

UrduBench: An Urdu Reasoning Benchmark using Contextually Ensembled Translations with Human-in-the-Loop

Muhammad Ali Shafique¹, Areej Mehboob¹, Layba Fiaz¹,
Muhammad Usman Qadeer¹, Hamza Farooq¹

¹Traversaal.ai, Correspondence: ali@traversaal.ai

Abstract

Recent advances in large language models (LLMs) have led to strong reasoning capabilities; however, evaluating such models in low-resource languages remains challenging due to the lack of standardized reasoning benchmarks. Urdu reasoning evaluation, in particular, is limited by machine translation sensitivity and an emphasis on general language tasks rather than reasoning benchmarks. We propose a contextually ensembled translation framework with human-in-the-loop validation that leverages multiple translation systems to develop Urdu reasoning benchmarks while preserving contextual and structural integrity. Using this framework, we translate widely adopted benchmarks, including MGSM, MATH-500, CommonSenseQA (CSQA), and OpenBookQA (OBQA), into Urdu, collectively referred to as UrduBench, and evaluate reasoning-oriented and instruction-tuned LLMs across multiple prompting strategies. Our analysis examines four datasets, five task difficulty levels, diverse model architectures, model scaling settings, and language consistency. We find that multi-step and symbolic reasoning tasks pose significant challenges in Urdu, and that stable language alignment is a critical prerequisite for robust reasoning. Overall, our work establishes a scalable methodology for standardized Urdu reasoning evaluation and provides insights into multilingual reasoning failures, with applicability to other low-resource languages. The code and datasets will be publicly released.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across reasoning, comprehension, and knowledge-intensive tasks. However, the benefits of these advancements are not evenly distributed across languages. Low-resource languages such as Urdu continue to face a significant gap in standardized evaluation resources. This disparity not only limits the development of

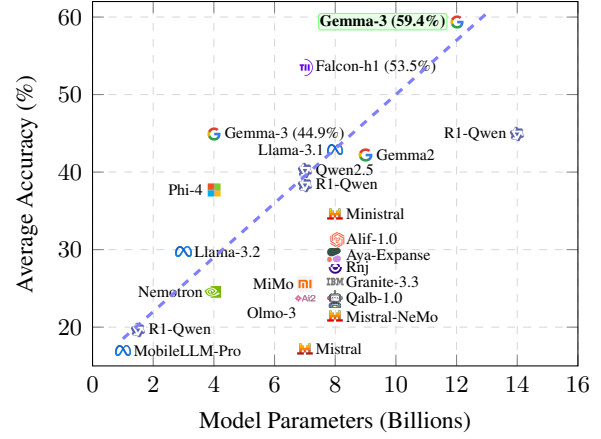


Figure 1: Performance scaling of state-of-the-art LLMs on Urdu benchmark datasets. Average accuracy is the average of MGSM (CoT), MATH-500 (CoT), CSQA (Direct), and OBQA (Direct) accuracies. Green highlights the top performing model in Urdu.

Urdu-capable LLMs but also restricts fair and rigorous comparison across multilingual models.

Existing Urdu evaluation resources discussed in Section 2 focus primarily on evaluating LLMs’ capabilities in Natural Language Processing (NLP) and Natural Language Understanding (NLU) tasks that include machine translation, sentiment analysis, basic classification, or extractive question answering. While valuable, these evaluations provide only a partial view of model capability and do not assess complex reasoning such as mathematical problem solving or common sense reasoning. In addition, regional initiatives, including TituLLM (Nahin et al., 2025) for Bangla, C²RBench (Wu et al., 2025) for Chinese, (Lamsiyah et al., 2025), (Hasanaath et al., 2025) for Arabic, and MasakhaNER for African languages (Adelani et al., 2021) further highlight local efforts to benchmark complex model behavior in various linguistic settings.

To address this gap, we introduce a reasoning-focused benchmark for evaluating LLMs in Urdu by translating four widely adopted datasets includ-

English question with choices	Translation without Context	Context Aware Translation (with human validation)
The sun is responsible for: a. puppies learning new tricks b. children growing up and getting old c. flowers wilting in a vase d. plants sprouting, blooming and wilting	سورج ذمے دار ہے: a. کتے کے بچوں نے تربیت سیکھ رہے ہیں b. بچے بڑے ہو رہے ہیں اور بوڑھے ہو رہے ہیں c. گھداں میں پھول مرجھا رہے ہیں d. پودے اک رہے ہیں، پھل رہے ہیں اور مرجھا رہے ہیں	سورج ذمے دار ہے: a. کتے کے بچوں کے نئے کرب سیکھنے کا b. بچوں کے بڑے ہونے اور بوڑھے ہونے کا c. گھداں میں پھولوں کے مرجھانے کا d. پودوں کے اگلنے، پھلنے اور مرجھانے کا

Table 1: Example with and without Context Translation

ing Multilingual Grade School Math (MGSM) (Shi et al., 2022), OBQA (Mihaylov et al., 2018), CSQA (Talmor et al., 2019) and MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024) into high-quality Urdu. These datasets span diverse and complementary reasoning abilities that have not been systematically evaluated in Urdu. MGSM measures mathematical and multi-step reasoning, OBQA assesses factual and retrieval-based reasoning, and CSQA evaluates commonsense knowledge and inference.

To ensure translation fidelity, we employ a contextual ensemble translation framework with human-in-the-loop validation as examples reported in Table 1 and Appendix B. Using the resulting datasets, we benchmark multiple state-of-the-art reasoning and non-reasoning LLMs under a unified LM Evaluation Harness protocol, enabling large-scale and standardized comparison of reasoning performance in Urdu as shown in Figure 1.

Key Findings The key findings of this study are summarized below:

- 1. Scalable Urdu Translation Framework (§3)**
A contextual ensemble translation pipeline with human-in-the-loop validation produces high-fidelity Urdu translations of MGSM, MATH-500, CSQA, and OBQA, enabling reproducible evaluation.
- 2. Comprehensive Model Evaluations (§4.3)**
We evaluate diverse reasoning and non-reasoning LLMs under multiple prompting strategies in a unified, reproducible setup.
- 3. Impact of Task Difficulty (§5.2)** MATH-500 accuracy declines monotonically with difficulty; non-instruction-tuned models degrade sharply, while DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025) and Phi-4-Mini-Reasoning (Xu et al., 2025) better sustain performance.
- 4. Scaling Effects (§5.3)** Larger models

(Gemma-3-12B-it (Team et al., 2025), Falcon-h1-7B-Instruct (Zuo et al., 2025)) generally perform better, but scale alone is insufficient, as Gemma-3-4B-it outperforms several 7B models.

2 Related Work

Prior work shows that automatically translated benchmarks, including Spanish MMLU, introduce substantial errors that distort evaluation outcomes (Plaza et al., 2024). Large-scale multilingual studies further confirm strong sensitivity to translation quality (Theilmann et al., 2024). Translation-following demonstrations can also improve cross-lingual alignment beyond monolingual instruction tuning (Ranaldi et al., 2024a). Machine translation research highlights metric limitations and instability, particularly for low-resource languages (López, 2023; Zhu et al., 2024). Recent benchmarks similarly reveal persistent performance gaps and the need for human validation (Bhowmik et al., 2025; Song et al., 2026). In contrast, we introduce a scalable, context-aware ensemble translation framework with human-in-the-loop validation for reliable Urdu reasoning evaluation.

