BuddhismEval: Bilingually Evaluating LLMs on Theravada Buddhist Ethics and Reasoning

Anonymous ACL submission

Abstract

Though large language models (LLMs) have achieved significant advancements, their capacity to grasp culturally nuanced wisdom from non-Western traditions, such as Buddhism, remains largely underexplored. This is especially pronounced in low-resource languages such as Sinhala, which is central to Theravada Buddhism. To address this gap, we propose BuddhismEval, the first bilingual evaluation benchmark to assess LLMs on Theravada Buddhist ethics and reasoning in both English and Sinhala. We establish a taxonomy that consists of six question categories for Buddhist knowledge and reasoning assessment in LLMs such as factual understanding, interpretative and philosophical analysis, ethical and moral reasoning, practical application for life advice, comparative and analytical reasoning, and cultural and historical question category. The benchmark packs 1,242 English and 1,023 Sinhala MCQs, chiefly from the Dhammapada, covering all six categories. We also release a parallel English-Sinhala Dhammapada verse corpus to support cross-lingual research. We zero-shot test a range of open-source and proprietary LLMs on all six categories. Experiment results demonstrate that BuddhismEval is a challenging benchmark for LLMs spanning from non-reasoning to the strongest reasoning capabilities, and we additionally observe that multilingual LLMs with cultural alignment and instruction tuning tend to outperform larger LLMs explicitly fine-tuned for reasoning, particularly in the culturally grounded, low-resource context.

1 Introduction

"All things are preceded by mind, led by mind, made by mind" (Dhammapada) (Buddharakkhita, 1985). This foundational verse in the Dhammapada reflects the central Buddhist view that cognition, ethics, and reality are inherently shaped by the human mind, which is also the creator of LLMs (from the perspective of both model developer and data provider). As LLMs have increasingly influenced how knowledge is accessed, reasoned (Zhu et al., 2024) about, and communicated, it becomes essential to ask: can these models comprehend and reflect traditions where moral (Kucuk and Kocyigit, 2023) and philosophical reasoning is deeply entwined with culture and spiritual practice? Despite remarkable progress in LLM capabilities from commonsense reasoning to multistep problem solving, their ability to engage with culturally embedded, low-resource, and non-Western philosophical systems remains critically underexplored (Smart et al., 2024). Nowhere is this limitation more visible than in domains such as Theravāda Buddhism (Visuddhimagga) (Buddhaghosa, 2020), which emphasizes contextual ethical judgment, symbolic language, and culturally grounded metaphysics (Hallisey and Hansen, 1996).

Theravāda Buddhism is not only a deeply influential spiritual tradition but also a philosophical system that shapes moral life and societal values (Hallisey and Hansen, 1996) across South and Southeast Asia. Sinhala, the primary language of Sri Lankan Buddhists (Guzmán et al., 2019), plays a central role in the preservation and propagation of these teachings. However, the language remains critically low-resource in the NLP landscape (Guzmán et al., 2019), with limited curated datasets, benchmarks, or tools for evaluating LLMs' behavior in this culturally nuanced space.

To bridge this gap, we curate BuddhismEval: the first bilingual evaluation benchmark designed to assess LLMs on Theravada Buddhist ethical reasoning and cultural knowledge across English and Sinhala. Our goal is to evaluate how well current LLMs comprehend, reason through, and respond to Buddhist doctrinal, philosophical, and practical questions under zero-shot conditions. We address the following research question: *Can current LLMs accurately interpret and reason through culturally*

embedded Buddhist ethical and philosophical questions, especially in a low-resource language such as Sinhala?

To answer this, we construct two high-quality multiple-choice question (MCQ) datasets (1,242 questions in English and 1,023 in Sinhala) covering six categories of Buddhist reasoning: **(1) factual understanding, (2) interpretative and philosophical analysis, (3) ethical and moral reasoning, (4) practical life advice, (5) comparative and analytical reasoning, and (6) cultural and historical embedding**. These questions are derived from canonical texts such as the Dhammapada, as well as Sri Lankan exam materials[1] and Abhidharma literature.[2] To further support cross-lingual research, we release an English–Sinhala verse-aligned Dhammapada parallel corpus.

We evaluate 13 LLMs under zero-shot setting using controlled short, long, and bilingual prompts (Yu et al., 2024b). Our results highlight stark disparities across models. While instruction-tuned and multilingual LLMs with cultural alignment demonstrate strong reasoning and transfer capabilities in Sinhala, other high-benchmark models like Qwen3 32B and GPT-3.5 Turbo underperform drastically, sometimes failing to generate valid answers altogether.

