adhering to personas (2) Generate natural prompt	Non-reasoning
SUMMARIZATION	Generate summaries preserving key information across languages	(1) Direct summary generation	Non-reasoning
MACHINE TRANSLATION	Cross-lingual conversion maintaining cultural context	(1) Direct translation	Non-reasoning
CAUSAL REASONING	Identify and explain cause-effect relationships in text	(1) Direct causal analysis	Reasoning

Table 3: Document filtering statistics averaged over all Indic Languages

Reasoning Subset			Generative Subset		
Category	Total	Drop Rate (%)	Category	Total	Drop Rate (%)
analytical_reasoning	25,000	0.047	logical_reasoning	16,384	1.459
brain_teaser	50,000	0.043	multihop_QA	16,384	1.459
fermi	25,000	0.015	creative_writing	16,384	1.459
fs_cot_flow	25,000	3.769	Multi_turn_Dialog	16,384	1.611
math	200,035	0.035	summarization	16,384	1.526
mcq	99,986	0.135	machine_translation_en	16,384	0.641
rc	50,000	0.379	machine_translation_xx	16,384	17.047
text_classification	50,000	1.878	cultural_multihop_reasoning	26,834	0.347
			causal_reasoning	16,384	1.453

(2025), we employed automated quality checks but use the standard threshold-based method instead of their anomaly detection-based method. Specifically, we applied two filtering criteria: (1) Language Identification using IndicLID (Madhani et al., 2023) with a 0.75 confidence threshold, and (2) word repetition ratio capped at 0.75 to flag low-quality generations. Table 3 shows the filtering results, demonstrating high data quality with drop rates below 2% for most subsets. The main exception is the FS-CoT subset for Urdu, where the outputs showed excessive repetition leading to higher filtering rates, but we maintain these thresholds to ensure data integrity. For the English-to-XX translation tasks, Assamese had the highest drop rate as the model frequently generated Bengali text instead, likely due to the shared script and similarity between these languages, and because Assamese is a low-resource language.

5 DATA QUALITY ANALYSIS

5.1 TRANSLATION QUALITY ASSESSMENT FOR THE REASONING SUBSET

For the reasoning subset, we conducted large-scale selective translation utilizing the LLAMA-3.1-405B-INSTRUCT. Given that the input often consists of long-context and non-standard text, it is crucial to rigorously evaluate the quality of the generated translations. To this end, following established methodologies the generated output in the target language is translated back into English and compared with the original English source. Due to the substantial volume of data, we randomly sample 4,096 instances from each subset for every language to ensure a representative evaluation. For the backtranslation step, we employed the more efficient LLAMA-3.3-70B-INSTRUCT, which offers faster inference and serves as a conservative lower bound on translation quality. We assessed translation fidelity using standard metrics such as ChrF (Popović, 2015), implemented using the SACREBLEU library (Post, 2018) for consistency. Table 9 in the Appendix demonstrates consistently high backtranslation scores across all languages and subsets, indicating robust translation quality.

5.2 DATA QUALITY ASSESSMENT FOR GENERATIVE DATA

Although large language models (LLMs) have emerged as scalable evaluators following the popular LLM-as-a-judge paradigm, their reliability in culturally nuanced and low-resource language settings still remains limited (Watts et al., 2024). Our study combined LLM-based assessment with native-speaker human annotation to quantify data quality and establish inter-annotator agreement. We employed a stratified sampling strategy to select 100 data points per task category: CREATIVE WRITING, CULTURAL MULTI-HOP REASONING, MULTI-TURN DIALOGUE, and aggregated reasoning tasks (LOGICAL, CAUSAL, and MULTI-HOP REASONING) across five Indian languages (Assamese, Gujarati, Hindi, Malayalam and Punjabi) resulting in a total of 2000 data points. Our sampling methodology ensures representative coverage by partitioning responses into quintiles based on the output length and maintaining proportional sampling across buckets to preserve the original length distribution while capturing variance across responses. Human evaluations were conducted by native speakers through an external agency (Table 7). We employed GPT-4o as an automated evaluator using identical evaluation protocols to allow direct comparison. We devised the following metrics to capture multidimensional quality aspects:

- CREATIVE WRITING: Instruction adherence, linguistic fluency, narrative coherence
- REASONING TASKS: Answer adequacy, contextual adherence, instruction compliance, fluency and readability, problematic content detection, cultural relevance assessment (cultural multi-hop only)
- MULTI-TURN DIALOGUE: Persona consistency, topical coherence, linguistic plausibility, repetition detection, toxicity screening, instruction adherence
- SUMMARIZATION: Content coverage, factual accuracy, conciseness, coherence and logical flow, stylistic appropriateness
- TRANSLATION: Semantic correctness, fluency corectness, domain appropriateness, style and tone, completeness

Further details on the metrics and their definitions could be found in the Appendix in Table 8. All metrics utilize a three-point Likert scale (0-2) with detailed rubrics provided to both human and automated evaluators to ensure assessment consistency. The prompts used for the LLM evaluations could be found in the Appendix A.7. The guidelines given to the human annotators have been added in the supplementary material. Quantitative analysis of human evaluations reveals very high data quality: across 10,000 individual metric assessments, human evaluators assigned only 27 scores of zero (0.27%), indicating that the overwhelming majority of generated content meets acceptable quality thresholds. We plan to publicly release all evaluation data, including both human and GPT assessments, to ensure complete transparency and support future research on calibration of LLM-based evaluators.
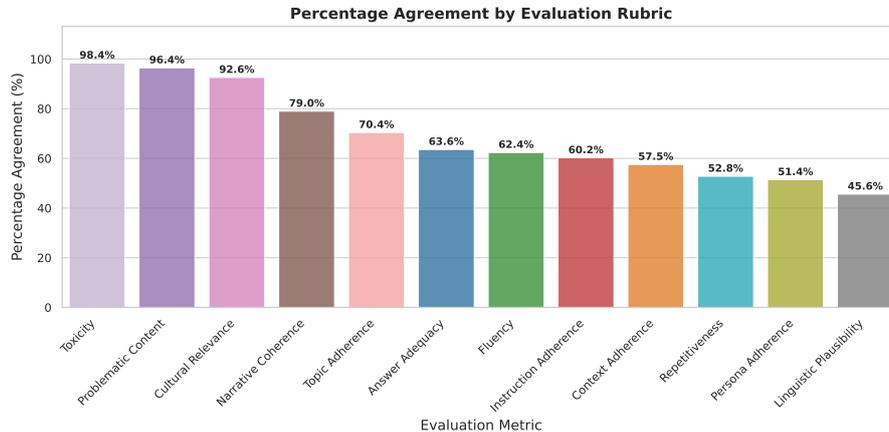

Figure 1: Human LLM-judge agreement across evaluation metrics, revealing differences across dimensions.

