TALIM: A Benchmark Dataset for Tajik Academic Language Intelligence Measurement

Azamat Rashidov, Anis Rashidov, Archit Narumanchi, Jong C. Park
School of Computing, KAIST
{azamat03, anisjon.rashidov, archit.narumanchi, jongpark}@kaist.ac.kr

Abstract

We introduce TALIM (**T**ajik **A**cademic **L**anguage **I**ntelligence **M**easurement), a large-scale academic benchmark for evaluating academic reasoning and language understanding in Tajik. TALIM spans nine subjects and three question formats, multiple-choice, matching, and numerical short-answer, sourced natively from the Tajik national curriculum rather than translated from existing benchmarks. We evaluate nine LLMs, both proprietary and open-weight models, finding that proprietary models substantially outperform open-weight models across all subjects and question types. Grammar is the weakest subject across all models, while physics is consistently the strongest. English translation substantially improves open-weight performance while Farsi transliteration yields only marginal gains, indicating that failures stem from data scarcity rather than script unfamiliarity. Chain-of-thought analysis reveals several systematic failure modes: hallucination, reasoning process failures, and grammatical category confusion. We hope TALIM will serve as a resource for assessing Tajik language capabilities of frontier models and motivate similar native benchmark construction efforts for other low-resource languages.

1 Introduction

Large language models (LLMs) have demonstrated remarkable progress across a wide range of reasoning and language understanding tasks (Brown et al., 2020; OpenAI, 2023), and benchmarks such as MMLU (Hendrycks et al., 2021), BIG-bench (Srivastava et al., 2023), AGIEval (Zhong et al., 2024), and GPQA (Rein et al., 2024) have become central tools for measuring model capability across mathematics, science, and humanities. However, the overwhelming majority of these benchmarks focus on English or high-resource multilingual settings, and much of the current evaluation ecosystem remains translation-centric (Singh et al., 2025a). Translated benchmarks often fail to preserve the natural syntax and morphology of the target language, rarely capture culturally grounded knowledge (Kabir et al., 2025; Gaim et al., 2023; Simonsen et al., 2025), and do not reflect how native speakers naturally formulate reasoning processes or educational problems.

Subject	MC	Matching	Free-resp.	Total
Mathematics	852	87	464	1,403
Chemistry	847	130	323	1,300
Physics	506	142	194	842
Geography	744	86	52	882
History	826	215	37	1,078
Biology	897	223	11	1,131
Grammar	789	152	—	941
Literature	819	200	—	1,019
Law	434	106	—	540
Total	**6,714**	**1,341**	**1,081**	**9,136**

Table 1: Question distribution by subject and format in TALIM.

Tajik, despite being spoken by more than 10 million native speakers across Central Asia and diaspora communities (Eberhard et al., 2023), remains heavily underrepresented in modern NLP benchmarks and multilingual evaluation resources (Joshi et al., 2020; Arabov, 2026a). Tajik belongs to the IndoEuropean language family and is written primarily in Cyrillic script, unlike closely related Persian variants written in Perso-Arabic script. Although Tajik shares substantial lexical and grammatical similarity with Persian, modern multilingual language models often demonstrate inconsistent performance due to script differences, training data imbalance, and limited exposure to Tajik-specific linguistic structures (Arabov, 2026a,b). These limitations open severe gaps for low-resource languages (Joshi et al., 2020).

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 35042–35056
October 24-29, 2026 ©2026 Association for Computational Linguistics

To address these gaps, we introduce TALIM[1] (Tajik Academic Language Intelligence Measurement), a benchmark for evaluating academic reasoning and language understanding in Tajik, sourced natively from the National Testing Center of Tajikistan. The benchmark name intentionally echoes the Tajik word *ta'lim*, meaning "education," reflecting the dataset's curriculum-oriented design philosophy.

TALIM contains 9,136 questions collected from educational materials aligned with the Tajik national curriculum. The dataset spans nine academic subjects, including mathematics, physics, chemistry, biology, geography, history, law, literature, and Tajik grammar, across three question formats: multiple-choice, matching, and numerical short-answer. While zehnlab (2026) recently released a collection of Tajik benchmarks covering humanities and factual QA, their STEM portion is a direct translation of MMLU, the same translation-centric approach this work aims to move beyond. TALIM addresses this gap by providing a fully natively-sourced dataset across all nine subjects, making it the only natively-sourced Tajik STEM benchmark currently available.

We evaluate nine state-of-the-art language models, including both proprietary frontier and open-weight models, across all nine subjects and three question formats. Beyond overall accuracy, we conduct several targeted analyses: We examine subject-level performance to identify where models consistently struggle, with particular attention to Tajik grammar and culturally grounded knowledge; we probe the role of script and language familiarity by evaluating models on Farsi-transliterated and English-translated versions of the benchmark; and we apply chain-of-thought prompting (Wei et al., 2022) to the best and worst performing open-weight models to identify reasoning failure modes in a low-resource language setting.

Our contributions include: (1) We introduce TALIM, a curriculum-grounded benchmark for Tajik academic reasoning and language understanding, (2) we provide extensive evaluation of nine frontier proprietary and open-weight language models across multiple question formats and subjects, (3) we analyze the impact of script representation and translation on low-resource language understanding, and (4) we conduct a chain-of-thought failure analysis for Tajik reasoning tasks, identifying systematic failure modes.

[1]Dataset available at: TALIM

2 Related Work

2.1 Academic Reasoning Benchmarks

Several language-specific academic benchmarks have followed MMLU's framework by constructing native evaluation resources rather than translating from English. C-Eval (Huang et al., 2023) and CMMLU (Li et al., 2024) cover Chinese academic and culturally grounded knowledge across dozens of subjects; KMMLU (Son et al., 2025) draws from Korean professional licensing exams and university entrance tests; Arabic MMLU (Koto et al., 2024) sources questions from Arabic educational materials; IndoMMLU (Koto et al., 2023) includes Indonesian-specific cultural and curricular knowledge; and BLUCK (Kabir et al., 2025) evaluates Bengali linguistic understanding and cultural knowledge. These efforts demonstrate the value of native benchmark construction, but remain concentrated in relatively higher-resource languages.

2.2 Existing Tajik NLP Resources

Like many low-resource languages, Tajik remains largely absent from modern multilingual evaluation ecosystems (Joshi et al., 2020; Arabov, 2026a). zehnlab (2026) recently released a collection of six benchmarks covering history, literature, language, and factual QA across grades 5 to 11, providing broad curricular coverage. However, the STEM portion of their collection is a direct translation of MMLU (Hendrycks et al., 2021), which raises the same concerns about translated evaluation discussed above. Prior work on Tajik NLP has also highlighted the challenges posed by script duality, training data scarcity, and the limited transferability of Persian-language resources to Cyrillic Tajik (Arabov, 2026a,b). TALIM addresses these gaps by providing a fully natively sourced evaluation resource for Tajik, spanning nine subjects and three question formats.

