ZEnQA: A Parallel Dzongkha-English Evaluation Suite for Foundation Models on Low-resource Multilingual Reasoning

Anonymous ACL submission

Abstract

Multilingual evaluation of foundation models still underrepresents ultra-low-resource languages, especially in educational reasoning. We introduce ZEnQA, the first parallel Dzongkha–English multiple-choice question answering benchmark derived from Bhutanese national examinations. ZEnQA contains 5,161 aligned STEM questions from grades 8, 10, and 12, with parallel question text, answer options, and a shared answer key, enabling controlled cross-lingual comparison. Using ZEnQA, we evaluate foundation models under zero-shot, few-shot, chain-of-thought, and translation-augmented prompting. Results show a persistent English–Dzongkha performance gap across models. Chain-of-thought prompting improves procedural and reasoning-heavy questions more reliably than factual questions, while appending English translations to Dzongkha prompts yields consistent gains. A prompt-language ablation further shows that the gap is driven primarily by question language rather than instruction language. We also find that translation augmentation transfers to Dzongkha BIG-Bench Hard. Beyond the benchmark, ZEnQA provides a reproducible protocol for constructing parallel evaluation resources for ultra-low-resource educational settings.

1 Introduction

Foundation models are increasingly evaluated as multilingual reasoning systems, yet most evaluation remains concentrated in English and a small set of high-resource languages. This imbalance obscures substantial variation in model reliability across languages, scripts, and domains. The issue is especially consequential in education: if models perform poorly in a student's language of instruction, their apparent reasoning ability may not translate into useful or equitable educational support.

Dzongkha, the national language of Bhutan, is a clear instance of this gap. Although it is used in education, government, and public life, it remains largely absent from mainstream NLP benchmarks. Existing multilingual evaluations have expanded coverage beyond English, but ultra-low-resource languages with non-Latin scripts, limited pretraining representation, and locally grounded educational terminology remain under-evaluated (Ponti et al., 2020; Doddapaneni et al., 2023; Lai et al., 2023). As a result, it is still unclear whether current foundation models can answer curriculum-based questions in Dzongkha with the same reliability often observed in English.

We introduce ZEnQA, the first parallel Dzongkha–English benchmark for educational question answering. ZEnQA contains 5,161 aligned multiple-choice STEM questions from Bhutanese national examinations for grades 8, 10, and 12, covering Biology, Chemistry, Physics, Mathematics, and general Science. Each item includes parallel Dzongkha and English question text, answer options, and a shared answer key, enabling controlled cross-lingual evaluation: the underlying content and correct answer remain fixed, while the evaluation language varies.

This parallel design supports a more diagnostic analysis than standard multilingual benchmark reporting. Beyond measuring whether a model answers correctly, ZEnQA enables us to examine why performance differs across languages. In particular, it helps disentangle factors that are often conflated in multilingual evaluation, including model capability, question language, prompt language, tokenization efficiency, access to scientific terminology,

and answer extraction. These factors are central to assessing model readiness in ultra-low-resource settings, where failure may reflect language-access bottlenecks rather than reasoning ability alone (Ahia et al., 2023; Petrov et al., 2023).

We study four questions:

1. **RQ1: Cross-lingual performance gap.** How large is the performance gap between English and Dzongkha on curriculum-grounded STEM question answering?

2. **RQ2: Model progress.** Do stronger proprietary and open-weight foundation models reduce this gap, or does it persist across model families?

3. **RQ3: Prompting effects.** How do chain-of-thought and translation-augmented prompting affect factual, procedural, and reasoning-heavy questions?

4. **RQ4: Failure diagnosis.** What do model failures reveal about low-resource language readiness, including prompt language, tokenization, terminology access, and answer extraction?

Our contributions are threefold. First, we introduce a locally grounded parallel Dzongkha–English QA benchmark for evaluating educational reasoning in an ultra-low-resource setting. Second, we provide a reproducible alignment protocol for constructing parallel corpora from examination sources whose language versions are not reliably ordered. The protocol combines translation-assisted candidate generation, embedding-based matching, answer-key consistency checks, and native-speaker verification, making it applicable to other low-resource educational corpora with weak metadata. Third, we present a diagnostic evaluation of foundation models across languages, model families, and prompting strategies, including chain-of-thought prompting and translation augmentation (Wei et al., 2023; Shi et al., 2022; Huang et al., 2023). Our results show a persistent English–Dzongkha performance gap across models: stronger systems improve absolute performance, but do not eliminate the gap, while translation-augmented prompting partially narrows it without task-specific fine-tuning.

Together, ZEnQA provides both a benchmark for Dzongkha and a methodological case study for evaluating foundation models in ultra-low-resource educational settings, where local relevance, script coverage, tokenization, and reproducibility are central to measuring real-world model readiness.

2 ZEnQA Benchmark

We introduce ZEnQA, a parallel Dzongkha–English multiple-choice benchmark for educational question answering. ZEnQA is constructed from Bhutanese national board examinations for grades 8, 10, and 12 and contains aligned STEM questions in both Dzongkha and English. Because the source materials are officially available in both languages, the benchmark supports controlled cross-lingual comparison in a locally grounded educational setting.

2.1 Benchmark Design

2.1.1 Dataset Curation

ZEnQA was constructed from publicly available Bhutanese national examination materials[1] in Dzongkha and English. The source exams cover grades 8, 10, and 12 and span Biology, Chemistry, Physics, Mathematics, and general Science. We collected both digital examination papers and physical answer-key materials. Three typists proficient in both Dzongkha and English independently digitized the questions and answer choices. Mathematical expressions and chemical formulae were preserved in LaTeX.

Each benchmark instance consists of a question stem, four answer options, and a single correct answer label. We excluded figure-dependent questions because the present benchmark focuses on text-only question answering. We also removed items whose Dzongkha and English versions could not be aligned reliably or whose answer keys differed across languages.

Answer keys were verified in three stages. First, digitized answer labels were cross-referenced against official Bhutan Council for School Examination and Assessment answer

[1] https://www.bcsea.bt/questions-download

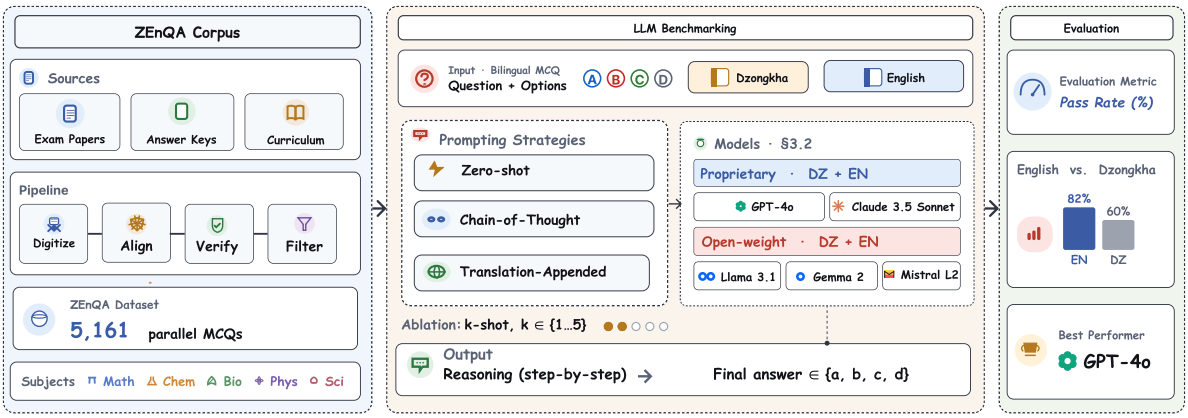

Figure 1: ZEnQA overview. The benchmark is built from Bhutanese examination sources across grades 8, 10, and 12, aligned into 5,161 parallel Dzongkha–English multiple-choice questions, and used to evaluate zero-shot, few-shot, chain-of-thought, and translation-augmented prompting.

keys. Second, discrepancies among the three typists were reconciled manually. Third, two secondary-school science teachers reviewed a random 10% sample to validate subject-level correctness. Items with unresolved answer-key inconsistencies were removed.