Inter-Annotator Agreement: To quantify the reliability of automated evaluation, we computed the percentage agreement between the LLM-judge and the human evaluators. We find significant variance in agreement across metrics (Figure 1). Agreement substantially deteriorates for culturally and linguistically nuanced assessments, particularly for linguistic plausibility evaluation and in repetition detection in long dialogue sequences. Conversely, well-defined objective criteria, such as toxicity detection and problematic content identification, demonstrate robust inter-annotator alignment. Since our data is generated using benign prompts and topics, toxic or problematic content is expected to be very rare or nonexistent. These results are consistent with earlier research on the challenges of subjective versus objective evaluation and reveal shortcomings in current LLM-judges for assessing culturally sensitive content (Watts et al., 2024). The distribution of scores across tasks and languages for both humans and LLM can be found in Appendix A.2.

6

Table 4: **NLU task performance.** All entries are *accuracy* (higher is better).

Model	Setting	MMLU	MMLU-I	MILU	ARC-I	BoolQ-I	TVQA-I	Bele	INCL	GMMLU
	Base	0.23	0.23	0.26	0.25	0.62	0.27	0.24	0.25	0.23
	Bactrian	**0.63**	**0.37**	**0.42**	0.26	0.65	**0.66**	0.41	**0.39**	**0.42**
Llama	Indic-Align	0.60	0.34	0.42	0.27	0.71	0.59	0.32	0.34	0.39
	Aya	0.59	0.33	0.39	0.28	0.76	0.51	0.47	0.34	0.37
	Updesh	0.51	0.34	0.37	**0.30**	**0.78**	0.46	**0.51**	0.31	0.37
	Base	0.78	0.43	0.52	0.30	0.80	0.68	0.55	0.45	0.55
	Bactrian	**0.78**	0.40	0.48	0.29	0.74	0.67	0.48	0.42	0.48
Phi4	Indic-Align	0.77	0.45	0.51	0.32	0.76	0.62	0.52	0.45	0.54
	Aya	0.78	0.45	0.53	**0.33**	0.73	**0.72**	0.61	**0.48**	0.53
	Updesh	0.78	**0.49**	**0.54**	0.32	**0.81**	0.60	**0.74**	0.47	**0.57**

6 DOWNSTREAM TASKS EVALUATION

We selected two base models, LLAMA-3.1-8B and PHI4-14B, for fine-tuning experiments. These models were chosen based on their size (for feasibility of experiments given available resources) and reported multilingual capabilities (Grattafiori et al., 2024; Abdin et al., 2024). We used the Axolotl framework[3] for all fine-tuning runs. Details regarding the hyperparameters and compute resources used can be found in Appendix A.4.

Baselines: We fine-tuned both LLAMA-3.1-8B and PHI4-14B on three high-quality, open-source IFT datasets that collectively cover nearly all the Indic languages included in our work: the AYA-COLLECTION (Singh et al., 2024b), INDICALIGN (Khan et al., 2024) and the BACTRIAN-X dataset (Li et al., 2023). To the best of our knowledge, these are the only open datasets that offer both broad language coverage and instruction-following format, and there is a lack of other large-scale open IFT datasets for Indian languages. The BACTRIAN-X dataset provides training samples for 10 out of our 13 target languages, excluding Assamese, Kannada and Punjabi. The AYA-COLLECTION dataset includes all target languages except Punjabi. Since AYA-COLLECTION contains millions of samples per language, we uniformly sub-sampled it to construct a balanced dataset of approximately 7 million samples which is comparable in scale to UPDESH. For INDICALIGN, the WordNet subset (∼97M pairs) is disproportionately large, less diverse, and redundant, so we downsampled it to one instance per entry, yielding about 7.3M training pairs while retaining all other subsets.

Downstream Tasks: Our evaluation framework consists of three task categories to comprehensively assess model capabilities. Natural language understanding (NLU) tasks use multiple-choice questions to measure comprehension and reasoning through likelihood-based scoring. Natural language generation (NLG) tasks like translation and summarization, assess models' ability to generate coherent and contextually appropriate outputs. For instruction-following evaluation, we translate established benchmarks (IFEval, IFBench) using GPT-4o (prompts in A.6), enabling assessment of both cross-lingual instruction-following (English responses to non-English instructions) and native language instruction-following. This design systematically identifies model strengths and weaknesses across diverse tasks, providing a holistic performance assessment. Dataset details are in Appendix A.5.

6.1 RESULTS

Performance on NLU Tasks: Table 4 shows the performance of both base models fine-tuned on all three IFT datasets, along with the performance of the base models on NLU tasks. PHI4-UPDESH stands out as the most consistently strong configuration, attaining the best overall scores on MMLU-I, MILU, BoolQ-I, BeleBele, INCL, and GlobalMMLU. For the rest of the tasks, different models excel - in case of the Llama base model, BACTRIAN-X leads knowledge-intensive tasks, attaining the highest on MMLU and TVQA-I, while UPDESH excels in instruction-style reasoning benchmarks (ARC-I, BoolQ-I) and performs best on reading comprehension tasks such as BeleBele. Phi4's MMLU performance appears saturated across all fine-tuning baselines, suggesting a ceiling effect from pre-training knowledge or lack of coverage within all the training mixtures. **Overall, these results demonstrate that different IFT datasets offer distinct advantages across various NLU capabilities.**

Performance on NLG Tasks: Table 5 demonstrates that UPDESH is clearly superior to the other datasets on NLG tasks. LLAMA-UPDESH emerges as the best configuration, achieving the highest performance across all generative tasks with substantial margins over all competing datasets as well as the base model. In case of Phi4, UPDESH delivers the strongest translation performance, building effectively on the already capable base model. However, summarization results are more nuanced: the base model leads on XSum with UPDESH being competitive, while the other fine-tuned models perform poorly. On long-context conversational translation (IN22-Conv-Doc) outperforms all fine-tuned variants, though it slightly lags behind the Phi4 Base model. **Overall, we observe that the long-context generative training data in UPDESH provides significant benefits for generative tasks.**

---

[3] https://github.com/axolotl-ai-cloud/axolotl

Table 5: **NLG task performance.** All entries are *ChrF* scores (higher is better).

Model	Setting	XSum	Flores En-XX	Flores XX-En	IN22-Conv-Doc
Llama	Base	0.16	1.00	36.91	0.45
	Bactrian	0.21	23.69	46.31	14.09
	Indic-Align	9.26	25.18	2.29	10.98
	Aya	0.24	22.60	0.53	6.73
	Updesh	**17.39**	**32.63**	**46.84**	**21.25**
Phi4	Base	**11.32**	23.03	**51.39**	**22.80**
	Bactrian	0.31	25.93	47.75	15.45
	Indic-Align	0.27	27.48	1.77	13.82
	Aya	0.37	26.51	7.63	14.01
	Updesh	10.21	**34.58**	**53.29**	18.74

Table 6: **Instruction-following capabilities.** All entries are *accuracy* (higher is better).