3 Dataset for Tajik Academic Language Intelligence Measurement

Table 1 presents the question distribution across subjects and formats. TALIM contains 9,136 questions spanning nine academic subjects across two difficulty levels: Level A (Grade 11) and Level B (Grade 9), both aligned with the Tajik national curriculum. Questions are distributed across



Figure 1: Example literature question from TALIM shown in four representations: original Tajik Cyrillic, Farsi transliteration, English translation, and romanization. ✓: correct answer. ×: incorrect answer.

three formats: multiple-choice (73.5%), matching (14.7%), and numerical short-answer (11.8%). Multiple-choice questions present four labeled options with a single correct answer; matching questions require simultaneous assignment of four item pairs; and numerical short-answer questions require generating an exact numerical answer.

3.1 Dataset Curation

Questions were sourced directly from the National Testing Center (NTC) of Tajikistan, the government body established to administer the Unified University Entrance Examinations and national student learning assessments. Following Zhong et al. (2024) in grounding evaluation in official standardized examinations, NTC materials represent questions that Tajik students actually encounter in educational settings, ensuring authentic linguistic and cultural grounding.

Source materials were provided as digitally typeset PDF documents, from which text was extracted via OCR using the dots.mocr system (Zheng et al., 2026), which produced structured JSON output containing bounding box coordinates, element category labels (e.g., list items, formulas, page footers), and extracted text with mathematical expressions preserved in LaTeX notation. The extracted output was subsequently parsed to reconstruct question structure, separate answer options, and discard non-question elements such as page footers. The text was then cleaned to remove formatting artifacts and corrupted Unicode characters introduced during extraction.

Each question was annotated according to a structured schema capturing five attributes: correct answer label, subject category, fine-grained topic label, difficulty level, and question format type. All questions, answer options, and annotations underwent manual verification by two native Tajik-speaking annotators. The initial extraction yielded 10,071 candidate items. Of these, 935 were excluded: Questions requiring visual information such as diagrams, graphs, or images essential to answering were removed to maintain a fully text-based evaluation setting, and questions with ambiguous, inconsistent, or unrecoverable content were discarded during manual verification. The final dataset contains 9,136 questions.

To quantify residual error rates in the final dataset, we conducted an independent verification audit on a random subsample of 200 items drawn from mathematics, physics, chemistry, and biology, the subjects most susceptible to extraction errors due to mathematical notation, chemical formulas, and scientific symbols in the OCR pipeline. Both annotators independently re-examined all items and agreed on all annotation attributes. Three subtle orthographic errors were identified by both annotators, yielding a residual error rate of 1.5%.

4 Experimental Setup

We evaluate a diverse set of widely-used proprietary and open-weight models spanning a range of scales and capabilities:

- **Open-weight models**: Qwen-2.5-72B (Qwen et al., 2024), Llama-3.3-70B (Grattafiori et al., 2024), Qwen-3-32B (Yang et al., 2025), Gemma-3-27B (Gemma Team et al., 2025), Qwen-3-14B (Yang et al., 2025), DeepSeek V3 (DeepSeek-AI et al.,

Format	Model	Phys.	Hist.	Geog.	Math	Law	Chem.	Bio.	Lit.	Gram.
Multiple-Choice	GPT-5.5	**98.6**	**98.3**	**98.2**	**99.4**	**96.4**	**99.3**	**96.0**	**91.6**	**93.7**
	Claude Sonnet 4.6	95.6	90.2	84.9	97.4	85.3	86.8	80.9	82.1	69.2
	GPT-4o	76.8	86.4	83.7	53.2	77.6	63.9	81.9	77.0	62.7
	Gemma-3-27B	72.3	69.4	66.0	56.1	63.4	49.9	57.1	59.6	45.0
	Qwen-3-32B	78.1	57.0	58.5	63.5	55.8	64.3	54.1	50.5	42.0
	DeepSeek V3	77.4	61.4	62.6	70.6	47.7	54.0	47.8	50.1	37.3
	Llama-3.3-70B	64.8	68.2	67.1	42.5	56.9	49.5	54.1	56.3	44.2
	Qwen-2.5-72B	65.2	60.4	59.9	55.6	44.0	49.7	46.4	54.5	40.2
	Qwen-3-14B	62.5	51.7	48.3	47.8	45.2	50.9	45.5	40.9	36.2
Matching	GPT-5.5	**97.6**	**95.6**	**87.8**	**97.5**	**92.7**	**100.0**	**92.0**	**95.9**	**91.7**
	Claude Sonnet 4.6	91.5	76.7	48.8	88.5	79.2	85.6	60.5	70.4	63.2
	GPT-4o	83.8	68.4	55.8	24.1	69.8	40.0	56.1	57.8	49.0
	DeepSeek V3	72.1	39.9	36.5	32.6	35.6	32.8	22.6	33.2	26.0
	Gemma-3-27B	64.8	37.2	33.7	12.6	50.0	17.7	26.5	35.5	27.0
	Qwen-3-32B	73.9	32.1	29.1	21.8	37.7	41.5	17.5	24.5	23.0
	Llama-3.3-70B	50.7	43.3	45.3	13.8	43.4	23.8	19.3	32.0	19.1
	Qwen-2.5-72B	47.9	29.3	26.7	13.8	16.0	20.8	10.8	20.5	12.5
	Qwen-3-14B	42.3	20.9	17.4	14.9	17.9	18.5	12.1	9.5	9.9
Numerical Short-Answer	Claude Sonnet 4.6	98.4	**70.4**	**92.3**	96.9	—	95.5	**81.8**	—	—
	GPT-5.5	**99.4**	62.1	88.1	**98.6**	—	**99.7**	66.7	—	—
	DeepSeek V3	87.1	32.4	86.5	38.4	—	24.5	45.5	—	—
	GPT-4o	87.6	35.1	63.5	32.3	—	25.7	63.6	—	—
	Qwen-3-32B	79.4	13.5	65.4	34.5	—	21.7	45.5	—	—
	Gemma-3-27B	68.0	22.2	67.3	27.4	—	13.7	54.5	—	—
	Qwen-2.5-72B	74.2	25.0	46.2	30.4	—	12.4	54.5	—	—
	Qwen-3-14B	67.0	10.8	63.5	26.9	—	17.3	54.5	—	—
	Llama-3.3-70B	65.5	27.0	50.0	22.4	—	10.6	18.2	—	—

Table 2: Model performance (%) across all three question formats. Each row reports a model's scores across the subjects available for that format. Subject abbreviations: Phys. = Physics, Hist. = History, Geog. = Geography, Math = Mathematics, Chem. = Chemistry, Bio. = Biology, Lit. = Literature, and Gram. = Grammar. **Bold**: highest score per column within each format. Underline: lowest score per model row. Law, Literature, and Grammar have no numerical short-answer items (—).

2024).