To support reproducibility, upon publication we will release the final dataset together with filtering logs, alignment candidates, similarity scores, native-speaker verification annotations, prompts, and the evaluation protocol.

2.1.2 Parallel Corpus Construction

Although Bhutanese national examinations are available in both Dzongkha and English, the two versions are not reliably aligned by question order. For example, a question that appears first in Dzongkha may appear much later in the English version. This makes direct metadata-based matching unreliable and motivates an explicit alignment procedure.

We first translate each English question into Dzongkha using the Google Translate API[2]. We then compute cosine similarity between the machine-translated Dzongkha question and the official Dzongkha version using OpenAI `text-embedding-ada-002`. Candidate matches are generated within each subject, after which two native Dzongkha speakers proficient in English manually verify the alignments. We further discard items whose Dzongkha and English versions have inconsistent ground-truth answers. Manual inspection suggests that these mismatches typically arise from annotation errors or intrinsic issues in the source material.

Some English questions contain minor grammatical errors or awkward phrasing, likely reflecting variation in translation quality. To test whether such artifacts affect model performance, we created grammar-corrected English variants with GPT-4 under human supervision. As shown in Appendix B, these corrections have little effect on benchmark performance.

[2] https://cloud.google.com/translate

Table 1: Statistics of the ZENQA dataset by subject and grade level.

Group	#Fields	#Instances (%)
Grouped by subject		
Biology	3	1004 (19.45%)
Chemistry	3	1160 (22.48%)
Physics	3	970 (18.79%)
Mathematics	5	1794 (34.76%)
Science	1	233 (4.51%)
Grouped by grade		
12th Grade	8	2857 (55.36%)
10th Grade	5	1857 (35.98%)
8th Grade	2	447 (8.66%)
Total	**15**	**5161 (100.00%)**

2.2 Dataset Properties

ZENQA contains 5,161 aligned question pairs. Of these, 2,857 questions (55.36%) come from grade 12, 1,857 (35.98%) from grade 10, and 447 (8.66%) from grade 8. The benchmark covers 15 subject partitions in total: Biology, Chemistry, Physics, Mathematics, and general Science, with grade-specific subdivisions where applicable. Table 1 summarizes the distribution across grades and subjects.

2.3 Question Categorization

The benchmark contains a diverse range of question types, which we group into three categories based on the primary skill required to answer them: *Factual Knowledge*, *Procedural & Application*, and *Reasoning*. We use GPT-4 with a fixed categorization prompt to assign questions to these categories, and we use these labels in later analyses of prompting effects and cross-lingual performance differences. Biology and Chemistry lean more heavily toward factual knowledge, while Mathematics contains a larger proportion of procedural and application questions. Physics shows a more balanced mix of factual and procedural items. These differences are important for interpreting subject-level and category-level performance trends in later sections. Because these category labels are assigned with GPT-4, we treat the resulting category-level analyses as diagnostic rather than as gold-standard annotation. To estimate the reliability of these diagnostic labels, we manually validated the GPT-4 categorization on a stratified sample of 500 questions (approximately 10% of the dataset) drawn across grades and subjects. Two bilingual annotators independently assigned one of the three categories, with disagreements resolved through adjudication. Human–human agreement was measured using Cohen's κ and was 0.88, while agreement between the adjudicated human labels and the GPT-4 labels, measured as exact-match agreement, was 0.74.

3 Methodology

3.1 Experimental Setup

We evaluate foundation models on ZENQA in both English and Dzongkha using a controlled cross-lingual setting: the underlying question content and gold answer remain fixed, while the evaluation language varies. Our primary benchmark uses zero-shot prompting, which allows us to compare model behavior across languages without introducing additional few-shot context that would substantially increase token cost, especially for Dzongkha (Petrov et al., 2023). We report the main benchmark results in this zero-shot setting and place 3-shot and 5-shot analyses in the appendix.

For the main English–Dzongkha comparison, we keep the system prompt in English for both languages, following prior work on multilingual prompting with CHATGPT (Lai et al., 2023). This choice keeps instruction language fixed while varying only the question language. To address the potential prompt-language confound directly, we additionally run a prompt-language ablation in which the system prompt is switched from English to Dzongkha for Dzongkha questions; we report those results separately in the Results section.

Table 2: Average ZENQA accuracy by model and language. Scores are averaged over the 15 subject partitions. English and Dzongkha averages are accuracies; the value in parentheses indicates the Dzongkha change relative to English.

Type	Model	English Avg.	Dzongkha Avg. (Δ)
Proprietary	GPT-4o	**91.20**	**76.57** (−14.63)
	Claude 3.5 S	89.04	73.55 (−15.49)
Open-weight	Llama 3.1 70B	72.63	35.26 (−37.37)
	Mistral L. 2	78.67	39.13 (−39.54)
	Gemma 2 27B	64.45	26.94 (−37.51)

3.2 Evaluated Models

We evaluate a mix of proprietary and open-weight language models. The full roster of evaluated models is summarized in Table 2. We retain earlier baselines for continuity with prior experiments and include stronger models in the revised evaluation to provide a more current picture of performance on ZENQA.

Because multilingual support for Dzongkha varies substantially across model families, we do not assume uniform low-resource language readiness. In the revised main comparison, we focus on models for which both English and Dzongkha evaluation was feasible and informative. For targeted ablations and prompting analyses, we use GPT-3.5 as the main analysis model in order to balance capability and inference cost, especially for Dzongkha, where tokenization inefficiency increases prompt length and cost (Ahia et al., 2023; Petrov et al., 2023). We therefore interpret the detailed prompting analyses as a focused case study on a practically usable multilingual model, rather than as a claim that all prompting effects generalize uniformly across model families.

3.3 Prompting Strategies

We study three prompting settings. First, we use a standard zero-shot prompting setup in which the model receives a multiple-choice question and its answer options and is asked to produce a single final answer. Second, we evaluate chain-of-thought (CoT) prompting, motivated by prior work showing that step-by-step reasoning can improve performance on reasoning-intensive tasks (Wei et al., 2023; Shi et al., 2022). Third, we test translation-augmented prompting, in which an English translation is appended to a Dzongkha input while the model is still required to answer in Dzongkha.

For Dzongkha CoT experiments, the model is explicitly instructed to reason and answer in Dzongkha. We do this because our target setting is educational support for Dzongkha-speaking users, and reasoning traces that default to English are less useful in practice. We therefore avoid a pipeline in which reasoning is first generated in English and then translated, since translation artifacts could distort the original explanation. Full prompts for all settings are provided in the appendix.

3.4 Evaluation

Our primary evaluation metric is accuracy (pass rate), computed from the final selected answer option. A prediction is counted as correct if the model's final answer matches the gold option label. To make this evaluation robust, we normalize model outputs to the answer space $\{a, b, c, d\}$ and manually inspect non-canonical responses when needed. This step is necessary because models do not always return a single letter: some output the full option text, include additional reasoning after the answer, or produce multiple candidate letters. Manual adjudication is used only to recover the intended final option in such cases; it does not change the underlying correctness criterion.

The correctness of intermediate reasoning steps is not part of the primary benchmark metric. In other words, a response is counted as correct if its final answer is correct, regardless of whether the intermediate CoT trace is fully valid. We analyze intermediate reasoning quality separately in a targeted audit, reported outside the main evaluation metric.

4 Experimental Results

4.1 Model Performance on ZEnQA

4.1.1 All-subject Performance

We first evaluate model performance across all subjects in ZEnQA. A summary of the updated model comparison is shown in Table 2, and the full subject-wise results are reported in Appendix F, Table 7. Three patterns are clear. First, proprietary models consistently outperform open-weight models on the English portion of the benchmark. Second, performance drops substantially when evaluation shifts from English to Dzongkha, showing that strong English performance does not transfer reliably to this ultra-low-resource setting. Third, the English–Dzongkha gap persists even for the strongest models, although stronger systems achieve higher absolute accuracy in both languages.

Performance is broadly stable across many subjects, but some subject partitions remain more difficult than others. In particular, selected biology and higher-grade mathematics subsets show larger drops, suggesting that benchmark difficulty is shaped not only by language, but also by question composition. This pattern is consistent with the category distribution of ZEnQA: mathematics contains more procedural and application-oriented questions, while biology contains a larger share of factual questions. The complete benchmark table is provided in Appendix F.