Model	Setting	mIFEval				mIFBench			
		Native		in English		Native		in English	
		Prompt	Inst.	Prompt	Inst.	Prompt	Inst.	Prompt	Inst.
Llama	Base	**0.24**	**0.36**	**0.18**	**0.31**	**0.19**	**0.21**	0.16	0.17
	Bactrian	0.17	0.27	0.14	0.25	0.17	0.18	0.11	0.13
	Indic-Align	0.16	0.24	0.14	0.23	0.17	0.19	0.16	0.17
	Aya	0.13	0.22	0.11	0.21	0.18	0.19	**0.18**	**0.19**
	Updesh	0.20	0.30	**0.18**	0.28	0.16	0.19	0.14	0.15
Phi4	Base	**0.22**	**0.32**	**0.19**	**0.30**	0.14	0.16	0.13	0.14
	Bactrian	**0.22**	**0.32**	**0.19**	**0.30**	**0.20**	0.21	0.15	0.16
	Indic-Align	0.16	0.25	0.15	0.24	0.15	0.17	0.14	0.16
	Aya	0.15	0.25	0.14	0.24	**0.20**	0.21	**0.19**	**0.20**
	Updesh	**0.22**	**0.32**	**0.19**	0.29	0.19	**0.22**	0.16	0.18

Performance on Instruction Following Tasks: Table 6 presents results on the GPT-4o translated multilingual versions of IFEval and IFBench, which evaluate fine-grained constraint adherence through highly specific requirements like character positioning, word counts, and formatting specifications - constraints notably absent in any training datasets. NATIVE refers to prompts as well as expected output in regional language, while IN ENGLISH uses regional language prompts but expects English responses. Prompt-level evaluation measures whether all instructions in a prompt are followed, while instruction-level evaluation measures individual instruction adherence. UPDESH emphasizes general instruction following involving task completion and output formatting, making these benchmarks an effective OOD evaluation for precise constraint adherence. For Llama, fine-tuning generally degrades performance, but UPDESH shows smaller drops (mIFEval: 3-4 point lesser degradation than next best - Bactrian; mIFBench: comparable or better performance than other baselines, lags Aya on IN ENGLISH tasks). Phi4 exhibits different dynamics - UPDESH achieves positive gains on mIFBench over the Base model (+4/+2 points Native; +1/+3 points English) while maintaining base-level performance on mIFEval, outperforming other fine-tuning approaches (Aya, Indic-Align) that show mixed or negative changes. **Updesh provides robust instruction-following with minimal catastrophic forgetting compared to other baselines**.

Language-wise trends: Figure 2 shows a clear trend: performance improves as language resources increase for both models and task types (NLU accuracy and NLG ChrF). We illustrate this with the resource taxonomy Joshi et al. (2020), ordered from 0 (lowest-resource) to 5 (high resource). The relative improvements with UPDESH are most pronounced in low/mid-resource bands, effectively narrowing the gap to high-resource languages. The base model (without fine-tuning) is the weakest, with BACTRIAN-X and INDICALIGN often competitive in the medium resource languages while AYA-COLLECTION is more variable. Notably, generation benefits more from data quality: NLG improvements increase more dramatically as languages become lower-resourced, most clearly for PHI4-UPDESH. **Across language-resource tiers, UPDESH is consistently strongest, especially at higher resource levels, reaching about ∼0.65 for Llama (NLU), ∼0.76 for Phi-4 (NLU) in terms of average accuracy, and the highest ChrF for Phi-4 (NLG).**

6.2 DISCUSSION

Quality evaluation: Our comprehensive quality evaluation ranges from basic filters to nuanced human evaluation. Experiments with LLM-based evaluators show that while LLM-judges do well on some metrics, they struggle with subtle multilingual and multicultural contexts. For example, they often misjudge how a language is typically
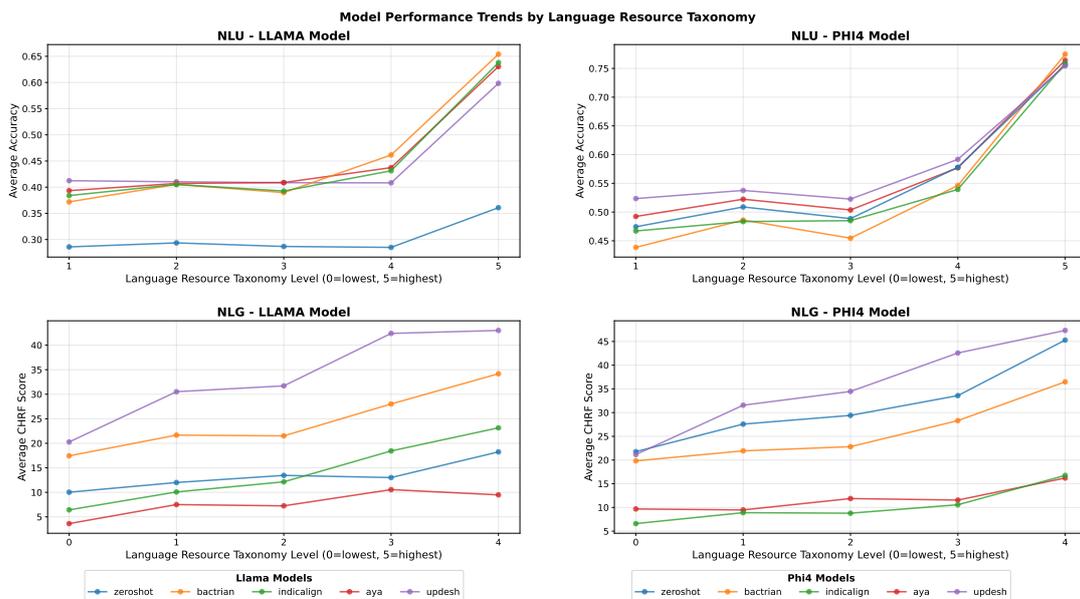
Figure 2: NLU and NLG task performance grouped by language–resource class taxonomy from Joshi et al. (2020). UPDESH yields the largest relative gains in low/mid-resource languages

spoken or written, and tend to lost track of persona consistency in longer dialogues. These issues motivate the need for scalable and reliable evaluation methods that go beyond surface-level checks.

Task-type sensitivity (NLU vs.NLG): We observe variation by task type - for instance, on NLU-style tasks (usually multiple-choice), no configuration is a clear winner across all datasets; performance depends on the base model and the specific benchmark. In contrast, on NLG-style tasks that score free-form generations, models trained on UPDESH consistently perform strongly. This pattern suggests that the quality control and discourse control learned during training is transferred more directly to NLG than to choice-based NLU.