We summarize our key contributions as follows:

1. **BuddhismEval**, the first bilingual benchmark for Buddhist ethical and philosophical reasoning across English and Sinhala, covering six culturally grounded question types.

2. **A Question Taxonomy**, a culturally grounded QA framework for benchmarking LLMs on ethical, moral, and philosophical reasoning.

3. **A parallel English–Sinhala Dhammapada corpus**, aligned at the verse level for future research in translation and semantic grounding.

4. **A comprehensive evaluation** of 13 LLMs under zero-shot conditions, across prompt types, question categories, sources, and linguistic domains.

5. **Empirical insights** into how LLMs handle Buddhist moral and cultural reasoning, revealing both the promise of instruction-tuned multilingual models and the limitations of high-parameter models lacking cultural grounding.

By releasing BuddhismEval and our full evaluation pipeline, we aim to encourage culturally inclusive LLM research and foster the development of AI systems that can engage with diverse religious and ethical traditions in meaningful and contextually appropriate ways, grounded in the belief that AI ethics is ultimately an ethics of the future of humanity (Kajiwara and Kawabata, 2024).

2 Related Work

Benchmarks for evaluating LLMs have rapidly matured from basic translation, classification, and syntactic/semantic tasks to sophisticated reasoning evaluations. Multilingual suites such as XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021), FLORES101 (Goyal et al., 2022), XQuAD (Santos et al., 2010), MLQA (Lewis et al., 2019), and TyDiQA (Clark et al., 2020) measure cross-lingual capabilities, including low-resource languages, while recent efforts like the Aya evaluation suite (Üstün et al., 2024), IndicNLP (Kakwani et al., 2020), IndicGenbench (Singh et al., 2024), and IndicSquad (Endait et al., 2025) target region-specific needs.

Reasoning-focused benchmarks probe deeper cognitive skills: GSM8K (Zeng et al., 2023) and MATHbench (Liu et al., 2024) assess multi-step mathematical reasoning; CommonsenseQA (Talmor et al., 2018), SocialIQA (Sap et al., 2019), PIQA (Bisk et al., 2020), and CausalBench (Wang, 2024) test commonsense, social/emotional, physical, and causal inference. The ETHICS benchmark (Hendrycks et al., 2020), Moral Stories (Emelin et al., 2020), and CMoralEval (Yu et al., 2024a) evaluate moral and ethical judgment across diverse philosophical frameworks.

Religious and culturally grounded datasets. Table 1 presents unique challenges. Digitization initiatives such as the Chinese Buddhist Canon (Wu, 2014), Pali Canon Corpora (Zigmond, 2021), and Tibetan Canon Corpora (Hackett, 2013) have created vast raw resources, but only recently have researchers produced aligned, task-oriented collections such as BibleQA (Zhao and Liu, 2018), QUQA (Alnefaie et al., 2023), QASiNa (Rizqullah et al., 2023), the Humanistic Buddhist Corpus (HBC) (Wong et al., 2024), the Nak Tham[3],

[1] https://www.doenets.lk/

[2] https://archive.org/details/AbhidhammaMulikaKarunu

[3] https://huggingface.co/datasets/biodatlab/thai_buddhist_studies_exam

Dataset	Category	Lang	#Inst	Task	Construction
BibleQA (Zhao and Liu, 2018)	Christianity (Bible)	En	1,001	Sent. selection	Trivia; answers aligned to verses
QUQA (Quran QA) (Alnefaie et al., 2023)	Islam (Quran)	Ar	2,189 (1,778 single; 411 multi)	QA benchmark	AQQAC & AyaTEC + five Islamic books
QASiNa (Rizqullah et al., 2023)	Sirah Nabawiyah	Id	500	Extractive QA	From Sirah Nabawiyah literature
ETHICS (Hendrycks et al., 2020)	General ethics	En	∼130k	MC classification	Crowdsourced moral scenarios (5 frameworks)
Moral Stories (Emelin et al., 2020)	Everyday norms/ethics	En	∼12k	Generative narrative	Crowdsourced branching dilemmas
BuddhismEval (ours)	Buddhism (doctrine, ethics)	Si & En	1,023 (Si); 1,242 (En)	MCQ (6 cats.)	Expert–AI MCQs from Dhammapada, exams, commentaries

Table 1: Comparison of religious/ethical QA benchmarks. Lang = language; #Inst = number of instances.

Buddhist Embeddings (Lugli et al., 2022), KG-Buddhism (Wu et al., 2017), AQQAC (Alqahtani and Atwell, 2018), and AYATEC (Malhas and Elsayed, 2020). These capture doctrinal, metaphorical, and symbolic language for machine translation, question–answering, and knowledge-graph construction.