4.1.2 Performance in English and Dzongkha

We next compare English and Dzongkha performance directly using the multilingual models summarized in Table 2; the full subject-wise benchmark is reported in Appendix F, Table 7. The results show a substantial and persistent accuracy gap between the two languages across subjects. While stronger models reduce this gap relative to weaker ones, they do not eliminate it.

This comparison is particularly informative because ZEnQA controls for task content: the same underlying question, answer options, and answer key are evaluated across

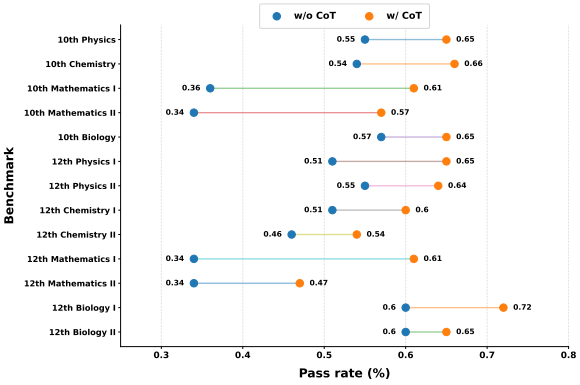

Figure 2: Subject-by-subject CoT performance in English.

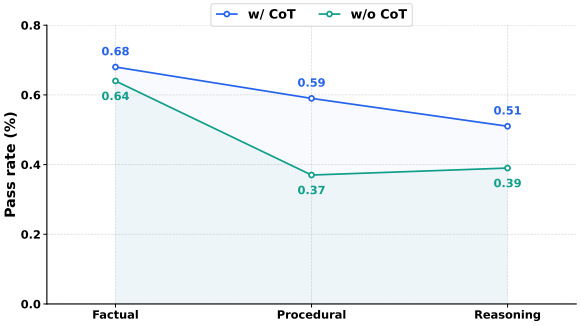

Figure 3: Performance summary in English by question type.

languages. The observed difference therefore reflects model readiness for Dzongkha rather than variation in benchmark difficulty across datasets. Taken together, the results suggest that tokenization, terminology access, and language-specific comprehension remain important bottlenecks in this setting.

4.2 Effect of Chain-of-Thought Prompting

We next examine whether chain-of-thought (CoT) prompting improves performance on ZENQA. Because CoT is expected to help most on questions that require intermediate reasoning, we analyze its effect across subjects and question categories.

4.2.1 Subject- and Category-level Effects

Figure 2 shows that CoT improves English performance for GPT-3.5 across subjects, but the gains are uneven: mathematics benefits the most, whereas biology shows much smaller improvement. This subject-level pattern is consistent with the category composition of the benchmark. Mathematics contains more procedural and application-based items, which benefit more from intermediate reasoning, while biology contains more factual questions, for which CoT offers limited advantage.

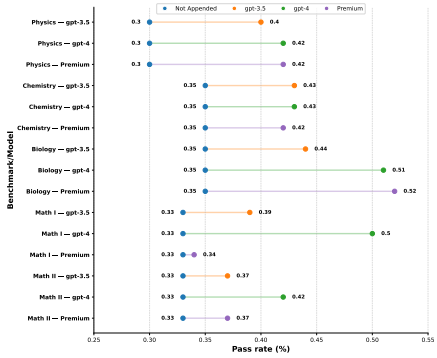

Figure 4: Performance of translation-augmented prompting with LLM-generated English translations. "Not Appended" denotes the setting without any English translation.

The category-level analysis in Figure 3 confirms this trend. CoT produces the largest gains on procedural and reasoning-heavy questions, but only limited improvement on factual questions. Even after these gains, procedural and reasoning questions remain harder than factual ones. This pattern suggests that reasoning support helps, but does not eliminate the underlying language and knowledge bottlenecks present in the benchmark.

4.2.2 Few-shot Results

Few-shot CoT prompting yields only modest gains over zero-shot prompting in both English and Dzongkha, while inference cost increases substantially, especially for Dzongkha because of inefficient tokenization. We therefore use zero-shot prompting as the primary evaluation setting in the main paper. Additional k-shot CoT analyses, Dzongkha-specific subject/category breakdowns, and the full few-shot benchmark table are reported in Appendix A.3 and Appendix F.

4.3 Translation-Augmented Prompting

A central question in this work is whether appending English translations can improve performance on Dzongkha inputs. Figure 4 shows that translation augmentation is consistently beneficial: appending an English trans-

Table 3: Prompt-language ablation for Dzongkha questions using GPT-3.5.

Prompt language	Avg. perf.
English	24.37
Dzongkha	25.12

Table 4: Average gains from translation augmentation on Dzongkha BIG-Bench Hard.

Translation source	Avg. gain
Premium English	↑6.52
GPT-4-generated English	↑6.05

lation improves performance across subjects, and LLM-generated English translations perform comparably to, and in some cases slightly better than, the original human translation.

In this experiment, we used translations produced by GPT-3.5 and GPT-4. We did not use Google Translate because it often corrupted mathematical expressions and equations. These results suggest that translation-augmented prompting is practical even when human-generated English translations are unavailable, and that the gains are not limited to a single translation source.

4.4 Prompt-language Ablation

To test whether the English–Dzongkha gap is partly caused by an instruction-language mismatch, we compare two settings for Dzongkha questions: an English system prompt and a Dzongkha system prompt. Table 3 shows that changing the prompt language produces only a negligible difference. This suggests that the dominant source of the gap is the language of the question itself rather than the language of the instruction.

This ablation is important because it rules out a simple confound in the main experimental setup. Holding the question fixed in Dzongkha while changing only the instruction language leaves performance almost unchanged, reinforcing the conclusion that question language, not prompt language, drives most of the observed cross-lingual gap.

4.5 Transfer to Dzongkha BIG-Bench Hard

We next test whether the translation-augmentation strategy transfers beyond ZENQA. For this purpose, we evaluate the same prompting idea on a Dzongkha version of BIG-Bench Hard constructed from 14 tasks selected for their relevance and translatability.

Table 4 shows that the gains transfer to this second benchmark: adding premium English translations yields an average improvement of 6.52 points, while GPT-4-generated translations yield an average improvement of 6.05 points. At the task level, translation augmentation is beneficial in seven of the fourteen tasks, marginally harmful in three, and essentially unchanged in four. These findings show that the translation-augmentation strategy is not specific to ZENQA, but also transfers to a broader reasoning benchmark in Dzongkha.

4.6 Error Analysis

To better understand model failures, we examine the relationship between intermediate CoT correctness and final-answer correctness through a targeted human audit. The dominant pattern is that incorrect CoT usually leads to an incorrect final answer, while correct CoT is strongly associated with a correct final answer. Answer extraction or option-selection mismatches do occur, but they are relatively rare compared with genuine reasoning failures. These findings suggest that the primary failure mode is not formatting noise, but rather a breakdown in intermediate reasoning. Limited access to Dzongkha scientific terminology may be one contributing factor. The full audit counts are reported in Appendix C, Table 5.

5 Related Work

A large body of work evaluates the reasoning ability of large language models using English benchmarks such as CommonsenseQA (Talmor et al., 2019), HellaSwag (Zellers et al., 2019), CosmosQA (Huang et al., 2019), ARC (Clark et al., 2018), MMLU (Hendrycks et al., 2020), MATH (Hendrycks et al., 2021), and BIG-Bench Hard (Suzgun et al., 2022). Multilingual evaluation efforts broaden this landscape through resources such as X-COPA (Ponti et al., 2020), Indic benchmarks (Doddapaneni et al., 2023), and multilingual evaluations of frontier models (Lai et al., 2023; Bang et al., 2023; Ahuja et al., 2023). However, ultra-low-resource languages remain sub-

stantially underrepresented in these resources. Dzongkha, in particular, is largely absent from mainstream multilingual benchmarks; although MGSM includes Dzongkha (Shi et al., 2022), it is limited to mathematics and does not provide the broader curriculum-grounded coverage needed to evaluate reasoning across multiple school subjects.