Multilingual Reasoning and Translation Safeguards Recent multilingual reasoning suites address translation artifacts via native authoring, expert post-editing, and pairwise audits: MATH-MIST (Sobhani et al., 2026) for math with code-switching analysis; FLORES-101 (Goyal et al., 2022) for adequacy/fluency-audited MT across 101 languages; and Indic-QA (Singh et al., 2025)/IndicParam (Maheshwari et al., 2025) for human-validated Indic QA. Cross-lingual Tree-of-Thoughts further improves multilingual multi-step reasoning through cross-lingual alignment and self-consistent reasoning paths (Ranaldi et al., 2024b). Language consistency diagnostic (§5.5) provides

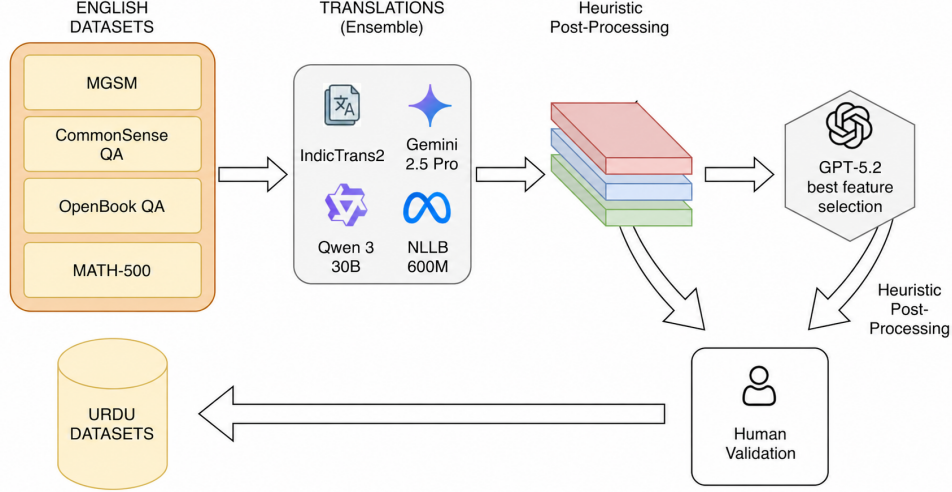


Figure 2: Contextual Ensembled Translations Framework with human-in-the-loop

complementary behavioral evidence in Urdu. We adopt Sprague et al. (2025)’s finding that CoT primarily benefits symbolic reasoning by restricting CoT/FS+CoT to MGSM and MATH-500.

Urdu Benchmarking Datasets Prior Urdu evaluations primarily target general NLP capabilities rather than structured reasoning. Existing work compares generalist and specialist models across Urdu tasks (Arif et al., 2024), studies cross-lingual QA and translation-induced errors (Kazi et al., 2025), and probes linguistic competence using minimal-pair benchmarks such as UrBLiMP (Adeeba et al., 2025). Other efforts address factual verification (Ahmad et al., 2025) or classical NLP tasks. However, these benchmarks do not systematically evaluate multi-step mathematical or commonsense reasoning. UrduBench addresses this gap by providing standardized, context-preserving, and human-validated Urdu translations of established reasoning benchmarks.

3 Contextually Ensembled Translation Method

Constructing reliable reasoning benchmarks in Urdu requires more than direct machine-translation, particularly for structured and multi-step reasoning tasks. We observe that multiple-choice datasets such as OBQA and CSQA are especially prone to contextual fragmentation when questions and answer options are translated independently, leading to semantically misaligned Urdu outputs. To mitigate this, we employ a contextually ensembled, human-in-the-loop translation pipeline that preserves contextual coherence while integrating

multiple translation models and quality controls as shown in Figure 2. A comparative analysis of translations generated with and without this contextual grounding is provided in Appendix B (Table 5), demonstrating a clear qualitative improvement in semantic alignment.

3.1 Contextual Translations

We translate MGSM, CSQA, OBQA, and MATH-500 into Urdu using a unified pipeline designed to preserve semantic fidelity, contextual coherence, and structural integrity. Each English instance is first translated independently using four complementary translation candidates: IndicTrans2 (Gala et al., 2023), NLLB (Koishekenov et al., 2023), Qwen-3-30B (Yang et al., 2025a), and Gemini-2.5-Pro (Comanici et al., 2025). These candidates are selected based on their demonstrated translation quality scores in prior multilingual evaluation studies (Basit et al., 2024; Nahin et al., 2025). Translation prompts used for these candidates are provided in Appendix A.1.

For structured multiple-choice datasets such as CSQA and OBQA, we adopt a context-preserving translation strategy by concatenating the question stem with all associated answer choices prior to translation. This prevents semantic fragmentation and ensures grammatical and logical alignment between the question and its answer options in Urdu. After translation, questions and answer choices are programmatically separated into their original components.

3.2 Candidate Translation Fusion

GPT-5.2 (OpenAI, 2025) is used to refine candidate translations by leveraging complementary strengths across models in lexical choice, grammar, and handling of technical content. It compares candidates’ proposed translations against the original English source to either synthesize an improved Urdu translation or select the best-performing output. Beyond quality improvement, this stage also serves as an automated judging step and limits the need for manual human validation. The prompt for best feature selection and representative examples are provided in Appendix A.2. The released artifacts are the human-validated Urdu strings rather than the fusion model’s outputs, ensuring downstream use of UrduBench does not require GPT-5.2 API access. Furthermore, our pipeline’s architecture is model-agnostic, meaning future iterations can readily substitute GPT-5.2 with emerging open-weight multilingual judges as they reach semantic parity.

3.3 Heuristic Post-Processing

All translations undergo automatic heuristic validation tailored to Urdu, as described in Algorithm 1. The validation detects the following issues in translated outputs:

- Empty or incomplete translations
- Missing Urdu questions or answer options
- Abnormally short translated questions
- Invalid or missing answer keys
- Excessive residual English content

Most translations pass these checks automatically; flagged cases are reviewed and reprocessed to ensure accurate and complete translations.

3.4 Human Validation

All candidate translations, including those produced by GPT-5.2, undergo final review by professional native Urdu-speaking annotators fluent in English. Annotators¹ follow detailed guidelines provided in Appendix C and select the most accurate and natural translation for each instance. This step is essential for resolving residual ambiguities, ensuring semantic fidelity, and accounting for cultural and linguistic nuances that automated systems may overlook. Using this framework, the best quality translation is selected for inclusion in the final benchmark, ensuring linguistic coherence and

¹Paid hourly; above min. wage of country of employment.

Dataset	Train	Val	Test
MGSM	8	—	250
CSQA	9,741	1,221	1,140
OBQA	4,957	500	500
MATH-500	—	—	500

Table 2: Statistics of the translated Urdu datasets. For MGSM the train column lists the 8 worked exemplars released with the benchmark and used as few-shot demonstrations rather than as a training set. CSQA and OBQA training splits are translated for completeness but not used as training data in this paper; all reported numbers in §4.3 are computed on the test splits (CSQA evaluation uses the standard validation split because the test labels are not public).

contextual fidelity suitable for evaluating LLM reasoning in Urdu. Dataset statistics for the resulting benchmarks are reported in Table 2. Human Validation Win–Loss Statistics Across Translation Candidates and human preference win–loss comparison between GPT-5.2 refined and human-edited Urdu translations across four datasets are shown in Table 9 in Appendix.

For CSQA and OpenBookQA, annotators additionally verify that the Urdu option corresponding to the English gold key remains uniquely correct under standard Urdu semantics; residual key-drift cases are corrected in-place before inclusion.

4 Evaluation of Models

In this section, we present a systematic evaluation of diverse language models on the proposed Urdu benchmarks, analyzing the effects of model design, prompting strategy, and task difficulty on reasoning performance.

