UrduMMLU: A Massive Multitask Benchmark for Urdu Language Understanding

Anonymous ACL submission

Abstract

Meaningful multilingual evaluation must test models in the target language and educational context. Urdu, spoken by more than 230 million people, lacks a broad MMLU-style benchmark built from native educational sources. We introduce UrduMMLU, a benchmark of 26,431 Urdu MCQs across 26 subjects and five domains, collected from native Urdu MCQ banks and public examination PDFs. Unlike translation-based resources, UrduMMLU covers both standard academic subjects and Urdu- and region-specific content. We label the exam-derived portion through dual human annotation with strict consensus filtering. We evaluate 30 LLMs under English and Urdu prompts, yielding 60 zero-shot evaluations, and further evaluate four open-source LLMs under multiple few-shot settings across both prompt languages. Gemini-3.5-Flash performs best, reaching 90.20% and 90.34% accuracy, while no other model exceeds 85%. The strongest open-source model trails by 7.79 and 8.92 points, and many models lose 25 to 40 points on Urdu-centered Humanities subjects compared with STEM. Few-shot prompting yields only modest gains. UrduMMLU shows that Urdu knowledge remains uneven in current LLMs, especially for regionally grounded content. We release our data and code at http://anonymous.for.review.

1 Introduction

Evaluating the knowledge and reasoning abilities of large language models (LLMs) has become central to natural language processing (NLP). Benchmarks such as MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024) are widely used for this purpose, but they are in English and largely reflect English-language educational and cultural contexts. This limits their ability to test whether model competence transfers across language, script, and regional knowledge.

The issue is especially important for Urdu, a language spoken by over 230 million people, with a long literary and educational tradition, but limited broad-coverage evaluation resources. Existing Urdu benchmarks focus mainly on reading comprehension, syntactic diagnostics, task-level NLP evaluation, or translated reasoning benchmarks (Kazi and Khoja, 2026; Kazi et al., 2025; Adeeba et al., 2025; Tahir et al., 2025; Shafique et al., 2026). Multilingual benchmarks that include Urdu, such as MMLU-ProX (Xuan et al., 2025), Global-MMLU (Singh et al., 2025), and IndicMMLU-Pro (KJ et al., 2025), also rely mainly on translated questions. As a result, they only partially capture knowledge grounded in Urdu-medium education, Urdu literature, local history, religious studies, and civic curricula.

Recent language-specific benchmarks such as ArabicMMLU (Koto et al., 2024), CMMLU (Li et al., 2024), IndoMMLU (Koto et al., 2023), KMMLU (Son et al., 2025), and KazMMLU (Togmanov et al., 2025) highlight the importance of evaluation grounded in local educational material. Following this direction, we introduce UrduMMLU, the first broad-coverage, natively written MMLU-style benchmark for Urdu. UrduMMLU contains 26,431 MCQs across 26 subjects and five domains, collected from Urdu MCQ banks and public SSC/HSSC examination PDFs, and combines answer-labeled questions with exam-derived questions annotated through dual human annotation and strict consensus filtering, and covers both standard academic subjects and Urdu- and region-specific content. Figure 1 summarizes the resulting subject distribution.

We evaluate 30 open-source and closed-source LLMs on UrduMMLU under English and Urdu prompts, yielding 60 zero-shot evaluations, and further evaluate four open-source LLMs in 1-, 3-, and 5-shot settings. Gemini-3.5-Flash (Google DeepMind, 2026) achieves the highest overall ac-

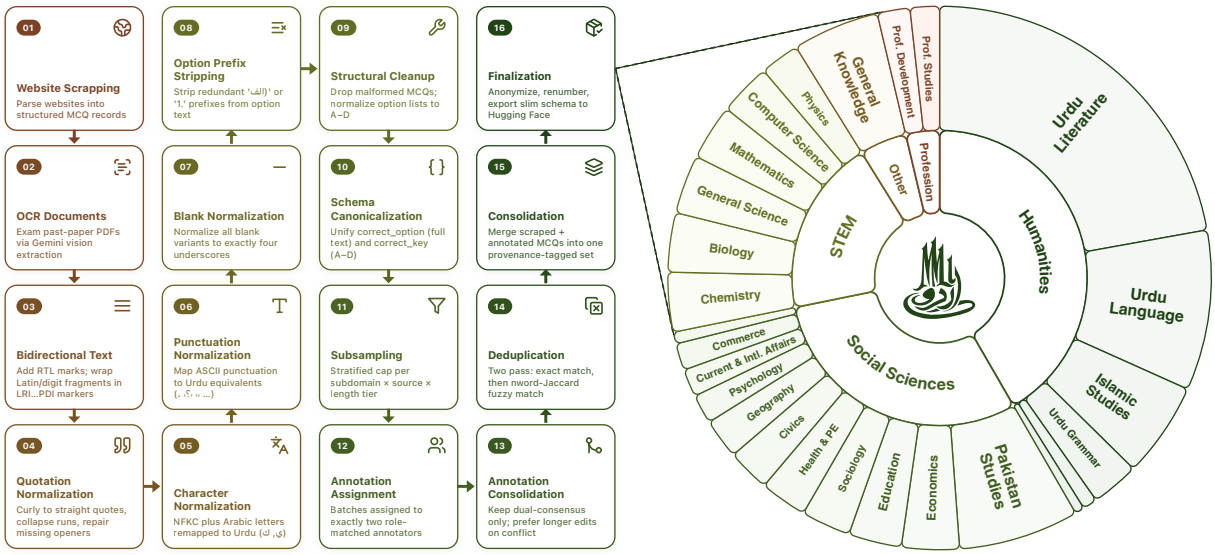


Figure 1: The 16-stage UrduMMLU construction pipeline (left) and the resulting 26,431-MCQ benchmark broken down by 5 domains and 26 subdomains (right); wedge size is proportional to MCQ count.

curacy at 90.20% and 90.34%, while the strongest open-source model trails by 7.79 and 8.92 points. Across models, performance remains substantially higher on STEM subjects than on Urdu-centered Humanities, with many systems losing 25 to 40 points on Urdu literature, Urdu language, and Islamic studies. These results show that strong English-centered benchmark performance does not reliably transfer to Urdu educational and cultural knowledge.

The main contributions of this work are:

- We introduce UrduMMLU, a natively written Urdu MMLU-style benchmark with 26,431 MCQs across 26 subjects and five domains, covering both standard academic subjects along with Urdu- and region-specific knowledge.
- We produce human-annotated gold answers for the exam-derived portion of the benchmark using dual annotation and strict consensus filtering.
- We conduct 60 zero-shot evaluations across 30 open-source and closed-source LLMs under English and Urdu prompt settings, and 24 additional few-shot evaluations across four open-source LLMs.
- We release the dataset and evaluation code to support future work on Urdu-capable language models.

2 Related Work

Urdu evaluation resources: Existing Urdu resources cover reading comprehension, cross-lingual question answering , syntax, and task-level NLP. UQuAD+ (Kazi and Khoja, 2026) provides annotated Urdu reading comprehension, while Kazi et al. (2025) study Urdu-English QA with UQuAD1.0 (Kazi and Khoja, 2021) and SQuAD2.0 (Rajpurkar et al., 2018). UrBLiMP (Adeeba et al., 2025) evaluates Urdu syntax via minimal pairs, and Tahir et al. (2025) benchmark models across Urdu NLP tasks. For reasoning, UrduBench (Shafique et al., 2026) translates MGSM (Shi et al., 2023), CommonsenseQA (Talmor et al., 2019), OpenBookQA (Mihaylov et al., 2018), and MATH-500 (Lightman et al., 2024) into Urdu, and UrduFactCheck (Ahmad et al., 2025) targets factual QA. These resources remain task-specific, diagnostic, or translation-derived. In contrast, UrduMMLU evaluates broad educational knowledge using questions originally written for Urdu-speaking educational settings.

Multilingual benchmarks: MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024) are widely used for evaluating general knowledge and reasoning. Several multilingual extensions adapt these benchmarks through translation. MMLU-ProX (Xuan et al., 2025) extends MMLU-Pro to 29 languages using LLM-based translation and expert review, while Global-MMLU (Singh et al., 2025) studies cultural and linguistic bias

in multilingual evaluation. IndicMMLU-Pro (KJ et al., 2025) adapts MMLU-Pro to nine Indic languages, including Urdu. Other multilingual exam-based resources, such as EXAMS (Hardalov et al., 2020), INCLUDE (Romanou et al., 2025), and MILU (Verma et al., 2025), collect examination questions across multiple languages and regions. However, Urdu still appears primarily in translated or cross-lingual settings rather than through a dedicated native benchmark.

Localized MMLU-style benchmarks: Recent work increasingly builds MMLU-style benchmarks from local educational material instead of translating English benchmarks. ArabicMMLU (Koto et al., 2024), CMMLU (Li et al., 2024), IndoMMLU (Koto et al., 2023), KMMLU (Son et al., 2025), and KazMMLU (Togmanov et al., 2025) show that language-specific curricula and regional cultural knowledge remain important for evaluating LLMs beyond English. UrduMMLU follows this direction for Urdu by combining regional SSC/HSSC examination material, native Urdu MCQ banks, human annotation for exam-derived questions, and broad coverage of both standard academic subjects and Urdu- and Pakistan-specific knowledge.

3 UrduMMLU

UrduMMLU is a broad-coverage benchmark for evaluating knowledge and reasoning in Urdu. Unlike translation-based multilingual benchmarks, UrduMMLU draws its questions directly from Urdu educational and examination material. The benchmark contains 26,431 MCQs across 26 subdomains and five domains, covering both standard academic subjects and Urdu- and region-specific content such as Urdu literature, Urdu language, Islamic studies, and Pakistan studies. Appendix A.1 and Figure 7 provide detailed benchmark statistics and subdomain distributions. We collect questions from Urdu MCQ banks and public SSC/HSSC examination PDFs, and produce gold answers for exam-derived questions through dual human annotation with strict consensus filtering. We design UrduMMLU around broad subject coverage, faithful representation of Urdu educational material, and reliable multiple-choice evaluation through clean text extraction, normalized metadata, and verified gold labels. Figure 1 summarizes the overall construction pipeline.

3.1 Data Sources

We collect candidate questions from two source families. The first consists of public SSC and HSSC examination PDFs from Pakistan covering school- and high school-level subjects such as mathematics, physics, chemistry, biology, computer science, Urdu, Islamic studies, Pakistan studies, civics, education, and economics. The second consists of native Urdu MCQ websites that publish answer-labeled questions for examination preparation. Together, these sources allow UrduMMLU to cover both globally shared academic subjects and region-specific educational content taught in Urdu-medium curricula. We treat all collected items as candidates and include them in the final benchmark only after cleaning, answer annotation or verification, deduplication, and release packaging.

3.2 Raw MCQ Extraction

For PDF-based sources, we use a multi-stage extraction pipeline to recover Urdu MCQs from heterogeneous examination layouts. We first convert each PDF into page images and use Claude Opus 4.7 (Anthropic, 2026) as OCR to classify each page, filtering out English-only pages, non-MCQ pages, answer keys, and unrelated material. For the remaining pages, we extract question stems, answer options, source metadata, and page-level provenance using a vision-language OCR procedure. We design the extraction prompt specifically for Urdu examination documents. The prompt preserves Urdu question text, answer options, poetry, quotations, and other context required to answer the question correctly. In bilingual material, we ignore English text unless it forms a structural part of the Urdu question, and we discard unreadable questions rather than reconstructing missing content. For web-based sources, we directly scrape question stems, answer options, category labels, and answer keys when available.

3.3 Metadata and Schema Normalization

The collected sources use heterogeneous category names, grade labels, and answer formats, so we normalize all examples into a unified representation. We map source-specific labels to a controlled set of subdomains. For example, we map variants such as Everyday Science and General Science to general science, and mathematics-related labels such as maths, General Mathematics, and riazi

to mathematics. For curriculum-derived material, we normalize grade labels into regional examination levels: Grade 9 to SSC-I, Grade 10 to SSC-II, Grade 11 to HSSC-I, and Grade 12 to HSSC-II. Table 13 in Appendix C.1 summarizes the final domain hierarchy, subdomains, acronyms, and examination levels covered in UrduMMLU.

We also canonicalize the MCQ schema to support consistent evaluation. Each released item stores a question, four answer options, normalized domain and subdomain labels, academic level, source metadata, and answer annotations. We remove ambiguous index-based answer fields because different sources follow different option-ordering conventions.

3.4 Cleaning and Quality Control

We apply several cleaning and validation steps to reduce noise from OCR, web scraping, and heterogeneous source formatting. First, we normalize Urdu text representation through right-to-left display normalization, punctuation and quote normalization, standardization of fill-in-the-blank markers, and Unicode normalization for visually similar Arabic and Urdu codepoints. We then enforce structural validity by removing items with missing, empty, duplicate, or malformed answer options, discarding examples with invalid option counts, and standardizing option fields into a consistent schema format.

Next, we deduplicate the candidate pool. We merge exact duplicates with consistent answers while preserving source provenance and discard duplicate groups with conflicting labels. To handle OCR and wording variations, we additionally apply conservative near-duplicate filtering based on high question-token overlap together with answer-option overlap. Finally, we remove residual non-Urdu artifacts, including a small number of English OCR artifacts that survived earlier filtering stages. We use the resulting cleaned pool for annotation, answer verification, and final benchmark construction.

3.5 Human Annotation

The exam-derived portion of UrduMMLU did not include answer keys, so we produced gold labels through human annotation. We organized annotation batches by subdomain and assigned each item to two annotators with relevant subject familiarity. Annotators selected the correct answer, marked questions as unsure, flagged problematic items, and could suggest light corrections to question text, answer options, and subdomain labels. Seventeen annotators participated in the process; 94.1% identified Urdu as their native language, and most held either a bachelor's degree (47.1%) or a master's degree (41.2%). Appendix B.1 reports full demographic and satisfaction details. We applied a strict consensus rule and retained an item only when both annotators selected the same valid answer without flags or unsure labels.

Provenance	Count	%
Answer-labeled web MCQs	13,672	51.7
Exam-extracted annotated MCQs	12,759	48.3
Total	**26,431**	**100.0**

Table 1: Composition of the final UrduMMLU benchmark by source type. Web-derived questions use human-validated published answers, while exam-derived questions use dual human annotation with strict consensus filtering.

In total, 17,565 exam-extracted MCQs entered annotation, and 14,459 satisfied the consensus criteria. The main exclusion reasons included answer disagreement (1,611 items), flags (1,247), unsure selections (243), and incomplete annotations (5). Annotators also corrected 141 domain labels during the process. Overall observed agreement reached 89.98%, with simplified Cohen's $\kappa = 0.8663$. After verification, deduplication, and release packaging, the final benchmark retained 12,759 human-annotated exam-derived questions. Annotators additionally verified the correctness of pre-existing answer labels for web-derived MCQs.