ZENQA addresses this gap by providing the first parallel Dzongkha–English benchmark built from Bhutan's national curriculum. Because it includes aligned question pairs, shared answer keys, and multiple STEM subjects across grade levels, it enables controlled cross-lingual evaluation in a setting that is both educationally grounded and methodologically reproducible. More broadly, our prompting analysis connects to prior work on chain-of-thought prompting (Wei et al., 2022, 2023; Shi et al., 2022) and multilingual prompting strategies that use English as an intermediate reasoning language (Ahuja et al., 2023; Huang et al., 2023). Our work studies a different trade-off: whether performance in Dzongkha can be improved by appending English translations while still requiring the model to reason and answer in Dzongkha, thereby preserving accessibility for Dzongkha-speaking users.

6 Conclusion

We introduced ZENQA, the first parallel Dzongkha–English benchmark for educational question answering, constructed from Bhutanese national examinations across grades 8, 10, and 12. Because the benchmark contains aligned question text, answer options, and shared answer keys in both languages, it enables controlled cross-lingual comparison in a locally grounded educational setting.

Our experiments reveal a persistent English–Dzongkha performance gap across models. Stronger models improve absolute performance, but they do not eliminate this gap. We further find that chain-of-thought prompting helps most on procedural and reasoning-heavy questions, while translation-augmented prompting consistently improves performance on Dzongkha inputs. Additional analyses show that prompt language has only a limited effect relative to question language, and that translation augmentation also transfers to Dzongkha BIG-Bench Hard. Taken together, these results suggest that current models remain limited in their readiness for ultra-low-resource educational settings, and that a substantial portion of the gap reflects language-access bottlenecks rather than reasoning ability alone. These findings should be interpreted as evidence about curriculum-grounded, text-only, multiple-choice educational reasoning, rather than as a complete account of multilingual reasoning in Dzongkha more broadly.

Limitations

ZENQA excludes figure-dependent questions and therefore focuses on text-only reasoning, which improves control but limits coverage of visually grounded educational problems. The benchmark is also based on multiple-choice questions, which may permit occasional shortcut behavior or answer-selection artifacts that would be less likely in free-response settings. Although ZENQA enables controlled comparison between Dzongkha and English, it remains focused on a single ultra-low-resource language. Finally, our evaluation primarily measures final-answer correctness; while we include targeted audits of intermediate reasoning, a larger-scale evaluation of reasoning quality remains an important direction for future research.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and 1 others. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, and 1 others. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hal-

lucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

Aleksandar Petrov, Emanuele La Malfa, Philip HS Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *arXiv preprint arXiv:2305.15425*.

Edoardo Maria Ponti, Goran Glava, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. 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, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and 1 others. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

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.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Further Tests

This appendix reports supplementary analyses that support the main findings in the paper. Because GPT-3.5 provides a practical balance between capability and inference cost for Dzongkha, we use it for the additional prompting analyses reported in this section.

A.1 Performance on Adding English Translation

One hypothesis for weaker Dzongkha performance is that models struggle with Dzongkha scientific terminology and with processing a low-resource non-Latin script (Lai et al., 2023). To test whether additional high-resource language context is helpful, we append English translations to Dzongkha questions and evaluate the resulting prompts on a subset of 10th-grade exam questions. We use 105 questions from each 10th-grade subject, since these questions lie between the easier 8th-grade and harder 12th-grade materials and therefore provide a useful middle-range test bed.

A.1.1 Findings

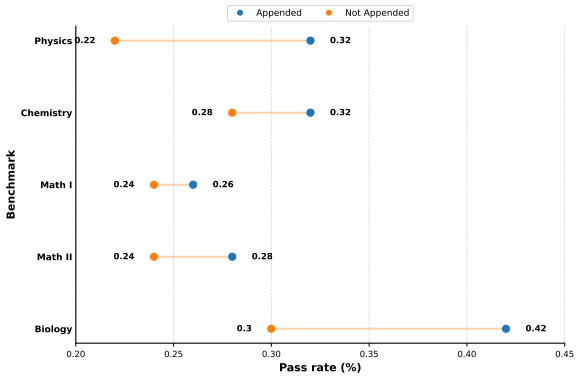

Figure 5: Answering questions in Dzongkha is made easier by including an English translation. Note that "Appended" denotes with and "Not Appended" denotes without including the translation. The model was requested to do CoT in Dzongkha.

Figure 5 shows that appending English translations improves performance across all tested subjects. Because the system prompt remains fixed, this experiment supports the interpretation that the gains come from the additional English question context rather than from changing the instruction language. The largest gains appear in Biology, where scientific terminology is especially prominent, while the improvements in Mathematics are smaller. We also tested LLM-generated translations in place of human translations; the main paper further shows that these translations perform comparably in practice.

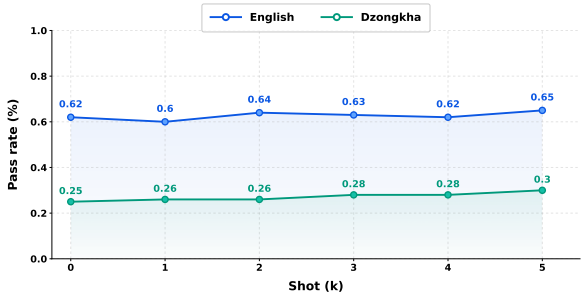

Figure 6: Dzongkha and English k-shot CoT prompting on GPT-3.5.

A.2 Performance of Translation-Augmented Prompting on Different Datasets

To test whether the translation-augmentation strategy generalizes beyond ZENQA, we extend the analysis to Dzongkha BIG-Bench Hard.

A.2.1 Dzongkha BIG-Bench Hard

We use GPT-4 to create a Dzongkha version of BIG-Bench Hard for this experiment. Because only a relatively small fraction of ZENQA consists of reasoning-heavy items, BIG-Bench Hard provides a complementary test of whether the same prompting strategy transfers to a broader reasoning benchmark. We exclude tasks whose core structure does not transfer naturally to Dzongkha, such as tasks that depend heavily on English alphabetical ordering. Prompts for each selected task were iteratively created by two native Dzongkha speakers.

When premium English translations are appended, the average gain is 6.52 points; GPT-4-generated English translations yield an average gain of 6.05 points. Additional task-level details are provided in Appendix D.

A.3 Additional CoT Analyses on ZEnQA

A.3.1 Few-shot CoT in English and Dzongkha

Figure 6 compares zero-shot and few-shot CoT prompting for GPT-3.5 in English and Dzongkha. The main pattern is that few-shot prompting yields only limited gains over zero-shot prompting in both languages. Because these improvements are modest while inference cost increases substantially, we use zero-

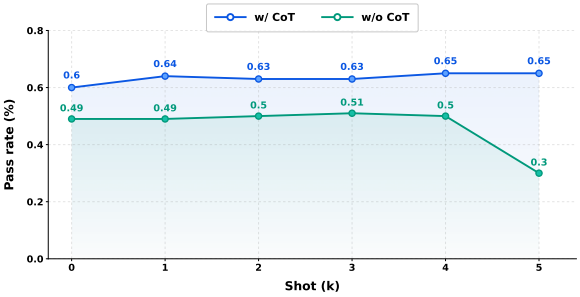

Figure 7: Impact of CoT reasoning on GPT-3.5 for English ZENQA across k-shot settings.

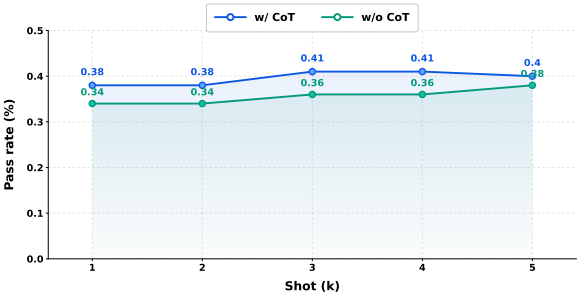

Figure 8: Impact of CoT reasoning on GPT-3.5 for Dzongkha across k-shot settings.

shot prompting as the primary evaluation setting throughout the paper. This choice is especially important for Dzongkha, where few-shot prompting becomes considerably more expensive because of inefficient tokenization.