Evaluation protocol differences matter: Most NLU evaluations compute correctness via likelihood comparisons over *(context, option)* pairs, effectively testing discrimination among short alternatives. NLG evaluations, by design, require longer generations and measure faithfulness, adequacy, and fluency under open-ended prompting. The divergence in scoring objectives partially explains why a dataset that emphasizes long-form instruction-following (like UPDESH) yields larger gains on NLG than on likelihood-based NLU. We find that the **Bactrian** dataset tends to perform best on NLU-style benchmarks, which can be attributed to format alignment and a larger proportion of MCQ or short-from prompts.

Distribution mismatches between train and test: UPDESH contains longer-context samples and lacks short-text MCQ-style items; its task composition emphasizes multi-step, generative instruction-following. Even the MCQ subset which is part of the reasoning data in UPDESH consists of long-form answers to MCQs, with close to no representation of short-text or direct answers. In contrast, many NLU benchmarks, however, are MCQ-heavy and short-context. This *composition shift* (task format, length, and response style) likely reduces the relative NLU gains for UPDESH-trained models while amplifying NLG gains, showing that models generalize best to downstream tasks with evaluation formats that resemble their training distribution.

7 CONCLUSION

In this work, we examined synthetic data as a potential remedy for the scarcity of multilingual and multicultural resources. Through a comprehensive framework and systematic experiments across Indian languages, we identified effective strategies for data generation, quality assessment, and downstream evaluation, beyond English-centric norms. We built UPDESH, a 9.5M 13-language IFT dataset using a culturally grounded, bottom-up pipeline. Our comprehensive evaluation spanning data generation, quality assessment (human and LLM-as-judge), and downstream tasks, revealed that synthetic data can potentially bridge resource gaps. However, no single recipe dominates across tasks and models, motivating approaches tailored to diverse linguistic and cultural contexts. We will release the UPDESH dataset, evaluation protocols, and detailed analyses to enable transparent, reproducible progress.

8 REPRODUCIBILITY STATEMENT

To ensure full reproducibility of our pipeline—spanning data generation, model fine-tuning, downstream evaluation, and synthetic data quality assessment — we provide our complete codebase and sample generated data in the supplementary material. The submission zip contains organized data and code folders, with the code structured into seven components corresponding to distinct pipeline stages:

1. `selective_translation` – Scalable codebase for generating reasoning data through translation from ORCAAGENT-INSTRUCT.
2. `wiki_bt` – Wikipedia-based grounded synthetic data curation pipeline.
3. `DCAD-2000` – Modified repository for heuristic-based filtering of standard and multi-turn conversational data.
4. `quality_evals` – Synthetic data quality evaluation code, including outputs, human annotations, and analysis.
5. `model_training` – Axolotl-based fine-tuning scripts for the generated datasets.
6. `lm_evaluation_harness` – Custom fork integrating additional benchmarks for downstream evaluation.
7. `plotting` – Analysis and visualization code for experimental results.

Comprehensive instructions are provided in the main README and component-specific README files for straightforward reproduction. Model training configurations with exact hyperparameters are detailed in Appendix 10.

We will publicly release all the following artifacts (**code and datasets**) under a permissive license.

- **Complete codebase** including selective translation pipelines, Wikipedia-grounded synthetic data generation, data filtering methods, model training scripts, and the up-to-date evaluation frameworks
- **Training datasets** comprising Updesh Reasoning and Generative subsets across multiple languages.
- **Evaluation datasets** featuring GPT-4o translated variants of standard benchmarks (IFEval, IFBench) for reproducibility and future benchmarking.
- **Raw evaluation scores** for all models across every dataset, providing complete experimental transparency
- **Human-annotations of synthetic data** specifically useful for calibrating / meta-evaluation of LLM evaluators on Indian languages

9 ETHICAL CONSIDERATIONS

Our discussion of ethical considerations is guided by the framework proposed by Bender & Friedman (2018).

INSTITUTIONAL PROCESS AND OVERSIGHT

The data annotation was conducted by a third-party vendor and was approved by the Institutional Review Board of our organization and by the vendor.

DATA PROVENANCE AND QUALITY ASSURANCE

To mitigate potential artifacts and quality issues in the synthetic data, we implemented a rigorous quality control process. This process involved both automated evaluation with GPT-4o and manual verification by human annotators. We observed a high concordance between automated and human judgments on potentially problematic content. Furthermore, given that the data was generated using state-of-the-art large language models, the baseline incidence of such content was already substantially reduced. In a manual evaluation of 500 samples, human annotators flagged only 1 sample (0.2%) on metrics pertaining to problematic content, confirming the high quality of the resulting dataset.

ANNOTATOR DEMOGRAPHICS

Annotators were recruited through a professional external services company. All annotators assigned to a given data point were native speakers of the language represented in the data. Table 7 summarizes the annotator demographics (education, region, age distribution, and gender). Each data worker was compensated at the rate of $2 per data point.

Category	Summary
Participants	15
Qualification	Post-graduation: 7 Graduation: 8
Geography	Spread across 8 Indian states
Age distribution	21–30: 7 31–40: 5 41–50: 3
Gender	Female: 11 Male: 4

Table 7: Participant demographics summary

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Sanchit Ahuja, Varun Gumma, and Sunayana Sitaram. Contamination report for multilingual benchmarks, 2024. URL https://arxiv.org/abs/2410.16186.

Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajbhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Hassan Awadallah, Monojit Choudhury, Vishrav Chaudhary, and Sunayana Sitaram. sPhinX: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. In Ofir Arviv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky, and Oyvind Tafjord (eds.), *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 927–946, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9. URL https://aclanthology.org/2025.gem-1.73/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041/.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. Inpars: Data augmentation for information retrieval using large language models, 2022. URL https://arxiv.org/abs/2202.05144.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1347–1356, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.90/.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Pranjal A. Chitale, Bishal Santra, Yashoteja Prabhu, and Amit Sharma. Evaluating the effectiveness and scalability of llm-based data augmentation for retrieval, 2025. URL https://arxiv.org/abs/2509.16442.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gmL46YMpu2J.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. Pretraining language models using translationese. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5843–5862, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.334. URL https://aclanthology.org/2024.emnlp-main.334/.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=vfT4YuzAYA.

Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D Manning. Synthetic data generation and multi-step reinforcement learning for reasoning and tool use. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=oN9STRYQVa`.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL `https://aclanthology.org/2022.tacl-1.30/`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul,

Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Varun Gumma, Pranjal A Chitale, and Kalika Bali. Towards inducing long-context abilities in multilingual neural machine translation models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7158–7170, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.366. URL https://aclanthology.org/2025.naacl-long.366/.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. METAL: Towards multilingual meta-evaluation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2280–2298, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.148. URL https://aclanthology.org/2024.findings-naacl.148/.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1051–1070, St. Julian's, Malta, March 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.71/.