- **Proprietary models**: Claude Sonnet 4.6 (Anthropic, 2026), GPT-4o (OpenAI et al., 2024), GPT-5.5 (OpenAI, 2026).

For the main evaluation, we use direct prompting without chain-of-thought (CoT) reasoning to ensure comparability with standard academic benchmarks such as MMLU (Hendrycks et al., 2021) and to measure model knowledge independently of reasoning strategy. CoT prompting is applied selectively to Gemma-3-27B and Qwen-3-14B on grammar questions, where it serves as a diagnostic tool to expose systematic failure modes rather than as a performance optimization.

4.1 Evaluation Protocol

Evaluation is performed separately across subject category, question format, language understanding, and country-specific knowledge. Multiple-choice questions are evaluated using exact-match accuracy. Matching questions require all four pair assignments to be simultaneously correct. Numerical short-answer questions are evaluated using strict exact-match against the reference answer. All prompts are presented in Tajik Cyrillic unless otherwise specified.

4.2 Transliteration and Translation Experiments

To investigate the role of script representation and language familiarity, we evaluate two additional conditions for open-weight models:

1. **Farsi transliteration**: Tajik Cyrillic is converted to Persian script using the `ParsTranslit` (Merchant and Tang, 2026) toolkit.

2. **English translation**: Questions are translated into English using Claude Sonnet 4.6 (Anthropic, 2026).

Since Tajik and Persian are closely related languages sharing substantial lexical and grammatical overlap (Windfuhr and Perry, 2009), the pri-

mary difference between the original Tajik Cyrillic and the transliterated Farsi condition is orthographic script. If script unfamiliarity is the dominant bottleneck, we would expect measurable performance gains after transliteration. If gains remain marginal, this would suggest that the bottleneck lies elsewhere, likely in the scarcity of Tajik-specific training data rather than script representation (Joshi et al., 2020). To further distinguish between low-resource language limitations and conceptual reasoning deficits, we additionally translate questions into English, removing both the script barrier and the language barrier while preserving question semantics. Large gains after English translation but not after Farsi transliteration would suggest that model failures are driven primarily by language representation rather than an inability to reason about the underlying content. Both conditions are evaluated on multiple-choice questions only.

4.3 Chain-of-Thought Analysis

To understand *why* models fail, we apply chain-of-thought prompting (Wei et al., 2022) to the best-performing open-weight model on multiple-choice questions (Gemma-3-27B) and the worst-performing open-weight model on multiple-choice questions (Qwen-3-14B). We deliberately select these two extremes to examine whether reasoning behavior differs systematically across the performance spectrum. CoT prompting is applied exclusively to grammar questions, as these subjects produce the lowest and most linguistically interesting failure patterns across all models. By eliciting explicit reasoning patterns, we aim to identify the specific points at which model reasoning breaks down.

5 Results

5.1 Overall Performance

Table 2 presents model performance across all three question formats. On multiple-choice, GPT-5.5 achieves the strongest overall performance, followed by Claude Sonnet 4.6, and GPT-4o among proprietary models. Among open-weight models, Gemma-3-27B ranks first on multiple-choice format, followed by Qwen-3-32B, DeepSeek V3, Llama-3.3-70B, Qwen-2.5-72B, and Qwen-3-14B.

5.2 Subject-Level Analysis

Physics consistently produces the highest scores across every format and model on average. By contrast, Tajik Grammar produces the lowest multiple-choice accuracy of any subject on average, which is lower than literature, which also carries cultural grounding. Even GPT-5.5 scores lowest on grammar (93.7%) and on literature (91.6%). For open-weight models, grammar scores cluster between 36–45%, barely exceeding random chance for a four-option question (25%). Notably, mathematics also underperforms relative to other STEM subjects: Despite strong symbolic reasoning ability, models struggle with matching and numerical short-answer math questions. Contrary to expectations, the performance gap between STEM subjects and culturally grounded subjects is smaller and less consistent than anticipated.

5.3 Format-Level Analysis

The three question formats reveal distinct aspects of model capability and produce notably different performance profiles. Multiple-choice performance is the highest across all models, benefiting from the ability to explicitly compare answer options and identify the correct answer (Robinson and Wingate, 2023). Matching questions are substantially more discriminative: The gap between proprietary and open-weight models widens considerably under this format, with GPT-5.5 achieving 94.5% while all open-weight models fall below 37%. Numerical short-answer questions are the most challenging format overall, with performance lagging multiple-choice for every model. This gap is particularly pronounced for quantitative subjects: DeepSeek V3 drops from 70.6% on mathematics MCQ to 38.4% on numerical short-answer, and GPT-4o from 53.2% to 32.3%. This mirrors findings from prior work showing that MCQ scaffolding can substantially overestimate model capability relative to generative evaluation (Singh et al., 2025b).

5.4 Tajik Language Understanding

Tajik grammar and literature represent the most challenging subjects across all models. Even frontier proprietary models show notable weaknesses in fine-grained grammar subtopics. As shown in Table 3, models exhibit persistent failure across morphology, syntax, phonetics and spelling, lexicology, classical literature, and modern litera-

Model	Morph.	Syntax	Phonetics	Lexicology	Classical Lit.	Modern Lit.	Folklore	Lit. Devices
GPT-5.5	**94.7**	**92.1**	**92.8**	**91.5**	**93.0**	**76.8**	**97.3**	**97.4**
Claude Sonnet 4.6	70.5	67.6	45.4	83.9	80.4	69.5	92.3	93.6
GPT-4o	59.3	51.9	47.4	81.1	76.8	60.5	88.5	85.9
Gemma-3-27B	44.7	31.1	35.1	55.6	56.0	49.7	73.1	66.7
Llama-3.3-70B	38.5	37.3	30.9	51.1	55.0	52.1	68.6	41.0
DeepSeek V3	30.3	34.7	28.9	43.9	49.7	40.1	59.0	51.3
Qwen-3-32B	32.8	37.8	32.5	47.8	47.5	47.3	60.3	52.6
Qwen-2.5-72B	35.2	31.1	29.4	45.0	57.0	42.5	63.5	47.4
Qwen-3-14B	29.1	30.7	27.8	42.2	40.5	36.5	46.2	41.0

Table 3: Model performance (%) on Tajik language and literature subtopics. All question formats are included. **Bold**: highest score per column. Underline: lowest score per model row. "Morph." denotes Morphology.

ture tasks. Morphology and phonetics/spelling are the hardest grammar subtopics: Qwen-3-14B scores 29% and 28%, respectively, while GPT-5.5 achieves 95% and 93%, respectively.

5.5 Culturally Grounded Knowledge

Performance on Tajikistan-specific knowledge falls substantially below STEM subjects for every open-weight model. As shown in Table 4, open-weight models range from 26% (Qwen-3-14B) to 41% (Llama-3.3-70B) on Tajik geography, while GPT-5.5 achieves 94%. The legal system and history subtopics similarly remain challenging, with open-weight models clustered between 36% and 61%.