Preparation of few-shot examples. When constructing few-shot examples, we ensured that the first three examples covered the three question categories defined in Section 2.3: factual knowledge, procedural & application, and reasoning. The full five-example set also included different question formats commonly found in examinations, including both multiple-choice and multi-selection items. In addition, examples were varied across subjects to better reflect the diversity of the benchmark.

A.3.2 Impact of CoT Across k-shot Settings

Figures 7 and 8 show the effect of CoT prompting across few-shot settings for English and Dzongkha, respectively. In English, CoT consistently outperforms the non-CoT baseline across all shot values. A similar qualitative trend holds for Dzongkha, where CoT also improves performance, although the absolute accuracy remains much lower. Together, these figures show that CoT provides a stable benefit across languages, but that its gains are not large enough to remove the underlying English–Dzongkha gap.

A.3.3 Dzongkha-specific Breakdown of CoT

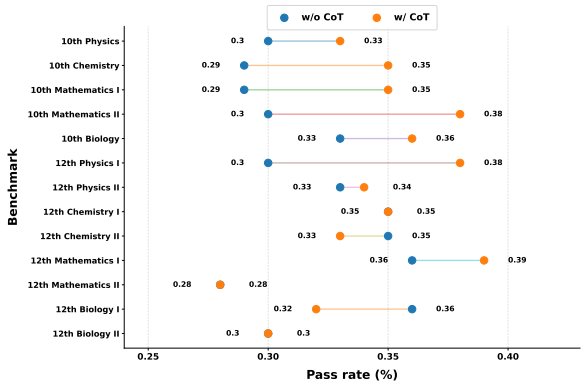

Figure 9: CoT performance summary in Dzongkha by subject.

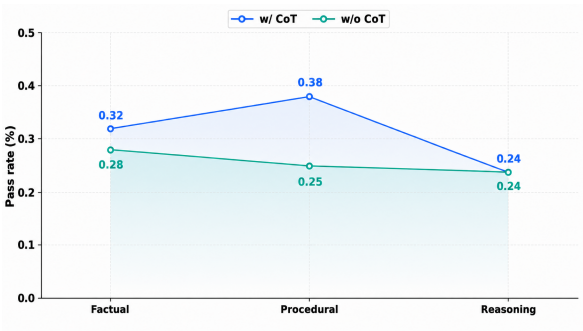

Figure 10: CoT performance summary in Dzongkha by question type.

Figure 9 provides the Dzongkha subject-wise breakdown of CoT performance for GPT-3.5. As in English, CoT improves performance in most subjects, but the gains are uneven. Figure 10 further breaks down Dzongkha performance by question type. The largest gains appear in procedural and application-oriented questions, while improvements for factual and reasoning questions are smaller.

Compared with English, Dzongkha shows a similar directional effect of CoT but starts from a much lower baseline. One possible explanation is that some reasoning-heavy subject partitions, such as physics, contain relatively smaller subsets of questions, making the resulting gains less stable. More broadly, the lower base accuracy suggests that Dzongkha perfor-

mance remains constrained not only by reasoning difficulty, but also by language-access bottlenecks.

B Grammar-correction Analysis on ZEnQA

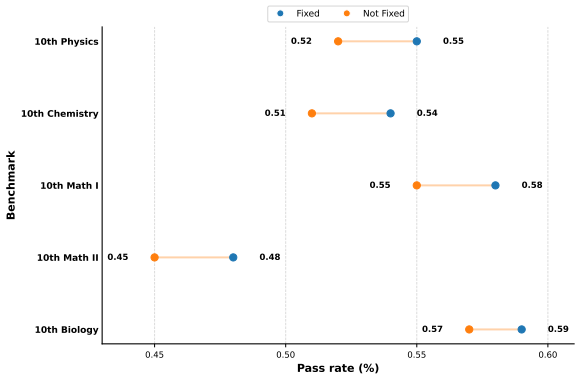

Figure 11: Grammatical errors' effects on GPT 3.5 performance. Note that "Fixed" denotes with and "Not Fixed" denotes without including grammar-fixed versions.

To assess whether minor grammatical errors or awkward phrasing in the English exam texts materially affect model performance, we select 115 questions from each 10th-grade subject and create grammar-corrected English variants using GPT-4. After automatic correction, a Dzongkha speaker fluent in English reviews the outputs and makes any necessary revisions.[3]

Figure 11 shows that the difference between the grammar-corrected and original versions is small. This suggests that minor grammatical issues in the English text are not a major driver of the benchmark results.

C Additional Notes on the Intermediate CoT Audit

The main paper reports a targeted human audit of intermediate chain-of-thought correctness and final-answer correctness. Table 5 reports the full audit counts. The key finding is that incorrect intermediate reasoning usually leads to an incorrect final answer, while correct intermediate reasoning is strongly associated with a correct final answer. We report the full audit table here for completeness.

[3]We did not apply this correction process to the full dataset because GPT-4 occasionally alters the original meaning of a question while fixing grammar.

Table 5: Manual audit of the relationship between intermediate chain-of-thought (CoT) correctness and final-answer correctness. "w/ translation" denotes prompts with appended English translation, and "w/o translation" denotes prompts without appended English translation.

Setting	Final ✓	Final ×
w/ translation, CoT ✓	25	3
w/ translation, CoT ×	2	28
w/o translation, CoT ✓	18	0
w/o translation, CoT ×	3	32

D Further Details on BIG-Bench Hard

D.1 Prompt Variants

We compare four prompting variants that differ only in how English is incorporated alongside Dzongkha. In the first setting, the model receives the question only in Dzongkha. In the remaining settings, we keep the Dzongkha prompt intact and append an English rendering of the same content: (i) Dzongkha + English (GPT-3.5), (ii) Dzongkha + English (GPT-4), and (iii) Dzongkha + English (Premium), which uses the premium English version from the source dataset.

D.2 Task Selection

From the 23 tasks in BIG-Bench Hard, we select 14 that can be transferred meaningfully to Dzongkha[4]. The selected tasks are: Reasoning About Colored Objects, Web of Lies, Multi-step Arithmetic, Navigate, Object Counting, Penguin in a Table, Causal Judgement, Date Understanding, Disambiguation QA, Formal Fallacies, and Logical Deductions with three, five, and seven choices.

We translate these tasks into Dzongkha with GPT-4, using three human-annotated examples as prompts for each task. Two Dzongkha speakers iteratively refine the prompts to match each task format.

D.3 Findings

The full task-level results are reported in Table 9. Translation augmentation improves performance in seven of the fourteen tasks, slightly harms performance in three, and

[4]Some BIG-Bench Hard tasks do not admit a natural Dzongkha equivalent, especially those that rely directly on English alphabet manipulation.

Prompt for Grammar-corrected ZENQA Questions

Instruction:
You are provided with a multiple-choice exam question written in English, including its answer choices. The question is generally well-constructed but may occasionally contain slight grammatical errors or awkward phrasing. Your task is to revise and polish the question to make it sound more natural and fluent.

Please follow this format:

- Revise the full question text only.
- Do not include the answer choices in your output.
- If there is an equation or formula, leave it unchanged.

 # Question:
`<question>`

 # Revised Question:
`<LLM Response>`

Figure 12: GPT-4 prompt for creating grammar-corrected questions.

Prompt for ZENQA Question Categorization

Instruction:
You are given a multiple-choice question, and your task is to determine what type of reasoning is necessary to solve it.

Available Categories:

- **Factual Knowledge**: The question relies on recalling basic facts, dates, events, or concepts.
- **Procedural & Application**: The question requires applying a procedure, formula, or set of steps to reach the answer.
- **Reasoning**: The question demands multi-step reasoning or logical deduction to find the solution.

Please follow this format:

- Output exactly one category.

 # Question:
`<question>`

 # Category:
`<LLM Response>`

Figure 13: ZEnQA dataset categorization prompt.

leaves performance essentially unchanged in four.

E Experimental Prompts

We use the prompt in Figure 12 to create grammar-corrected English question variants with GPT-4.