3.6 Final Benchmark

The final release of UrduMMLU contains 26,431 Urdu MCQs after cleaning, annotation, answer verification, deduplication, and release packaging. Answer-labeled web sources contribute 13,672 questions, while exam-derived sources contribute 12,759 questions annotated through dual human labeling and strict consensus filtering (Table 1). Appendix A reports statistics for the larger cleaned candidate pool before annotation and final selection. UrduMMLU spans 26 subdomains grouped into five domains: STEM, Humanities, Social Sciences, Profession, and Other. Humanities and Social Sciences constitute the largest portions of the benchmark, reflecting strong coverage of Urdu language, Urdu literature, Islamic studies, Pakistan

	Group	Question	Answer
Domain	Humanities	41.8	10.2
	Social Sciences	43.0	8.6
	STEM	46.0	7.9
	Other	48.7	9.7
	Profession	41.3	8.2
Level	SSC-I	43.5	9.3
	SSC-II	44.0	9.0
	HSSC-I	41.8	9.2
	HSSC-II	42.8	9.1

Table 2: Average character length of questions and correct answers in UrduMMLU, grouped by domain and academic level. Values denote mean character counts.

Domain	Questions	Subdomains
STEM	5,113	6
Humanities	11,010	6
Social Sciences	7,968	11
Profession	975	2
Other	1,365	1
Total	26,431	26

Table 3: Distribution of questions and subdomains across the five domains in UrduMMLU.

studies, and related educational content. Table 3 summarizes the domain-level composition, while Table 13 lists the corresponding subdomains and academic levels. Table 2 reports average question and answer lengths, and Appendix C describes the dataset schema.

4 Experiments

We evaluate UrduMMLU with generation-based protocols that require each model to select an answer option for an Urdu MCQ. We run a large zero-shot evaluation across 30 open- and closed-source LLMs using both English and Urdu instruction prompts. We also run a focused few-shot study on four open-source LLMs using 1-, 3-, and 5-shot settings in both prompt languages. All evaluations use the same benchmark format and accuracy metric, which allows direct comparison across model families, prompt languages, and shot settings.

4.1 Models, Prompting, and Decoding

We evaluate 30 LLMs spanning a broad range of model sizes, access regimes, and training backgrounds, including proprietary API systems, open-weight multilingual instruction-tuned models, compact models, mixture-of-experts architectures, reasoning-oriented variants, and Urdu- or regionally specialized models. Table 14 in Appendix D.1 lists the full model roster. This setup allows us to compare open-source and closed-source systems and examine transfer to native Urdu educational content.

We evaluate each model with English and Urdu prompt templates while keeping the Urdu question stem, answer options, and response format fixed. The two settings differ only in the instruction language and field labels. Appendix D.2 provides the prompt templates in Figures 16 and 17. We use temperature 0 whenever deterministic decoding is available and otherwise follow provider-specific reasoning settings. We set the maximum output length to 4096 tokens, batch API requests with a concurrency of 10, and decode locally evaluated Hugging Face models greedily.

4.2 Evaluation Protocols

Zero-shot evaluation: We use a generation-based zero-shot protocol in which each input contains the domain, subdomain, academic level, Urdu question, and labeled answer options. We evaluate all 30 models under both English and Urdu prompt templates, resulting in 60 zero-shot runs. Since the question and answer options remain unchanged across settings, this protocol isolates the effect of instruction language on the same Urdu MCQs.

Few-shot evaluation: We conduct a controlled few-shot study on four open-source LLMs under 1-, 3-, and 5-shot settings with both English and Urdu prompts, yielding 24 runs. We reserve 200 validated MCQs as a demonstration pool for lm-evaluation-harness (Gao et al., 2024), which we use only for prompt construction, demonstration sampling, and execution management. All models generate structured answers that we parse and compare against the gold labels.

4.3 Evaluation Measure

We use accuracy as the primary evaluation measure by comparing the generated answer with the gold label. Alongside accuracy, we report the invalid-output rate, defined as the percentage of unparsable, malformed, or error outputs. We further analyze results by domain, subdomain, academic level, prompt language, and model category to examine performance differences across standard academic and Urdu- or region-specific subjects.

Model	English Prompt (Accuracy % ↑)							Urdu Prompt (Accuracy % ↑)						
	STEM	H	SS	P	O	Overall	Inv% ↓	STEM	H	SS	P	O	Overall	Inv% ↓
Open-source Models: > 25B Parameters														
DeepSeek-V4-Flash	**97.49**	**68.71**	**90.05**	**89.85**	**86.44**	**82.41**	0.1	**97.57**	**67.32**	**88.95**	**89.44**	**84.88**	**81.42**	0.3
Gemma-4-26B-A4B-IT	85.92	57.43	73.83	75.49	70.55	69.23	<0.1	87.21	57.73	75.41	77.82	71.79	70.23	<0.1
Gemma-4-31B-IT	93.33	63.62	81.70	83.69	79.49	76.38	<0.1	93.86	63.25	82.06	84.10	78.39	76.39	<0.1
LLaMA-3.3-70B	81.32	56.30	75.45	76.51	73.77	68.56	0	78.39	56.10	73.43	71.28	71.65	67.00	<0.1
Qwen3.6-27B	90.46	55.95	73.19	74.77	69.60	69.22	0	91.12	55.71	74.55	74.46	70.55	69.70	0
Qwen3.6-35B-A3B	88.77	56.81	73.52	75.49	69.82	69.39	0	96.32	58.12	84.43	84.50	77.99	75.46	0.4
LLaMA-4-Scout-17B-16E	84.14	57.03	71.20	74.15	69.52	67.82	0	85.59	56.55	72.49	72.62	69.01	68.20	<0.1
LLaMA-4-Maverick-17B-128E	91.98	63.27	80.11	81.64	80.66	75.47	<0.1	92.38	63.25	81.30	79.59	80.81	75.83	<0.1
Open-source Models: ≤ 25B Parameters														
BLOOMZ-1.1B	23.52	27.21	24.39	26.06	25.63	25.52	0.5	24.53	25.83	25.34	23.49	24.47	25.27	<0.1
BLOOMZ-1.7B	28.94	27.70	33.33	37.17	31.43	30.19	2.5	28.74	28.76	33.84	32.60	27.51	30.57	24.8
BLOOMZ-3B	27.89	30.25	33.26	32.34	28.39	30.68	6.5	26.56	27.70	30.35	32.64	26.75	28.62	74.2
BLOOMZ-7B	27.75	27.81	33.56	31.83	28.11	29.73	11.2	29.24	30.88	33.35	33.45	30.73	31.36	34.6
Gemma-2-9B-IT	67.08	**47.21**	58.40	**60.00**	**54.58**	**55.28**	0	69.02	**48.08**	60.53	**62.15**	55.82	56.80	0
Gemma-3-4B-IT	49.97	37.87	47.68	47.79	45.57	43.93	0	51.79	38.27	48.69	50.15	46.23	44.88	0
LLaMA-3.2-3B	37.12	26.78	38.06	36.00	35.60	32.98	0	37.24	29.32	39.97	38.36	38.02	34.85	0
LLaMA-3.1-8B	46.96	36.38	49.66	48.51	44.54	43.30	0	46.49	37.61	49.85	49.54	44.98	43.84	0
Ministral-3-3B	55.04	43.69	48.99	49.08	47.91	47.90	<0.1	57.25	43.07	52.07	52.26	48.86	49.16	<0.1
Ministral-3-8B	67.81	45.27	**58.99**	59.59	54.43	54.77	0	71.37	45.74	**62.74**	61.54	**57.00**	**56.99**	0
Phi-4-mini	37.07	28.85	38.10	40.41	32.67	33.85	<0.1	37.08	28.70	38.97	38.67	35.85	34.15	<0.1
Phi-3.5-mini	33.76	22.89	30.90	32.21	28.28	28.03	0	33.83	27.25	31.75	34.22	30.57	30.31	0.4
Qwen3-4B-Instruct-2507	68.61	42.33	51.48	52.62	47.84	50.84	0	68.75	43.00	53.30	53.23	47.69	51.70	0
Qwen3-8B	**70.70**	39.21	53.99	56.62	50.18	50.97	0	**74.37**	30.87	56.54	57.48	49.38	48.97	0.5
Proprietary Models														
Claude-Haiku-4.5	90.49	58.57	75.86	75.90	74.14	71.40	0.1	91.96	59.31	77.06	78.26	74.29	72.45	<0.1
Claude-Sonnet-4.6	96.34	72.69	87.36	87.18	86.01	82.91	0	96.26	72.69	87.53	87.18	85.86	82.94	0
Gemini-3.1-Flash-Lite	96.85	74.20	90.01	90.26	86.08	84.56	<0.1	97.09	74.38	90.10	90.36	85.57	84.68	<0.1
Gemini-3.5-Flash	**97.75**	**84.98**	**92.15**	**92.10**	**91.43**	**90.20**	0.1	**97.81**	**85.31**	**92.14**	**91.38**	**91.72**	**90.34**	<0.1
GPT-5.4-mini	88.34	62.82	77.52	79.08	75.24	73.43	0	88.25	62.35	78.45	79.59	75.09	73.51	0
GPT-5.4	95.13	69.29	86.35	85.85	84.10	80.81	0	97.40	74.82	89.37	87.08	83.74	84.53	0.4
Urdu Models														
Qalb-1.0-8B	38.18	**29.99**	37.68	**40.31**	39.56	**34.77**	0	**36.26**	**32.72**	**37.78**	**39.34**	42.55	**35.52**	11.3
Alif-1.0-8B	**41.09**	25.74	**41.40**	39.87	**41.04**	34.72	0.6	33.27	29.00	36.26	37.67	**42.93**	32.68	12.6

Table 4: **Model performance on UrduMMLU.** Accuracy (%) under English and Urdu prompts across five domains and overall average. **Inv%** denotes the percentage of unparsable or malformed outputs (lower is better). Boxed values mark the best overall score per column, while bold values indicate the best score within each model group.

5 Results

We first analyze overall zero-shot performance across all evaluated models and then examine how performance changes across domains, prompt languages, model scales, and few-shot settings. Table 4 summarizes zero-shot accuracy on UrduMMLU under English and Urdu prompts, together with invalid-output rates. We focus on four main findings: a small set of models performs strongly, STEM transfers much better than Urdu-centered Humanities, prompt language has limited effect for most models, and few-shot prompting gives modest but insufficient gains. Appendix F provides per-subdomain and per-level results.

5.1 Overall Model Performance

Gemini-3.5-Flash achieves the best overall performance, reaching 90.20% under the English prompt and 90.34% under the Urdu prompt. No other model exceeds 85% accuracy. The next strongest group includes Gemini-3.1-Flash-Lite (Google, 2026a), GPT-5.4 (Singh et al., 2026), Claude-Sonnet-4.6 (Anthropic, 2026), and DeepSeek-V4-Flash (DeepSeek-AI, 2026), with DeepSeek-V4-Flash giving the strongest open-source result at 82.41/81.42. Even so, it trails Gemini-3.5-Flash by 7.79 and 8.92 points. Performance drops sharply outside this top-tier. In the ≤ 25B open-source group, Gemma-2-9B-IT (Team et al., 2024) and Ministral-3-8B (Liu et al., 2026) lead at roughly 55–57%, while Qwen3-4B and Qwen3-8B (Yang et al., 2025) remain near 50%. BLOOMZ (Muennighoff et al., 2023) models stay close to the 25% random baseline despite multilingual pretraining that includes Urdu. The two Urdu-specific models, Qalb-1.0-8B (Hassan et al., 2026) and Alif-1.0-8B (Shafique et al., 2025), also remain below 36%, showing that Urdu-focused tuning alone does not produce strong broad-coverage Urdu knowledge.

5.2 Domain-Level Performance

Domain-level results reveal the clearest pattern in UrduMMLU. Nearly every model that performs above chance scores highest on STEM and lowest on Humanities. Under the Urdu prompt, Gemini-3.5-Flash scores 97.81% on STEM and

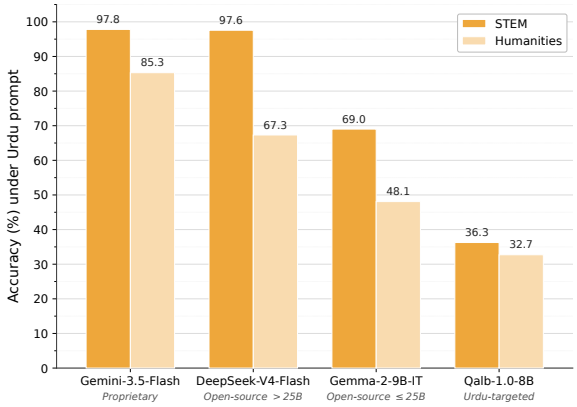


Figure 2: STEM and Humanities accuracy on UrduMMLU under the Urdu prompt for top representative models from each model group. All models score lower on Humanities.

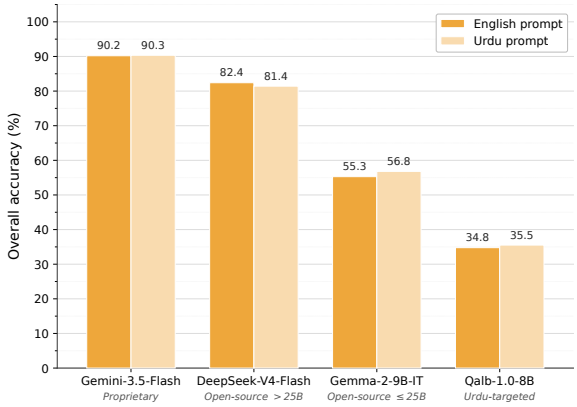


Figure 3: Overall accuracy on UrduMMLU under English and Urdu prompts for representative models from each model group. Prompt language has only a small effect on overall performance.

85.31% on Humanities, a gap of 12.50 points, while DeepSeek-V4-Flash drops from 97.57% to 67.32%. GPT-5.4 and Claude-Sonnet-4.6 lose more than 22 points, and several Qwen models lose more than 35 points between the two domains. Figure 2 illustrates this trend for representative top-performing models from each section of Table 4.

This pattern highlights the main challenge that UrduMMLU exposes. STEM questions rely on scientific and mathematical concepts that transfer more consistently across languages, whereas the Humanities domain requires stronger coverage of Urdu literature, Urdu language, Islamic studies, ethics, and other culturally grounded subjects. Many models can process Urdu well enough to answer science questions, but they struggle on Urdu literary, linguistic, and religious content. Social Sciences generally falls between STEM and Humanities, reflecting a mix of globally shared and region-specific knowledge.

5.3 Prompt-Language Effects

Changing the prompt language usually has little effect on overall accuracy. Figure 3 compares representative models from each group in Table 4. The English and Urdu prompt results nearly overlap for all four: Gemini-3.5-Flash changes by $+0.14$ points, DeepSeek-V4-Flash by -0.99, Gemma-2-9B-IT by $+1.52$, and Qalb-1.0-8B by $+0.75$. The full table shows the same general pattern. Most models move by less than one point when the prompt changes from English to Urdu. A few exceptions show model-specific behavior: Qwen3.6-35B-A3B gains 6.07 points with the Urdu prompt, and GPT-5.4 gains 3.72 points, while Qwen3-8B and Alif-1.0-8B lose about two points. These shifts remain much smaller than the STEM-Humanities gaps for capable models. We therefore attribute the main difficulty of UrduMMLU to the content of the Urdu questions rather than to the language of the instruction wrapper.