Religious chatbots powered by AI. BuddhaBot (Pataranutaporn et al., 2020) and DharmaGPT[4] demonstrate the application of NLGs to spiritual guidance, raising challenges in domain adaptation, interpretability, user trust, and underscoring the need for ethically and culturally sensitive deployment.

Sinhala NLP remains underdeveloped compared with high-resource languages, despite large raw corpora like Sinhala Wikipedia[5] and OSCAR.[6] Key resources include the FLoRES evaluation dataset (Guzmán et al., 2019), SiTSE (Ranathunga et al., 2025), NSina (Hettiarachchi et al., 2024), CC100-Sinhala[7], the Sōḷd corpus (Ranasinghe et al., 2024), Dakshina Dataset (Roark et al., 2020), and Dolly Sinhala.[8] Recent Sri Lankan initiatives (De Silva, 2019), the National AI Strategy[9], and cross-lingual transfer work with models like BERTifying Sinhala (Dhananjaya et al., 2022) and XLM-R (Rajapaksha et al., 2023) address data gaps; yet dialectal variation, tokenization, and annotated resources continue to limit progress.

To our knowledge, there are no bilingual Sinhala-English evaluation datasets that focus on religious, ethical, or philosophical understanding and religious reasoning of LLM, especially those grounded in canonical Buddhist texts like the Dhammapada. This absence creates a significant gap in domain-specific evaluation and restricts the ability of LLMs to engage meaningfully with Sinhala Buddhist content, which has historically served as a moral compass for the Sinhala-speaking population (Smart et al., 2024).

3 Data Curation

Data curation plays a critical role in the creation of evaluation datasets, ensuring their integrity, authenticity, and relevance to Buddhist teachings. As illustrated in Figure 1, for this purpose, we employ a meticulous and structured approach that involves question taxonomy design, careful selection of limited and reliable sources, MCQ and answer creation, and quality control.

3.1 BuddhismEval Question Taxonomy

To develop our taxonomy, we have digged into some of the standard question taxonomies that exist in the world which are directly or indirectly related to our work, ranging from Sri Lankan Buddhist education assessments and traditional Buddhist question-types, to Bloom's Taxonomy and contemporary AI evaluation benchmarks. Further explanations and detailed comparisons are presented in the Appendix.

Inspired by the investigated principles and frame-

[4] https://www.dharmagpt.net/
[5] https://lukelindemann.com/wiki_corpus.html
[6] https://huggingface.co/datasets/oscar-corpus/OSCAR-2301#data-splits-sample-size
[7] https://autonlp.ai/redirect/cc100-sinhala
[8] https://huggingface.co/Suchinthana/Sinhala-Translate-and-Dolly-Llama-7b
[9] https://mot.gov.lk/assets/files/National_AI_strategy_for_Sri_Lanka-08a88c78d541a0746aeb8c71ed312231.pdf

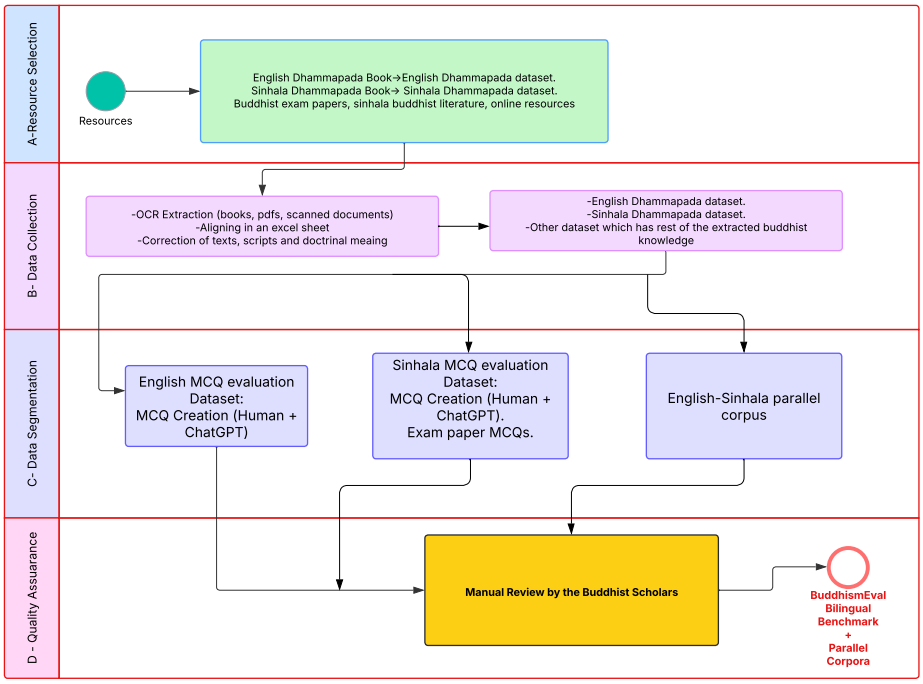

Figure 1: End-to-end pipeline for creating our Sinhala and English MCQ datasets and the parallel Dhammapada corpus.