4.1 Experimental Setup and Datasets

We evaluate models on the Urdu-translated reasoning and question-answering benchmarks introduced in this work: MGSM (arithmetic reasoning, evaluated with Direct, CoT, and three Few-Shot + CoT prompting), MATH-500 (symbolic mathematics across five difficulty levels, L1–L5), CSQA (commonsense reasoning), and OBQA (knowledge-based scientific reasoning). Together, these benchmarks span arithmetic, formal mathematics, commonsense understanding, and factual knowledge translated in Urdu. All evaluations use lm-evaluation-harness (Gao et al., 2024) with a vLLM (Kwon et al., 2023) backend.

4.2 Evaluated Models

We evaluate a broad set of open-source and open-weight LLMs, including instruction-tuned and ex-

Model	MGSM			CSQA	OBQA	M500	Avg
	Direct	CoT	FS+CoT	Direct	Direct	CoT	%
<i>Non-Reasoning-oriented Open-source LLM</i>							
MobileLLM-Pro (1B)	0.40±0.0	0.00±0.0	0.40±0.4	20.8±1.2	27.4±2.0	2.80±0.8	17.0±1.0
Llama-3.2-3B-Instruct	28.4±2.9	40.0±3.0	35.6±3.0	37.1±1.4	29.4±2.0	12.6±1.5	29.7±1.9
Gemma-3-4B-it	54.8±3.2	64.0±3.0	44.0±3.2	40.5±1.4	31.2±2.1	44.2±2.2	44.9±2.2
Nemotron-Mini-4B	18.8±2.5	22.4±2.6	25.6±2.8	38.4±1.4	32.0±2.1	5.00±1.0	24.5±1.8
Falcon-h1-7b-instruct	71.2±2.9	72.4±2.8	70.4±2.9	59.3±1.4	32.8±2.1	49.6±2.2	53.5±2.1
MiMo-7B-Base	29.6±2.9	20.8±2.6	56.4±3.1	37.4±1.4	28.8±2.0	15.4±1.6	25.6±2.2
Mistral-7B-Instruct	0.40±1.2	4.40±1.3	11.6±2.0	27.5±1.3	30.2±2.1	6.60±1.1	17.2±1.5
Olmo-3-7B-Instruct	21.6±2.6	23.6±2.7	24.0±2.7	27.9±1.3	26.6±2.0	20.6±1.8	24.7±1.9
Qwen2.5-7B-Instruct	53.2±3.2	57.6±3.1	60.4±3.1	30.4±1.3	31.6±2.1	41.2±2.2	40.2±2.2
Alif-1.0-8B-Instruct	21.6±2.6	30.4±2.9	43.6±3.1	51.3±1.4	33.4±2.1	10.6±1.4	31.4±1.9
Aya-Expanse-8B	22.0±2.6	30.8±2.9	28.8±2.8	37.5±1.4	31.6±2.1	17.2±1.7	29.2±2.0
Granite-3.3-8b-instruct	8.00±1.7	33.6±3.0	30.4±2.9	28.2±1.3	29.6±0.0	12.4±1.5	25.9±1.4
Qalb-1.0-8B-Instruct	19.6±2.5	17.6±2.4	7.60±1.7	38.4±1.4	32.2±2.1	6.00±1.1	23.6±1.6
Llama-3.1-8B-Instruct	62.0±3.1	62.8±3.1	54.4±3.2	52.8±1.4	33.8±2.1	22.4±1.9	42.9±1.7
Ministral-8B-Instruct-2410	55.2±3.2	58.4±3.1	50.8±3.2	29.6±1.3	30.8±2.1	19.4±1.8	34.6±2.1
Mistral-NeMo-Minitron-8B-Instruct	4.00±1.2	0.40±0.4	0.40±0.4	39.2±1.4	29.0±2.0	17.0±1.7	21.4±1.4
Rnj-1-instruct (8B)	22.4±2.6	13.2±2.2	9.20±1.8	27.9±1.3	26.6±2.0	42.6±2.2	27.6±1.9
Gemma2-9B-it	74.0±2.8	73.6±2.8	75.2±2.7	25.3±1.2	32.8±2.1	37.0±2.2	42.2±2.1
Gemma-3-12B-it	80.4±2.5	80.8±2.5	82.8±2.4	49.4±1.4	37.4±2.2	70.0±2.1	59.4±2.1
<i>Reasoning-oriented Open-source LLM</i>							
DeepSeek-R1-Distill-Qwen-1.5B	8.40±1.8	6.40±1.6	9.60±1.9	20.6±1.2	30.0±2.1	21.4±1.8	19.6±1.7
Phi-4-mini-reasoning (4B)	38.4±3.1	32.4±2.9	29.2±2.9	20.4±1.2	32.4±2.1	65.8±2.1	37.7±2.1
DeepSeek-R1-Distill-Qwen-7B	39.6±3.1	42.0±3.1	36.0±3.0	22.9±1.2	31.2±2.1	57.0±2.2	38.3±2.2
DeepSeek-R1-Distill-Qwen-14B	54.4±3.2	59.2±3.1	68.8±2.9	19.4±1.1	31.8±2.1	69.2±2.1	44.9±1.2

Average considers MGSM (cot), MATH-500 (cot) CSQA (direct), and OBQA (direct).

Table 3: Evaluation Across All Urdu Benchmark Datasets. Metric: Normalized Accuracy

plicitly reasoning-oriented models, under a fixed evaluation protocol as shown in Table 3. The models span parameter scales from 1B to 14B+, diverse multilingual pretraining regimes, and different reasoning recipes, including distillation- and RL-based approaches (e.g., DeepSeek-R1, Phi-4-Mini-Reasoning) as shown in Table 8. This di-

versity enables controlled analysis of how various model architectures, model scaling, training strategy, prompting methods, and language consistency, influence Urdu reasoning performance, particularly under increasing task difficulty as demonstrated in Figure 3. All evaluation prompts, configurations and model details are provided in Appendix D.

4.3 Results

We report normalized accuracy across all Urdu benchmarks in Table 3, covering arithmetic reasoning (MGSM, MATH-500) and commonsense question answering (CSQA, OBQA). Results are aggregated across multiple prompting strategies to analyze prompt sensitivity, architectural effects, and overall model robustness in a low-resource language setting.

Prompt Sensitivity Prompting strategy is a major determinant of performance on MGSM. Direct prompting consistently underperforms reasoning-aware prompts, indicating that most models struggle to implicitly perform multi-step reasoning in Urdu without explicit guidance. Introducing Chain-of-Thought (CoT) leads to substantial gains across nearly all architectures, often yielding large relative improvements over Direct prompting. However, the benefits of Few-Shot + CoT are uneven: while it enables peak performance for some models (e.g., Falcon-h1-7B-Instruct (Zuo et al., 2025)), it introduces instability or regressions for others, particularly smaller models. This suggests that Urdu samples can impose additional distributional shift rather than uniformly improving generalization. Overall, CoT emerges as the most reliable prompting strategy, while few-shot gains remain highly model-dependent.

Model Architecture Model architecture and training objectives strongly influence Urdu reasoning performance. Models with stronger multilingual pretraining and instruction tuning exhibit more reliable transfer across tasks. Gemma-family models, particularly Gemma-3-12B-it and Gemma-3-4B-it, achieve strong performance across both reasoning-intensive and commonsense benchmarks, indicating effective cross-lingual abstraction. In contrast, compact or efficiency-oriented architectures (e.g., MobileLLM-Pro (Huber et al., 2025), Mistral-NeMo-Minitron (Muralidharan et al., 2024; NVIDIA, 2024a)) perform low on MGSM and degrade sharply on MATH-500, reflecting limited capacity for sustained symbolic reasoning in Urdu. Reasoning-oriented distilled models (e.g., DeepSeek-R1 variants) show moderate but inconsistent improvements, suggesting that reasoning distillation alone does not fully compensate for limited multilingual grounding.