5.4 Invalid-Output Rates

Invalid-output rates provide a useful complement to accuracy. Most modern proprietary and open-source models follow the required response format, with invalid-output rates below 0.1%. This includes Gemini, Claude, GPT, Gemma, LLaMA (Meta, 2025), Ministral, and most larger Qwen models. Smaller and weaker models show a different pattern. Under the Urdu prompt, BLOOMZ-3B returns invalid answers for 74.2% of examples, BLOOMZ-7B for 34.6%, and BLOOMZ-1.7B for 24.8%. The Urdu-targeted models also degrade under the Urdu prompt: Qalb-1.0-8B reaches an invalid-output rate of 11.3%, and Alif-1.0-8B reaches 12.6%. These failures matter because accuracy over parseable outputs can hide severe formatting breakdowns. Reporting invalid outputs separately shows which models can both answer Urdu questions and follow Urdu evaluation instructions reliably; Appendix G gives one real example of each failure mode.

5.5 Few-Shot Performance

Table 5 and Figure 4 summarize the few-shot evaluation for LLaMA-3.1-8B (Grattafiori et al., 2024), Gemma-3-4B-IT, Qwen3-8B, and Qwen3-4B-Instruct-2507. We evaluate each model at 1-, 3-

Model	English Prompt (Accuracy % ↑)				Urdu Prompt (Accuracy % ↑)			
	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot
LLaMA-3.1-8B	43.30	44.93 (+1.63)	46.09 (+2.79)	46.59 (+3.29)	43.84	45.08 (+1.24)	46.32 (+2.48)	46.10 (+2.26)
Gemma-3-4B-IT	43.93	45.91 (+1.98)	46.07 (+2.14)	46.27 (+2.34)	44.88	46.40 (+1.52)	46.36 (+1.48)	45.99 (+1.11)
Qwen3-8B	50.97	50.23 (−0.74)	52.27 (+1.30)	53.21 (+2.24)	48.97	51.86 (+2.89)	53.24 (+4.27)	53.49 (+4.52)
Qwen3-4B-Instruct-2507	50.84	52.56 (+1.72)	53.40 (+2.56)	53.65 (+2.81)	51.70	52.06 (+0.36)	52.89 (+1.19)	52.94 (+1.24)
Mean	47.26	48.41 (+1.15)	49.46 (+2.20)	49.93 (+2.67)	47.35	48.85 (+1.50)	49.70 (+2.35)	49.63 (+2.28)

Table 5: **Few-shot performance on UrduMMLU.** Accuracy (%) at 0-, 1-, 3-, and 5-shot settings under English and Urdu instruction prompts. Coloured deltas in parentheses are relative to the 0-shot baseline of the same model under the same prompt; green indicates a gain and red indicates a loss. The **Mean** row aggregates the four evaluated models per shot setting.

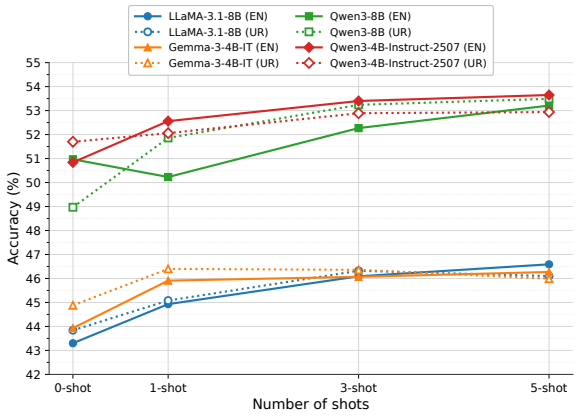


Figure 4: Few-shot accuracy on UrduMMLU for four open-source models under English (solid) and Urdu (dotted) prompts. Few-shot prompting improves most settings, but the gains remain modest.

, and 5-shot under both English and Urdu prompts using validated demonstrations from the held-out few-shot pool. Few-shot prompting improves almost every setting: 23 of the 24 few-shot configurations outperform their corresponding zero-shot baselines. Under the English prompt, the mean gains reach +1.15, +2.20, and +2.67 points at 1-, 3-, and 5-shot, while the Urdu prompt reaches gains of +1.50, +2.35, and +2.28. Qwen3-8B under the Urdu prompt shows the largest improvement, increasing from 48.97% at zero-shot to 53.49% at five-shot.

Despite these gains, few-shot prompting does not change the overall ranking. Even at five-shot, all four models remain well below the $\geq$ 25B open-source tier and far behind the proprietary models. Few-shot prompting also reduces prompt-language differences, with every English-Urdu gap remaining within 0.71 points at five-shot. Overall, demonstrations improve format consistency and calibration, but they do not compensate for missing Urdu-specific knowledge.

6 Conclusion and Future Work

We introduced UrduMMLU, a broad-coverage, natively written MMLU-style benchmark for Urdu with 26,431 MCQs across 26 subjects and five domains, collected from Urdu MCQ banks and public SSC/HSSC examination PDFs. The benchmark combines standard academic subjects with Urdu- and region-specific content and uses dual human annotation with strict consensus filtering for exam-derived questions. Evaluating 30 open-source and closed-source LLMs under English and Urdu prompts reveals a clear gap in current model capability. Gemini-3.5-Flash performs best at 90.20% and 90.34% accuracy, while the strongest open-source model trails by 7.79 and 8.92 points. Models perform substantially better on STEM than on Urdu-centered Humanities, often losing 25 to 40 points on Urdu literature, Urdu language, and Islamic studies. Prompt language has limited effect for most models, and few-shot prompting yields only modest gains. Overall, UrduMMLU shows that strong English-centered benchmark performance does not ensure reliable Urdu educational and cultural knowledge and provides a stronger foundation for evaluating Urdu-capable LLMs.

Future work can extend UrduMMLU beyond MCQ-based evaluation through open-ended generation, summarization, and translation tasks. Expanding the benchmark to include Indian Urdu curricula, undergraduate material, professional examinations, and dialectal content would further broaden its scope. Psychometrics also remains difficult for all evaluated models, motivating future Urdu reasoning benchmarks focused on analogies, logical patterns, and aptitude-style tasks. Finally, the weak performance of Urdu-targeted models highlights the need for stronger continued pretraining and instruction tuning on native Urdu educational and literary material.

Limitations

Curriculum and source scope: UrduMMLU focuses on the Pakistani SSC/HSSC curriculum and a limited set of Urdu MCQ websites targeting the same educational setting. Strong performance therefore reflects competence on Pakistani secondary-school material rather than Urdu in its full linguistic diversity. The benchmark does not cover undergraduate content, Indian Urdu curricula, dialectal variation, or Urdu–English code-switching. Although we reduce source skew through deduplication, annotation, and balancing, Ustad 360 still contributes 58.8% of the cleaned candidate pool.

Format and ceiling effects: UrduMMLU uses a four-option multiple-choice format and therefore does not evaluate open-ended writing, summarization, translation quality, long-form reasoning, or conversational ability. Psychometrics partially offsets this limitation by introducing reasoning-heavy questions; however, no model exceeds 60% accuracy on this subdomain. Future work should extend evaluation toward more open-ended Urdu tasks.

Prompt-language and few-shot effects are limited: English and Urdu instruction wrappers, together with 1-, 3-, and 5-shot prompting, change accuracy by only a few points and rarely alter model rankings. We also do not evaluate option-order robustness, chain-of-thought prompting, or specialized reasoning modes. Our setup therefore prioritizes consistency and comparability over fully optimized prompting configurations.

Ethical Statement & Broad Impact

We develop UrduMMLU to support more inclusive multilingual evaluation for Urdu, a widely spoken but underrepresented language in NLP research. The benchmark draws from publicly available educational and examination material and aims to improve evaluation coverage beyond English-centered benchmarks.

Transparency and Reproducibility: We release the dataset, evaluation code, and prompting protocols to support reproducible research and transparent comparison across models. We also document the dataset construction pipeline, annotation procedure, and evaluation setup in detail.

Annotation and Data Quality: We use dual human annotation with strict consensus filtering for exam-derived questions and additionally verify answer labels for web-derived items. We further apply cleaning, deduplication, and normalization procedures to reduce OCR noise, malformed questions, and metadata inconsistencies.

Bias and Scope Limitations: UrduMMLU primarily reflects the Pakistani SSC/HSSC curriculum and the educational content available through Urdu MCQ resources. As a result, it may not fully represent other Urdu-speaking communities, dialects, or educational systems. The benchmark also contains culturally and regionally grounded subjects such as Islamic studies and Pakistan studies that reflect the underlying curriculum sources.

Broader Impact: We hope UrduMMLU supports the development of stronger Urdu-capable language models and more representative multilingual evaluation. At the same time, benchmark scores should not be interpreted as complete measures of reasoning ability, factual reliability, or cultural understanding beyond the educational scope represented in the dataset.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Farah Adeeba, Brian Dillon, Hassan Sajjad, and Rajesh Bhatt. 2025. UrBLiMP: A benchmark for evaluating the linguistic competence of large language models in Urdu.

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 *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22788–22802, Suzhou, China. Association for Computational Linguistics.

Anthropic. 2025. Claude Haiku 4.5. https://www.anthropic.com/claude/haiku. Accessed 2026-05-23.

Anthropic. 2026. Claude Opus 4.7 system card. https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-05-26.

Anthropic. 2026. Claude Sonnet 4.6 System Card. https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf. Accessed 2026-05-23.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence.

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.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Google. 2026a. Gemini 3.1 Flash-Lite. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite. Accessed 2026-05-23.

Google. 2026b. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Accessed 2026-05-23.

Google DeepMind. 2026. Gemini 3.5 Flash Model Card. https://deepmind.google/models/model-cards/gemini-3-5-flash/. Accessed 2026-05-23.

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, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2601.08141.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Samreen Kazi and Shakeel Khoja. 2021. UQuAD1.0: Development of an Urdu question answering training data for machine reading comprehension. *arXiv preprint arXiv:2111.01543*.

Samreen Kazi and Shakeel Khoja. 2026. UQuAD+: Benchmark dataset for Urdu machine reading comprehension. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 25(2).

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, Abu Dhabi. Association for Computational Linguistics.

Sankalp KJ, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. 2025. IndicMMLU-Pro: Benchmarking Indic large language models on multi-task language understanding. *Preprint*, arXiv:2501.15747.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. 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 *The Twelfth International Conference on Learning Representations*.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Meta. 2024a. Llama 3.2 3b instruct model card. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct. Accessed 2026-05-23.

Meta. 2024b. Llama 3.3 70b instruct model card. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed 2026-05-23.

Meta. 2025. Llama 4 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md. Accessed 2026-05-23.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-Mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs. *Preprint*, arXiv:2503.01743.

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, Brussels, Belgium. Association for Computational Linguistics.

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. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Muhammad Shafique, Areej Mehboob, Layba Fiaz, Muhammad Qadeer, and Hamza Farooq. 2026. UrduBench: An Urdu reasoning benchmark using contextually ensembled translations with human-in-the-loop.

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, Suzhuo, China. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. OpenAI GPT-5 System Card. *Preprint*, arXiv:2601.03267.

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, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *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)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Munief Hassan Tahir, Sana Shams, Layba Fiaz, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking the performance of pre-trained LLMs across Urdu NLP tasks. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 17–34, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Mukhammed Togmanov, Nurdaulet Mukhituly, Diana Turmakhan, Jonibek Mansurov, Maiya Goloburda, Akhmed Sakip, Zhuohan Xie, Yuxia Wang, Bekassyl Syzdykov, Nurkhan Laiyk, Alham Fikri Aji, Ekaterina Kochmar, Preslav Nakov, and Fajri Koto. 2025. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14403–14416, Vienna, Austria. Association for Computational Linguistics.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. In *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)*, pages 10076–10132, Albuquerque, New Mexico. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Candidate Pool Analysis

We construct UrduMMLU in two stages. First, an automatic preprocessing pipeline collects and cleans multiple-choice questions from Pakistani examination boards and Urdu MCQ websites to produce a candidate pool. Second, annotation, verification, deduplication, and balancing transform this pool into the final benchmark used in all evaluations. This appendix analyzes both stages and shows how the dataset composition changes throughout the construction process.

Figure 5 shows the distribution of UrduMMLU items across the four Pakistani examination levels: SSC-I, SSC-II, HSSC-I, and HSSC-II. The left panel reports absolute item counts, while the right panel reports the within-level domain distribution. The Figure highlights two consistent trends, first, Humanities dominates the SSC levels, where language and literature subjects occupy a larger portion of the curriculum. Second, STEM and Social Sciences become more prominent at the HSSC levels, where students specialize into science, commerce, and humanities tracks. The level distribution therefore reflects the structure of the Pakistani curriculum rather than collection artifacts.

We also analyze question length because stem length can influence model performance and varies across domains. Figure 6 summarizes the overall length distribution and the domain-wise split between short and long stems using a 9-word threshold. Most UrduMMLU stems are short, but the dataset retains a substantial long-question tier. STEM has the most balanced short/long distribution, while Humanities and Profession contain relatively more short stems.

A.1 Source and Domain Distributions

Tables 6 and 7 compare the cleaned candidate pool (**Raw**) with the released benchmark (**Final**). This comparison distinguishes the initial collection distribution from the curated benchmark used in evaluation.

Source distribution: The cleaned candidate pool contains 40,427 items collected from nine Pakistani examination and MCQ-bank sources, of which 26,431 survive into the final benchmark. Table 6 shows that the raw pool is heavily concentrated in a few large sources. Ustad 360 alone contributes 23,788 raw items, while the four largest sources together account for more than 90% of the pool. The final benchmark is substantially less skewed. Annotation, deduplication, and balancing reduce the relative share of the largest sources, while smaller sources such as MCQTimes, TestPointPK, and ETest contribute proportionally more to the released benchmark. BISE Multan 2025 shows the largest reduction because of a high duplicate rate against other examination sources. Figure 7 expands the domain-level statistics from Table 7 to the subdomain level. Humanities is dominated by Urdu Literature and Urdu Language, whereas Social Sciences and STEM distribute more evenly across multiple medium-sized subdomains.

Source	Raw	Final	%	Share
Ustad 360	23,788	11,068	41.9%	
MCQTimes	6,099	5,918	22.4%	
TestPointPK	3,619	3,502	13.2%	
ETest	3,102	2,783	10.5%	
FBISE	2,406	1,459	5.5%	
ExamAunty	643	540	2.0%	
GoTest	566	515	1.9%	
PakMCQs	434	414	1.6%	
BISE Multan 2025	440	232	0.9%	
Total	**40,427**	**26,431**	**100.0**	

Table 6: Source distribution of the cleaned candidate pool (**Raw**) and the released UrduMMLU benchmark (**Final**). Percentages and share bars correspond to the final benchmark distribution.