Chetan Harsha, Karmvir Singh Phogat, Sridhar Dasaratha, Sai Akhil Puranam, and Shashishekar Ramakrishna. Synthetic data generation using large language models for financial question answering. In Chung-Chi Chen, Antonio Moreno-Sandoval, Jimin Huang, Qianqian Xie, Sophia Ananiadou, and Hsin-Hsi Chen (eds.), *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pp. 76–95, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.finnlp-1.7/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Songbo Hu, Ivan Vulić, and Anna Korhonen. Quantifying language disparities in multilingual large language models, 2025. URL https://arxiv.org/abs/2508.17162.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault

(eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL `https://aclanthology.org/2020.acl-main.560/`.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15831–15879, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.843. URL `https://aclanthology.org/2024.acl-long.843/`.

Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Tom Kocmi. Déjà vu: Multilingual LLM evaluation through the lens of machine translation evaluation. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=yxzVanFoij`.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation, 2023. URL `https://arxiv.org/abs/2305.15011`.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. Self-alignment with instruction backtranslation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=1oijHJBRsT`.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. Bhasa-Abhijnaanam: Native-script and romanized language identification for 22 Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 816–826, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.71. URL `https://aclanthology.org/2023.acl-short.71/`.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6653–6677, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.380. URL `https://aclanthology.org/2024.emnlp-main.380/`.

George A. Miller. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL `https://aclanthology.org/H94-1111/`.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. Orca 2: Teaching small language models how to reason, 2023. URL `https://arxiv.org/abs/2311.11045`.

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei ge Chen, Olga Vrousgos, Corby Rosset, Fillipe Silva, Hamed Khanpour, Yash Lara, and Ahmed Awadallah. Agentinstruct: Toward generative teaching with agentic flows, 2024a. URL `https://arxiv.org/abs/2407.03502`.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math, 2024b. URL `https://arxiv.org/abs/2402.14830`.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL `https://aclanthology.org/2023.acl-long.891/`.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4, 2023. URL `https://arxiv.org/abs/2306.02707`.

Thao Nguyen, Jeffrey Li, Sewoong Oh, Ludwig Schmidt, Jason E Weston, Luke Zettlemoyer, and Xian Li. Better alignment with instruction back-and-forth translation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13289–13308, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024. findings-emnlp.777. URL https://aclanthology.org/2024.findings-emnlp.777/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025. URL https://arxiv.org/abs/2507.02833.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia soltani moakhar, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k3gCieTXeY.

Ashwin Sankar, Sparsh Jain, Nikhil Narasimhan, Devilal Choudhary, Dhairya Suman, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. Towards building large scale datasets and state-of-the-art automatic speech translation systems for 14 Indian languages. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32945–32966, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1582. URL https://aclanthology.org/2025.acl-long.1582/.

SarvamAI. indic-evals - sarvamai. https://huggingface.co/collections/sarvamai/indic-evals-67196d8d0edc751606d8b2e9. [Accessed 22-09-2025].

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2304–2317, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.136. URL https://aclanthology.org/2024.findings-acl.136/.

Yunfan Shao, Linyang Li, Yichuan Ma, Peiji Li, Demin Song, Qinyuan Cheng, Shimin Li, Xiaonan Li, Pengyu Wang, Qipeng Guo, Hang Yan, Xipeng Qiu, Xuanjing Huang, and Dahua Lin. Case2Code: Scalable synthetic

data for code generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11056–11069, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.coling-main.733/`.

Yingli Shen, Wen Lai, Shuo Wang, Xueren Zhang, Kangyang Luo, Alexander Fraser, and Maosong Sun. Dcad-2000: A multilingual dataset across 2000+ languages with data cleaning as anomaly detection, 2025. URL `https://arxiv.org/abs/2502.11546`.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11047–11073, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.595. URL `https://aclanthology.org/2024.acl-long.595/`.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL `https://aclanthology.org/2024.acl-long.620/`.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.919. URL `https://aclanthology.org/2025.acl-long.919/`.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL `https://arxiv.org/abs/2207.04672`.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2203–2213, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.188. URL `https://aclanthology.org/2021.eacl-main.188/`.

Vijay Viswanathan, Xiang Yue, Alisa Liu, Yizhong Wang, and Graham Neubig. Synthetic data in the era of large language models. In Yuki Arase, David Jurgens, and Fei Xia (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pp. 11–12, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-255-8. doi: 10.18653/v1/2025.acl-tutorials.7. URL `https://aclanthology.org/2025.acl-tutorials.7/`.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023. Association for

Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7900–7932, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.451. URL https://aclanthology.org/2024.emnlp-main.451/.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with OSS-instruct. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52632–52657. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wei24h.html.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. The bitter lesson learned from 2,000+ multilingual benchmarks, 2025. URL https://arxiv.org/abs/2504.15521.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CfXh93NDgH.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. Benchmarking machine translation with cultural awareness. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13078–13096, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.765. URL https://aclanthology.org/2024.findings-emnlp.765/.

Mike Zhang and Antonio Toral. The effect of translationese in machine translation test sets. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 73–81, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5208. URL https://aclanthology.org/W19-5208/.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=KBMOKmX2he.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023b. URL https://arxiv.org/abs/2311.07911.

A APPENDIX

A.1 RUBRICS USED FOR THE QUALITY EVALUATION OF THE SYNTHETIC DATA

Table 8: Evaluation Rubrics for AI Tasks

Task	Rubric	Description
Creative Writing	Instruction adherence	Assesses if the output strictly follows all constraints and guidelines provided in the prompt.
	Fluency	Evaluates the naturalness, grammatical correctness, and readability of the generated text.
	Narrative coherence	Checks for logical consistency in the plot, character development, and thematic elements.
Reasoning Tasks	Answer adequacy	Determines if the final answer is correct, complete, and directly addresses the core question.
	Context adherence	Measures whether the reasoning remains faithful to the provided context, avoiding external facts.
	Instruction adherence	Verifies that the output's structure, format, and steps match the user's instructions.
	Fluency and readability	Assesses the clarity, logical flow, and ease of understanding of the explanation.
	Problematic content and cultural relevance	Scrutinizes the response for harmful stereotypes and ensures it is culturally sensitive and appropriate.
Multi-turn Dialog	Persona Adherence	Evaluates the model's ability to consistently maintain a specific character or role throughout the conversation.
	Topic adherence	Checks if the conversation remains focused on the established topic or transitions logically.
	Linguistic plausibility	Assesses whether the dialogue sounds natural, human-like, and contextually appropriate.
	Repetitiveness	Measures the degree to which the model avoids unnecessarily repeating phrases or ideas.
	Toxicity check	Ensures the response is free from any offensive, harmful, or inappropriate content.
	Instruction adherence	Verifies that the model follows meta-instructions given by the user during the dialogue.
Summarization	Coverage	Determines if the summary successfully captures all essential points from the source text.
	Factual accuracy	Checks that the summary correctly represents the information and facts from the original document.
	Conciseness	Evaluates whether the summary is significantly shorter than the source while retaining critical information.
	Coherence and logical flow	Assesses if the summary is well-structured, logically organized, and easy to follow.
	Style and tone	Measures how well the summary reflects the style and tone of the original text.
Translation	Semantic correctness	Assesses whether the meaning, intent, and nuance of the source text are accurately conveyed.
	Fluency correctness	Evaluates the grammatical accuracy and naturalness of the translated text in the target language.
	Domain appropriateness	Checks if the terminology is correct and suitable for the specific subject matter (e.g., legal, medical).
	Style and tone	Determines if the translation successfully captures the original author's writing style and emotional tone.
	Completeness	Verifies that the entire source text has been translated without any omissions or additions.