As described in Section 2.3, we use GPT-4 to categorize questions into *Factual Knowledge*, *Procedural & Application*, and *Reasoning*. Figure 13 shows the prompt used for this zero-shot categorization procedure. Figure 14 shows the prompt used for proprietary-model zero-shot benchmarking on ZENQA. Figure 15 shows the few-shot prompt template used for proprietary-model benchmarking.

In the translation-augmentation experiment, as described in Appendix D, we provide an English translation for additional context while still requiring the model to reason and answer in Dzongkha. Figure 16 shows the corresponding prompt.

E.1 Prepared ZEnQA Questions with Correct Grammar

We use the prompt in Figure 12 to create grammar-corrected English question variants with GPT-4.

E.2 Dataset Categorization for ZEnQA

As described in Section 2.3, we use GPT-4 to categorize questions into *Factual Knowledge*, *Procedural & Application*, and *Reasoning*. Figure 13 shows the prompt used for this zero-shot categorization procedure.

We apply this categorization prompt independently to each question in the dataset. The resulting subject-wise category counts are reported in Table 6.

E.3 Prompting CoT for ZEnQA Zero-shot Benchmark

Figure 14 shows the prompt used for proprietary-model zero-shot benchmarking on ZENQA.

Prompting CoT for ZEnQA Few-shot Benchmark

Figure 15 shows the few-shot prompt template used for proprietary-model benchmarking.

E.4 Translation-Appended Benchmark Prompt

In the translation-augmentation experiment, as described in Appendix D, we provide an English translation for additional context while still requiring the model to reason and answer

Prompt for ZEnQA Zero-shot Benchmarking

Instruction:
You are provided with a multiple-choice question and its options in English or Dzongkha. Your task is to correctly answer the question.

Please follow this format:

- Think and reason step by step in either English or Dzongkha.
- After reasoning, provide the final answer using only one letter: `a`, `b`, `c`, or `d`.
- Do not include the full option text in the final answer.
- There is always exactly one correct answer.

 # Question:
`<question>`

 # Response:
`<LLM Response>`

Figure 14: Proprietary models prompt for ZEnQA zero-shot benchmarking.

Prompt for ZEnQA Few-shot Benchmarking

Instruction:
You are provided with a multiple-choice question in either English or Dzongkha. Your task is to solve the question step by step and then give the correct answer.

Please follow this format:

```
{Question 1}
{Reasoning and Explanation}
{Answer 1}
{Question 2}
{Reasoning and Explanation}
{Answer 2}
————————————
{Question}
```

Now, follow the same process for the given question.

 # Response:
`<LLM Response>`

Figure 15: Proprietary models prompt for ZEnQA few-shot benchmarking.

in Dzongkha. Figure 16 shows the corresponding prompt.

F Full Benchmark Tables

This section reports the full benchmark tables used in the paper.

F.1 ZEnQA Zero-shot Benchmark

The full zero-shot benchmark results on ZEnQA are reported in Table 7.

Prompt for Translation-Augmented Benchmarking

Instruction:
You are given a situation along with its possible reason or answer in Dzongkha. For your convenience, an English translation is also provided. Your task is to identify the correct reason or answer for the given situation.

Please follow this format:

- Use the English translation only as additional context.
- Explain your reasoning step by step in Dzongkha.
- Provide the final answer in Dzongkha.

 # Question:
`<question>`

 # Response:
`<LLM Response>`

Figure 16: Zero-shot translation-appended prompt.

F.2 ZEnQA Few-shot Benchmark with and without CoT

The full few-shot benchmark results on ZEnQA, with and without chain-of-thought prompting, are reported in Table 8.

F.3 BIG-Bench Hard Zero-shot Benchmark

The task-level zero-shot benchmark results on Dzongkha BIG-Bench Hard are reported in Table 9.

Table 6: Question-count distribution in ZEnQA by subject and reasoning category.

Subject	Category	Questions	Instances (%)
	12th Grade Subjects		
12th Bio I	Factual Knowledge	288	5.58%
	Procedural & Application	12	0.23%
	Reasoning	15	0.29%
12th Bio II	Factual Knowledge	297	5.75%
	Procedural & Application	8	0.16%
	Reasoning	28	0.54%
12th Chem I	Factual Knowledge	229	4.44%
	Procedural & Application	85	1.65%
	Reasoning	58	1.12%
12th Chem II	Factual Knowledge	185	3.59%
	Procedural & Application	145	2.81%
	Reasoning	64	1.24%
12th Phy I	Factual Knowledge	115	2.23%
	Procedural & Application	161	3.12%
	Reasoning	32	0.62%
12th Phy II	Factual Knowledge	168	3.25%
	Procedural & Application	143	2.77%
	Reasoning	27	0.52%
12th Math I	Factual Knowledge	13	0.25%
	Procedural & Application	368	7.13%
	Reasoning	20	0.39%
12th Math II	Factual Knowledge	24	0.46%
	Procedural & Application	327	6.34%
	Reasoning	45	0.87%
	10th Grade Subjects		
10th Bio	Factual Knowledge	308	5.97%
	Procedural & Application	21	0.41%
	Reasoning	27	0.52%
10th Phy	Factual Knowledge	178	3.45%
	Procedural & Application	119	2.31%
	Reasoning	27	0.52%
10th Math I	Factual Knowledge	45	0.87%
	Procedural & Application	267	5.17%
	Reasoning	73	1.41%
10th Math II	Factual Knowledge	24	0.46%
	Procedural & Application	316	6.12%
	Reasoning	58	1.12%
10th Chem	Factual Knowledge	268	5.19%
	Procedural & Application	91	1.76%
	Reasoning	35	0.68%
	8th Grade Subjects		
8th Math	Factual Knowledge	30	0.58%
	Procedural & Application	139	2.69%
	Reasoning	45	0.87%
8th Sci	Factual Knowledge	194	3.76%
	Procedural & Application	26	0.50%
	Reasoning	13	0.25%
Total	**All**	**5161**	**100.00%**

Table 7: ZENQA zero-shot benchmark results across all models in English and Dzongkha. Best results are highlighted in Orange **bold**, and second-best results in Blue underline.

Language	Subject	GPT-4o	Claude 3.5 S.	Llama 3.1 70B	Mistral L. 2	Gemma 2 27B	GPT 3.5	Claude 2.1	LLaMA2 13B	LLaMA2 7B	Mistral 7B
					English						
English	12th Bio I	**92.1**	89.7	76.4	81.2	68.3	69.33	59.52	38.24	30.61	31.59
English	12th Bio II	**90.8**	88.5	73.9	79.6	65.1	64.11	55.44	33.82	24.79	38.27
English	12th Chem I	**91.3**	89.2	72.1	78.4	63.7	57.51	52.18	24.29	19.47	27.23
English	12th Chem II	**89.7**	87.9	70.8	77.1	62.4	56.24	51.89	23.75	14.63	22.51
English	12th Phy I	**90.5**	88.3	74.6	80.3	66.9	62.33	48.41	29.28	17.32	25.39
English	12th Phy II	**89.2**	87.1	71.5	76.8	61.2	59.63	37.95	23.52	25.61	27.96
English	12th Math I	**93.4**	91.8	68.2	75.6	59.4	59.27	58.12	10.37	6.35	11.38
English	12th Math II	**91.8**	89.5	65.7	73.2	57.8	53.45	56.73	17.05	6.95	15.28
English	10th Bio	**91.2**	88.9	75.3	80.7	67.2	65.03	51.91	35.02	25.17	36.68
English	10th Phy	**90.6**	88.1	73.4	78.9	64.5	62.63	53.19	29.07	19.53	26.92
English	10th Math I	**92.7**	90.4	71.8	79.3	65.6	61.31	48.11	13.35	12.33	12.84
English	10th Math II	**89.7**	87.6	69.5	76.4	62.1	63.17	49.68	9.41	10.67	15.41
English	10th Chem	**92.4**	90.1	74.9	81.5	68.7	62.59	56.97	28.83	22.24	28.58
English	8th Math	**93.1**	91.3	78.6	83.2	70.4	70.24	56.23	26.34	21.19	21.67
English	8th Sci	**89.5**	87.2	72.8	77.9	63.5	63.27	54.24	42.23	23.78	37.08
					Dzongkha						
Dzongkha	12th Bio I	**79.6**	76.3	38.7	42.5	29.8	28.57	32.14	–	–	–
Dzongkha	12th Bio II	**78.1**	74.9	36.4	40.1	27.6	25.34	28.93	–	–	–
Dzongkha	12th Chem I	**77.9**	74.2	35.8	39.7	26.9	24.16	27.53	–	–	–
Dzongkha	12th Chem II	**76.3**	72.8	34.2	38.3	25.4	22.63	25.82	–	–	–
Dzongkha	12th Phy I	**78.4**	75.1	37.9	41.6	29.1	26.43	29.75	–	–	–
Dzongkha	12th Phy II	**76.8**	73.5	35.6	39.2	27.3	23.85	26.27	–	–	–
Dzongkha	12th Math I	**74.2**	70.9	32.4	36.8	24.7	21.53	24.86	–	–	–
Dzongkha	12th Math II	**72.5**	69.3	30.1	34.5	22.9	18.94	22.15	–	–	–
Dzongkha	10th Bio	**78.4**	75.6	36.9	40.8	28.5	25.67	29.27	–	–	–
Dzongkha	10th Phy	**76.9**	74.1	35.2	38.9	26.8	23.17	26.85	–	–	–
Dzongkha	10th Math I	**75.3**	72.7	33.7	37.4	25.9	20.83	23.57	–	–	–
Dzongkha	10th Math II	**72.3**	69.8	31.5	35.2	24.1	19.47	21.96	–	–	–
Dzongkha	10th Chem	**76.9**	74.5	36.2	40.3	28.2	24.37	27.64	–	–	–
Dzongkha	8th Math	**79.8**	77.2	39.5	43.1	30.6	27.28	30.17	–	–	–
Dzongkha	8th Sci	**75.1**	72.4	34.8	38.6	26.3	22.75	25.48	–	–	–