Overall Performance Across Benchmarks. Despite improvements from prompting and archi-

tecture, performance in Urdu remains uneven across tasks, underscoring the challenges of low-resource evaluation. Gemma-3-12B-it (Team et al., 2025) and Falcon-h1-7B-Instruct (Zuo et al., 2025) achieve the highest overall averages, demonstrating that strong instruction tuning combined with multilingual exposure transfers most effectively. Taken together, the results indicate that reliable Urdu reasoning requires a combination of explicit reasoning prompts, sufficient model capacity, and multilingual-aware training, with no single factor being sufficient on its own.

5 Analyses and Discussion

In this section, we analyze the results to isolate how dataset properties, task difficulty, model scale, reasoning specialization, and language consistency influence Urdu reasoning performance.

5.1 Impact of Dataset

How does dataset choice affect model performance in Urdu?

Dataset choice has a strong, task-dependent effect on model performance in Urdu. Reasoning-intensive benchmarks such as MGSM and MATH-500 exhibit the highest variance across models, indicating that multi-step and symbolic reasoning is particularly sensitive to linguistic transfer and translation noise in a low-resource setting as shown in Figure 5 in Appendix E. In contrast, commonsense and knowledge-based tasks such as CSQA and OBQA remain comparatively stable, with tighter performance distributions. Overall, these results indicate that reasoning depth and symbolic complexity are primary drivers of evaluation difficulty in Urdu, while semantic and knowledge-oriented tasks are more robust to cross-lingual transfer.

5.2 Impact of Task Difficulty

How does task difficulty affect mathematical reasoning in Urdu?

Task difficulty has a systematic impact on performance in MATH-500-Urdu. As shown in Figure 3 and Table 7 in the appendix, accuracy declines steadily with increasing problem difficulty, with the largest drops at higher levels (L4–L5). While most models perform well on low-difficulty problems, they degrade sharply on more complex subsets, indicating limited robustness to deep symbolic and multi-step reasoning in Urdu. Larger and reasoning-oriented models such as Gemma-3-12B-

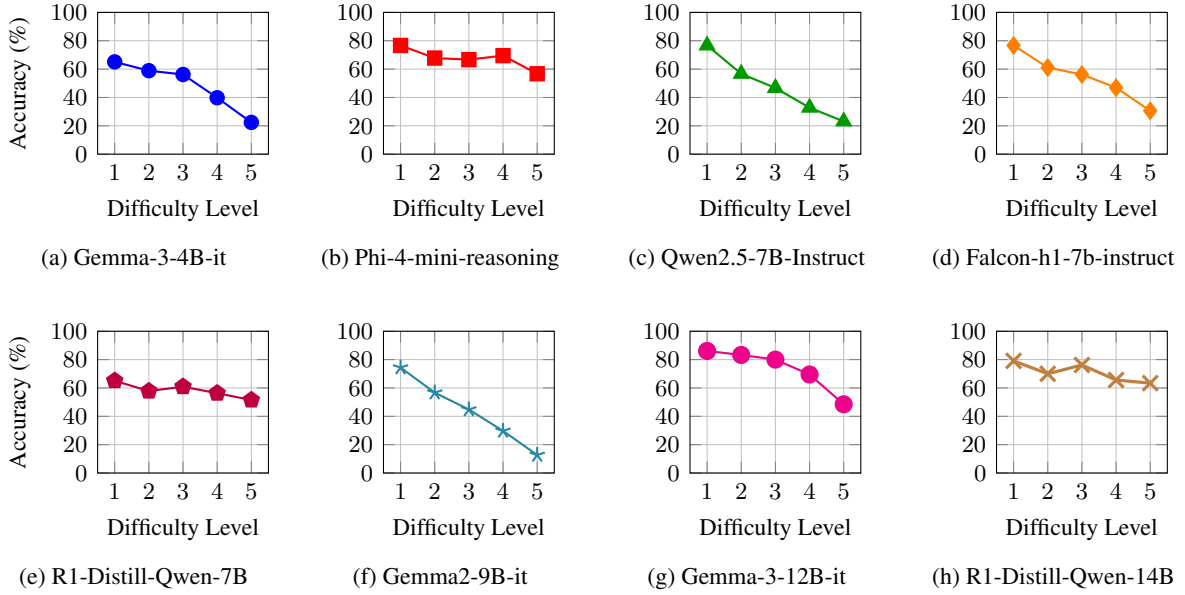


Figure 3: Performance of different language models across difficulty levels (L1–L5) on the MATH-500-Urdu benchmark. Each subplot shows the accuracy degradation pattern as problem difficulty increases.

it and DeepSeek-R1-Distill-Qwen-14B retain comparatively stronger performance at higher difficulty levels, but the overall trend shows that linguistic transfer amplifies the inherent challenges of formal mathematics.

5.3 Impact of Model Size

How does model scale influence performance in Urdu?

Model scale contributes to performance on Urdu benchmarks, but it is not sufficient on its own. While larger models such as Gemma-3-12B-it achieve strong overall results, scale does not guarantee superiority. Notably, Gemma-3-4B-it outperforms several larger models, including Llama-3.1-8B-Instruct and DeepSeek-R1-Distill-Qwen-7B as shown in Table 3 and in Figure 1, highlighting the importance of multilingual training and Urdu exposure. Similarly, Gemma-3-12B-it surpasses the larger DeepSeek-R1-Distill-Qwen-14B on reasoning-intensive benchmarks, indicating that reasoning-oriented scaling alone does not ensure better low-resource language performance. Overall, these results show that model size is only one factor, with targeted multilingual coverage playing a critical role in Urdu reasoning quality.

5.4 Impact of Reasoning and Non-Reasoning Models

How do reasoning-specialized models compare to general instruction-tuned models in Urdu?

A central goal of UrduBench is to evaluate whether explicitly reasoning-oriented models outperform general instruction-tuned models in Urdu. As shown in Figure 1, reasoning-specialized models such as Phi-4-Mini-Reasoning and DeepSeek-R1-Distill variants exhibit clear gains on MGSM, indicating that explicit reasoning supervision partially transfers to low-resource languages. However, these advantages are less consistent on CSQA and OBQA, where large instruction-tuned models such as Gemma-3-12B-it (Team et al., 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) often outperform reasoning-specialized counterparts. This pattern suggests that reasoning supervision alone is insufficient without strong multilingual grounding, and that performance on commonsense and factual reasoning tasks in Urdu depends more on pretraining diversity and language alignment than on explicit reasoning objectives alone.

5.5 Language Confusion Test on Reasoning Models

How well do models maintain Urdu language consistency in their responses?

Language confusion test measures whether a model produces its final answers consistently in the target language (Urdu) without unintended code-switching. For each generated response, special tokens and mathematical notation are removed, and language identification is applied to the cleaned text. An answer is counted as language-consistent

Model	Acc. (%)	Cons. (%)
Gemma-3-4B-it	64.0	98.8
Falcon-h1-7b-instruct	72.4	89.2
Qwen2.5-7B-Instruct	57.6	100.0
Llama-3.1-8B-Instruct	62.8	100.0
Ministral-8B-Instruct-2410	58.4	98.8
Gemma2-9B-it	73.6	99.6
Gemma-3-12B-it	80.4	100.0
DeepSeek-R1-Distill-Qwen-14B	59.2	82.0

Table 4: Language Confusion Test on MGSM Urdu CoT. Accuracy versus Answer language consistency for Urdu. Higher consistency indicates lower language confusion or codeswitching

only if Urdu is the sole detected language; very short responses are excluded to avoid unreliable detection.