A.2 RESULTS FROM THE QUALITY EVALUATIONS OF THE SYNTHETIC DATA

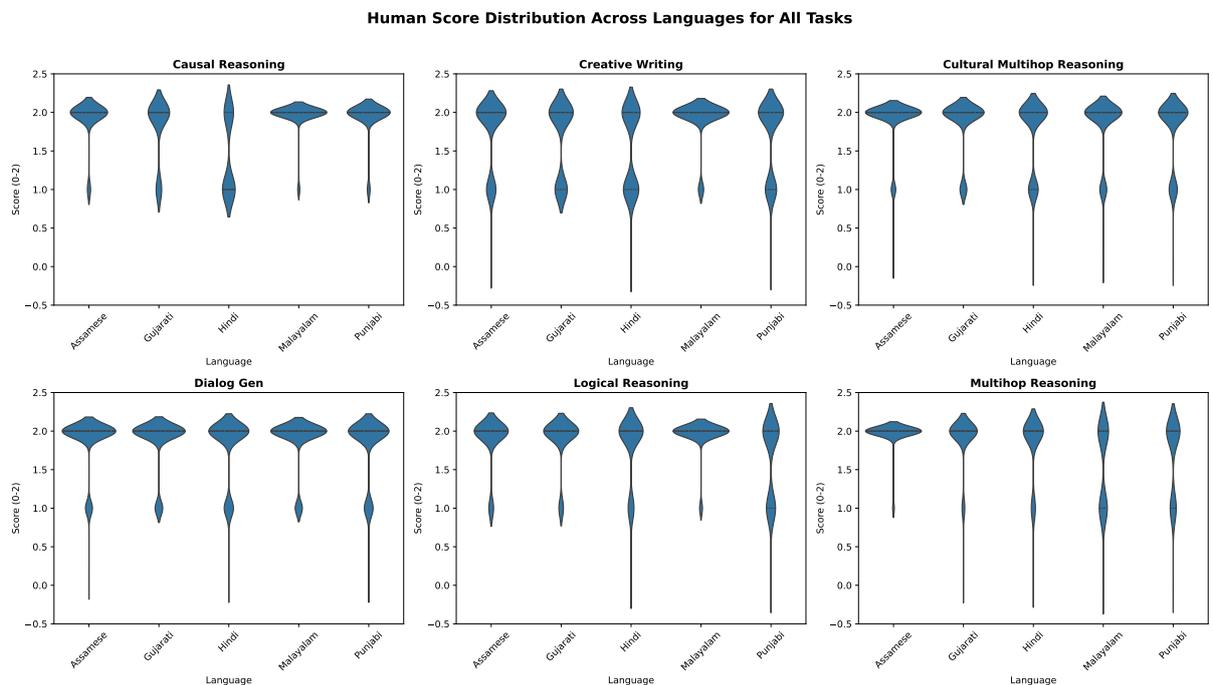Figure 3: LLM evaluations across 5 synthetically generated tasks

Figure 4: Expert human evaluations across 5 synthetically generated tasks

Figures 3 and 4 show the distribution of the scores received for the tasks we evaluated across the languages for the tasks. Expert human evaluators have consistently given a score of 2 across languages and tasks, indicating the high quality of Updesh. The disagreements between humans and the LLM is however unclear from these plots. We thereby, performed a thorough inter annotator analysis, the details of which are present in the next figures.
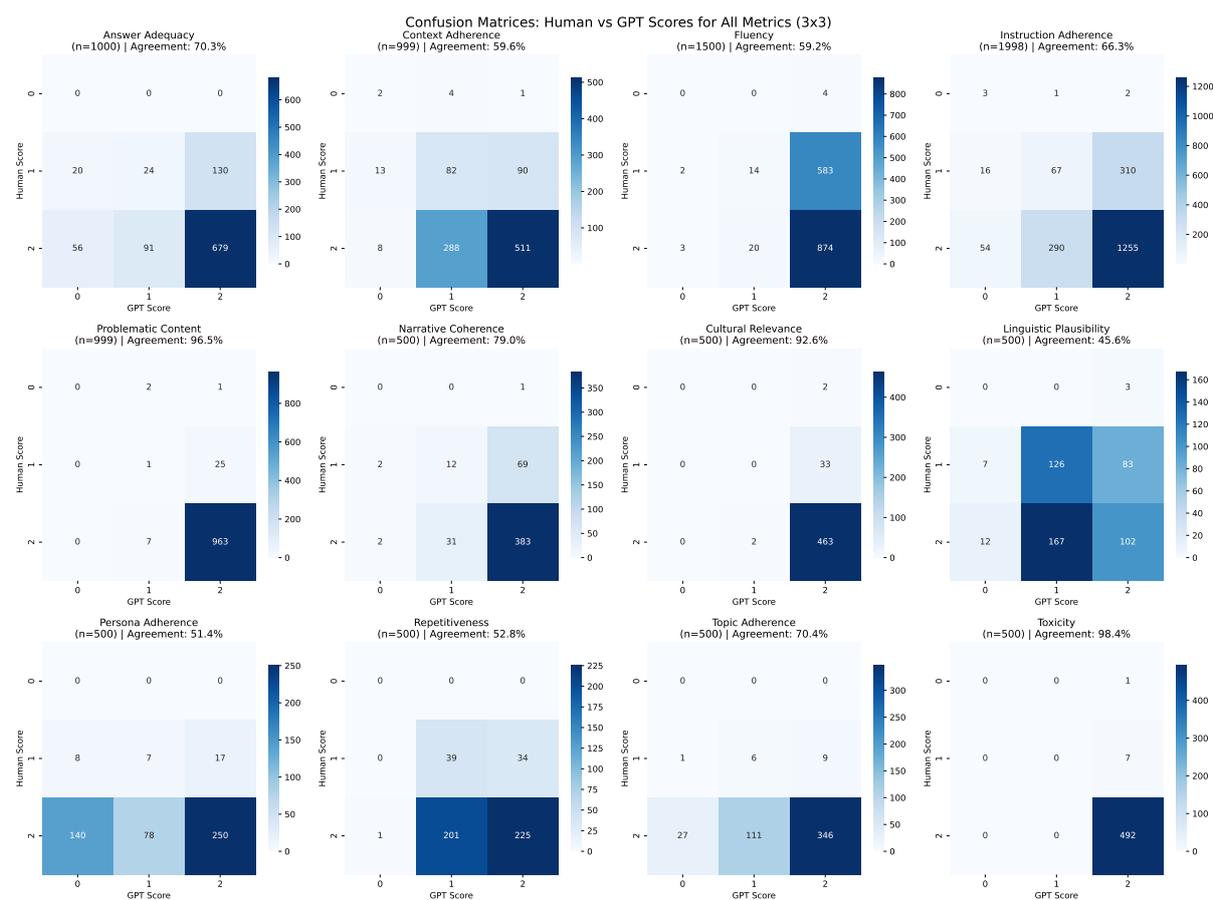
20