Model	History	Geography	Legal System
GPT-5.5	**97.4**	**93.7**	**95.7**
Claude Sonnet 4.6	83.8	53.6	84.1
GPT-4o	77.8	61.7	76.1
Gemma-3-27B	53.7	33.9	60.7
Llama-3.3-70B	53.8	41.1	54.3
DeepSeek V3	49.5	32.0	45.4
Qwen-3-32B	42.2	30.6	52.2
Qwen-2.5-72B	46.3	33.1	38.5
Qwen-3-14B	36.1	25.8	39.8

Table 4: Model performance (%) on Tajikistan country-specific knowledge subtopics. All question formats are included. **Bold**: highest score per column. Underline: lowest score per model row.

6 Transliteration and Translation Analysis

6.1 Transliteration into Farsi Script

Figure 2 (left) shows performance of open-weight models on Farsi-transliterated data, with deltas relative to original Tajik. Script familiarity provides negligible benefit for Tajik language understanding. Gains across all open-weight models are minimal, ranging from -0.9 pp (Qwen-3-14B) to $+1.2$ pp (Llama-3.3-70B), and remain consistent across all subjects. While residual transliteration errors may introduce minor noise, the consistency of near-zero deltas across all models and subjects confirms that script representation is not the primary bottleneck for Tajik language understanding.

6.2 Translation into English

Figure 2 (right) presents results on English-translated data, translated using Claude Sonnet 4.6. English translation yields substantially larger gains than Farsi transliteration for all five open-weight models. Average improvements range from $+1.4$ pp (Gemma-3-27B) to $+14.3$ pp (Qwen-3-14B). Qwen-3-14B shows the largest gains, improving by $+17.9$ pp on biology, $+17.7$ pp on physics, and $+17.0$ pp on literature. Grammar is the subject that benefits least from English translation (average gain $+2.3$ pp), substantially below other subjects which average $+8$–$+13$ pp, indicating that grammar failures reflect structural language understanding deficits rather than data scarcity alone.

6.3 Language and Knowledge Deficits

Section 6.2's subject-level gains conflate two sources of error: the language barrier, where the model cannot parse Tajik input, and the knowledge deficit, where it lacks the information in any language. To separate them we compare English-translated and original Tajik versions of three Tajikistan-specific subtopics: History of the Tajik People, Tajik geography, and the legal system. Law is wholly Tajikistan-specific, so no subtopic filter was applied. Physics serves as a contrast subject, universally represented in pretraining and not culturally grounded. All results are multiple-choice only; Tajik baselines therefore differ from Table 4, which pools formats.

Table 5 shows that translation yields +7–10 pp on national history and the legal system, but mod-

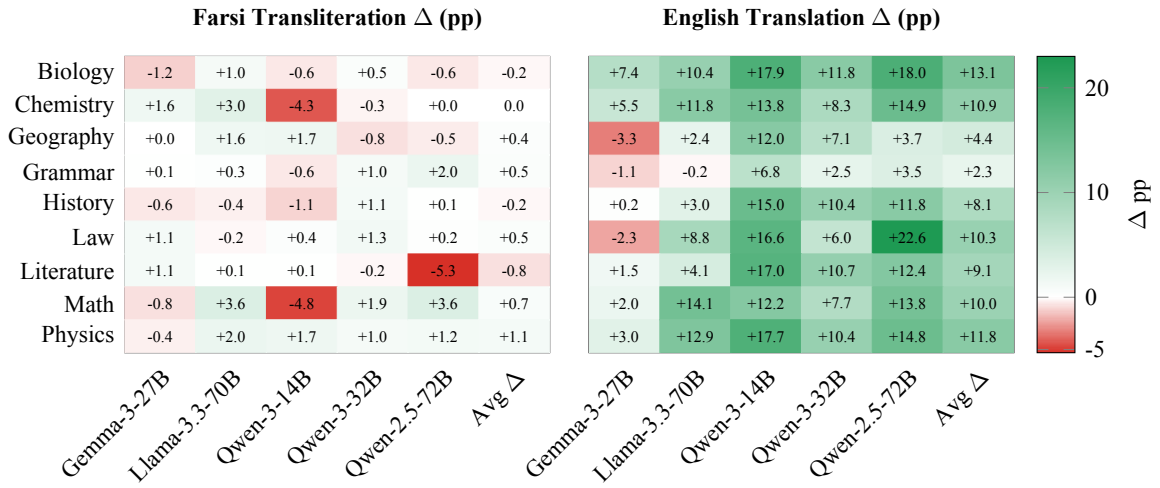


Figure 2: Performance delta (pp) relative to Tajik Cyrillic baseline under Farsi transliteration (left) and English translation (right), on a shared scale. The rightmost column shows the average delta across all five models per subject.

els still average only 61–63% under English. Tajik geography gains nothing from translation (−2.8 pp) and remains at 37%. Physics gains +11.8 pp and reaches 80.4%, a level no Tajikistan-specific subtopic approaches in either language.

Culturally grounded content is not uniformly constrained. The legal system and national history have some representation in English-language sources, so models hold partial knowledge that the language barrier blocks, hence removing it yields accuracy gains. Tajik geography (administrative divisions, regional economic data, local toponyms) is likely underrepresented even in English-language pretraining corpora, leaving little knowledge for translation to unlock.

6.4 Implications

The near-zero transliteration deltas confirm that script representation is not a meaningful bottleneck for Tajik language understanding. The translation experiments reveal that the remaining constraint is not uniform. For universally represented content, Tajik-language data scarcity is the dominant limitation and translation largely closes the gap. Culturally grounded content splits further: Topics with some cross-lingual presence, such as the legal system and national history, show substantial translation gains but retain a knowledge floor, while topics with little cross-lingual presence, such as Tajik geography, show no translation benefit at all.

The magnitude of the language component is inversely related to Tajik-language proficiency. Models with weaker Tajik performance gain the most from translation, with average deltas ranging from +1.4 pp for Gemma-3-27B to +14.3 pp for Qwen-3-14B. This spread is amplified on culturally grounded subtopics: On the legal system, Qwen-2.5-72B gains +22.6 pp and Qwen-3-14B +16.6 pp, while Gemma-3-27B loses 2.3 pp. The weaker models possess relevant knowledge but are blocked from accessing it by the language barrier; Gemma can already access what knowledge it has through Tajik input.

7 Chain-of-Thought Analysis

To understand how models reason on linguistically demanding tasks, we manually inspect chain-of-thought outputs from Gemma-3-27B (best-performing open-weight model on multiple-choice questions) and Qwen-3-14B (lowest-performing open-weight model on multiple-choice questions) on grammar questions. We focus on this subject as it requires native Tajik linguistic knowledge that cannot be recovered through general reasoning, making it the most effective diagnostic for identifying language-specific failure modes. Through qualitative analysis of reasoning traces, we identify recurring failure patterns that are consistent across both models.