Formally, the Answer Language Consistency formula is taken from the study (Wang et al., 2026) as:

$$\text{Lang. Consistency (\%)} = \frac{\# \text{ answers in Urdu}}{\# \text{ total answers}} \times 100$$

In our evaluation, we measure this by checking whether the generated answer contains only Urdu; higher consistency therefore indicates less language confusion or code-switching. The results in Table 4 reveal a notable issue with DeepSeek-R1-Distill-Qwen-14B, which achieves only 82.0% Urdu consistency on MGSM-CoT, substantially lower than the other evaluated models, most of which exceed 98%. Importantly, language consistency should not be interpreted as a direct predictor of reasoning accuracy. For example, Qwen2.5-7B-Instruct achieves 100% Urdu consistency but reaches only 57.6% reasoning accuracy, while DeepSeek-R1-Distill-Qwen-14B has lower consistency (82.0%) but slightly higher reasoning accuracy (59.2%). Falcon-h1-7B achieves only 89.2% Urdu consistency, it attains the second-highest reasoning accuracy (72.4%). This observation is consistent with the study (Wang et al., 2026), which shows that reasoning models can exhibit substantial language inconsistency, including DeepSeek-R1-671B, which also shows notable language consistency issues. The similar behavior observed in DeepSeek-R1-Distill-Qwen-14B suggests that this language inconsistency may persist in models distilled from DeepSeek-R1. Taken together, these results highlight language consistency as an important diagnostic of unintended language switching, while also showing that language inconsistency

does not necessarily translate into lower reasoning accuracy.

6 Conclusion

In this study, we introduced UrduBench, to the best of our knowledge, the first standardized reasoning-focused benchmark suite for Urdu, comprising translations of MGSM, OBQA, CSQA, and MATH-500. Our contextually ensembled human-in-the-loop framework combines four diverse translation systems with LLM-based feature integration and native human validation. By processing questions and answer choices as unified contextual units, the framework preserves linguistic coherence and task structure during translation.

Through comprehensive evaluation of state-of-the-art reasoning and non-reasoning models, we provide the first large-scale standardized comparison of LLM reasoning capabilities in Urdu across mathematical reasoning, commonsense knowledge, and factual understanding. Our results highlight key challenges in multilingual reasoning and language alignment. Beyond advancing Urdu NLP, our methodology provides a scalable framework for adapting standardized reasoning benchmarks to other low-resource languages.

Limitations

Scope. UrduBench targets Modern Standard Urdu across MGSM, MATH-500, Common-SenseQA, and OpenBookQA. Dialectal, code-mixed, and Urdu-origin reasoning corpora are outside the present scope. We evaluate 20+ open-source/open-weight LLMs from 1B–14B; proprietary and frontier-scale models are excluded.

Pipeline-side proprietary dependency. Candidate fusion uses GPT-5.2, while Gemini-2.5-Pro often provides the strongest base candidate (Table 9). However, downstream use of the released human-validated benchmark requires no API access; only pipeline reconstruction does. A fully open-weight pipeline can be made in future work.

Methodological constraints. Validation relied on expert consensus and win-loss tracking rather than formal Inter-Annotator Agreement (IAA). We provide qualitative translation-failure examples (Appendix B).

Acknowledgments

ChatGPT 5.2 was used for grammatical refinement. The authors take full responsibility for the technical content of this paper.

References

- Farah Adeeba, Brian Dillon, Hassan Sajjad, and Rajesh Bhatt. 2025. [Urblimp: A benchmark for evaluating the linguistic competence of large language models in urdu](#).
- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel Dsouza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, and 1 others. 2021. [Masakhaner: Named entity recognition for african languages](#). *Transactions of the Association for Computational Linguistics*, 9:1116–1131.
- Sarfraz Ahmad, Hasan Iqbal, Momina Ahsan, Numaan Naeem, Muhammad Ahsan Riaz Khan, Arham Riaz, Muhammad Arslan Manzoor, Yuxia Wang, and Preslav Nakov. 2025. [Urdufactcheck: An agentic fact-checking framework for urdu with evidence boosting and benchmarking](#). In *30th Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*, pages 22788–22802. Association for Computational Linguistics (ACL).
- Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza, and Awais Athar. 2024. [Generalists vs. specialists: Evaluating large language models for urdu](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7263–7280.
- Abdul Basit, Abdul Hameed Azeemi, and Agha Ali Raza. 2024. [Challenges in urdu machine translation](#). In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 44–49.
- Shimanto Bhowmik, Tawsif Tashwar Dipto, Md Sazad Islam, Sheryl Hsu, and Tahsin Reasat. 2025. [Evaluating llms’ multilingual capabilities for bengali: Benchmark creation and performance analysis](#). *arXiv preprint arXiv:2507.23248*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *arXiv preprint arXiv:2507.06261*.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#).
- Jay Gala, Pranjal A Chitale, Raghavan Ak, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. [Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages](#). *arXiv preprint arXiv:2305.16307*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, MarcAurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). volume 10, pages 522–538. MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA.
- IBM Granite Team. 2025. [Granite 3.3 language models](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Ahmed Abul Hasanaath, Aisha Alansari, Ahmed Ashraf, Salmane Chafik, Hamzah Luqman, and Saad Ezzini. 2025. [AraReasoner: Evaluating reasoning-based LLMs for Arabic NLP](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18898–18914, Suzhou, China. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the math dataset](#). *arXiv preprint arXiv:2103.03874*.
- Patrick Huber, Ernie Chang, Wei Wen, Igor Fedorov, Tarek Elgamal, Hanxian Huang, Naveen Suda, Chinadhurai Sankar, Vish Vogeti, Yanghan Wang, and 1 others. 2025. [Mobilellm-pro technical report](#). *arXiv preprint arXiv:2511.06719*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,

- Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. *Mistral 7b*. Preprint, arXiv:2310.06825.
- Samreen Kazi, Maria Rahim, and Shakeel Ahmed Khoja. 2025. *Crossing language boundaries: Evaluation of large language models on urdu-english question answering*. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 141–151.
- Yeskendir Koishchenov, Alexandre Berard, and Vasilina Nikoulina. 2023. *Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Salima Lamsiyah, Kamyar Zeinalipour, Samir El amary, Matthias Brust, Marco Maggini, Pascal Bouvry, and Christoph Schommer. 2025. *ArabicSense: A benchmark for evaluating commonsense reasoning in Arabic with large language models*. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11, Abu Dhabi, UAE. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. *Let's verify step by step*. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. *Ministral 3*. arXiv preprint arXiv:2601.08584.
- Álvaro López. 2023. *Machine translation evaluation metrics benchmarking: From traditional mt to llms*.
- Ayush Maheshwari, Kaushal Sharma, Vivek Patel, and Aditya Maheshwari. 2025. *Indicparam: Benchmark to evaluate llms on low-resource indic languages*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. *Can a suit of armor conduct electricity? a new dataset for open book question answering*. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2381–2391.
- Saurav Muralidharan and 1 others. 2024. Compact language models via pruning and knowledge distillation. arXiv preprint arXiv:2407.14679.
- Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj Alam. 2025. *Titulms: A family of bangla llms with comprehensive benchmarking*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24922–24940.
- NVIDIA. 2024a. *Mistral-nemo-minitron-8b-instruct*. <https://huggingface.co/nvidia/Mistral-NeMo-Minitron-8B-Instruct>.
- NVIDIA. 2024b. *Nemotron-mini-4b-instruct*. <https://huggingface.co/nvidia/Nemotron-Mini-4B-Instruct>.
- Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. *Olmo 3*.
- OpenAI. 2025. Introducing gpt-5.2. <https://openai.com/index/introducing-gpt-5-2/>. Accessed: 2025-12-26.
- Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. *Spanish and llm benchmarks: is mmlu lost in translation?* arXiv preprint arXiv:2406.17789.
- Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024a. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973.
- Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024b. A tree-of-thoughts to broaden multi-step reasoning across languages. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241.
- Muhammad Ali Shafique, Kanwal Mehreen, Muhammad Arham, Maaz Amjad, Sabur Butt, and Hamza Farooq. 2025. *Alif: Advancing urdu large language models via multilingual synthetic data distillation*. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 271–284.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. *Language models are multilingual chain-of-thought reasoners*.
- Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025. *INDIC QA BENCHMARK: A multilingual benchmark to evaluate question answering capability of LLMs for Indic languages*. In *Findings of the Association for Computational Linguistics*:

- NAACL 2025, pages 2607–2626, Albuquerque, New Mexico. Association for Computational Linguistics.
- Mahbub E Sobhani, Md Faiyaz Abdullah Sayeedi, Muhammad Tasnim Mohiuddin, Md Mofijul Islam, and Swakkhar Shatabda. 2026. [Mathmist: A parallel multilingual benchmark dataset for mathematical problem solving and reasoning](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2524–2550.
- Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Tegawendé F Bissyandé, Jacques Klein, and 1 others. 2026. [Are small language models the silver bullet to low-resource languages machine translation?](#) In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 1–26.
- Zayne Sprague, Fangcong Yin, Juan Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, volume 2025, pages 94118–94162.
- Muhammad Taimoor Hassan, Jawad Ahmed, and Muhammad Awais. 2026. [Qalb: Largest state-of-the-art urdu large language model for 230m speakers with systematic continued pre-training](#). *arXiv e-prints*, pages arXiv–2601.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [Commonsenseqa: A question answering challenge targeting commonsense knowledge](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and 1 others. 2024. [Towards multilingual llm evaluation for european languages](#). *arXiv preprint arXiv:2410.08928*.
- Ashish Vaswani, Mike Callahan, Adarsh Chaluvareja, Aleksa Gordi, Devaansh Gupta, Yash Jain, Divya Mansingka, Philip Monk, Khoi Nguyen, Mohit Parmar, Michael Pust, Tim Romanski, Peter Rushton, Ali Shehper, Divya Shivaprasad, Somanshu Singla, Kurt Smith, Saurabh Srivastava, Anil Thomas, and 4 others. 2025. [Rnj-1-Instruct](#). Instruction-tuned model release.
- Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, and 1 others. 2026. [Polymath: Evaluating mathematical reasoning in multilingual contexts](#). *Advances in Neural Information Processing Systems*, 38.
- Sajana Weerawardhena, Paul Kassianik, Blaine Nelson, Baturay Saglam, Anu Vellore, Aman Priyanshu, Supriti Vijay, Massimo Auffero, Arthur Goldblatt, Fraser Burch, and 1 others. 2025. [Llama-3.1-foundationai-securityllm-8b-instruct technical report](#).
- Junru Wu, Tianhao Shen, Linxi Su, and Deyi Xiong. 2025. [C²RBench: A Chinese complex reasoning benchmark for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21031–21050, Vienna, Austria. Association for Computational Linguistics.
- LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, and 1 others. 2025. [Mimo: Unlocking the reasoning potential of language model—from pretraining to posttraining](#).
- Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, and 1 others. 2025. [Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math](#). *arXiv preprint arXiv:2504.21233*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual machine translation with large language models: Empirical results and analysis](#). In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.
- Jingwei Zuo, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, Hamza Yous, Brahim Farhat, Ibrahim

Khadraoui, and 1 others. 2025. [Falcon-h1: A family of hybrid-head language models redefining efficiency and performance](#). *arXiv preprint arXiv:2507.22448*.

A Prompts and Results Used in the Translation Framework

This appendix provides the exact prompts used in our Contextual Ensemble Translation framework with human-in-the-loop. All translations in the benchmark were generated using the following LLM prompts.

A.1 Qwen 3 and Gemini 2.5 Pro Urdu Contextual Translation Prompt

The prompt below was used to generate Urdu translations directly from English source texts for MGSM.

Translate given text into Urdu.
DO NOT respond the text. ONLY
output the translation of the
text. Maintain natural, fluent,
and cultural context in Urdu
translation.

For OBQA, and CSQA, following prompt was used using Context-aware strategy:

You are a translation expert.
Analyze the given translations
against the original text and
create one improved translation
by combining the best features
from all of them.

Translate the following English
text into Urdu. Return **only** a
valid JSON object with two keys:

1. "question": the translated question
2. "choices": an array of translated choice texts (maintain the same order)

Maintain natural, fluent, and culturally appropriate Urdu. Do **not** include any explanation or extra text.

Question: {question_stem}

Choices:
{choices_formatted}

Return format:
{ "question": "translated

question", "choices":
["choice1", "choice2", ...]}

For MATH-500, following prompt was used:

You are an expert translator
specializing in translating
mathematical content from
English to Urdu while preserving
all LaTeX mathematical notation.

Task: Translate the following
from English to Urdu.

Critical Rules:

1. NEVER translate or modify ANY LaTeX math expressions keep all content inside \$ or \$\$ EXACTLY unchanged
2. NEVER translate mathematical symbols, variables, or formulas
3. Only translate natural language text
4. Preserve exact structure, line breaks, and spacing
5. Use fluent, formal academic Urdu
6. Preserve punctuation and special characters

A.2 GPT-5.2 Prompt for Best Feature Selection

GPT-5.2 was used to merge, refine, and improve candidate translations generated by IndicTrans2, Qwen-3-30B, Gemini 2.5 Pro, and NLLB.

You are a translation expert.
Analyze the given translations
against the original text and
create one improved translation
by combining the best features
from all of them.

A.3 Translation and Heuristic Processing Algorithm

The translation and heuristic processing pipeline applied to all translated samples is described in Algorithm 1.

B Context-Aware Translation Examples

This appendix provides illustrative examples demonstrating how context-aware translation improves the grammatical coherence and semantic

alignment of Urdu multiple-choice questions. Table 5 compares translations generated with and without contextual information, highlighting the benefits of translating the full questionchoice sequence as a unified input.

C Human Annotation Guidelines

The rubrics provided in Table 6 were given to native Urdu-speaking annotators to guide their evaluation and selection of the best translation candidate from the ensemble outputs.

D Evaluation Prompts and Configurations

All evaluation prompts are fixed across models and prompting strategies, including Direct, Chain-of-Thought (CoT), and Few-Shot + CoT (FS+CoT), to ensure comparability. Evaluations are conducted using lm-eval-harness with a vLLM backend and tensor parallelism set to 4. Figure 4 shows evaluation prompts for Urdu MGSM, MATH-500, CSQA, and OBQA.