works, we define a question-answer taxonomy which grounds in Theravada Buddhism to evaluate how good the current LLMs are in giving Buddhist religious advice to lay people. Our BuddhismEval question taxonomy is created upon six question categories. **(1) Factual Understanding** evaluates the model's ability to recall and understand the basic factual elements of Buddhist teachings. Questions in this category assess comprehension, memorization and foundational knowledge. **(2) Interpretative and philosophical analysis** category tests the model's ability to interpret complex Buddhist concepts and extract deep meanings from metaphors or allegories. Questions from this category assess abstract reasoning, critical thinking and philosophical interpretation. **(3) Ethical and moral reasoning** questions test the model's moral judgment and ability to recognize ethical dilemmas. This shows how well LLMs can reason through ethical principles and apply them in different contexts, assessing high-order moral reasoning. **(4) Practical application for life advice** question category aims to assess how well LLMs can apply Buddhist principles to modern situations and provide practical solutions based on Buddhist wisdom. **(5) Comparative and analytical reasoning** question category evaluates the model's ability to compare Buddhist teachings with models societal values or other religious and philosophical systems. **(6) Cultural and Historical** question category measures how well the LLMs understand and interpret local cultural and historical context of Buddhism, particularly in Sri Lanka. Please check the Appendix for these 6 question categories.

3.2 Data Collection and Question Generation

BuddhismEval is a bilingual benchmark dataset which consists of an English MCQ dataset and a Sinhala MCQ dataset. Both are manually and separately created by extracting Buddhist knowledge from limited number of credible sources rather than machine translated versions of each other. The main data collection source is the Dhammapada (Buddharakkhita, 1985) which is considered to be one of the pure Buddhist compendiums.

The process of dataset creation begins with comprehensive scraping of Dhammapada verses and their corresponding meanings in both English and Sinhala, using OCR tools for extraction. We create two distinct Dhammapada datasets: one for English and the other for Sinhala. These datasets are cleaned, aligned, and organized to ensure the accurate extraction of meaningful content.

As shown in Figure 1, for the English evalua-

tion dataset, the MCQs directly derive from the English version of the Dhammapada dataset. The Sinhala dataset has been made not only from the Sinhala Dhammapada dataset, but also from other culturally relevant sources, including Sri Lankan government-issued A/L and O/L Buddhism examination papers, Sinhala Buddhist literature, Abhidharma treatises, and Sinhala commentarial texts. Only the Buddhist MCQ exam paper questions directly qualifies to go to the finalize Sinhala MCQ dataset. We also extract Buddhism knowledge from other sources apart from exam papers using an OCR tool and align those in a separate dataset. With the help of ChatGPT (web version) (Huang and Xiong, 2023), we create MCQ questions and answers by prompting it to generate questions which can fall into one of the BuddhismEval question categories by using only the knowledge we provide. And we decide which question category that the question and the answer should belongs to. The same question and answer generation process applies to the English MCQ dataset.

Once we cover all the Buddhist knowledge we have gathered by making suitable questions and answers, we complete the question answer generation. Then we proofread and fact check the generated questions and answers in both datasets by consulting two Sri Lankan Theravada Buddhist knowledge experts. If they find any unreasonable or unsuitable questions and answers, they manually edit or update them. This ensures that the questions are rooted in both the core text and the specific cultural and religious practices of the Sinhalese Buddhist tradition by involving generative AI (Huang and Xiong, 2023) to improve the diversity and creativity in the process, yet still keeping humans as the top verifiers and controllers of this question-answer design process.

After constructing the MCQ evaluation datasets, a parallel corpus has also been created by aligning the English and Sinhala Dhammapada (Buddharakkhita, 1985) verses. This parallel corpus facilitates machine translation tasks, cross-lingual question answering, and semantic alignment (Hettiarachchi et al., 2024) studies, thus enhancing the scope and application of these datasets.