Domain	Raw	Final	%	Share
Humanities	11,539	11,010	41.7%	
Social Sciences	14,626	7,968	30.2%	
STEM	11,590	5,113	19.3%	
Other	2,030	1,365	5.2%	
Profession	642	975	3.7%	
Total	**40,427**	**26,431**	**100.0**	

Table 7: Domain distribution before and after final benchmark selection. **Raw** denotes the cleaned candidate pool, while **Final** denotes the released UrduMMLU benchmark.

Domain distribution: Table 7 compares the cleaned candidate pool and the final benchmark across domains. The candidate pool distributes relatively evenly across Humanities, Social Sciences, and STEM, while Other and Profession remain much smaller. The final benchmark shifts toward Humanities, which grows from 28.5% to 41.7%, while Social Sciences and STEM decrease to 30.2% and 19.3%, respectively. Profession is the only domain whose absolute count increases during balancing (642 $\rightarrow$ 975), which improves cov-

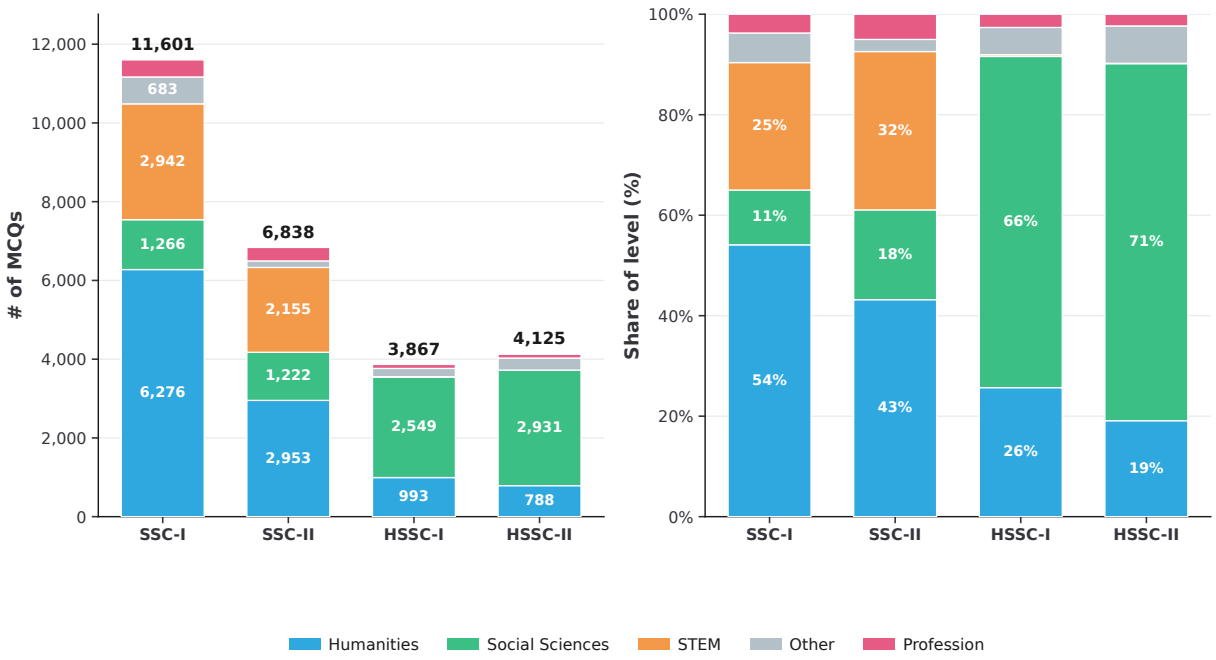
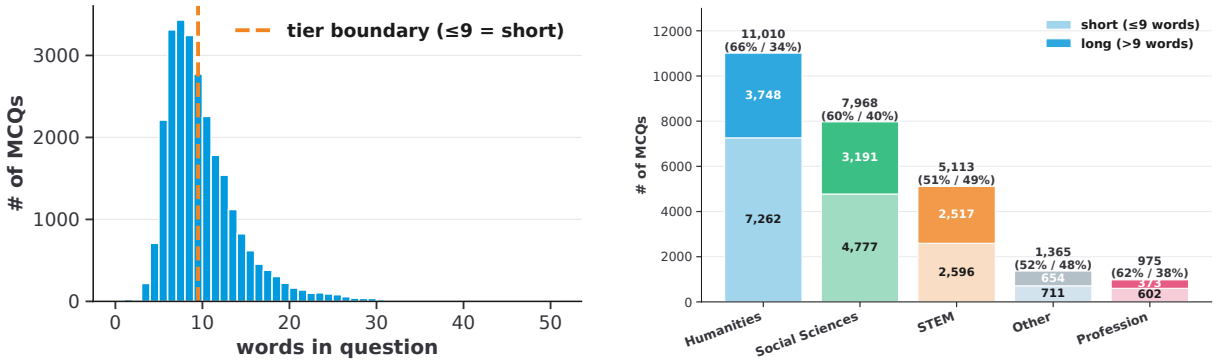


Figure 5: Distribution of UrduMMLU items across Pakistani examination levels, grouped by domain. **Left:** absolute item counts per level. **Right:** within-level domain distribution. Humanities dominates SSC-I and SSC-II, while STEM and Social Sciences become more prominent at the HSSC levels, reflecting the structure of the Pakistani secondary-school curriculum.



(a) Distribution of question lengths in words. The dashed vertical line at 9 words marks the short/long boundary.

(b) Domain-wise counts of short and long questions, with within-domain percentages shown in parentheses.

Figure 6: Question-length analysis for UrduMMLU. **Left:** histogram of question lengths. **Right:** domain-wise counts of short ($\leq$ 9 words) and long ($>$ 9 words) questions. STEM is closest to a balanced split, while Humanities and Profession skew shorter.

erage for reliable domain-level evaluation. These changes reflect a deliberate balancing step rather than artifacts from preprocessing or cleaning. We down-sample overrepresented STEM and Social Sciences items and preserve underrepresented Profession items to better align the benchmark with the structure of the Pakistani SSC and HSSC curriculum shown in Figure 5. The share bars in the rightmost column visualize the final domain distribution.

B Annotation Details

The exam-derived portion of UrduMMLU came from Pakistani examination boards and MCQ sources that did not provide answer keys. To produce reliable gold labels, we recruited 17 Urdu-fluent annotators and ran a dual-annotator consensus process supported by a custom dashboard and written guidelines. This appendix documents the annotator pool, the annotation guidelines and dashboard, the inclusion and edit-resolution rules, and the resulting agreement statistics.

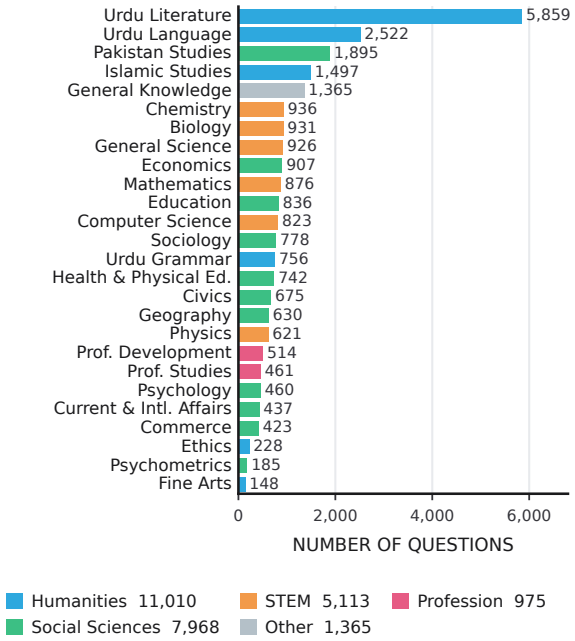


Figure 7: Final UrduMMLU item counts by subdomain, grouped by domain. Urdu Literature and Urdu Language contribute the largest shares, while Social Sciences and STEM distribute across a larger number of medium-sized subdomains.

B.1 Annotator Demographics and Feedback

The annotation pool consisted of 17 annotators recruited for native Urdu fluency and familiarity with the Pakistani school curriculum. Table 8 summarizes the demographic profile of the pool. The annotators were approximately gender-balanced (52.9% female, 47.1% male), predominantly native Urdu speakers (94.1%), and concentrated in the 18–34 age range. Educationally, 88.3% held at least a bachelor's degree and 41.2% held a master's degree, which is important for a benchmark that targets SSC- and HSSC-level subject content. Most annotators also reported between one and six years of professional experience.

After completing their assigned batches, all annotators filled out a post-task satisfaction survey. Table 9 summarizes the responses. Feedback was consistently positive, and no annotator selected *Disagree* or *Strongly disagree* for any statement. Instruction clarity, compensation fairness, guideline usefulness, and overall satisfaction received entirely positive responses. Task enjoyment received a smaller number of neutral responses (23.5%), suggesting that annotators found the process clear and manageable even if not inherently engaging.

Attribute	Count	%
Gender		
Female	9	52.9
Male	8	47.1
Native Urdu speaker		
Yes	16	94.1
No	1	5.9
Age range		
18–24	7	41.2
25–34	10	58.8
Highest completed education		
High school diploma	1	5.9
Some college / vocational	1	5.9
Bachelor's degree	8	47.1
Master's degree	7	41.2
Professional work experience		
Less than 1 year	5	29.4
1–3 years	6	35.3
4–6 years	4	23.5
7–9 years	2	11.8
Total annotators	**17**	**100.0**

Table 8: Demographic profile of the UrduMMLU annotator pool ($n = 17$; identities anonymised).

Statement	SA	A	N	Pos.
Instructions clearly explained the task	82.4	17.6	0.0	100.0
Task was enjoyable and engaging	47.1	29.4	23.5	76.5
Paid fairly for the task	76.5	23.5	0.0	100.0
Examples / guidelines were helpful	76.5	23.5	0.0	100.0
Overall glad to have completed the task	76.5	23.5	0.0	100.0

Table 9: Post-task satisfaction survey results ($n = 17$, values in %). SA = Strongly agree, A = Agree, N = Neutral. No annotator selected *Disagree* or *Strongly disagree* on any item, so those columns are omitted. *Pos.* is the share of *Agree* plus *Strongly agree*.

B.2 Annotation Guidelines

Before annotation began, we held a live online onboarding session in which we walked through the task end-to-end, demonstrated each flag and edit category on real items, and answered annotator questions. The full written guidelines were also embedded as an always-available help page inside the annotation dashboard so that annotators could re-check policies during their work, and admins remained reachable by email throughout the annotation period for cases not covered by the written guidelines. This subsection summarizes the core policy and illustrates the flag and edit cases with screenshots from the help page.

Task overview. Annotators were asked to verify the answer to each MCQ by selecting the *single best* option from A/B/C/D. When multiple options

looked plausible, annotators were instructed to select the most precise or directly relevant answer rather than guessing.

Look-up and abstention policy. Annotators were encouraged to consult Google or Wikipedia for fact-based questions (dates, authors, capitals, scientific terms, historical events) rather than relying on memory, with a target pace of 15–30 seconds per question including verification. Annotators were asked to mark an item as *unsure / skip* rather than submit a confident guess in any of the following cases: (i) the answer could not be resolved within roughly a minute of search, (ii) two or more options remained equally plausible after verification, or (iii) the question required specialist context (e.g., niche fiqh details or obscure regional history) that they could not quickly acquire.

Flag vs. edit. Annotators were given a single rule of thumb to choose between the two actions: *edit* when the issue could be fixed in-place by changing text (a typo, missing space, duplicated word, wrong subdomain label), and *flag* when the issue required admin review and could not be repaired by text correction (no correct answer, multiple correct answers, ambiguity, missing visual, out-of-scope content). Annotators were explicitly instructed not to rewrite question semantics, not to "fix" wrong distractors into correct ones, and not to flag a question solely because they had edited a typo in it.

When to flag. Figure 8 illustrates the five flag categories covered in the guidelines. These are: (a) two or more options are simultaneously correct; (b) none of the listed options is the correct answer; (c) the question is ambiguous, vague, or under-specified; (d) the question references a diagram, image, or chart that is not included in the text; and (e) the question is out of scope for the benchmark (hyper-local trivia, sectarian content, opinion questions). For each case, annotators were asked to attach a short free-text note explaining the issue. Flagging was independent of answer selection, and annotators could flag with or without picking an option.

When to edit. Figure 9 illustrates the three most common edit categories: (a) spelling fixes and missing diacritics, where the intended word is clear from context; (b) duplicated words and other scraping artifacts; and (c) subdomain reassignment, where the original subdomain label is clearly inconsistent with the question content. Beyond these, the guidelines also permitted spacing corrections, removal of stray punctuation, stripping of redundant in-text option-letter prefixes (e.g., A., B., or their Urdu equivalents) already shown by the option badge, and translation of stray English option text when an unambiguous Urdu equivalent existed. Technical English terms, HTML/CSS tags, proper names, and brand names were left in English.

Check	Drop if
Dual annotation	Only one assigned annotator submitted an annotation.
Flag	Either annotator flagged the item.
Unsure	Either annotator selected the unsure/skip option.
Valid pick	Either annotator did not select one of A/B/C/D.
Consensus	The two annotators selected different answer keys.
Assignment	The annotation came from an annotator who was not assigned to that batch.

Table 10: Inclusion rules for the final annotated pool. An item is dropped if any rule fires; only items that pass every check enter the gold-labelled set.

Workflow. Annotators worked in batches of approximately 50 MCQs, with accuracy prioritized over speed.

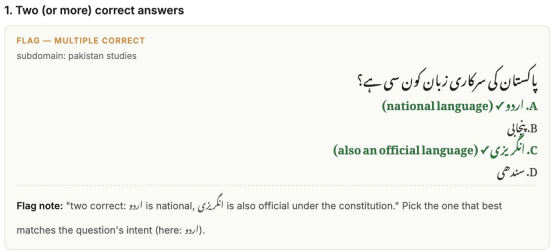
B.3 Inclusion Rules

We applied a deterministic consensus filter (Table 10): an item was retained only when both assigned annotators independently submitted the same valid answer choice and neither flagged nor abstained. The agreed answer was stored as the final gold label.

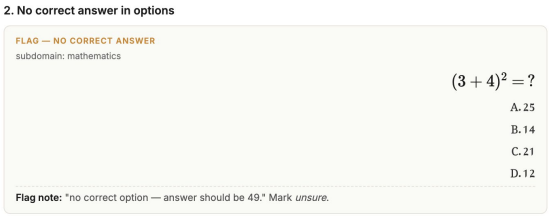
Two edge cases require clarification. First, when both annotators selected an option such as "none of these", we treated it as a valid agreed answer because such options commonly appear in Pakistani MCQ examinations. Second, only the explicit *unsure / skip* action counted as abstention. Missing annotations triggered the incomplete-annotation rule instead, so abstentions always reflected deliberate annotator decisions.