Table 8: Grade 10 ZENQA few-shot results with and without chain-of-thought prompting. "w/" denotes with CoT and "w/o" denotes without CoT. Best results are highlighted in Orange **bold**, and second-best results in Blue underline.

Language	Dataset	CoT	GPT 3.5		LLaMA2 (7b)		LLaMA2 (13b)	
			5-shot	3-shot	5-shot	3-shot	5-shot	3-shot
				English				
English	10th Bio	w/o	58.39	**59.27**	48.85	43.63	43.63	39.28
		w/	67.95	**69.63**	49.72	49.72	48.85	43.63
English	10th Phy	w/o	56.67	**61.02**	37.52	39.28	29.72	27.98
		w/	66.24	**70.58**	40.15	38.41	38.41	43.63
English	10th Math I	w/o	41.89	**43.63**	30.58	25.37	34.06	28.85
		w/	66.24	**69.63**	28.85	26.24	37.52	33.19
English	10th Math II	w/o	**46.24**	37.52	21.89	21.89	40.15	42.76
		w/	**62.76**	58.39	29.72	27.98	25.37	30.58
English	10th Chem	w/o	**58.39**	**58.39**	36.67	37.52	27.11	25.37
		w/	**68.85**	63.63	44.50	42.76	43.63	42.76
				Dzongkha				
Dzongkha	10th Bio	w/o	**28.14**	27.42	–	–	–	–
		w/	**31.12**	28.89	–	–	–	–
Dzongkha	10th Phy	w/o	**35.52**	30.32	–	–	–	–
		w/	**39.21**	33.28	–	–	–	–
Dzongkha	10th Math I	w/o	22.94	**27.42**	–	–	–	–
		w/	29.65	**35.52**	–	–	–	–
Dzongkha	10th Math II	w/o	28.14	**28.85**	–	–	–	–
		w/	34.78	**38.47**	–	–	–	–
Dzongkha	10th Chem	w/o	**35.52**	**35.52**	–	–	–	–
		w/	34.78	**40.68**	–	–	–	–

Table 9: Task-level Dzongkha BIG-Bench Hard zero-shot results under English-only, Dzongkha-only, and translation-augmented prompting. Best results are highlighted in orange **bold**, and second-best results in blue underline.

Dataset	English Only	Dzongkha Only	Translation Append Premium	Translation Append GPT
Causal Judgement	**57.2**	43.2	53.0	51.4
Date Understanding	**56.0**	20.2	36.8	28.1
Disambiguation QA	45.2	32.0	**47.4**	45.9
Formal Fallacies	**55.2**	42.9	42.8	41.4
Logical Deductions Five	**35.2**	17.3	15.1	16.6
Logical Deductions Seven	**30.8**	13.1	18.2	13.4
Logical Deductions Three	**51.6**	28.2	30.7	31.2
Multistep Arithmetic	**68.5**	39.2	34.8	–
Navigate	**60.4**	44.5	49.7	50.8
Object Counting	**83.6**	27.8	34.9	37.4
Penguin In A Table	**61.0**	20.8	29.1	29.1
Reasoning About Colored Objects	**42.4**	19.2	25.9	23.0
Temporal Sequences	**37.6**	18.9	18.2	17.0
Web of Lies	**52.8**	17.0	39.0	38.4

Physics Sample Questions from ZEnQA

Physics

Factual Knowledge
Which one is the fundamental unit?
(a) Joule
(b) Newton
(c) Candela
(d) Pascal

གཞི་རྩའི་ཀྱི་རྩ་བའི་ཚད་གཞི་གང་ཡིན་རྣམ།
(a) ཇུལ།
(b) ནིའུ་ཏཱན།
(c) ཀེན་ཌེ་ལ།
(d) པ་སཀཱལ།

Procedural & Application
The atmospheric pressure of a certain place is 93296 Pa. The density of kerosene is 800 kg/m^3, and the density of benzene is 980 kg/m^3. Which one of the following is correct?
(a) The height of the mercury column is 76 cm.
(b) The height of the water column is 9.52 m.
(c) The height of the kerosene column is 9.71 m.
(d) The height of the benzene column is 11.9 m.

གནས་ཀྱི་གཤོན་ཀློང་གི་ཚད་རྒྱས་པ་དེ་ 93296 Pa ཡིན། ཀེ་རོ་སིན་གྱི་ཚད་རྒྱས་པ་དེ་ 800 kg/m^3 དང། བེན་ཛིན་གྱི་ཚད་རྒྱས་པ་དེ་ 980 kg/m^3 ཡིན། གཤམས་གྱི་དཔེ་རབས་རྣོར་གང་རུང་གི་ནང་ཞིག་གི་ཚོས་གཞི་ཡིན་རྣམ།
(a) དངུལ་རྩི་གཤུ་རིམ་གྱི་མཐོ་ཚད་དེ་ 76 cm ཡིན།
(b) རྩི་གཤུ་རིམ་གྱི་མཐོ་ཚད་དེ་ 9.52 m ཡིན།
(c) ཀེ་རོ་སིན་གྱི་གཤུ་རིམ་གྱི་མཐོ་ཚད་དེ་ 9.71 m ཡིན།
(d) བེན་ཛིན་གྱི་གཤུ་རིམ་གྱི་མཐོ་ཚད་དེ་ 11.9 m ཡིན།

Reasoning
As someone goes up from the sea level:
i. The weight of the atmosphere increases.
ii. The density of air decreases.
iii. The pressure of air decreases.
Which one is correct?
(a) i and ii
(b) i and iii
(c) ii and iii
(d) i, ii, and iii

མཚོ་རོགས་ནས་མཐོ་གླང་དུ་སོང་བའི་གླགས་ལུ:
i. གནས་ཀྱི་གཤོན་ཀློང་གི་ཚད་རྒྱས་པ་ཆེར་འགྱོ་བ།
ii. གནས་ཀྱི་ཚད་རྒྱས་པ་རྩུར་དུར་འགྱོ་བ།
iii. གནས་ཀྱི་གཤོན་ཀློང་གི་ཚད་རྒྱས་པ་རྩུར་དུར་འགྱོ་བ།

གང་ཞིག་ཚོས་གཞི་ཡིན་རྣམ།
(a) i དང་ ii
(b) i དང་ iii
(c) ii དང་ iii
(d) i, ii, དང་ iii

Figure 17: Physics sample questions for the 10th grade in ZEnQA by category.