Buddhist Scholar Requirements. We recruited two Buddhist experts from Sri Lanka that matches with the eligibility criteria defined by us. (1) An expert must be a Sri Lankan Buddhist from birth; (2) Their parents, grandparents must be Buddhists; (3) They must go to a government Buddhist school in Sri Lanka; (4) Must attend Sunday Dhamma school since their childhood and must complete their final dharma expert examination (conducted by the department of examinations in Sri Lanka that considered to be the highest academic qualification that a student of Dhamma School can achieve. It remains a recognized and proper qualification to become a Dhamma School teacher.); (5) Have been studying Buddhism and Buddhist philosophy in both English and Sinhala for more than 20 years. One thing to note is that both these Buddhist experts are lawyers who have completed the final bar exam offered by the Sri Lankan Law College and are currently practicing law. Both experts are properly briefed and actively participated in releasing the finalized BuddhismEval bilingual benchmark datasets and assessing the LLMs' behaviors over the answers provided by them.

Data Quality Control. To ensure educational and evaluative integrity, we make sure that the MCQs are designed based on several principles. (1) Answers are drawn from reliable sources to maintain doctrinal accuracy; (2) questions are worded to reflect Buddhist teachings while being accessible to modern users; (3) the dataset balances linguistic complexity and cognitive demand, challenging both surface-level understanding and deeper philosophical reasoning. It also emphasizes ethical and moral grounding, encouraging reflection on Buddhist virtues.

Also, to ensure the quality of the dataset collection, after data has been collected and curated in English Dhammapada, Sinhala Dhammapada and other resource datasets, we properly proofread them. The Buddhist scholars' expertise in both the language and philosophy of Buddhism is crucial to maintaining the original intent and meaning of the source material. They make sure that the final datasets are properly polished and accurate.

3.3 Data Statistics

Two evaluation datasets are constructed, one in English containing 1,242 instances, and one in Sinhala containing 1,023 instances. Each instance is crafted and structured as an MCQ with two incorrect answers and one correct answer. The English dataset labels based on the question type it belongs to. According to our labelling, most of the questions and answers are falling into factual understanding, ethical reasoning and practical ap-

plication for life question categories. The Sinhala dataset labels are based on the question category that each question falls into, the source and the location of the source whether the source exists online or offline. This labelling helps us to analyze the model responses from a broader perspective. In the Sinhala dataset most of the MCQ questions fall under the interpretative and philosophical analysis, cultural and historical context and ethical, moral reasoning question categories. Check Appendix for category distribution.

Token-level analysis was conducted to understand the linguistic complexity and compatibility of the datasets with popular language models. For the English dataset, a total of 58,081 tokens were recorded, including 45,998 input tokens (questions and options) and 12,083 output tokens (answers). The tokenizer used was EleutherAI/pythia-70m[10], chosen for its optimization on English corpora and open-source availability. The Sinhala dataset recorded 57,412 tokens, with 48,113 input tokens and 9,299 output tokens, tokenized using NLLB-200.[11] This tokenizer was selected due to its effective sub-word segmentation for low-resource languages Sinhala and its broad multilingual capabilities.

4 Experiment

We evaluated 13 well-known LLMs on our bilingual benchmark dataset. This evaluation focused on how well these models interpret, process, and respond to questions grounded in Buddhist teachings, philosophy, and ethical reasoning in both Sinhala and English.

4.1 Evaluation Setup

LLMs. We selected a diverse set of LLMs that encompass both open-source and proprietary models, covering a range of reasoning capabilities and training paradigms. The LLMs to be evaluated are: BERT, RoBERTa, Mistral 7B Instruct v0.3, Qwen3 32B, Grok-3 Mini (beta), Command R7B, Command R, GPT-3.5 Turbo, Claude 3 Haiku, Gemini Flash 1.5, Phi 4 Reasoning Plus, Gemini 2.5 Flash Preview (thinking), DeepSeek R1.

Prompts. To ensure consistency and variety across evaluations, we used various prompts (Yu et al., 2024b) tailored to each language. Sinhala questions were introduced with three types of prompts such as short prompts, long prompts and bilingual prompts. The English questions were introduced with two types of prompts: short and long. The short prompt considered as the base prompt would look like "Choose the correct answer based on the given Buddhism question and output the final answer (1, 2, or 3)". In contrast, the long prompt is longer and more advanced than the short prompt. With the long prompt, we want the tested LLMs to follow a step to think internally and output the right answer. It goes like: "You are a Theravada Buddhist scholar with deep knowledge of the Dhammapada. Carefully analyze the following question based on Lord Buddha's teachings. Think step-by-step internally to find the right answer, but only output the final answer (1, 2, or 3)".

The bilingual prompt is applied only when we evaluate LLMs using the Sinhala dataset. The same long prompt was scripted in both English and Sinhala to see if the models are showing any distinctive results. Check Appendix for prompt samples.

Each question was followed by three answer options. The LLMs were instructed to return a single digit (1, 2, or 3) indicating their selected answer. We used a constant temperature of 0.2 and allowed a maximum of 150 tokens to encourage brief, focused outputs while leaving room for internal reasoning.