B.4 Edit Resolution

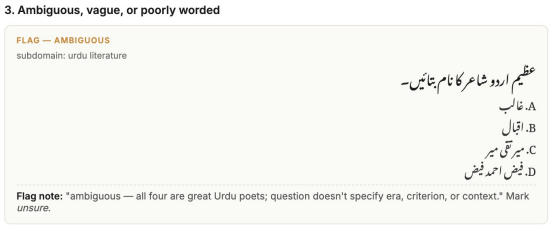
In addition to selecting answers, annotators could suggest edits to question text, answer options, or subdomain labels. These edits targeted minor extraction and metadata issues such as OCR errors,



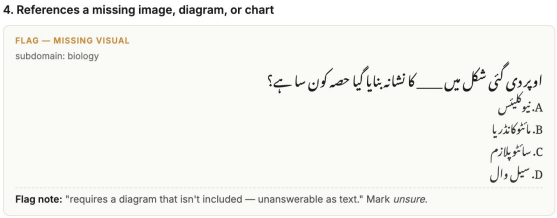
(a) Two or more options are simultaneously correct.



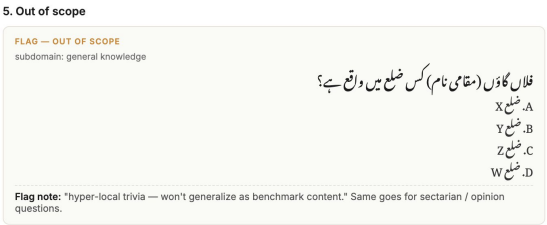
(b) No option in the list is the correct answer.



(c) Question is ambiguous, vague, or under-specified.



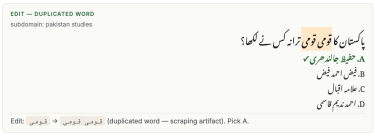
(d) Question references a missing image, diagram, or chart.



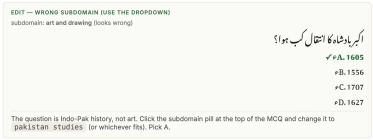
(e) Hyper-local, sectarian, or opinion content that is out of scope for the benchmark.

Figure 8: Examples of the five flag categories used in the annotation guidelines. Annotators were asked to flag the item and attach a short free-text note for each case.

(a) Spelling fix: a missing letter is restored from context.



(b) Duplicated word from a scraping artifact is removed.



(c) Wrong subdomain (art and drawing) is reassigned via the dropdown to pakistan studies.

Figure 9: Examples of the most common edit categories permitted by the annotation guidelines. Edits are restricted to OCR, scraping, formatting, and metadata corrections; annotators do not rewrite question semantics or modify answer correctness.

dropped diacritics, malformed option labels, and incorrect subdomain assignments rather than substantive question rewrites. We resolved all edits deterministically so that the final benchmark could be reconstructed directly from the raw annotations. Table 11 summarizes the resolution rules. The resolution policy prioritizes agreed edits when both annotators propose the same correction and otherwise prefers the more conservative or informative revision. For subdomain edits, we recompute the corresponding domain label from the corrected subdomain to preserve consistency between the two fields in the released benchmark.

B.5 Annotation Dashboard

While Section B.2 covered *when* to flag or edit, this subsection illustrates *how* the dashboard supports these actions. The interface supported four independent actions for each item: selecting an answer option, marking the item as *unsure / skip*, editing question or option text in place, and flagging the item for review with a free-text explana-

Case	Resolution
Only one annotator edited a field	Use that edit.
Both annotators made the same edit	Use the agreed edit.
Both annotators edited the same field differently	Use the longer edit.
No edit was suggested	Keep the original field.
Subdomain was edited	Recompute the domain from the corrected subdomain.

Table 11: Edit-resolution rules for annotated MCQs. The rules are applied per field, and the resolved values are written back to the item before the inclusion rules in Table 10 are evaluated.

tion. This separation allowed annotators to distinguish answer selection from quality-control feedback and text correction. Figures 10, 11, and 12 illustrate the annotation workflow.

Picking an answer. Figure 10 shows the standard annotation workflow. Annotators view the Urdu question stem, four labeled answer options, and metadata describing the subdomain, academic level, length tier, and item identifier (Figure 10a). Selecting an option highlights the choice but does not automatically advance to the next item (Figure 10b); annotators must explicitly confirm the selection before proceeding, which reduces accidental submissions. Keyboard shortcuts (1–5 for option selection, arrow keys for navigation) support efficient batch traversal.

Editing an item. Figure 11 illustrates the in-place editing UI. The example contains OCR and formatting artifacts in the answer options (Figure 11a). After entering edit mode, annotators modify the question text and options through inline editable fields (Figure 11b). The interface records all changes and attaches revision tags to each edited field for later review, which can be reverted with a single click (Figure 11c).

Flagging an item. Figure 12 shows the flagging UI. In the illustrated case, OCR corruption removes superscript formatting from a physics question, making all answer options invalid (Figure 12a). The annotator marks the item as *unsure / skip* and submits a flag with a free-text explanation (Figure 12b). The dashboard visually highlights flagged items so admins can review them, and the inclusion rules in Section B.3 automatically remove flagged items from the consensus pool.

B.6 Annotation Outcomes

Outcome	Count
Input annotated MCQs	17,565
Retained after consensus filtering	14,459
Dropped: annotator disagreement	1,611
Dropped: flagged by annotator	1,247
Dropped: unsure/skip selected	243
Dropped: single annotated	5
Domain corrections	141

Table 12: Annotation outcomes for the exam-derived portion of UrduMMLU. Each excluded item appears under a single exclusion rule.

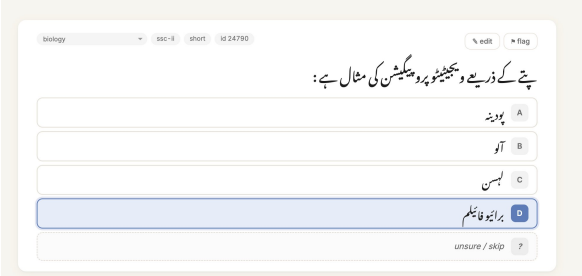
A total of 17,565 exam-derived MCQs entered annotation, of which 14,459 were retained after applying the edit-resolution and inclusion rules from Tables 11 and 10 (an overall yield of 82.3%). Table 12 breaks down the 3,106 exclusions. Answer disagreement is the dominant cause (51.9% of all drops), reflecting questions where two qualified Urdu annotators could not converge on a defensible answer and which are therefore unsuitable for evaluation under a strict consensus policy. Flagged items form the second-largest category and predominantly contain OCR corruption or malformed options similar to Figure 12a.

Inter-annotator agreement was correspondingly high. Across all annotated items, observed agreement reached 89.98%, with a simplified Cohen's κ of 0.8663. Figure 13 further breaks agreement down by annotator pair. Each cell reports the simplified Cohen's κ together with the number of shared retained items, while blank cells indicate annotator pairs without overlap. Most populated cells exceed $\kappa = 0.85$, showing that agreement remains consistently strong across annotator pairs rather than depending on a small subset of annotators. Lower-agreement cells correspond mainly to pairs with relatively few shared items and therefore have limited influence on the aggregate statistic.

C Dataset Format

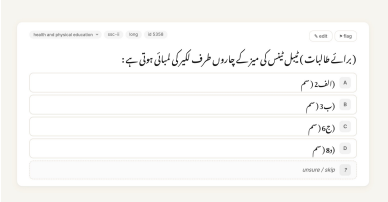
Each UrduMMLU example is stored as a multiple-choice item containing a question, four answer options, a gold answer label, domain and subdomain labels, academic level information, and source metadata. The evaluation pipeline uses the question, options, and gold answer fields during infer-

(a) Question view with answer options, metadata, and annotation controls.



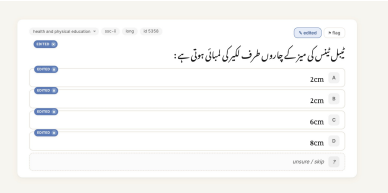
(b) Answer selection interface before advancing to the next item.

Figure 10: Annotation dashboard workflow for answer selection.



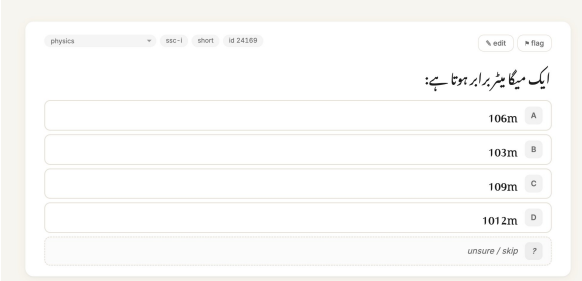
(a) Original OCR-extracted question and options.

(b) In-place editing of question and option text.

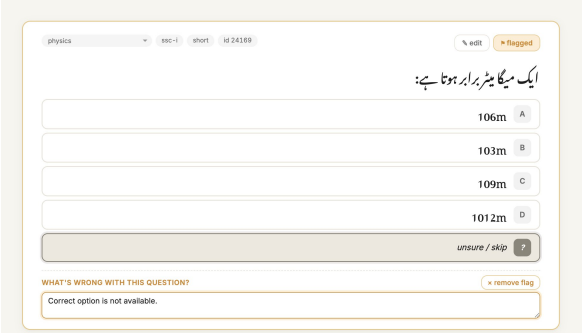


(c) Saved edits with editable revision markers.

Figure 11: Annotation dashboard workflow for text correction and normalization.



(a) Problematic OCR example marked as *unsure / skip*.



(b) Flagged item with an attached review reason.

Figure 12: Annotation dashboard workflow for flagging problematic items.

ence, while the remaining metadata supports analysis, filtering, and reproducibility. Figure 14 reports two dataset-level sanity checks: item-length distribution and answer-key distribution. The median item length is 76 characters, and the answer labels remain close to uniformly distributed across A–D, reducing the risk of prompt-length or answer-position bias during evaluation. Figure 15 shows the JSON schema used for each released benchmark item.

C.1 Subject Acronyms and Education Levels

Table 13 lists the full names, acronyms, and education levels for all 26 UrduMMLU subdomains, grouped by domain. We use these acronyms in the per-subdomain results tables (Tables 16 and 17).

D Evaluation Details

This appendix provides the full model roster and prompt templates used in the UrduMMLU experiments.

D.1 Model Roster

Table 14 lists the 30 models evaluated in this work. We group models by family for readability, while the main paper discusses them using broader categories such as proprietary API models, open-weight multilingual models, compact models, mixture-of-experts models, reasoning-oriented variants, and Urdu- or regionally special-

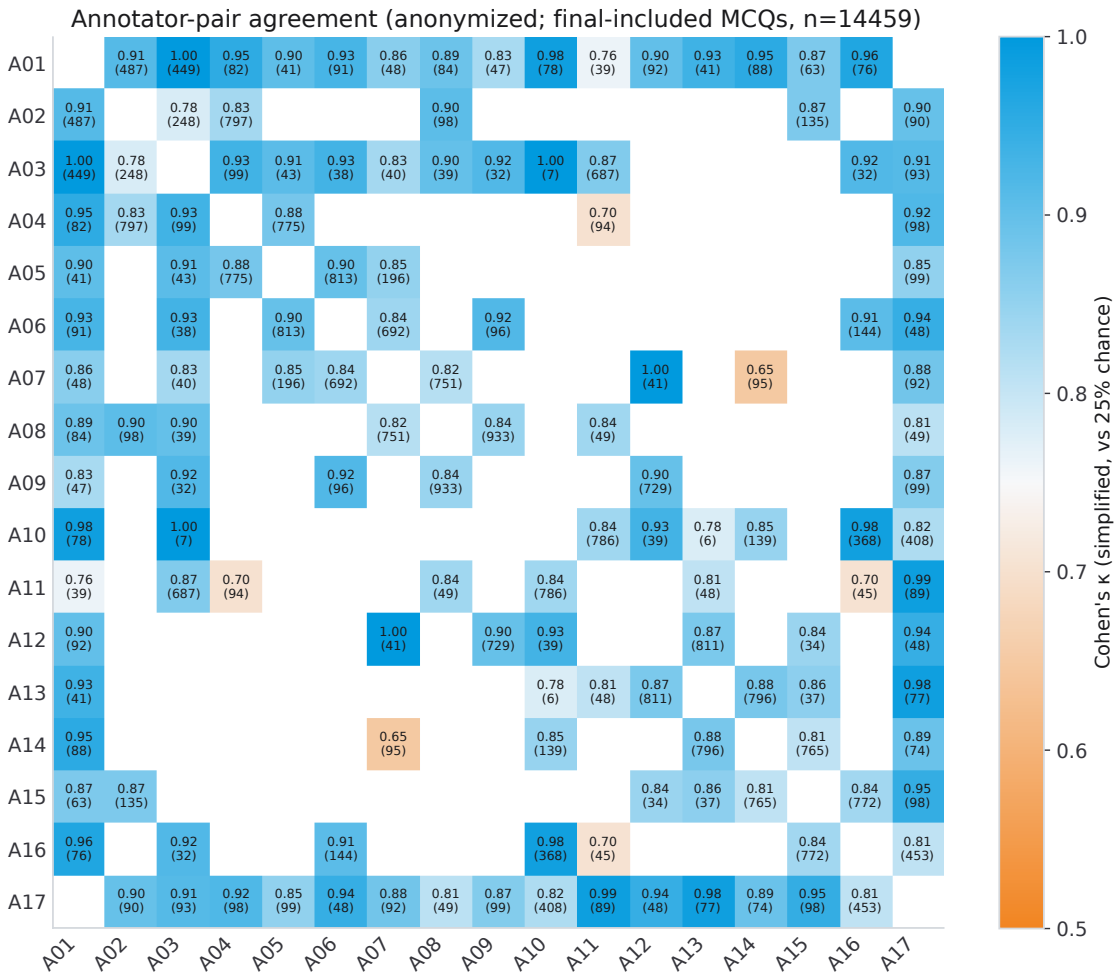


Figure 13: Pairwise annotator agreement on final-included MCQs. Each cell reports simplified Cohen's κ, with the number of shared items shown in parentheses. Blank cells indicate annotator pairs with no shared final-included items.

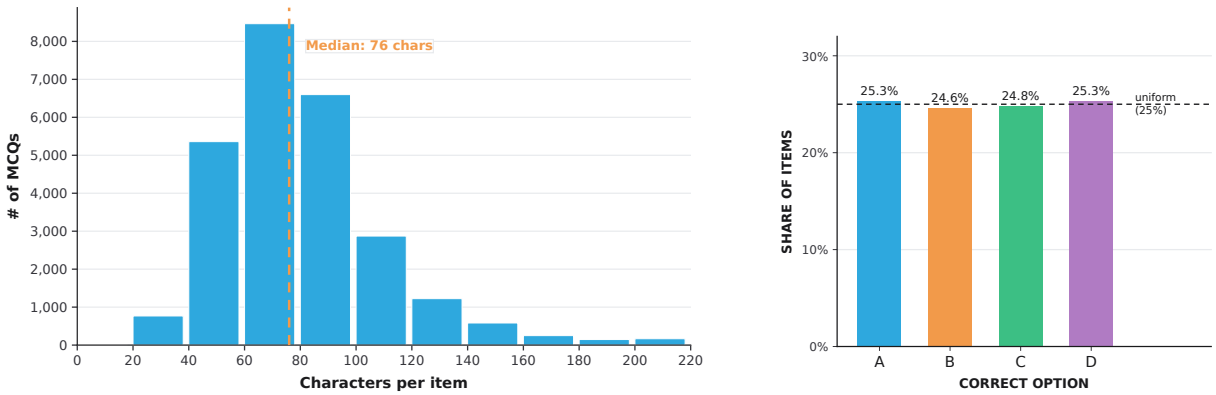
ized models.