Chemistry Sample Questions from ZEnQA

Chemistry

Factual Knowledge

In which of the following compounds is the latent valency of iron zero?

(a) $FeSO_4$
(b) $Fe(NO_3)_2$
(c) $Fe_2(SO_4)_3$
(d) $FeCO_3$

གང་ཞིག་གི་ནང་དུ་ལྷན་གས་ཀྱི་གཏན་ཚོགས་ཀྱི་འཕེལ་བ་ཚད་རྒྱས་པ་གླུ་དགོངས་ཡིན་རྟག།
(a) $FeSO_4$
(b) $Fe(NO_3)_2$
(c) $Fe_2(SO_4)_3$
(d) $FeCO_3$

Procedural & Application

A is a hydrocarbon with the formula C_nH_{2n+1}-CH=CH_2. What is the name of 'A' when $n = 2$?

(a) 1-Butene
(b) 2-Butene
(c) 3-Butene
(d) Butyne

A རི་ཅ་ཀི་ཏྲོ་ཀར་བྱས་ཕྱི་ཐ་བའི་ཚད་གཞི་ C_nH_{2n+1}-CH=CH_2 ཡིན། $n = 2$ ཡིན་པའི་སྐབས་སུ་ 'A' ཡི་མིང་གང་ཡིན་རྟག།
(a) 1-ཐྱུ་ཏེན།
(b) 2-ཐྱུ་ཏེན།
(c) 3-ཐྱུ་ཏེན།
(d) ཐྱུ་ཏེན།

Reasoning

The proton numbers of P, Q, R, and X are 5, 9, 11, and 12, respectively. [Here, P, Q, R, and X are not actual symbols of elements.] Which element has the largest atomic size?

(a) P
(b) Q
(c) R
(d) X

P, Q, R, དང་ X ཡི་རྡོ་རྗེ་ཀྱི་ཨང་གི་རིམ་པ་ 5, 9, 11, དང་ 12 ཡིན། [འདིར་ P, Q, R, X རྣམ་པར་ལྷན་ཀྱི་མཚོན་རྟགས་མིན།] གང་ཞིག་གི་ཚེ་རྡོ་རྗེ་ཀྱི་ཚད་རྒྱས་པ་ཆེ་མོས་ཡིན་རྟག།
(a) P
(b) Q
(c) R
(d) X

Figure 18: Chemistry sample questions for the 10th grade in ZEnQA by category.

Biology Sample Questions from ZEnQA

Biology

Factual Knowledge
Which blood group is called the universal blood donor?
(a) A
(b) B
(c) AB
(d) O

གང་ཞིག་གི་ཁྲག་གི་རིགས་ཀྱི་མིང་ལ་སྤྱི་རབས་ཏུར་གྱི་ཁྲག་སྦྱོར་པ་ཞེས་པ་ཡིན་རམ།
(a) A
(b) B
(c) AB
(d) O

Procedural & Application
In the formation of female gametophyte, four haploid cells are produced from a reproductive mother cell. The upper three cells degenerate, and the lower one divides. Which process is responsible for the formation of these four cells?
(a) Mitosis
(b) Meiosis
(c) Amitosis
(d) Binary fission

མོ་རྒྱལ་ཕྱི་ཕྱི་ཕྲེང་གི་ཐུག་པར་སྐྱབས་ལ་སྐྱེ་བ་ཕྱི་ཕྲེང་བ་ནས་བཟུང་སྟེ་བཞི་སྟེ་རྒྱུ་དཔེ་ཁྲག་ཕྲེང་བ་བཟོ་བ་ཡིན། སྟེང་གི་ཕྲེང་བ་གསུམ་དེ་ཉམས་པ་དང་། ཐོག་གི་ཕྲེང་བ་གཅིག་དེ་ཕྱེ་བ་ཡིན། གང་ཞིག་གི་ལས་སྒྲུ་པ་དེ་བཞི་སྟེ་རྒྱུ་དཔེ་ཁྲག་ཕྲེང་བ་བཟོ་བ་ཡིན་རམ།
(a) མི་དཀོ་སིས།
(b) མེ་ཡོ་སིས།
(c) ཞ་མི་དཀོ་སིས།
(d) གཉིས་སྟེ་ཕྱེ་བ།

Reasoning
Grasshopper → Martin → Snake
What does the flowchart indicate?
(a) Struggle with the environment
(b) Interspecies struggle
(c) Intraspecies struggle
(d) Natural selection

ཚོ་དཔྱང། → མར་དེན། → སྦྲུལ།
གང་ཞིག་གི་རྒྱུན་ལས་དེ་གང་ཞིག་གི་དོན་དུ་བཀོད་པ་ཡིན་རམ།
(a) ཁོར་ཡུག་དང་འབྲེལ་བའི་འཐབ་རྩོད།
(b) རིགས་ཀྱི་བར་གྱི་འཐབ་རྩོད།
(c) རིགས་ཀྱི་ནང་གི་འཐབ་རྩོད།
(d) རང་བཞིན་གྱི་འདེམས་པ།

Figure 19: Biology sample questions for the 10th grade in ZEnQA by category.

Mathematics Sample Questions from ZEnQA

Mathematics

Factual Knowledge
The centre of the circumscribing circle of what type of triangle is situated on the longest side of the triangle?
(a) Equilateral
(b) Acute-angled
(c) Obtuse-angled
(d) Right-angled

གང་ཞིག་གི་རིགས་ཀྱི་རྒྱ་བཞིའི་ཐོན་དཀྱུ་སྒོར་བའི་སྒྲོར་དཔྱིས་ཀྱི་ཐོག་གས་ཆེ་མོས་སུ་གནས་པ་ཡིན་ནས།
(a) མཉམ་ཐོག་གས་རྒྱ་བཞི།
(b) མཐོ་ཚད་ཀྱི་རྒྱ་བཞི།
(c) དམའ་ཚད་ཀྱི་རྒྱ་བཞི།
(d) གཡས་ཀྱི་རྒྱ་བཞི།

Procedural & Application
If the length of a perpendicular on the chord is 3 cm from the centre of a circle with a radius of 5 cm, what is the length of that chord?
(a) 16 cm
(b) 8 cm
(c) 4 cm
(d) 2 cm

གལ་ཏེ་སྒྲོར་དཔྱིས་ཀྱི་ཐོག་གས་ཆེ་མོས་ནས་ཐོག་གི་ཐོག་གས་ཆུང་ངུ་ལ་ཟླག་པའི་ཟླག་པའི་ཚད་དེར་ཚད་ 3 cm ཡིན་པ་དང་། སྒྲོར་དཔྱིས་ཀྱི་ཐོག་གས་ཆེ་མོས་ཀྱི་ཚད་དེར་ཚད་ 5 cm ཡིན་ན། དེའི་ཐོག་གས་ཀྱི་ཐོག་གས་ཆུང་ངུའི་ཚད་དེར་ཚད་ཀར་ཡིན་ནས།
(a) 16 cm
(b) 8 cm
(c) 4 cm
(d) 2 cm

Reasoning
If $x + \frac{1}{x} = 5$, then:
(i) $(x - \frac{1}{x})^2 = 21$
(ii) $x^2 - 5x + 1 = 0$
(iii) $x^3 + \frac{1}{x^3} = 25$

Which one is correct?
(a) i and ii
(b) ii and iii
(c) i and iii
(d) i, ii, and iii

གལ་ཏེ་ $x + \frac{1}{x} = 5$, དེ་ལས་:
(i) $(x - \frac{1}{x})^2 = 21$
(ii) $x^2 - 5x + 1 = 0$
(iii) $x^3 + \frac{1}{x^3} = 25$

གང་ཞིག་ཆོས་གཞི་ཡིན་ནས།
(a) i དང་ ii
(b) i དང་ iii
(c) ii དང་ iii
(d) i, ii, དང་ iii

Figure 20: Mathematics sample questions for the 10th grade in ZEnQA by category.