Evaluation Metrics. The evaluation metrics included (1) Accuracy or the overall correctness rate to measure the overall accuracy of the model performed based on the given prompt. (2) Category-wise Accuracy Breakdown Metrics: to analyze LLM behavior more insightfully, we computed accuracy across question categories (e.g., factual, philosophical, ethical etc.), source (e.g. textbooks, government exams etc.), and source location (whether the source has taken from online or offline). Question category is common to both datasets, but the source and source location only applies in the Sinhala dataset, because the Sinhala dataset is made by information taken from multiple sources and source locations whereas the English dataset questions are made only from the Dhammapada book. (3) Prompt sensitivity metrics: in further analyses, we assessed whether prompt length affects performance by comparing accuracy between: Short/simple prompts

[10] https://huggingface.co/EleutherAI/pythia-70m
[11] https://huggingface.co/facebook/nllb-200-3.3B

LLM	ENG-short(%)	ENG-long (%)	SN-short (%)	SN-long (%)	SN-bi (%)
BERT	22.62	31.48	42.62	42.72	42.91
RoBERTa	38.33	40.90	19.55	17.11	17.69
Mistral 7B Instruct v0.3	81.48	80.35	14.08	19.75	16.62
Qwen3 32B	0.08	0.00	0.59	0.00	0.78
Grok-3 Mini (beta)	0.00	0.00	20.82	32.45	32.45
Command R7B	36.88	73.51	14.08	11.73	16.23
Command R	82.29	81.64	66.47	**73.80**	**77.91**
GPT-3.5 Turbo	82.29	**82.69**	6.16	17.40	0.10
Claude 3 Haiku	**82.93**	0.00	65.40	69.21	69.40
Gemini Flash 1.5	81.69	82.05	**69.52**	**73.80**	75.60
Phi 4 Reasoning Plus	0.00	0.00	0.00	2.74	0.39
Gemini 2.5 Flash Preview (thinking)	1.37	1.21	13.20	22.48	33.82
DeepSeek R1	10.14	0.00	22.48	12.02	2.93

Table 2: Overall zero-shot **Digit Exact-Match (EM)** (%) under short/long prompts in English (ENG) and Sinhala (SN), plus bilingual Sinhala (SN-bi). Only a single digit (1/2/3) is accepted as correct; any non-compliant output is counted as incorrect. For reference, the Digit-Valid Rate (share of items where the model returned a single valid digit) on SN-bi was: Qwen3 32B=6%, GPT-3.5 Turbo=4%, Phi-4 Reasoning+=11%, DeepSeek-R1=18%; top models were $\geq$98%.

vs. long/informative prompts for English dataset, Short/simple prompts vs. long/informative prompts Vs. long/informative/bilingual prompt for Sinhala dataset. (4) Time per LLM: We also measured the time taken to finish evaluating a model.

Evaluation Process. In the beginning, we noticed a few mistakes made by the model. It used to miss questions without answering much often, used to provide 1 as the correct answer for all the questions, it used to provide more than one digit. To avoid making such errors we added the think-internally (invoking CoT) but only output the correct answer either 1, 2 or 3. Overall, to standardize interpretation, the following post-processing steps were applied:

1. If a single-digit output (1, 2, or 3) was returned, it was directly used.

2. If multiple digits or verbose outputs appeared, we selected the most frequent valid digit.

3. After three attempts, if the final results are similar as before or nothing more has changed, we accept the results we got from the third attempt.

Output-format compliance. We report *Digit Exact-Match (EM)*: the model must output exactly one digit (1/2/3). Non-compliant outputs (e.g., multiple digits, free text, or empty) are marked incorrect. Several near-zero EM scores coincide with low "digit-valid" rates, especially in Sinhala/bilingual prompts.

5 Results

Overall. As shown in the Table 4.1, performance varies widely across the 13 tested LLMs, with proprietary, instruction-tuned multilingual models (e.g., Command R, Claude 3 Haiku, Gemini Flash 1.5) consistently leading in both English and Sinhala. According to Table 4.1, in English (short prompts), these three and GPT-3.5 Turbo rank the highest, while several models (e.g., Qwen3 32B, Phi 4 Reasoning Plus) fail to produce valid answers.

Under Sinhala prompts, lightweight models like BERT and RoBERTa hold steady, but many high-performers from English drop sharply; Command

R remains the most robust across short, long, and bilingual prompts. Bilingual prompts particularly benefit Gemini Flash 1.5 and Claude 3 Haiku, whereas GPT-3.5 Turbo and some open-source models exhibit extreme prompt sensitivity.