7.1 Reasoning Behavior

Despite their performance gap, both models adopt the same fundamental strategy: translate or paraphrase the question into English before reasoning. Although the resulting reasoning chains often ap-

Model	TJ	EN	Δ
National History (n=603)			
Gemma-3-27B	61.4	61.4	0.0
Llama-3.3-70B	60.5	63.0	+2.5
Qwen-3-14B	43.0	57.7	+14.8
Qwen-3-32B	48.1	57.5	+9.5
Qwen-2.5-72B	54.2	64.0	+9.8
Mean	53.4	60.7	+7.3
Tajik Geography (n=200)			
Gemma-3-27B	42.0	36.0	−6.0
Llama-3.3-70B	48.5	41.0	−7.5
Qwen-3-14B	32.0	33.5	+1.5
Qwen-3-32B	37.0	39.0	+2.0
Qwen-2.5-72B	41.0	37.0	−4.0
Mean	40.1	37.3	−2.8
Legal System (n=434)			
Gemma-3-27B	63.4	61.1	−2.3
Llama-3.3-70B	56.9	65.7	+8.8
Qwen-3-14B	45.2	61.8	+16.6
Qwen-3-32B	55.8	61.8	+6.0
Qwen-2.5-72B	44.0	66.6	+22.6
Mean	53.0	63.4	+10.3
Physics (n=506)			
Gemma-3-27B	72.3	75.3	+3.0
Llama-3.3-70B	64.8	77.7	+12.8
Qwen-3-14B	62.5	80.2	+17.8
Qwen-3-32B	78.1	88.5	+10.5
Qwen-2.5-72B	65.2	80.0	+14.8
Mean	68.6	80.4	+11.8

Table 5: Language–knowledge decomposition on Tajikistan-specific subtopics (multiple-choice only). TJ: original Tajik Cyrillic; EN: English translation; Δ = EN − TJ (pp).

pear structurally coherent, this strategy produces three systematic failure modes.

Hallucination is the dominant failure mode in both models (Ji et al., 2023). Both Gemma-3-27B and Qwen-3-14B frequently hallucinate the lexical properties of Tajik words, fabricating meanings, spellings, or grammatical categories, and then constructing internally consistent arguments on that false foundation. The distinction between models is one of framing rather than of kind: Gemma-3-27B tends to preface its errors with generic linguistic rule statements before misapplying them, while Qwen-3-14B more directly invents word definitions and reasons from them. In both cases the reasoning chain is logically coherent; only its factual basis is wrong.

Reasoning Process Failures Beyond hallucination, two additional failure modes emerge. In *self-reversal*, the model correctly identifies the right answer early in the reasoning chain, then talks itself into switching to an incorrect prediction through additional flawed reasoning. A third variant, *no-answer resolution failure*, occurs when none of the options satisfies the model's stated rule: Rather than revising the rule, the model arbitrarily selects among the options while acknowledging the inconsistency.

Grammatical Category Confusion Both models systematically fail on morphological classification questions by hallucinating the lexical identity of the target words. The models correctly know the classification rule but fabricate what the word in question means or what part of speech it belongs to, applying the rule correctly to that wrong identification. The result is a grammatically coherent but factually wrong classification: The reasoning is valid, but its input is invented.

8 Conclusion

We introduced TALIM, a curriculum-grounded benchmark for Tajik academic reasoning and language understanding, comprising 9,136 questions across nine subjects and three question formats. Our experiments demonstrate that current multilingual language models continue to struggle substantially with low-resource language understanding, culturally grounded reasoning, and Tajik-specific linguistic structure. GPT-5.5 achieves over 90% average multiple-choice accuracy while the best open-weight model (Gemma-3-27B) reaches slightly above 70% at best. English translation improves open-weight performance by up to 14.3 pp on average, but the benefit is uneven: Translation recovers substantial gains on universally represented content while Tajikistan-specific subtopics retain large residual deficits even under English, indicating that both language representation and knowledge coverage contribute to model failures on culturally grounded content. Chain-of-thought analysis reveals three systematic failure modes: hallucination, reasoning process failures, and grammatical category confusion. We hope TALIM will serve as a foundation for future research on low-resource multilingual evaluation, culturally grounded reasoning, and robust language understanding beyond translation-centric benchmarks.

Limitations

TALIM questions are sourced from official NTC documents that may be partially accessible online,

so pretraining contamination for proprietary models cannot be ruled out. However, open-weight models exhibit large performance gaps relative to proprietary models across all nine subjects and all three question formats, and these gaps cannot be explained by contamination of the proprietary models alone. The comparative findings and failure analyses reported in this paper therefore remain robust even if proprietary saturation numbers are partially inflated. A related concern is that our English translations were produced by Claude Sonnet 4.6, which may introduce subject-level variation in translation quality that inflates or deflates observed gains. This caveat is particularly relevant for grammar, where the small translation benefit (+2.3 pp) may partly reflect the difficulty of preserving morphological and syntactic distinctions under translation rather than a genuine structural deficit. While frontier proprietary models approach saturation on multiple-choice questions, open-weight models continue to struggle substantially across all formats, and grammar and culturally grounded subjects remain challenging even for the strongest models. Finally, our chain-of-thought analysis is qualitative in nature, based on manual inspection of reasoning traces for two models on a single subject; the identified failure modes are illustrative rather than exhaustive, and their relative frequencies have not been formally quantified. A portion of TALIM's questions may overlap with the zehnlab (2026) benchmark collection, as both draw from Tajik national curriculum materials originating from the same upstream source, the National Testing Center. However, given that zehnlab's STEM content is translated from MMLU rather than sourced from NTC materials, any overlap is likely confined to the humanities subjects and is not expected to be substantial.

Ethical Considerations

TALIM is released under the CC BY-NC 4.0 license and is intended solely for non-commercial research purposes; derivative works should respect the same access conditions. All questions were sourced from publicly administered examination materials of the National Testing Center of Tajikistan. No personally identifiable information was collected. Annotators were fairly compensated; all received payment exceeding the average wage in their region. We did not conduct explicit bias audits of the source materials and cannot guarantee that the dataset is free from cultural, pedagogical, or ideological biases inherent to standardized educational assessments. The authors acknowledge the use of AI-assisted tools for grammar and language refinement during manuscript preparation. These tools were not involved in research design, data collection, analysis, or implementation.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00208054).

References

Anthropic. 2026. Claude sonnet 4.6 system card. https://www.anthropic.com/claude-sonnet-4-6-system-card. Accessed: May 2026.

Mullosharaf K. Arabov. 2026a. Benchmarking parameter-efficient fine-tuning of large language models for low-resource tajik text generation with the tajik web corpus. *Preprint*, arXiv:2605.03742.

Mullosharaf Kurbonovich Arabov. 2026b. TajPersLexon: A Tajik–Persian lexical resource and hybrid model for cross-script low-resource NLP. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 29–37, Rabat, Morocco. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. Ethnologue: Languages of the world. https://www.ethnologue.com.

Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong Park. 2023. Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In *Proceedings of the 61st*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11857–11870, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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.

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*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Daeen Kabir, Minhajur Rahman Chowdhury Mahim, Sheikh Shafayat, Adnan Sadik, Arian Ahmed, Eunsu Kim, and Alice Oh. 2025. BLUCK: A benchmark dataset for Bengali linguistic understanding and cultural knowledge. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 161–179, Mumbai, India. Association for Computational Linguistics.

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.

Rayyan Merchant and Kevin Tang. 2026. ParsTranslit: Truly versatile Tajik-Farsi transliteration. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1431–1443, Rabat, Morocco. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2026. Introducing GPT-5.5. `https://openai.com/index/introducing-gpt-5-5/`. Accessed: May 2026.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

Annika Simonsen, Dan Saattrup Nielsen, and Hafsteinn Einarsson. 2025. FoQA: A Faroese

question-answering dataset. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 48–57, Tallinn, Estonia. University of Tartu Library, Estonia.

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. 2025a. 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.

Shrutika Singh, Anton Alyakin, Daniel Alexander Alber, Jaden Stryker, Ai Phuong S Tong, Karl Sangwon, Nicolas Goff, Mathew De La Paz, Miguel Hernandez-Rovira, Ki Yun Park, and 1 others. 2025b. The pitfalls of multiple-choice questions in generative ai and medical education. *Scientific reports*, 15(1):42096.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Gernot Windfuhr and John R. Perry. 2009. Persian and tajik. In Gernot Windfuhr, editor, *The Iranian Languages*, pages 416–544. Routledge, London and New York.

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.

zehnlab. 2026. Tajik benchmarks. `https://huggingface.co/collections/zehnlab/tajik-benchmarks`. Accessed: May 2026.

Handong Zheng, Yumeng Li, Kaile Zhang, Liang Xin, Guangwei Zhao, Hao Liu, Jiayu Chen, Jie Lou, Qi Fu, Rui Yang, Shuo Jiang, Weijian Luo, Weijie Su, Weijun Zhang, Xingyu Zhu, Yabin Li, Yiwei ma, Yu Chen, Yuqiu Ji, and 6 others. 2026. Multimodal ocr: Parse anything from documents. *Preprint*, arXiv:2603.13032.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

A Direct Prompting

We adopt a direct prompting strategy with no chain-of-thought reasoning across all models. Prompts are written in the target language and instruct the model to respond in a structured format without explanation. Figure 3 provides concrete examples of the prompt templates used for each question type.

Multiple-Choice

Автобуси мусофиркаш аз шаҳри Душанбе ба шаҳри Вахдат дар $t = 20$ дақиқа бо суръати миёнаи $\Theta = 15$ м/с рафта расид. Масофаи миёнаи байни ин шаҳрҳо чӣ қадар аст?
A: 18 км
B: 35 км
C: 9 км
D: 17,5 км

Танҳо ҷавоби дурустро дар ин шакл нависед: «A», «B», «C» ё «D». Шарҳ надиҳед.

Matching

Мувофиқатро муайян кунед:

A: кӯчиш	1: дарозии қитъаи масир (траектория), ки ҷисм тай кардааст
B: тағйирёбии бузургии физикӣ	2: объекти дилхоҳ интихобшуда, ки нисбат ба он мавқеи ҷисмҳои ҳаракаткунанда муайян мешавад
C: ҷисми сарҳисоб	3: вектори, ки аз мавқеи аввалаи ҷисм ба мавқеи охираи он гузаронда шудааст
D: қонуни ҳаракат	4: фарқи қимати охира ва аввалаи бузургии физикӣ
	5: вобастагии радиусвектор ё координатаҳо аз вақт

Ҷавобро танҳо дар ин шакл нависед: {``A'': ``1'', ``B'': ``2'', ``C'': ``3'', ``D'': ``4''}. Шарҳ надиҳед. Як вариантро ду маротиба истифода набаред.

Numerical Short-Answer

Ҳаракати дучархарон бо муодилаи $x = 150 - 4t$ (м) дода шудааст. Координатаи аввалаи дучархарон чӣ қадар аст?

Ҷавобро танҳо дар ин шакл нависед: `<answer> = … </answer>`. Шарҳ надиҳед.

Figure 3: Example prompt templates (Tajik Cyrillic) for each question format. Instruction lines are shown in *italics*.

All models are evaluated in a zero-shot setting.

Prompt

You are solving a Tajik grammar multiple-choice question. The question and answer choices are written in Tajik.

Дар кадом калимаҳо ба ҷойи ӯ ҳарфи у навишта шудааст?
A: кух, обру(й)
B: муъмин, муътабар
C: муҳра, муҳсин
D: буҳрон, буҳтон

Think step by step: identify the grammar rule being tested, evaluate each answer choice, and explain which one is correct.

End your response with exactly: "Answer: X" (where X is one of A, B, C, or D).

Figure 4: CoT example prompt template on a Tajik grammar question.

Model	Subj.	TJ	EN	Δ
Gemma-3-27B	Gram.	45.0	45.2	+0.3
Gemma-3-27B	Phys.	72.3	72.3	+0.0
Qwen-3-14B	Gram.	36.2	37.3	+1.0
Qwen-3-14B	Phys.	62.5	67.0	+4.5

Table 6: Instruction-language ablation (MCQ accuracy, %). Questions and options are Tajik Cyrillic in both conditions; only the instruction line differs. TJ/EN: Tajik/English instruction. Δ = EN − TJ (pp). n = 789 (grammar), 506 (physics).

No in-context examples or demonstrations are provided, and models are given only the question and the format instruction.

B Instruction-Language Ablation

The main evaluation (Appendix A) uses Tajik instructions while the chain-of-thought analysis (Section 7) uses English instructions. To verify that instruction language does not confound either evaluation, we compare the two instruction conditions in the direct-prompting setup on grammar and physics multiple-choice questions for Gemma-3-27B and Qwen-3-14B. The question and answer options remain in Tajik Cyrillic in both conditions; only the instruction line differs.

Table 6 shows that three of four conditions show deltas below 1 pp, and McNemar's test confirms that the discordant pairs are not significant ($p > 0.4$). The exception is Qwen-3-14B on physics, which gains +4.5 pp under English instructions ($p = 0.001$). Format compliance is identical across

both conditions for all cells, with both models producing a parseable answer on every item, so the gain does not reflect improved instruction following. Overall, instruction language has negligible impact on direct-prompting accuracy, supporting the use of English instructions in the chain-of-thought analysis.