E Extended Results

Algorithm 1 Multi-Model Translation with Heuristic Validation and Consensus

Require: Dataset D ,
translation models $\mathcal{M}=\{\text{IndicTrans2, NLLB, Gemini-2.5-Pro, Qwen-3}\}$,
maximum retries $K = 5$,
minimum length $L_{min} = 10$,
english content ratio threshold $\tau = 0.3$

Ensure: For each sample x , translations $\mathcal{T}_x = \{t^{(1)}, t^{(2)}, t^{(3)}, t^{(4)}, t^{GPT}\}$

```

1: for each sample  $x \in D$  do
2:    $T \leftarrow \emptyset$  {Validated model translations}
3:   for each model  $m \in \mathcal{M}$  in parallel do
4:      $attempt \leftarrow 0$ 
5:      $valid \leftarrow false$ 
6:     while  $attempt < K$  and  $valid = false$  do
7:        $x_m \leftarrow \text{TRANSLATE}(x, m, \text{prompt})$ 
8:       if required fields missing in  $x_m$  then
9:         Record missing field error
10:      end if
11:       $q^{ur} \leftarrow$  Urdu question in  $x_m$ 
12:      if  $q^{ur}$  is empty or  $|q^{ur}| < L_{min}$  then
13:        Record invalid Urdu question
14:      end if
15:      if no Urdu characters in  $q^{ur}$  or English ratio  $> \tau$  then
16:        Record mixed-language question
17:      end if
18:       $C^{ur} \leftarrow$  Urdu answer choices in  $x_m$ 
19:      if  $C^{ur}$  missing or number of choices incorrect then
20:        Record invalid Urdu choices
21:      end if
22:      for each choice  $c \in C^{ur}$  do
23:        if no Urdu characters in  $c$  or English ratio  $> \tau$  then
24:          Record invalid Urdu choice
25:        end if
26:      end for
27:      if answer key missing or invalid then
28:        Record answer key error
29:      end if
30:      if no errors recorded then
31:         $T \leftarrow T \cup \{x_m\}$ 
32:         $valid \leftarrow true$ 
33:      else
34:         $attempt \leftarrow attempt + 1$ 
35:      end if
36:    end while
37:    if  $valid = false$  then
38:       $T \leftarrow T \cup \{x_m\}$ 
39:    end if
40:  end for
41:   $t^{GPT} \leftarrow \text{GPT-5.2}(D, T, \text{prompt})$ 
42:   $\mathcal{T}_x \leftarrow T \cup \{t^{GPT}\}$ 
43:   $\text{Save}(\mathcal{T}_x)$ 
44: end for
```

English question with choices	Translation without Context	Context Aware Translation (with human validation)
The sun is responsible for: a. puppies learning new tricks b. children growing up and getting old c. flowers wilting in a vase d. plants sprouting, blooming and wilting	سورج ذمے دار ہے: a. کتے کے بچوں نے تربیت سیکھ رہے ہیں b. بچے بڑے ہو رہے ہیں اور بوڑھے ہو رہے ہیں c. گلدان میں پھولیں مرجھ رہے ہیں d. پودوں کے اگلے، کھلنے اور مرجھانے کا	سورج ذمے دار ہے: a. کتے کے بچوں کے نئے کرتب سیکھنے کا b. بچوں کے بڑے ہونے اور بوڑھے ہونے کا c. گلدان میں پھولوں کے مرجھانے کا d. پودوں کے اگلے، کھلنے اور مرجھانے کا
Squirrels spend their fall: a. looking for pretty leaves to collect b. stockpiling rocks for fighting in the winter c. stockpiling pecans for the frigid months d. collecting twigs to keep warm	گھری خزاں کا موسم گزارتی ہیں: a. جمع کرنے کے لیے خوبصورت پتے تلاش کر رہی ہیں b. موسم سرما میں لڑنے کے لیے پرندوں کا ذخیرہ کرنا c. شدید سردی کے مہینوں کے لیے پکٹوں کا ذخیرہ کرنا d. گرم رکھنے کے لیے تھیں دار لباس جمع کرنا	گھریاں خزاں کا موسم گزارتی ہیں: a. خوبصورت پتے جمع کرتے ہوئے b. سردیوں میں لڑائی کے لیے ہتھیار ذخیرہ کرتے ہوئے c. شدید سردی کے مہینوں کے لیے پکٹوں کا ذخیرہ کرتے ہوئے d. گرم رہنے کے لیے پٹیاں جمع کرتے ہوئے
Sean was in a rush to get home, but the light turned yellow and he was forced to do what? a. take time b. dawdle c. go slowly d. ocean e. slow down	شان کو گھر پہنچنے کی جلدی تھی، لیکن ٹریفک سگنل کی جی پٹی ہو گئی اور وہ کس چیز پر مجبور ہو گیا؟ a. وقت ضائع کرنا b. ٹال مٹول کرنا c. ڈھیل ڈالنا d. سمندر e. آہستہ چلو	شان کو گھر پہنچنے کی جلدی تھی، لیکن ٹریفک سگنل کی جی پٹی ہو گئی اور وہ کس چیز پر مجبور ہو گیا؟ a. وقت لینا b. وقت ضائع کرنا c. رفتار آہستہ کرنے پر d. سمندر e. رفتار کم کرنا
The runner was in third place, but he pushed harder and thought he might be able to reach second. What was beginning to do? a. near finish line b. finish c. get tired d. gain ground e. trip over	وہ رنر تیسرے نمبر پر تھا، لیکن اُس نے اور زور لگایا اور سوچا کہ وہ شاید دوسرے نمبر پر پہنچ سکے۔ وہ کیا کرنے لگا تھا؟ a. خط اختتام کے قریب b. ختم کرو / مکمل کرو c. تھک جانا d. مقبول ہونا e. ٹھوکر کھانا	وہ رنر تیسرے نمبر پر تھا، لیکن اُس نے اور زور لگایا اور سوچا کہ وہ شاید دوسرے نمبر پر پہنچ سکے۔ وہ کیا کرنے لگا تھا؟ a. اختتامی لکیر کے قریب پہنچنا b. ختم کرنا c. تھک جانا d. سبقت لے جانا e. ٹھوکر کھانا
A person wants to start saving money so that they can afford a nice vacation at the end of the year. After looking over their budget and expenses, they decide the best way to save money is to a. make more phone calls b. quit eating lunch out c. buy less with monopoly money d. have lunch with friends	ایک شخص سال کے آخر میں ایک اچھی چھٹی گزارنے کے لیے پیسے بچانا شروع کرنا چاہتا ہے۔ اپنے بجٹ اور اخراجات کا جائزہ لینے کے بعد، وہ فیصلہ کرتا ہے کہ پیسے بچانے کا بہترین طریقہ یہ ہے کہ a. زیادہ فون کالز کرو b. باہر دوپہر کا کھانا کھانا چھوڑ دینا c. مونوپولی کے پیسوں سے کم خریداری کریں d. دوستوں کے ساتھ دوپہر کا کھانا کھائیں	ایک شخص سال کے آخر میں ایک اچھی چھٹی گزارنے کے لیے پیسے بچانا شروع کرنا چاہتا ہے۔ اپنے بجٹ اور اخراجات کا جائزہ لینے کے بعد، وہ فیصلہ کرتا ہے کہ پیسے بچانے کا بہترین طریقہ یہ ہے کہ a. زیادہ فون کالز کریں b. باہر دوپہر کا کھانا کھانا چھوڑ دیں c. مونوپولی کے پیسوں سے کم خریدیں d. دوستوں کے ساتھ دوپہر کا کھانا کھائیں