In terms of efficiency, smaller models completed evaluation in less than a minute, while some large models required several hours such as Command R offering the best speed-accuracy trade-off. For detailed time wise accuracy by each model and prompt type, check the Appendix.

Further Analysis. A category-wise breakdown shows that top models like Command R and GPT-3.5 Turbo maintain high accuracy (>80%) across all English question types, while mid-ranked models such as Mistral 7B Instruct v0.3 excel in factual and analytical tasks but lag in practical application. In Sinhala, Command R remains the most balanced, scoring above 70% in every reasoning category, with Gemini Flash 1.5 and Claude 3 Haiku also strong, particularly on cultural questions.

Source-based results reinforce these trends: Command R and Claude 3 Haiku lead on both Dhammapada and Abhidharma materials, whereas GPT-3.5 Turbo and BERT perform poorly across sources. Performance is also stable across online versus offline sources for top models, but more variable for others. Please refer the Appendix.

Prompt sensitivity reveals that long and bilingual prompts generally boost accuracy: Command R rises from 66.5% (short) to 77.9% (bilingual) in Sinhala, while some models (e.g., GPT-3.5 Turbo, Claude 3 Haiku) show erratic behavior under different prompt styles.

Additional Analyses: Reasoning & Knowledge Gaps. Our benchmark spans six cognitive categories while highlighting large gaps in LLM performance on Buddhist-context tasks.

- **Factual Understanding.** Top English performers (Command R, GPT-3.5 Turbo, Mistral 7B: ∼88%) drop in Sinhala (Command R, Claude 3 Haiku: 73%; GPT-3.5 Turbo: 6.6%). BERT (45%) and RoBERTa (34%) outperform some larger models (e.g., Qwen3 32B: 0.7%).

- **Interpretative & Philosophical.** Command R retains 47.9% (vs. 80.2% English); others fall below 20%, showing weak abstract reasoning.

- **Ethical & Moral.** Command R (84% → 72%), Gemini Flash 1.5 (70%), and Claude 3 Haiku (68%) lead; GPT-3.5 Turbo (6.7%), Command R7B (13.4%) struggle in Sinhala.

- **Practical Application.** Gemini Flash 1.5 (56%), Claude 3 Haiku (58%), and Command R (45%) generalize; GPT-3.5 Turbo (4.2%) and Mistral 7B (23.9%) falter.

- **Comparative & Analytical.** Command R (88% → 76%) excels; Claude 3 Haiku (48%) and Gemini Flash 1.5 (56%) hold up; others near 0%.

- **Cultural–Historical (Sinhala only).** Gemini Flash 1.5 (82%) and Command R (73%) succeed; Claude 3 Haiku (64%) is moderate; others score 0%.

6 Conclusion

We present BuddhismEval, the first bilingual benchmark for evaluating LLMs on Buddhist ethical reasoning in English and Sinhala. Using high-quality MCQs from the Dhammapada, Sri Lankan exams, and other Buddhist texts, we offer a rigorous framework for assessing LLMs on non-Western moral and philosophical understanding.

Our taxonomy-driven design, grounded in Theravāda Buddhist epistemology, enables multi-dimensional evaluation across six reasoning categories: factual understanding, interpretative and philosophical analysis, ethical and moral reasoning, practical application, comparative and analytical reasoning, and cultural–historical grounding. Unlike general benchmarks, BuddhismEval requires models to move beyond surface-level reasoning and engage with the internal coherence and contextual subtleties of Buddhist thought.

Empirical results reveal large gaps: instruction-tuned multilingual models (e.g., Command R, Claude 3 Haiku) excel, but many top models struggle on Sinhala ethical and cultural questions. This underscores that model size and benchmarks alone don't ensure cultural competence—alignment and prompt design matter, and diverse training boosts cross-lingual performance.

In sum, BuddhismEval exposes critical gaps in the cultural and moral reasoning capacities of current LLMs and provides a concrete step toward more inclusive, spiritually aware, and ethically grounded AI.

Limitations

While BuddhismEval provides a rigorous bilingual benchmark, it has several caveats. First, a few models (e.g. Claude 3 Haiku, DeepSeek R1) consistently returned no valid answers under certain prompt formats, likely due to parsing/tokenization quirks. Second, prompt style proved sensitive: some models only improved when given longer or bilingual prompts—so our fixed templates may not suit every architecture. Third, we capped outputs at 150 tokens; longer chains of reasoning may have been truncated. Finally, our evaluation focuses on accuracy and category-level breakdowns; richer diagnostics (e.g. attention analysis or explicit rationales) would be needed to confirm true "reasoning" vs. memorization.