C Chain-of-Thought Prompting

Model Response *(Gemma-3-27B)*

The question asks in which words the letter "у" is written instead of "ӯ". This tests knowledge of when to use "ӯ" versus "у" in Tajik orthography. "ӯ" generally appears in native Tajik words and loanwords that have been naturalized. "у" is ***often found in Arabic and Persian loanwords****.*

Let's analyze each option:

A: кух, обру(ӣ) *— "кух" (mountain peak) uses "у" correctly. "обру(ӣ)" (reputation, honor) also uses "у" correctly. This is a strong candidate.*

B: муъмин, муътабар *— Both "муъмин" (believer) and "муътабар" (reliable) use "у" correctly. These are Arabic loanwords where "у" is standard. This is also a strong candidate.*

C: мухра, мухсин *— "мухра" (seal) and "мухсин" (benefactor) both should use "ӯ" instead of "у". These are words where "ӯ" is expected.*

D: бухрон, бухтон *— "бухрон" (crisis) and "бухтон" (slander) both should use "ӯ" instead of "у". These are words where "ӯ" is expected.*

Comparing A and B, both are correct in using "у". However, the question asks where "у" is correctly used instead of "ӯ". Options C and D are incorrect because they should be written with "ӯ". Options A and B are both correct, but the question implies only one correct answer. Considering the common usage and the origin of the words, option A is slightly more representative of the rule being tested.

Answer: A

Figure 5: Full reasoning trace of Gemma-3-27B on the grammar question from Figure 4. The hallucinated rule is highlighted in **red**.

In addition to direct prompting, we conduct a Chain-of-Thought (CoT) evaluation on a subset of models on tajik grammar. Unlike the zero-shot direct prompting setup, CoT prompts are written in English and explicitly instruct the model to reason step by step before producing a final answer. This cross-lingual prompting strategy of putting English instructions over Tajik content is intentional, as it leverages the stronger instruction-following capabilities of models in high-resource languages while still requiring comprehension of the Tajik question.

CoT evaluation is restricted to multiple-choice questions. The final answer is extracted by matching the last occurrence of "Answer: X" in the model output. The prompt template is described in Figure 4.

Figure 5 shows the chain of thought for Gemma-3-27B to the question in Figure 4. This question tests Tajik orthographic knowledge of when to write **у** versus **ӯ**. The model's reasoning illustrates a common failure mode: It fabricates a plausible but incorrect rule, stating that **у** is often found in Arabic and Persian loanwords, when in fact the reverse is closer to the truth. Despite this hallucinated rule, the model arrives at the correct answer (A) through elimination, demonstrating that correct predictions do not necessarily reflect correct linguistic understanding.

D Fine-Grained Topic Breakdown by Subject

Subject	Topic	Cl.S.4.6	DS-V3	GPT-4o	GPT-5.5	Ge-27B	Ll-70B	Qw3-32B	Qw2.5-72B	Qw3-14B
Biology	Botany	72.2	40.5	78.9	96.4	48.9	43.8	44.9	38.8	41.0
	Ecology	95.1	63.4	90.2	97.5	75.6	70.7	78.0	58.5	58.5
	General Biology & Ecology	90.2	63.7	85.0	94.9	76.7	71.2	79.1	66.0	65.1
	Human Anatomy, Physiology & Hygiene	83.1	46.0	86.8	95.2	55.0	54.0	51.9	45.0	41.3
	Zoology	76.1	38.8	77.4	97.5	45.9	45.9	35.9	34.9	32.5
	Zoology: Tajikistan	66.7	33.3	0.0	100.0	0.0	0.0	33.3	33.3	0.0
Chemistry	Basic Concepts & Laws	91.2	37.1	51.4	97.1	40.0	40.0	55.7	41.4	38.6
	Chemistry of the Elements	83.7	48.6	61.0	100.0	40.4	50.7	62.3	42.5	52.7
	Organic Chemistry	76.6	44.6	52.0	99.3	43.2	39.9	56.1	42.6	41.9
	Periodic Table & Atomic Structure	93.8	50.0	70.8	100.0	52.1	54.2	70.8	62.5	52.1
	Solutions	83.1	66.2	73.2	96.9	64.8	57.7	66.2	54.9	59.2
Geography	Economic & Social World Geography	88.9	75.6	80.0	100.0	68.9	75.6	64.4	60.0	64.4
	Econ. & Social Geo. (Countries)	95.5	60.6	90.8	97.7	75.8	78.0	70.5	71.2	50.8
	General Physical Geography	93.9	72.8	90.4	100.0	70.2	66.7	69.3	64.0	53.5
	Geography of Continents & Oceans	92.9	75.2	89.1	99.2	77.3	75.2	63.4	67.6	53.8
	Physical & Economic Geo. of Tajikistan	62.5	39.5	70.0	95.7	42.0	48.5	37.0	41.0	32.0
Grammar	Lexicology & Phraseology	83.9	43.9	81.1	91.5	55.6	51.1	47.8	45.0	42.2
	Literature	75.5	43.4	67.9	98.0	39.6	45.3	47.2	37.7	30.2
	Morphology	75.0	34.0	65.6	95.6	50.0	43.4	35.8	40.1	33.5
	Phonetics, Phonemic Phenomena & Spelling	43.2	29.0	45.1	93.5	37.0	34.6	37.0	35.2	32.1
	Syntax	69.2	40.1	55.5	92.5	37.4	46.7	46.2	40.7	39.0
History	General History	100.0	83.6	100.0	100.0	94.5	90.4	80.8	72.6	79.5
	History of the Tajik People	87.6	54.7	82.4	97.9	61.4	60.5	48.1	54.2	43.0
	World History	95.6	79.4	97.1	98.5	91.9	89.7	83.8	80.9	75.0
Literature	Artistic Devices	93.6	51.3	85.9	97.4	66.7	41.0	52.6	47.4	41.0
	Classical Literature	80.4	49.7	76.8	93.0	56.0	55.0	47.5	57.0	40.5
	Folklore. Literary Concepts	88.1	61.2	85.1	96.8	80.6	73.1	62.7	65.7	59.7
	Folklore. Literary Concepts. Life	95.5	57.3	91.0	97.7	67.4	65.2	58.4	61.8	36.0
	Modern Literature	69.5	40.1	60.5	76.8	49.7	52.1	47.3	42.5	36.5
	Vocabulary	95.0	65.0	95.0	95.0	80.0	80.0	55.0	60.0	45.0
Mathematics	Algebra & Number Sense	98.4	67.9	51.9	100.0	51.1	42.0	68.7	54.2	52.7
	Applied Math & Data Interpretation	98.3	84.7	54.2	100.0	64.4	54.2	71.2	55.9	52.5
	Equations & Inequalities	97.7	56.7	37.8	98.9	44.4	36.7	47.8	40.0	36.7
	Functions, Trigonometry & Calculus	98.8	81.2	57.1	98.8	55.3	41.8	57.6	59.4	45.3
	Geometry	94.3	62.9	60.0	100.0	66.7	38.1	72.4	61.0	48.6
Physics	Atomic & Nuclear Physics	93.2	71.2	72.6	98.6	61.6	61.6	69.9	56.2	63.0
	Electrodynamics	98.7	78.9	75.0	98.7	69.7	55.3	81.6	65.8	68.4
	Mechanics	100.0	81.2	69.1	100.0	70.1	66.0	82.5	64.9	59.8
	Molecular Physics & Thermodynamics	100.0	84.1	72.7	100.0	77.3	65.9	84.1	61.4	65.9
	Optics	87.9	81.8	75.0	97.0	72.7	72.7	72.7	69.7	60.6

Table 7: Fine-grained topic-level multiple-choice accuracy (%) for all models across all subjects and topics. Model abbreviations: Cl.S.4.6 = Claude Sonnet 4.6, DS-V3 = DeepSeek V3, Ge-27B = Gemma-3-27B, Ll-70B = Llama-3.3-70B, Qw3-32B = Qwen-3-32B, Qw2.5-72B = Qwen-2.5-72B, Qw3-14B = Qwen-3-14B.