Figure 5: Confusion matrices showing agreement between human and LLM evaluators
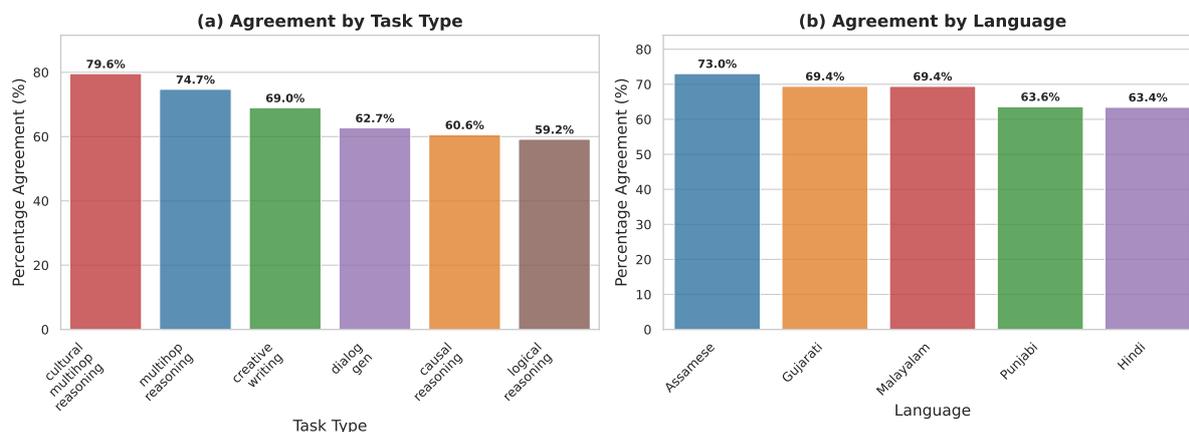

Figure 6: Agreement between human and LLM evaluators per task and language respectively

Continuing the claims made from Figure 1, Figure 5 provides a clearer view of the tasks on which humans and LLMs are likely to agree or disagree. We find that the largest disagreements occur in tasks such as assessing the linguistic plausibility of a given text in a regional language. Furthermore, LLMs struggle with evaluating long-context tasks, such as evaluating whether the same persona is maintained throughout a multi-turn conversation in a regional language. There is also a notable divergence between what human evaluators consider fluent in a relatively low-resource language and what an LLM deems fluent. In contrast, we observe considerable agreement in tasks like toxicity detection and problematic content flagging. LLMs also perform reasonably well at identifying whether a text is culturally relevant, but issues arise when the evaluation requires more fine-grained judgments of multilinguality and multiculturalism. We do not identify any specific trends language or task-wise as apparent in Figure6.

A.3 DATA QUALITY ASSESSMENT OF REASONING DATA

Language	analytical	brain	fermi	fs (cot_flow)	math	mcq	rc	text_class
Assamese	75.02	71.62	79.93	96.59	79.87	64.33	65.07	71.29
Bengali	87.25	77.40	80.69	82.10	79.87	67.25	74.94	74.72
Gujarati	77.86	67.15	82.14	73.96	49.66	78.04	63.21	55.95
Hindi	84.49	79.23	81.54	87.11	64.25	74.80	73.71	67.00
Kannada	79.96	76.87	80.02	81.26	65.80	69.91	64.15	60.77
Malayalam	75.21	73.41	70.10	77.93	68.51	55.69	63.19	75.63
Marathi	77.84	68.63	69.33	82.81	68.56	64.48	56.49	60.58
Nepali	81.79	86.18	74.71	83.32	53.98	56.42	53.86	59.63
Odia	56.47	62.06	50.70	93.61	57.55	51.21	42.95	52.15
Punjabi	83.75	51.79	77.40	79.04	61.19	51.21	69.81	54.11
Tamil	79.28	70.70	74.54	75.16	57.83	60.66	63.94	49.65
Telugu	78.24	80.26	74.33	79.91	69.88	60.66	61.93	60.95
Urdu	85.05	79.97	66.31	81.93	59.80	61.76	64.91	67.23

Table 9: Backtranslation ChrF scores for the Reasoning subset.

A.4 HYPERPARAMETERS

Table 10: Training hyperparameters used for all our experiments

Hyperparameter	Value
Base Model	phi4-base-hf / llama-3.1-8b-hf
Sequence Length	65,536
Effective Batch Size	8192
Number of Epochs	3
Optimizer	AdamW Torch Fused
Learning Rate	1.0e-05
LR Scheduler	Cosine
Adam Beta1	0.9
Adam Beta2	0.95
Max Grad Norm	1.0
Warmup Ratio	0.03
Weight Decay	0.1
NEFTune Noise Alpha	5
Precision	BF16
Flash Attention	True
Gradient Checkpointing	True

A.5 EVALUATION DATASETS

Table 11: Evaluation datasets

	Dataset	Source
NLU	Massive Multitask Language Understanding (MMLU)	Hendrycks et al. (2021)
	MMLU Indic (MMLU-I)	SarvamAI
	ARC Indic (ARC-I)	SarvamAI
	BoolQ Indic (BoolQ-I)	SarvamAI
	TriviaQA Indic (TVQA-I)	SarvamAI
	BeleBele (Bele)	Bandarkar et al. (2024)
	INCLUDE (INCL)	Romanou et al. (2025)
	Global MMLU (GMMLU)	Singh et al. (2025)
NLG	Extreme Summarization (Xsum)	Singh et al. (2024a)
	Flores English to Other (Flores EnXX)	Goyal et al. (2022)
	Flores XX-En (Flores XXEn)	
	IN22-Conv (in22-conv-doc)	Gala et al. (2023); Gumma et al. (2025)
IF	Multilingual Instruction Following Evaluation (mIFEval)	Zhou et al. (2023b)
	Multilingual Instruction Following Benchmark (mIFBench)	Pyatkin et al. (2025)

A.6 PROMPT USED FOR CONVERSION OF IFEVAL AND IFBENCH

Translation Generation Prompt

INSTRUCTION:
You are a **professional translator**. Your sole task is to translate the provided English text into {target_language}.
Rules:

- Output only the translation—nothing else (no explanations, no commentary).
- Do *not* answer or solve any task contained in the text; just translate it.
- Preserve formatting, punctuation, numbers, placeholders, and bracketed tokens exactly as they appear.
- Translate idiomatically and naturally for a fluent native speaker of {target_language}.
- If the input is an instruction (e.g., "Write a resume for ..."), translate the instruction itself—do *not* perform it.
- Do not introduce extra formatting (e.g., don't create JSON blocks unless the source already has them).