D.2 Prompt Templates

We use separate English and Urdu prompt templates for zero-shot evaluation. Both templates present the same Urdu question and answer options while changing only the instruction language and field labels. The output format remains identical in both settings to support automatic parsing and consistent evaluation.

English prompt: Figure 16 shows the English prompt template, which combines a fixed system prompt with a per-item user prompt. The system prompt instructs the model to answer in a strict two-line format consisting of an Answer key and Answer text, without additional explanation or formatting. This structure supports deterministic answer extraction and allows invalid or malformed outputs to be measured consistently across models. The user prompt fills the placeholders {domain}, {subdomain}, {level}, {question}, and {A}–{D} directly from the dataset schema described in Appendix C, preserving the Urdu question content across both prompt-language settings.

Urdu prompt: The Urdu prompt template mirrors the English template while translating the system instructions, user-field labels (مضمون, سطح, سوال), and surrounding instructional text into Urdu. The Urdu question stem and answer options remain unchanged across both settings. We also preserve the same two-line response structure using the English fields Answer key: and Answer text:, which allows a single parser to process outputs under both prompt languages. Figure 17 shows the full Urdu prompt template. This minimal-difference setup makes the prompt-language comparison in Section 5 directly interpretable, since any performance change comes from the instruction language



(a) Character-length distribution per item.

(b) Gold answer-key distribution.

Figure 14: Dataset-level sanity checks for UrduMMLU. Most questions remain compact enough for standard MCQ prompting, while the gold answer keys remain close to uniformly distributed across A–D.

UrduMMLU JSON Schema

```
{
  "id": "...",
  "question": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "correct_key": "...",
  "correct_option": "...",
  "domain": "...",
  "subdomain": "...",
  "level": "...",
  "source": [
    {
      "name": "...",
      "url": "..."
    }
  ]
}
```

Figure 15: JSON schema for individual UrduMMLU question items.

Zero-shot English Prompt

System Prompt: You are an expert multiple-choice question answering assistant. Read the question carefully and select the single best answer. Respond in EXACTLY this two-line format, with no extra text:

- Answer key: <one of A, B, C, D>
- Answer text: <verbatim text of the chosen option, copied character-for-character>

Do not add explanations, preambles, markdown, or punctuation outside of the format. The Answer text must match the option text exactly so the response can be parsed programmatically.

User Prompt:

Subject: {domain} – {subdomain}
Level: {level}
Question: {question}
A) {A}
B) {B}
C) {C}
D) {D}
Answer key:
Answer text:

Figure 16: Prompt for multiple-choice question answering with strict answer formatting requirements.

rather than changes in question content or evaluation logic.

D.3 Few-Shot Evaluation Setup

We evaluate using the lm-evaluation-harness framework (Gao et al., 2024). Each item is formatted as a four-option multiple-choice question, with the answer choices labeled A through D and the gold label stored as an integer index in $\{0, 1, 2, 3\}$. The benchmark comprises 26,431 items spanning all subject-level splits. We report accuracy (acc) and length-normalized accuracy (acc_norm) under 0-shot, 1-shot, 3-shot, and 5-shot conditions.

D.4 Implementation Details

All evaluations use a fixed random seed of 42. For locally loaded open-weight models, we use bfloat16 precision, greedy decoding, and a batch size of 10, with automatic device placement across available hardware. We evaluate instruction-tuned models using their corresponding chat templates. For API-based systems, including OpenAI, Anthropic, Google Gemini, and Hugging Face Inference API models, we use the same prompt format and decoding configuration whenever provider

Domain	Subdomain	Acronym	Levels
Humanities	Ethics	ETH	SSC-I, SSC-II, HSSC-I
Humanities	Fine Arts	FNA	SSC-I, SSC-II, HSSC-I
Humanities	Islamic Studies	ISL	SSC-I, SSC-II, HSSC-I, HSSC-II
Humanities	Urdu Grammar	UGR	SSC-I, SSC-II, HSSC-I, HSSC-II
Humanities	Urdu Language	ULG	SSC-I, SSC-II, HSSC-I, HSSC-II
Humanities	Urdu Literature	ULT	SSC-I, SSC-II, HSSC-I, HSSC-II
Other	General Knowledge	GKN	SSC-I, SSC-II, HSSC-I, HSSC-II
Profession	Professional Development	PRD	SSC-I, SSC-II, HSSC-I, HSSC-II
Profession	Professional Studies	PRS	SSC-I, SSC-II
STEM	Biology	BIO	SSC-I, SSC-II, HSSC-I, HSSC-II
STEM	Chemistry	CHM	SSC-I, SSC-II, HSSC-I
STEM	Computer Science	CSC	SSC-I, SSC-II, HSSC-I
STEM	General Science	GSC	SSC-I, SSC-II, HSSC-I
STEM	Mathematics	MTH	SSC-I, SSC-II, HSSC-I
STEM	Physics	PHY	SSC-I, SSC-II, HSSC-II
Social Sciences	Civics	CIV	SSC-I, SSC-II, HSSC-I, HSSC-II
Social Sciences	Commerce	COM	SSC-I, SSC-II, HSSC-I, HSSC-II
Social Sciences	Current & International Affairs	CIA	SSC-I, SSC-II, HSSC-I, HSSC-II
Social Sciences	Economics	ECO	SSC-I, SSC-II, HSSC-I, HSSC-II
Social Sciences	Education	EDU	SSC-I, SSC-II, HSSC-I, HSSC-II
Social Sciences	Geography	GEO	SSC-I, SSC-II, HSSC-I, HSSC-II
Social Sciences	Health & Physical Education	HPE	SSC-I, SSC-II, HSSC-II
Social Sciences	Pakistan Studies	PKS	SSC-I, SSC-II, HSSC-I, HSSC-II
Social Sciences	Psychology	PSY	HSSC-I, HSSC-II
Social Sciences	Psychometrics	PMT	SSC-I, SSC-II, HSSC-I, HSSC-II
Social Sciences	Sociology	SOC	HSSC-I, HSSC-II

Table 13: UrduMMLU domains, subdomains, acronyms, and corresponding education levels.

constraints allow it. Each API request is retried up to five times in the event of failure. To avoid silent large-scale evaluation failures, the evaluation pipeline terminates automatically after five consecutive unsuccessful responses.

E Detailed Results

This appendix expands the main results from Table 4. Table 15 reports the STEM–Humanities accuracy gap for each model under the Urdu prompt, sorted by STEM accuracy. Across nearly all model families, performance on STEM substantially exceeds performance on Humanities, and the gap generally widens as overall capability decreases. The gap becomes small for the BLOOMZ family and the two Urdu-targeted models, but for different reasons. BLOOMZ checkpoints remain close to the random baseline on both domains, while the Urdu-targeted models show similarly low performance on STEM and Humanities because their STEM accuracy is already far below that of comparably sized general-purpose models.

F Per-Subdomain Results

This appendix expands the domain-level results from Table 4 to all 26 subdomains. Table 16 reports accuracy under the English prompt, while Table 17 reports accuracy under the Urdu prompt. Both tables follow the same ordering, with subdomains grouped by domain and sorted by dataset size.

F.1 Subject-Wise Behavior

The subdomain results sharpen the main finding from Section 5: STEM subjects transfer much more reliably than Urdu-centered Humanities subjects. The strongest models approach saturation on several STEM subdomains. Under the English prompt, Gemini-3.5-Flash reaches 97.86% on chemistry, 98.60% on biology, and 98.86% on mathematics, while DeepSeek-V4-Flash reaches 98.71% on physics. These scores remain nearly unchanged under the Urdu prompt, and in some cases increase slightly. The consistency across prompt languages suggests that scientific and mathematical concepts transfer relatively cleanly once the

Model	Size	Family	License	Ref.
large-traversaal/Alif-1.0-8B-Instruct	8B	Alif	Apache 2.0	(Shafique et al., 2025)
mistralai/Ministral-3-3B-Instruct-2512	3B	Ministral	Apache 2.0	(Liu et al., 2026)
mistralai/Ministral-3-8B-Instruct-2512	8B	Ministral	Apache 2.0	(Liu et al., 2026)
enstazao/Qalb-1.0-8B-Instruct	8B	Qalb	Apache 2.0	(Hassan et al., 2026)
Qwen/Qwen3-4B-Instruct-2507	4B	Qwen 3	Apache 2.0	(Yang et al., 2025)
Qwen/Qwen3-8B	8B	Qwen 3	Apache 2.0	(Yang et al., 2025)
Qwen/Qwen3.6-27B	27B	Qwen 3.6	Apache 2.0	(Yang et al., 2025)
Qwen/Qwen3.6-35B-A3B	36B	Qwen 3.6	Apache 2.0	(Yang et al., 2025)
bigscience/bloomz-1b1	1.1B	BLOOMZ	Bigscience Bloom Rail 1.0	(Muennighoff et al., 2023)
bigscience/bloomz-1b7	1.7B	BLOOMZ	Bigscience Bloom Rail 1.0	(Muennighoff et al., 2023)
bigscience/bloomz-3b	3B	BLOOMZ	Bigscience Bloom Rail 1.0	(Muennighoff et al., 2023)
bigscience/bloomz-7b1-mt	7B	BLOOMZ	Bigscience Bloom Rail 1.0	(Muennighoff et al., 2023)
deepseek-ai/DeepSeek-V4-Flash	158B	DeepSeek	DeepSeek License	(DeepSeek-AI, 2026)
google/gemma-3-4b-it	4B	Gemma 3	Gemma	(Gemma Team, 2025)
google/gemma-2-9b-it	9B	Gemma	Gemma	(Team et al., 2024)
google/gemma-4-26B-A4B-it	27B	Gemma	Gemma	(Google, 2026b)
google/gemma-4-31B-it	31B	Gemma	Gemma	(Google, 2026b)
meta-llama/Llama-3.2-3B-Instruct	3B	LLaMA 3.2	LLaMA License	(Meta, 2024a)
meta-llama/Llama-3.1-8B-Instruct	8B	LLaMA 3.1	LLaMA License	(Grattafiori et al., 2024)
meta-llama/Llama-4-Scout-17B-16E-Instruct	109B	LLaMA 4	LLaMA License	(Meta, 2025)
meta-llama/Llama-4-Maverick-17B-128E-Instruct	402B	LLaMA 4	LLaMA License	(Meta, 2025)
meta-llama/Llama-3.3-70B-Instruct	70B	LLaMA 3.3	LLaMA License	(Meta, 2024b)
microsoft/Phi-4-mini-instruct	3B	Phi-4	MIT	(Microsoft et al., 2025)
microsoft/Phi-3.5-mini-instruct	4B	Phi-3.5	MIT	(Abdin et al., 2024)
claude-haiku-4-5	N/D	Claude	Proprietary	(Anthropic, 2025)
claude-sonnet-4-6	N/D	Claude	Proprietary	(Anthropic, 2026)
gemini-3.1-flash-lite	N/D	Gemini	Proprietary	(Google, 2026a)
gemini-3.5-flash	N/D	Gemini	Proprietary	(Google DeepMind, 2026)
gpt-5.4-mini	N/D	GPT	Proprietary	(Singh et al., 2026)
gpt-5.4	N/D	GPT	Proprietary	(Singh et al., 2026)

Table 14: Language models evaluated in this study. Model sizes are reported when publicly disclosed; N/D denotes not disclosed.

model can process Urdu input.

Humanities presents a much harder challenge. Islamic studies and Urdu grammar remain accessible for the strongest models, with Gemini-3.5-Flash reaching 94.25% on Islamic studies and 88.34% on Urdu grammar under the English prompt. In contrast, Urdu literature remains difficult across the entire model suite. Even the strongest model reaches only 80.35% under the English prompt and 80.81% under the Urdu prompt. Most other proprietary and open-source models perform substantially worse, often trailing by another 10 to 20 points. Urdu language occupies an intermediate position, with top scores near 89%. Across nearly all capable models, the same ordering persists: Islamic studies > Urdu grammar > Urdu language > Urdu literature. This consistency suggests that the differences reflect genuine variation in subject difficulty rather than isolated model behavior.

Social Sciences contains both highly accessible and consistently difficult subdomains. Geography, civics, sociology, psychology, and commerce all exceed 93% accuracy for the strongest models. Pakistan studies also remains relatively strong despite its large size. In contrast, current and international affairs and psychometrics stand out as the two hardest Social Sciences subdomains. Current and international affairs peaks at roughly 78% under both prompts, likely because many questions depend on time-sensitive world knowledge beyond model pretraining cutoffs. Psychometrics is even more difficult: no model in the evaluation exceeds 60% accuracy under either prompt language.

The smaller Profession and Other domains follow patterns similar to Social Sciences, with proprietary models reaching the low 90s and smaller open-source models trailing substantially behind. These domains do not introduce failure modes beyond those already visible in the aggregate results.

The subdomain breakdown also clarifies the behavior of smaller open-source models. Gemma-2-9B-IT gives the strongest sub-25B performance across nearly all Humanities subjects, including

Zero-shot Urdu Prompt

System Prompt:

سوال کو غور سے پڑھیں اور آپ ایک ماہر MCQ جوابدہ معاون ہیں۔
جواب بالکل اس دو سطری شکل میں دیں، کوئی صحیح آپشن منتخب کریں۔
اضافی متن شامل نہ کریں:
- Answer key: < A, B, C یا D میں سے ایک >
- Answer text : <منتخب آپشن کا متن من و عن، حرف بہ حرف نقل کریں>
کوئی وضاحت، تمہید، مارک ڈاؤن، یا اضافی الفاظ نہ لکھیں۔
Answer text کو آپشن کے متن سے بالکل مماثل ہونا چاہیے
تاکہ پروگرام اسے درست طور پر پارس کر سکے۔

User Prompt:

مضمون: {domain} – {subdomain}
سطح: {level}
سوال: {question}
(A : {A}
(B : {B}
(C : {C}
(D : {D}
Answer key:
Answer text:

Figure 17: Urdu prompt for multiple-choice question answering with strict answer formatting requirements.