Subject	Topic	Cl.S.4.6	DS-V3	GPT-4o	GPT-5.5	Ge-27B	Ll-70B	Qw3-32B	Qw2.5-72B	Qw3-14B
Biology	Botany	41.5	3.8	45.3	90.5	5.7	3.8	5.7	3.8	3.8
	Ecology	66.7	34.8	58.3	90.0	54.2	33.3	20.8	29.2	20.8
	General Biology & Ecology	65.3	36.7	51.0	89.5	38.8	36.7	34.7	24.5	32.7
	Human Anatomy, Physiology & Hygiene	84.2	47.4	84.2	88.2	39.5	31.6	31.6	7.9	10.5
	Zoology	56.4	5.5	49.1	98.0	12.7	3.6	1.8	0.0	0.0
Chemistry	Chemistry of the Elements	87.5	41.7	50.0	100.0	25.0	29.2	58.3	29.2	25.0
	Organic Chemistry	65.5	28.6	36.1	100.0	13.9	16.7	30.6	11.1	8.3
Geography	Economic & Social World Geography	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66.7
	Econ. & Social Geo. (Countries)	84.6	80.8	88.5	100.0	80.8	84.6	65.4	65.4	50.0
	Geography of Continents & Oceans	100.0	77.8	100.0	100.0	55.6	100.0	33.3	33.3	0.0
	Geography of Tajikistan	38.9	0.0	27.8	85.7	0.0	22.2	5.6	0.0	0.0
	Physical & Economic Geo. of Tajikistan	3.3	0.0	26.7	73.3	0.0	3.3	3.3	0.0	0.0
Grammar	Literature	96.6	62.1	89.7	100.0	79.3	62.1	72.4	58.6	34.5
	Morphology	40.6	6.2	16.1	86.4	9.4	6.2	12.5	3.1	0.0
	Phonetics, Phonemic Phenomena & Spelling	56.2	28.1	59.4	89.7	25.0	12.5	9.4	0.0	6.2
	Syntax	62.7	17.5	40.7	90.9	11.9	8.5	11.9	1.7	5.1
History	General History	100.0	83.3	100.0	100.0	83.3	83.3	83.3	72.2	61.1
	History of the Tajik People	68.7	28.4	59.3	94.7	22.7	26.7	18.7	14.7	8.7
	World History	97.7	60.5	90.7	100.0	69.8	86.0	58.1	65.1	46.5
Mathematics	Algebra & Number Sense	84.2	22.2	8.1	100.0	5.4	8.1	13.5	8.1	8.1
	Geometry	100.0	33.3	83.3	100.0	33.3	16.7	50.0	33.3	50.0
Physics	Electrodynamics	94.9	79.5	87.2	94.4	69.2	56.4	74.4	43.6	41.0
	Mechanics	100.0	86.2	100.0	100.0	75.9	58.6	79.3	69.0	55.2
	Molecular Physics & Thermodynamics	85.7	65.0	81.0	100.0	57.1	38.1	66.7	38.1	38.1
	Optics	78.6	30.8	57.1	100.0	28.6	28.6	57.1	14.3	14.3

Table 8: Fine-grained topic-level matching question accuracy (%) for all models across all subjects and topics. Model abbreviations: Cl.S.4.6 = Claude Sonnet 4.6, DS-V3 = DeepSeek V3, Ge-27B = Gemma-3-27B, Ll-70B = Llama-3.3-70B, Qw3-32B = Qwen-3-32B, Qw2.5-72B = Qwen-2.5-72B, Qw3-14B = Qwen-3-14B.

Subject	Topic	Cl.S.4.6	DS-V3	GPT-4o	GPT-5.5	Ge-27B	Ll-70B	Qw3-32B	Qw2.5-72B	Qw3-14B
Biology	General Biology	80.0	40.0	60.0	62.5	50.0	20.0	40.0	50.0	50.0
Chemistry	Basic Concepts & Laws	98.1	23.2	23.2	100.0	16.4	14.5	25.0	7.1	16.1
	Chemistry of the Elements	94.4	19.1	12.8	100.0	8.5	6.4	12.8	12.8	10.6
	Organic Chemistry	100.0	23.2	14.3	100.0	14.3	8.9	12.5	5.4	16.1
	Periodic Table & Atomic Structure	78.6	21.2	33.3	100.0	24.2	6.1	24.2	18.2	15.2
	Solutions	95.0	17.4	32.6	100.0	10.9	6.5	21.7	10.9	15.2
Mathematics	Algebra & Number Sense	97.1	18.8	12.5	97.9	10.4	29.2	10.4	8.3	12.5
	Applied Math & Data Interpretation	100.0	44.7	48.9	100.0	31.9	17.0	42.6	42.6	34.0
	Equations & Inequalities	97.2	37.9	33.9	99.1	24.1	15.5	31.9	30.2	24.1
	Functions, Trigonometry & Calculus	98.2	35.7	28.0	100.0	28.6	21.4	34.1	32.5	26.2
	Geometry	92.6	41.7	27.8	100.0	27.8	25.0	41.7	27.8	19.4
Physics	Atomic & Nuclear Physics	93.3	93.3	85.7	100.0	80.0	80.0	86.7	93.3	86.7
	Electrodynamics	100.0	87.8	92.7	100.0	68.3	58.5	82.9	73.2	68.3
	Mechanics	97.7	88.4	83.7	100.0	65.1	74.4	83.7	76.7	69.8
	Molecular Physics & Thermodynamics	100.0	78.3	78.3	100.0	52.2	65.2	56.5	73.9	47.8
	Optics	92.9	57.1	78.6	100.0	64.3	42.9	64.3	42.9	42.9

Table 9: Fine-grained topic-level numerical short-answer accuracy (%) for all models across all subjects and topics. Model abbreviations: Cl.S.4.6 = Claude Sonnet 4.6, DS-V3 = DeepSeek V3, Ge-27B = Gemma-3-27B, Ll-70B = Llama-3.3-70B, Qw3-32B = Qwen-3-32B, Qw2.5-72B = Qwen-2.5-72B, Qw3-14B = Qwen-3-14B.