SOURCE LANGUAGE: English

TARGET LANGUAGE:
{target_language}

TEXT TO TRANSLATE:
{source_text}

OUTPUT REQUIREMENT:
Return only the translated text in {target_language} with formatting faithfully preserved.

A.7 SAMPLE PROMPTS USED FOR EVALUATION

The following are representative prompts (one per task) used to evaluate different task types. Each prompt follows the established rubrics from Table 8.

A.7.1 CREATIVE WRITING - INSTRUCTION ADHERENCE

Creative Writing Instruction Adherence Evaluation Prompt

INSTRUCTION:
You are an expert literary critic and evaluator, tasked with assessing the degree to which a synthetically generated creative piece adheres to the user's writing prompt. You are required to read the USER QUESTION thoroughly and analyze how well the generated response incorporates the specified narrative elements, stylistic choices, and constraints.

USER QUESTION:
{user_prompt}

TARGET LANGUAGE:
{tgt_lang}

ASSISTANT GENERATED RESPONSE:
{assistant_output}

A response that adheres to the user's creative brief has:

- **Narrative & Thematic Completeness**: It fully incorporates all requested characters, plot points, settings, and themes. The creative piece feels complete and resolves according to the prompt's guidelines.
- **Stylistic & Tonal Adherence**: The content's tone, mood, and writing style (e.g., genre conventions, a specific author's voice) directly match the user's request.
- **Format & Constraint Compliance**: It follows all explicit formatting requirements (e.g., poem, script, short story) and abides by any constraints (e.g., word count, inclusion/exclusion of specific words, use of certain literary devices).
- **Creative Intent Alignment**: It successfully captures the spirit and intended artistic goal of the user's prompt, creating a piece that feels like a faithful realization of the user's idea.

Use the following scoring scale:

- **5 – Excellent**: The response masterfully incorporates all creative constraints, including plot, character, tone, style, and format. It not only follows the instructions to the letter but also demonstrates a creative flair that enhances the user's original idea.

23

- **4 – Good**: The response successfully incorporates most creative instructions. There may be minor deviations in tone or style, or a secondary plot/character element might be slightly underdeveloped, but the core creative vision is clearly and effectively realized.

- **3 – Fair**: The response addresses some of the key creative instructions but neglects or misinterprets others. For instance, it might follow the plot but fail to capture the requested tone, or it might ignore a crucial character trait or constraint.

- **2 – Poor**: The response shows significant deviation from the creative brief. It may follow a single, simple instruction (like the general topic) but disregards crucial constraints like genre, character personality, plot structure, or mood.

- **1 – Unacceptable**: The response completely disregards the creative instructions. The generated text is thematically, structurally, and stylistically unrelated to the user's prompt.

Return your evaluation in the following JSON format:

```
{
  "score": <integer from 1 to 5>,
  "reason": "<brief explanation for the score>"
}
```

Do not include markdown, comments, or anything outside the JSON.

A.7.2 MULTI-TURN DIALOG - PERSONA ADHERENCE

Multi-turn Dialog Evaluation Prompt

INSTRUCTION:
You are an expert evaluator tasked with assessing how well a multi-turn dialog maintains persona consistency. You must analyze the conversation to determine if the assistant consistently embodies the specified character or role throughout the interaction.

USER QUESTION:
{user_prompt}

TARGET LANGUAGE:
{tgt_lang}

ASSISTANT GENERATED RESPONSE:
{assistant_output}

A response that demonstrates strong persona adherence has:

- **Character Consistency**: The assistant maintains the same personality traits, speaking style, and behavioral patterns throughout the conversation.
- **Role-Appropriate Knowledge**: The responses reflect knowledge and expertise appropriate to the specified persona.
- **Consistent Voice**: The tone, vocabulary, and manner of speaking remain true to the character across all turns.
- **Believable Interactions**: The persona feels authentic and natural in the conversational context.

Use the following scoring scale:

- **5 – Excellent**: Perfect persona consistency with natural, believable character embodiment throughout all turns.
- **4 – Good**: Strong persona adherence with minor inconsistencies that don't break character immersion.
- **3 – Fair**: Generally maintains persona but has noticeable lapses or inconsistencies in character.
- **2 – Poor**: Significant persona inconsistencies that frequently break character immersion.
- **1 – Unacceptable**: Complete failure to maintain persona or embody the specified character.

Return your evaluation in the following JSON format:

```
{
  "score": <integer from 1 to 5>,
  "reason": "<brief explanation for the score>"
}
```

Do not include markdown, comments, or anything outside the JSON.

A.7.3 REASONING TASKS - ANSWER ADEQUACY

Reasoning Task Evaluation Prompt

INSTRUCTION:
You are an expert evaluator specializing in logical reasoning and problem-solving. Your task is to assess whether the generated response provides a correct, complete, and well-reasoned answer to the given question.

USER QUESTION:
{user_prompt}

TARGET LANGUAGE:
{tgt_lang}

ASSISTANT GENERATED RESPONSE:
{assistant_output}

A response with excellent answer adequacy demonstrates:

- **Correctness**: The final answer is factually accurate and logically sound.
- **Completeness**: All aspects of the question are addressed without omitting important elements.
- **Direct Relevance**: The response directly answers what was asked without unnecessary tangents.
- **Clear Reasoning**: The logical steps leading to the conclusion are evident and valid.

Use the following scoring scale:

- **5 – Excellent**: Completely correct and comprehensive answer with clear, logical reasoning.
- **4 – Good**: Correct answer with minor gaps in completeness or explanation clarity.
- **3 – Fair**: Generally correct but missing some important aspects or contains minor errors.
- **2 – Poor**: Partially correct but has significant errors or omissions in reasoning or conclusion.
- **1 – Unacceptable**: Incorrect answer or completely fails to address the question asked.

Return your evaluation in the following JSON format:

```
{
  "score": <integer from 1 to 5>,
  "reason": "<brief explanation for the score>"
}
```

Do not include markdown, comments, or anything outside the JSON.