Model	STEM	Hum.	Gap
Gemini-3.5-Flash	97.81	85.31	12.50
DeepSeek-V4-Flash	97.57	67.32	30.25
GPT-5.4	97.40	74.82	22.58
Gemini-3.1-Flash-Lite	97.09	74.38	22.71
Qwen3.6-35B-A3B	96.32	58.12	38.20
Claude-Sonnet-4.6	96.26	72.69	23.57
Gemma-4-31B-IT	93.86	63.25	30.61
LLaMA-4-Maverick-17B	92.38	63.25	29.13
Claude-Haiku-4.5	91.96	59.31	32.65
Qwen3.6-27B	91.12	55.71	35.41
GPT-5.4-mini	88.25	62.35	25.90
Gemma-4-26B-A4B-IT	87.21	57.73	29.48
LLaMA-4-Scout-17B-16E	85.59	56.55	29.04
LLaMA-3.3-70B	78.39	56.10	22.29
Qwen3-8B	74.37	30.87	43.50
Ministral-3-8B	71.37	45.74	25.63
Gemma-2-9B-IT	69.02	48.08	20.94
Qwen3-4B	68.75	43.00	25.75
Ministral-3-3B	57.25	43.07	14.18
Gemma-3-4B-IT	51.79	38.27	13.52
LLaMA-3.1-8B	46.49	37.61	8.88
LLaMA-3.2-3B	37.24	29.32	7.92
Phi-4-mini	37.08	28.70	8.38
Qalb-1.0-8B	36.26	32.72	3.54
Phi-3.5-mini	33.83	27.25	6.58
Alif-1.0-8B	33.27	29.00	4.27
BLOOMZ-7B	29.24	30.88	–1.64
BLOOMZ-1.7B	28.74	28.76	–0.02
BLOOMZ-3B	26.56	27.70	–1.14
BLOOMZ-1.1B	24.53	25.83	–1.30

Table 15: STEM–Humanities accuracy gap under the Urdu prompt. Models with gaps near zero either score at chance on both domains (BLOOMZ) or have an unusually low STEM score for their scale (Qalb-1.0-8B, Alif-1.0-8B). Values are taken directly from Table 4.

Urdu language, Urdu grammar, ethics, and fine arts, while Qwen3-8B performs best on most STEM subjects such as chemistry, mathematics, computer science, and physics. The pattern mirrors the domain-level results: Qwen3-8B retains relatively strong scientific knowledge but struggles on Urdu-centered humanities content, whereas Gemma-2-9B-IT trades some STEM performance for more balanced cross-subdomain behavior. The Urdu-targeted models, Qalb-1.0-8B and Alif-1.0-8B, do not lead any individual subdomain and remain below similarly sized general-purpose models across literary and linguistic subjects. The BLOOMZ checkpoints remain close to the random baseline on most subdomains and should be interpreted together with the high invalid-output rates reported in Section 5.4.

F.2 The Psychometrics Gap

Psychometrics is the most difficult subdomain in our evaluation. The best English-prompt accuracy reaches only 57.30% (Gemini-3.5-Flash), while the best Urdu-prompt accuracy reaches 52.97% (Claude-Sonnet-4.6). No model exceeds 60% under either prompt setting, in contrast with the 90–98% accuracies achieved on many STEM and Social Sciences subjects.

The difficulty appears specific to the content rather than the prompt language. English- and Urdu-prompt results remain close, and model rankings on psychometrics largely mirror their overall rankings. Psychometrics questions in Urdu SSC/HSSC curricula frequently emphasize analogies, number series, logical patterns, and aptitude-style reasoning tasks that require abstract structure recognition rather than factual recall. These results suggest that current LLMs still struggle on reasoning-heavy Urdu educational content even when they perform strongly on factual subjects.

F.3 Urdu Tuning Fails on Literature

Urdu literature is the largest subdomain in UrduMMLU, with 5,859 items, and contains content with limited overlap with English-dominated pretraining corpora, including classical poetry,

prosody, and literary history. Figure 18 compares two Urdu-targeted 8B models, Qalb-1.0-8B and Alif-1.0-8B, against two general-purpose 8B-class instruction-tuned models, Qwen3-8B and Ministral-3-8B.

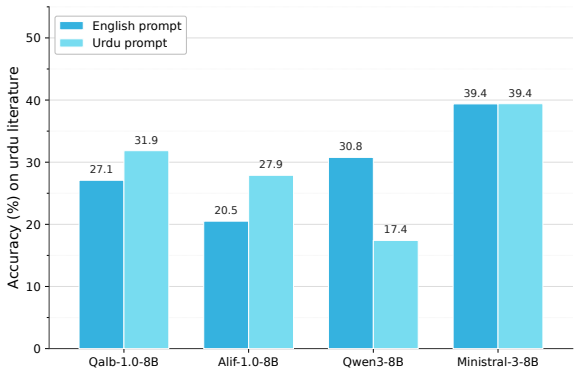


Figure 18: Urdu literature accuracy for four 8B-class instruction-tuned models under English and Urdu prompts. Ministral-3-8B performs best under both settings, while Qwen3-8B shows the largest prompt-language drop.

The Urdu-targeted models do not outperform the general-purpose baselines on this subdomain. Ministral-3-8B achieves the strongest performance under both prompts at 39.4%, while Qalb-1.0-8B and Alif-1.0-8B remain below 32%. Qwen3-8B performs competitively under the English prompt (30.8%) but drops sharply under the Urdu prompt (17.4%). In contrast, both Urdu-targeted models improve under the Urdu prompt, suggesting that Urdu-specific tuning improves instruction following more than literary knowledge acquisition. Overall, the results indicate that Urdu literature remains a challenging knowledge domain even for Urdu-targeted LLMs.

F.4 English-Prompt Subdomain Accuracy

Table 16 reports per-subdomain accuracy for all 30 models under the English prompt. The table groups subdomains by domain and orders them by dataset size within each group, so earlier columns contribute more strongly to the corresponding domain-level scores in Table 4. Acronym expansions appear directly in the table header. The results provide a fine-grained view of model behavior across subjects: Gemini-3.5-Flash remains consistently strong across nearly all subdomains, DeepSeek-V4-Flash approaches proprietary-level performance on STEM subjects but drops on Urdu language and literature, and the BLOOMZ models remain close to the random baseline across most subjects.

F.5 Urdu-Prompt Subdomain Accuracy

Table 17 reports the same per-subdomain breakdown under the Urdu prompt. The table follows the same structure and ordering as Table 16, which allows direct comparison between the two prompt settings. Most differences remain small, reinforcing the main finding from Section 5 that the difficulty of UrduMMLU comes primarily from the question content rather than the instruction language. For most proprietary models and the Gemma family, English- and Urdu-prompt accuracies remain nearly identical across the majority of subdomains. A few model-specific shifts become clearer at the subdomain level. Qwen3.6-35B-A3B improves substantially under the Urdu prompt, driven mainly by STEM subjects, where several subdomain scores rise into the mid-90s. In contrast, Qwen3-8B loses accuracy primarily on Humanities subjects, especially Urdu language and Urdu literature, which explains its large drop in overall Humanities performance under the Urdu prompt. The Urdu-targeted models also show modest gains on several Humanities subdomains under the Urdu prompt, although these improvements do not substantially change their overall ranking. Together, these patterns further support the conclusion that prompt language plays a secondary role compared with the underlying educational and cultural knowledge required by the benchmark.

G Invalid-Output Examples

Section 5.4 reports invalid-output rates across the model suite; this appendix provides representative examples of the corresponding failure modes. Each example is drawn from an actual model prediction under the Urdu prompt setting. We organize the examples by failure type in order to highlight recurring decoding behaviors and illustrate how invalid generations manifest in practice across different models.

Repetition collapse. In some cases, the model enters a degenerate decoding loop and repeatedly emits the same token sequence without producing a meaningful or valid answer. Example G illustrates this behavior for BLOOMZ-7B, which repeatedly generates the token "Question:" dozens of times instead of producing a task-relevant response.

Model	STEM						Humanities						Social Sciences											Profession		Other
	CHM	BIO	GSC	MTH	CSC	PHY	ULT	ULG	ISL	UGR	ETH	FNA	PKS	ECO	EDU	SOC	HPE	CIV	GEO	PSY	CIA	COM	PMT	PRD	PRS	GKN
Open-source Models: > 25B Parameters																										
DeepSeek-V4-Flash	**97.65**	**98.06**	**96.54**	**98.40**	**95.86**	98.71	**56.71**	**77.11**	**91.24**	**79.58**	**88.60**	**87.16**	**91.40**	**93.94**	**85.27**	95.12	**92.17**	**94.06**	**94.29**	**94.78**	**71.56**	95.04	40.22	**89.30**	**90.46**	**86.44**
Gemma-4-26B-A4B-IT	85.47	83.57	85.64	85.05	88.46	88.41	45.75	66.10	78.02	71.16	71.93	71.62	72.72	73.10	74.64	88.82	71.16	78.37	79.52	81.09	42.56	77.54	44.32	71.60	79.83	70.55
Gemma-4-31B-IT	91.88	93.45	92.33	93.49	93.44	96.46	51.63	71.93	84.64	77.25	84.21	82.43	80.05	84.34	82.06	90.75	81.13	87.26	87.62	90.00	54.92	86.05	40.54	79.57	88.29	79.49
LLaMA-3.3-70B	79.06	78.41	86.50	77.40	85.54	81.32	46.25	62.49	75.55	64.68	79.82	75.00	76.83	75.74	77.39	86.76	67.65	82.96	82.86	81.52	47.60	77.07	29.19	71.79	81.78	73.77
LLaMA-4-Scout-17B-16E	84.19	83.24	86.29	78.08	85.78	88.57	47.79	62.77	75.02	65.21	73.25	76.35	67.81	73.98	72.85	86.25	65.63	77.93	75.08	77.83	48.51	72.81	39.46	68.09	80.91	69.52
LLaMA-4-Maverick-17B-128E	91.99	90.23	91.79	91.89	92.70	94.04	52.47	70.26	83.37	74.07	83.77	81.76	81.27	83.35	77.87	88.43	80.19	87.41	84.76	86.96	52.40	79.43	34.59	78.79	84.82	80.66
Qwen3.6-27B	91.88	86.57	89.31	89.61	92.95	93.72	45.49	61.70	74.68	69.71	77.19	79.73	66.28	78.72	76.32	88.82	69.54	77.93	81.90	83.70	39.13	78.25	**47.57**	65.18	85.47	69.60
Qwen3.6-35B-A3B	88.14	86.25	89.42	86.64	92.47	90.66	46.13	62.37	78.49	67.46	77.63	79.05	68.18	78.50	75.12	88.05	68.73	80.00	80.63	83.70	43.02	78.96	41.08	70.04	81.56	69.82
Open-source Models: ≤ 25B Parameters																										
BLOOMZ-1.1B	24.49	24.06	19.03	24.11	25.64	24.32	27.90	27.45	25.10	25.96	28.07	21.38	24.91	24.26	21.65	24.42	24.53	24.00	23.65	25.33	25.93	26.48	24.86	25.10	27.11	25.63
BLOOMZ-1.7B	26.76	25.98	32.58	27.88	34.90	24.79	27.54	27.86	29.93	23.44	32.00	23.57	31.46	32.45	38.32	42.29	32.25	34.82	29.73	27.67	29.10	31.26	36.07	29.26	45.62	31.43
BLOOMZ-3B	25.27	30.66	29.68	28.16	26.72	26.18	30.08	29.19	31.88	30.65	31.56	31.43	30.54	35.94	34.05	39.43	31.89	39.19	29.39	33.48	28.60	32.59	24.55	27.42	38.79	28.39
BLOOMZ-7B	26.46	27.94	29.66	27.02	28.97	26.04	27.15	27.04	29.15	30.45	32.70	29.46	31.48	32.85	39.79	40.78	34.58	31.66	30.99	33.68	27.21	32.25	28.89	26.40	38.24	28.11
Gemma-2-9B-IT	65.28	**64.02**	**71.06**	64.16	71.57	66.67	40.98	**51.90**	55.58	**54.37**	**65.35**	**64.86**	**52.30**	58.99	**65.07**	**75.71**	51.35	64.00	62.86	61.30	**40.96**	**63.12**	30.81	49.81	**71.37**	**54.58**
Gemma-3-4B-IT	46.37	46.29	55.83	43.49	60.27	47.67	33.96	40.64	43.22	41.14	50.44	54.73	42.53	46.42	51.08	64.78	41.64	58.07	53.33	46.09	31.58	49.65	23.78	41.25	55.10	45.57
LLaMA-3.2-3B	32.48	34.05	40.28	38.47	42.41	35.10	21.42	31.72	32.53	35.98	41.67	27.03	33.83	36.38	43.42	48.33	35.85	48.89	42.70	32.17	21.74	36.88	31.89	34.82	37.31	35.60
LLaMA-3.1-8B	44.98	42.53	51.94	44.86	52.86	44.28	28.21	44.13	46.43	46.30	53.07	49.32	47.81	49.61	55.02	65.17	43.53	56.89	52.70	47.61	26.77	47.75	30.81	41.63	56.18	44.54
Ministral-3-3B	49.79	48.98	59.40	54.68	60.39	58.94	**41.25**	44.09	48.73	46.30	50.44	58.78	42.27	49.72	55.74	62.47	46.63	51.26	55.17	50.87	37.07	49.88	28.65	42.50	56.40	47.91
Ministral-3-8B	61.43	62.41	70.52	65.41	**77.52**	71.98	39.38	47.22	**57.31**	52.51	64.47	56.76	52.19	**61.08**	62.68	74.55	**53.91**	**67.26**	**65.40**	**63.70**	37.30	59.57	**42.70**	**51.75**	68.33	54.43
Phi-4-mini	35.47	28.46	36.43	45.21	41.88	35.48	25.57	32.87	32.46	30.29	41.67	26.35	33.61	39.36	43.95	54.88	31.85	42.67	37.62	37.39	21.74	39.01	29.19	34.63	46.85	32.67
Phi-3.5-mini	33.44	28.57	29.70	43.72	34.51	33.01	19.54	28.15	26.25	21.43	35.09	20.27	27.02	32.52	34.69	38.82	31.27	33.48	31.75	30.22	18.76	34.75	20.00	26.65	38.39	28.28
Qwen3-4B-Instruct	**66.88**	57.79	68.68	71.12	77.40	72.14	38.11	43.54	49.43	51.59	48.68	60.14	43.91	52.04	56.70	67.99	49.33	57.19	56.83	55.65	36.84	50.35	29.73	43.19	63.12	47.84
Qwen3-8B	**66.88**	60.47	70.63	**75.57**	**77.52**	**76.01**	30.77	46.07	50.50	51.59	60.09	46.62	45.01	56.45	62.68	71.08	52.43	59.85	57.46	59.35	29.98	57.92	30.27	45.91	68.55	50.18
Proprietary Models																										
Claude-Haiku-4.5	90.17	89.26	87.69	91.98	91.98	92.91	47.64	64.51	81.27	68.92	78.07	77.55	71.59	80.26	75.72	87.66	74.90	81.78	82.06	85.43	48.24	79.67	40.76	69.84	82.65	74.14
Claude-Sonnet-4.6	96.26	96.78	94.38	97.83	95.26	98.07	61.90	81.05	90.98	84.39	87.72	89.19	88.28	91.40	83.85	92.80	83.96	92.74	92.54	94.35	65.45	91.49	52.43	85.60	88.94	86.01
Gemini-3.1-Flash-Lite	97.76	97.10	96.00	96.46	96.23	97.75	63.53	83.06	90.78	87.04	92.11	85.14	91.98	94.16	88.04	93.70	90.70	94.37	94.60	95.43	67.05	94.56	38.92	87.55	93.28	86.08
Gemini-3.5-Flash	97.86	98.60	96.76	98.86	95.95	**98.55**	80.35	88.69	94.25	88.34	89.47	87.16	94.61	94.27	88.54	**93.70**	93.40	94.96	95.87	95.00	78.49	94.80	57.30	92.22	91.97	91.43
GPT-5.4-mini	90.28	89.26	89.52	80.59	90.77	90.02	51.87	71.41	82.10	72.22	80.70	79.05	78.05	81.15	76.32	89.85	67.52	83.11	83.02	84.78	56.75	79.43	35.68	73.74	85.03	75.24
GPT-5.4	96.26	96.56	94.06	92.81	94.90	96.46	57.91	78.15	90.38	78.17	84.65	86.49	87.07	88.20	82.30	91.26	86.93	91.56	91.43	91.74	70.02	91.25	42.70	83.66	88.29	84.10
Urdu Models																										
Alif-1.0-8B	**36.04**	**37.59**	**48.58**	**36.87**	**47.67**	**40.03**	20.51	**31.03**	33.98	26.06	40.53	35.14	**40.10**	**40.84**	**42.86**	**52.52**	**36.07**	**47.21**	**47.19**	**40.39**	26.44	**40.28**	**25.41**	33.27	**47.24**	**41.04**
Qalb-1.0-8B	35.26	34.26	45.68	33.68	43.38	36.71	**27.09**	31.01	**35.67**	**32.41**	**40.79**	**41.22**	36.57	36.38	36.12	47.69	28.44	47.11	43.49	35.22	**30.89**	38.06	24.32	**37.74**	43.17	39.56

Table 16: **Subdomain-level model performance on UrduMMLU under the English prompt.** Accuracy (%) across all 26 subdomains grouped by domain. Subdomains are ordered within each domain by dataset size (descending); see Table 13 for acronym expansions. Boxed values mark the best overall score per column, while bold values indicate the best score within each model group.