Table 5: Examples of with and without Context Translation

Marking Criteria	Evaluation Question	Correct Example	Incorrect Example
1. Response Completeness	Does the translation fully capture the source text without omitting any part?	Original: The cat sat on the mat. Translation: بلی چٹائی پر بیٹھی تھی۔	Original: The cat sat on the mat. Translation 1: بلی بیٹھی تھی۔ (Omitting "on the mat") Translation 2: I don't know (No response)
2. Translation vs Generation	Is the output a translation and not new content generation?	Original: The boy reads a book. Translation: لڑکا کتاب پڑھتا ہے۔	Original: The boy reads a book. Translation: لڑکا کتاب کے بارے میں سوچتا ہے۔ (New content generated)
3. Grammar and Structure	Is the grammar and structure of the Urdu translation accurate?	Original: She went to the market. Translation: وہ بازار گئی۔ Original: I bought an iPhone today. Translation: میں نے آج ایک iPhone خریدا۔	Original: She went to the market. Translation: وہ بازار گیا۔ (Incorrect gender agreement) Original: I bought an iPhone today. Translation: -iphone ایک آج میں نے خریدا (Incorrect placement of English equivalent)
4. Number and Date Formats	Are the number and date formats preserved as in the original?	Original: The event is on 12/12/2024. Translation: یہ تقریب 12/12/2024 کو ہے۔	Original: The event is on 12/12/2024. Translation: یہ تقریب ۱۲/۱۲/۲۰۲۴ کو ہے۔ (Converted to Arabic numerals)
5. Cultural Considerations	Are idioms translated using cultural equivalents, not literal translations?	Original: The ball is in your court. Translation: فیصلہ اب آپ کے ہاتھ میں ہے۔ (Cultural equivalent used)	Original: The ball is in your court. Translation: گیند اب آپ کے کورٹ میں ہے۔ (Literal translation)
6. Common Error Patterns	Does the translation avoid direct transliteration and ensure meaningful translation?	Original: He is an experienced teacher. Translation: وہ تجربہ کار استاد ہے۔	Original: He is an experienced teacher. Translation: وہ ایک ایکسپیرینسڈ ٹیچر ہے۔ (Direct transliteration)
7. Style and Register	Is the translation's tone and formality consistent with the source text?	Original: Please submit your documents at the earliest convenience. Translation: براہ کرم اپنی دستاویزات جلد از جلد جمع کرائیں۔	Original: Please submit your documents at the earliest convenience. Translation: دستاویزات دے دیں۔ (Casual tone used instead of formal)

Table 6: Overview of the Urdu-translated datasets refinement guidelines.

MGSM (Arithmetic Reasoning)

Direct Prompt

مسئلہ:

{Urdu math problem}

براہ کرم صرف آخری عددی جواب دیں۔

Chain-of-Thought (CoT)

مسئلہ:

{Urdu math problem}

مرحلہ وار حل فراہم کریں اور آخر میں جواب دیں۔
میں لکھیں۔ \boxed{} نوٹ: براہ کرم آخری جواب کو

Few-Shot + CoT (FS+CoT)

ذیل میں چند مثالیں دی گئی ہیں۔ انہی کی طرز پر مسئلہ حل کریں۔

{Few-shot Urdu examples}

مسئلہ:

{Urdu math problem}

مرحلہ وار حل لکھیں۔
نوٹ: براہ کرم آخری جواب کو میں لکھیں۔ \boxed{}

MATH-500 (Symbolic Mathematics)

Chain-of-Thought (CoT)

مسئلہ:

{Urdu math problem}

مرحلہ وار حل لکھیں۔
نوٹ: براہ کرم آخری جواب کو میں لکھیں۔ \boxed{}

CommonSenseQA and OpenBookQA

Direct Prompt

سوال:

{Urdu question}

اختیارات:

A) {Option A}

B) {Option B}

C) {Option C}

D) {Option D}

E) {Option E}

درست جواب کا حرف لکھیں۔

Figure 4: Evaluation prompts for Urdu MGSM, MATH-500, CSQA, and OBQA

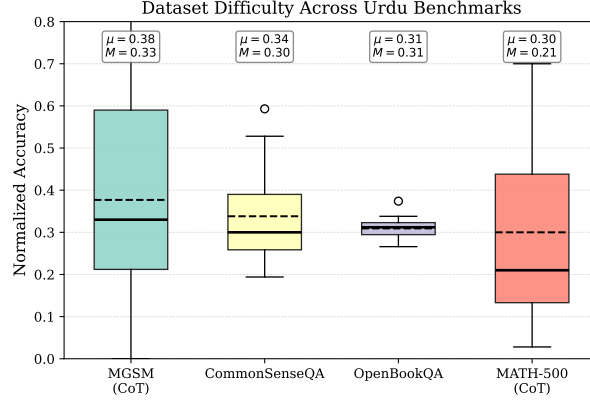


Figure 5: Distribution of model performance across four Urdu benchmarks. Box plots revealing MATH-500 and MGSM as the most challenging tasks with lower median scores and wider performance variance compared to CSQA and OBQA.

Model	L1	L2	L3	L4	L5	Avg
<i>Top-performing LLMs in Math500-Urdu</i>						
Gemma-3-4B-it	65.1	58.9	56.2	39.8	22.4	44.2
Phi-4-mini-reasoning (4B)	76.7	67.8	66.7	69.5	56.72	65.8
Qwen2.5-7B-Instruct	76.7	56.7	46.7	32.8	23.1	41.2
Falcon-h1-7b-instruct	76.7	61.1	56.2	46.9	30.6	49.6
DeepSeek-R1-Distill-Qwen-7B	65.1	57.8	60.9	56.3	51.5	57.0
Gemma2-9B-it	74.4	56.7	44.8	29.7	12.7	37.0
Gemma-3-12B-it	86.1	83.3	80.0	69.5	48.5	70.0
DeepSeek-R1-Distill-Qwen-14B	79.1	70.0	76.2	65.6	63.4	69.2

Lx correspond to increasing MATH-500 difficulty levels.
All results are zero-shot, CoT based prompts.

Table 7: MATH-500 Urdu Performance by Difficulty Level

<7B	7B	8B	>8B
MobileLLM-Pro (1B) (Huber et al., 2025)	Falcon-h1-7B-Instruct (Zuo et al., 2025)	Alif-1.0-8B-Instruct (Shafique et al., 2025)	Gemma2-9B-it (Team et al., 2024)
LLaMA-3.2-3B-Instruct (Grattafiori et al., 2024)	MiMo-7B-Base (Xiaomi et al., 2025)	Aya-Expanse-8B (Dang et al., 2024)	Gemma-3-12B-it (Team et al., 2025)
Gemma-3-4B-it (Team et al., 2025)	Mistral-7B-Instruct (Jiang et al., 2023)	Granite-3.3-8B-Instruct (Granite Team, 2025)	R1-Distill-Qwen-14B (Guo et al., 2025)
Nemotron-Mini-4B (NVIDIA, 2024b)	OLMo-3-7B-Instruct (Olmo et al., 2025)	LLaMA-3.1-8B-Instruct (Weerawardhena et al., 2025)	
R1-Distill-Qwen-1.5B (Guo et al., 2025)	Qwen2.5-7B-Instruct (Yang et al., 2025b)	Ministral-8B-Instruct (Liu et al., 2026)	
Phi-4-Mini-Reasoning (4B) (Xu et al., 2025)	R1-Distill-Qwen-7B (Guo et al., 2025)	Mistral-NeMo-8B-Instruct (NVIDIA, 2024a)	
		RNJ-1-Instruct (8B) (Vaswani et al., 2025)	
		Qalb-1.0-8B-Instruct (Taimoor Hassan et al., 2026)	

Table 8: Categorization of evaluated models by parameter scale

Split	IndicTrans2	NLLB-200	Qwen-3-30B	Gemini-2.5-Pro	GPT-5.2 fused	Human-edited	Total
MGSM (test, n=250)	3	2	1	125	97	22	250
CSQA (val, n=1,221)	74	52	17	826	205	47	1,221
OBQA (test, n=500)	36	27	9	313	86	29	500
MATH-500 (test, n=500)	3	2	51	312	67	65	500

Table 9: Human Validation Win–Loss Statistics Across Translation Candidates. Breakdown of the final Urdu translation sources for splits subjected to the human review pass. The table distinguishes between direct wins by the four base translators, selections of the GPT-5.2 fused candidate, and instances requiring a human-authored edit. Row totals correspond exactly to the split sizes. Held-out CSQA/OBQA training splits are excluded, as they underwent only automatic validation (Algorithm 1).