Model	STEM						Humanities						Social Sciences											Profession		Other
	CHM	BIO	GSC	MTH	CSC	PHY	ULT	ULG	ISL	UGR	ETH	FNA	PKS	ECO	EDU	SOC	HPE	CIV	GEO	PSY	CIA	COM	PMT	PRD	PRS	GKN
Open-source Models: > 25B Parameters																										
DeepSeek-V4-Flash	**97.75**	**97.63**	**96.65**	**98.17**	96.71	98.87	**54.79**	**76.29**	**91.03**	**78.38**	85.53	**85.81**	**89.74**	**93.04**	**85.73**	93.82	**90.27**	**94.50**	**93.62**	**94.10**	**67.13**	**94.09**	39.56	**87.94**	**91.11**	**84.88**
Gemma-4-26B-A4B-IT	87.91	84.32	86.93	85.27	89.67	90.34	46.58	65.38	78.09	69.84	74.56	75.34	71.19	78.50	76.91	89.59	73.72	81.33	80.95	84.35	44.04	81.09	41.76	72.76	83.48	71.79
Gemma-4-31B-IT	92.52	93.98	93.63	93.61	94.41	95.65	51.44	71.25	84.24	76.32	82.46	85.62	80.53	84.67	82.30	91.13	81.81	88.30	86.98	90.43	53.78	86.76	41.76	79.57	89.15	78.39
LLaMA-3.3-70B	74.04	77.12	83.69	72.26	83.72	80.52	47.31	60.15	74.47	64.29	77.09	75.68	73.72	75.08	74.28	85.48	65.63	80.15	78.89	78.26	48.51	76.83	35.14	65.37	77.87	71.65
LLaMA-4-Scout-17B-16E	85.15	85.82	86.50	79.63	89.19	88.08	47.26	61.18	75.42	65.61	77.63	75.68	69.71	76.07	72.73	87.79	64.96	79.56	78.10	81.74	46.68	74.70	36.22	67.32	78.52	69.01
LLaMA-4-Maverick-17B-128E	93.26	91.19	92.44	91.48	91.86	94.69	53.27	70.18	81.83	72.09	83.33	76.35	81.64	84.79	77.87	89.72	82.48	88.59	85.56	88.70	54.82	82.27	36.76	76.65	82.86	80.81
Qwen3.6-27B	92.95	88.83	89.42	88.93	93.07	94.85	45.47	61.02	75.02	67.33	80.26	78.38	66.91	81.48	75.96	89.33	71.43	81.19	84.29	85.00	40.96	81.09	**43.78**	64.59	85.47	70.55
Qwen3.6-35B-A3B	97.22	96.33	93.95	97.03	96.59	97.10	46.42	63.97	81.33	71.05	**85.90**	75.68	82.88	89.85	81.99	91.51	87.13	91.06	93.15	91.70	49.20	89.83	43.24	81.66	87.64	77.99
Open-source Models: ≤ 25B Parameters																										
BLOOMZ-1.1B	26.18	24.60	19.98	25.34	26.61	24.80	25.95	25.47	25.52	26.59	28.95	21.62	26.02	24.48	23.68	25.71	26.68	24.59	23.97	25.22	27.00	26.24	24.86	22.37	24.73	24.47
BLOOMZ-1.7B	28.95	26.57	28.78	29.03	34.85	22.77	28.31	28.73	31.31	28.57	31.71	21.93	33.50	32.27	38.78	39.66	35.04	33.63	30.13	27.93	30.50	29.38	37.57	28.97	36.75	27.51
BLOOMZ-3B	26.21	26.37	27.65	25.24	25.88	28.29	29.45	21.96	29.56	29.89	29.03	22.92	29.42	34.97	31.45	38.97	28.27	30.43	27.37	32.26	27.60	34.21	28.12	30.00	34.95	26.75
BLOOMZ-7B	27.00	30.25	30.41	28.78	30.83	28.52	30.55	30.50	30.95	29.92	37.27	38.53	32.42	31.42	34.35	37.48	33.73	40.21	33.84	30.89	29.81	33.64	21.62	28.52	37.70	30.73
Gemma-2-9B-IT	66.35	66.17	73.33	64.38	74.97	69.57	**41.92**	**52.02**	56.58	**56.48**	65.79	**68.92**	52.30	63.73	64.35	80.08	52.83	69.04	68.25	64.13	**42.33**	62.65	32.43	50.78	**74.84**	55.82
Gemma-3-4B-IT	46.79	49.52	58.32	44.86	62.21	48.95	33.38	40.84	46.43	44.84	50.00	54.05	42.85	48.07	51.91	67.74	44.47	56.74	53.81	47.61	29.75	51.06	29.19	43.39	57.70	46.23
LLaMA-3.2-3B	32.80	35.98	39.52	37.56	41.31	36.55	24.75	33.62	34.13	36.24	40.79	35.14	36.89	36.49	45.69	53.34	35.71	49.63	42.54	34.57	23.11	39.24	34.59	33.27	44.03	38.02
LLaMA-3.1-8B	42.20	45.97	51.51	43.38	51.76	43.64	32.09	42.35	43.75	44.84	57.02	46.62	47.02	49.28	58.25	68.89	42.45	56.30	53.33	44.35	27.00	47.75	30.27	41.05	59.00	44.98
Ministral-3-3B	52.35	50.38	62.85	55.02	64.64	59.90	39.84	43.44	49.70	48.54	54.82	51.35	43.06	53.36	59.69	71.98	45.82	56.00	58.57	53.91	37.99	53.90	32.97	45.03	60.30	48.86
Ministral-3-8B	66.67	**69.60**	73.43	66.21	79.10	75.04	39.41	47.58	**58.32**	53.70	**71.93**	56.76	**54.62**	**66.48**	**66.15**	**80.33**	**58.09**	**69.63**	**70.48**	**67.17**	38.67	**65.72**	**44.32**	**54.28**	69.63	**57.00**
Phi-4-mini	36.54	32.44	32.72	45.32	41.92	33.33	23.84	35.69	32.46	32.54	39.04	28.38	33.88	40.02	45.57	55.01	35.71	45.33	39.21	34.13	21.28	40.43	28.11	35.21	42.52	35.85
Phi-3.5-mini	33.37	28.91	30.09	42.59	34.51	34.47	25.40	30.85	27.91	26.79	32.74	26.53	29.15	31.82	36.53	37.98	31.49	32.34	31.37	33.04	21.74	37.91	16.76	30.54	38.34	30.57
Qwen3-4B-Instruct	66.45	56.82	69.55	71.69	76.31	74.72	39.00	44.61	49.50	51.32	46.05	60.81	43.43	57.22	58.49	72.11	50.54	60.74	56.83	56.30	37.07	55.32	30.81	45.14	62.26	47.69
Qwen3-8B	**71.14**	64.16	**75.49**	**77.92**	**79.75**	**80.78**	17.41	42.86	50.52	44.84	64.32	42.47	46.48	58.82	65.27	74.94	54.77	63.69	61.72	62.61	28.38	61.61	34.59	42.61	74.55	49.38
Proprietary Models																										
Claude-Haiku-4.5	91.77	92.05	89.42	92.91	92.95	93.23	47.84	66.10	82.10	71.16	77.19	78.38	72.88	81.15	76.20	87.15	76.82	83.70	84.76	87.83	48.60	82.27	38.80	70.82	86.55	74.29
Claude-Sonnet-4.6	96.69	96.56	94.38	97.26	95.50	97.58	62.13	80.65	90.92	85.05	87.72	84.46	88.07	91.29	84.33	**93.70**	85.04	91.70	92.70	93.91	66.82	91.49	52.97	84.63	90.02	85.86
Gemini-3.1-Flash-Lite	97.65	98.07	96.76	96.35	**96.23**	97.42	64.00	83.26	90.98	85.32	90.35	85.81	92.08	94.16	87.92	93.44	91.24	94.81	95.08	94.13	68.42	94.80	37.30	87.55	93.49	85.57
Gemini-3.5-Flash	97.86	98.60	97.08	98.97	96.10	98.23	80.81	88.89	94.19	88.76	89.04	**89.19**	94.56	94.82	88.76	**93.70**	93.67	94.96	95.71	95.65	77.80	95.27	52.43	91.44	91.32	91.72
GPT-5.4-mini	89.64	91.19	91.04	78.20	90.16	89.21	51.03	70.82	82.63	72.35	82.02	79.73	77.47	82.14	77.15	89.72	69.81	83.11	81.75	89.57	59.73	85.11	36.76	75.29	84.38	75.09
GPT-5.4	97.44	97.85	95.79	99.20	95.87	**98.55**	67.09	78.98	90.71	80.95	89.47	89.86	90.24	93.94	86.96	92.42	92.32	94.22	94.29	93.48	69.34	92.43	40.00	84.05	90.46	83.74
Urdu Models																										
Alif-1.0-8B	30.10	29.75	34.54	33.29	40.41	**32.94**	27.89	29.94	31.48	28.30	31.52	33.80	35.09	33.24	34.58	**44.93**	34.42	42.25	40.80	**36.57**	29.97	**35.87**	22.31	37.28	38.21	**42.93**
Qalb-1.0-8B	**33.33**	**36.80**	**37.76**	**34.49**	**44.32**	31.82	**31.85**	**32.27**	**34.99**	**32.90**	**38.12**	**42.86**	**36.69**	**37.98**	**39.97**	44.58	**34.73**	**42.64**	**41.94**	36.05	**30.82**	32.61	**30.26**	**38.20**	**40.70**	42.55

Table 17: **Subdomain-level model performance on UrduMMLU under the Urdu prompt.** Accuracy (%) across all 26 subdomains grouped by domain. Subdomains are ordered within each domain by dataset size (descending); see Table 13 for acronym expansions. Boxed values mark the best overall score per column, while bold values indicate the best score within each model group.

Example: repetition collapse (BLOOMZ-7B, Urdu literature, gold = B)

Question: خواجہ دل محمد کس مضمون کے استاد تھے
Options:

- A. ان میں سے کوئی نہیں
- B. ریاضی *(gold)*
- C. اردو
- D. پنجابی

Model output:
Question: Question: Question: Question: ...

Prompt echo. The model copies part of the user prompt instead of answering the question. Example G shows BLOOMZ-1.7B reproducing the beginning of the question prompt and terminating before generating a valid answer.

Example: prompt echo (BLOOMZ-1.7B, Islamic studies, gold = B)

Question: مومنو! اگر تم میری راہ میں لڑنے اور میری خوشنودی طلب کرنے کے لیے نکلے ہو تو میرے اور اپنے دشمنوں کو _____ مت

بناؤ۔
Options:
- A. بھائی
- B. دوست *(gold)*
- C. دشمن
- D. رازدار

Model output:
مضمون: Humanities – islamic studies | سطح: SSC-II
سوال: مومنو! اگر تم میری راہ میں لڑنے اور میری خوشنودی طلب کر

Refusal or clarification request. Instead of selecting an answer, the model returns a conversational clarification request. Example G shows Qalb-1.0-8B treating the MCQ as an ambiguous user query.

Example: refusal (Qalb-1.0-8B, mathematics, gold = A)

Question: قطاروں اور کالموں کی تعداد کسی قالب میں _____ کو ظاہر کرتی ہے۔
Options:
- A. مرتبہ *(gold)*
- B. مقطع
- C. قطاریں
- D. کالم

Model output:
آپ کا سوال واضح نہیں ہے۔ کیا آپ اسے دوبارہ بیان کر سکتے ہیں؟

System-prompt echo. The model reproduces the system prompt instead of answering the question. Example G shows Alif-1.0-8B repeating the assistant role description without generating an answer.

Example: system-prompt echo (Alif-1.0-8B, chemistry, gold = B)

Question: الکحلز کی ڈی ہائیڈریشن _____ کے ساتھ کی جا سکتی ہے۔
Options:
- A. NaOH
- B. H_2SO_4 *(gold)*
- C. HCl
- D. KOH

Model output:
آپ ایک ماہر MCQ جوابدہ معاون ہیں۔

Empty or placeholder output. The model emits a nearly empty response, often copied directly from a blank marker in the question. Example G shows BLOOMZ-3B returning only a placeholder token.

Example: empty output (BLOOMZ-3B, chemistry, gold = A)

Question: _____ فریکشن بطور جیٹ فیول استعمال ہوتی ہے۔
Options:
- A. کیروسین آئل *(gold)*
- B. لبریکیٹنگ آئل
- C. فیول آئل
- D. ڈیزل آئل

Model output:
_____

Discussion. All five examples produce outputs that cannot be mapped to a valid answer option and therefore contribute to the invalid-output rate rather than to model accuracy. The failures arise from different causes: repetition collapse and empty outputs reflect decoding instability, prompt and system-prompt echoes reflect instruction-following failures, and clarification requests reflect conversational misalignment with the MCQ format. These behaviors are not unique to UrduMMLU, but their concentration under the Urdu prompt for weaker models motivates reporting invalid-output rates alongside accuracy.