Ethics Statement

Our benchmark aims to improve LLMs' handling of Buddhist ethical and philosophical content. All source texts are publicly available (no private data), and expert reviewers ensured doctrinal fidelity. We commit to openly release all datasets, code, and metadata under a CC BY 4.0 license.

Disclosure of Interests

The authors have no competing interests to declare relevant to the content of this paper.

References

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023. HAQA and QUQA: Constructing two Arabic question-answering corpora for the Quran and Hadith. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 90–97, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mohammad Alqahtani and Eric Atwell. 2018. Annotated corpus of arabic al-quran question and answer.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Bhadantacariya Buddhaghosa. 2020. *The path of purification: Visuddhimagga*. Pariyatti.

Acharya Buddharakkhita. 1985. The dhammapada. *Kandy: Buddhist Publication Society*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Nisansa De Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. Bertifying sinhala–a comprehensive analysis of pre-trained language models for sinhala text classification. *arXiv preprint arXiv:2208.07864*.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*.

Sharvi Endait, Ruturaj Ghatage, Aditya Kulkarni, Rajlaxmi Patil, and Raviraj Joshi. 2025. Indicsquad: A comprehensive multilingual question answering dataset for indic languages. *arXiv preprint arXiv:2505.03688*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Paul Hackett. 2013. Digital resources for research and translation of the tibetan buddhist canon. In *Thirteenth Seminar of the International Association for Tibetan Studies (IATS-XIII), Ulaanbaatar, Mongolia*.

Charles Hallisey and Anne Hansen. 1996. Narrative, sub-ethics, and the moral life: Some evidence from theravāda buddhism. *The Journal of Religious Ethics*, pages 305–327.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Uyangodage, and Tharindu Ranasinghe. 2024. Nsina: A news corpus for sinhala. *arXiv preprint arXiv:2403.16571*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Yufei Huang and Deyi Xiong. 2023. Cbbq: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. *arXiv preprint arXiv:2306.16244*.

Yusuke Kajiwara and Kouhei Kawabata. 2024. Ai literacy for ethical use of chatbot: Will students accept ai ethics? *Computers and Education: Artificial Intelligence*, 6:100251.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Eyup Engin Kucuk and Muhammed Yusuf Kocyigit. 2023. Western, religious or spiritual: An evaluation of moral justification in large language models. *arXiv preprint arXiv:2311.07792*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

Ligeia Lugli, Matej Martinc, Andraž Pelicon, and Senja Pollak. 2022. Embeddings models for buddhist sanskrit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3861–3871.

Rana Malhas and Tamer Elsayed. 2020. Ayatec: building a reusable verse-based test collection for arabic question answering on the holy qur'an. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(6):1–21.

Pat Pataranutaporn, Bank Ngamarunchot, Korakot Chaovavanich, Sornchai Chatwiriyachai, Potiwat Ngamkajornwiwat, Nutchanon Ninyawee, and Werasak Surareungchai. 2020. Buddha bot: The exploration of embodied spiritual machine in chatbot. In *Proceedings of the Future Technologies Conference (FTC) 2019: Volume 2*, pages 589–595. Springer.

Nirupama Rajapaksha, Supunmali Ahangama, and Shalinda Adikari. 2023. Fine-tuning xlm-r for the detection of sinhala hate speech content on twitter and youtube. In *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, pages 19–23. IEEE.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2024. Sold: Sinhala offensive language dataset. *Language Resources and Evaluation*, pages 1–41.

Surangika Ranathunga, Rumesh Sirithunga, Himashi Rathnayake, Lahiru De Silva, Thamindu Aluthwala, Saman Peramuna, and Ravi Shekhar. 2025. Sitse: Sinhala text simplification dataset and evaluation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(5):1–19.

Muhammad Razif Rizqullah, Ayu Purwarianti, and Alham Fikri Aji. 2023. Qasina: Religious domain question answering using sirah nabawiyah. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–6. IEEE.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing south asian languages written in the latin script: the dakshina dataset. *arXiv preprint arXiv:2007.01176*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and 1 others. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Rodrygo LT Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. 2010. Explicit search result diversification through sub-queries. In *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings 32*, pages 87–99. Springer.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: a multilingual benchmark to evaluate generation capabilities of llms on indic languages. *arXiv preprint arXiv:2404.16816*.

Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, and 1 others. 2024. Socially responsible data for large multilingual language models. *arXiv preprint arXiv:2409.05247*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Zeyu Wang. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151.

Youheng W Wong, Natalie Parde, and Erdem Koyuncu. 2024. Humanistic buddhism corpus: A challenging domain-specific dataset of english translations for classical and modern chinese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8406–8417.

Jiang Wu. 2014. The chinese buddhist canon. *The Wiley Blackwell Companion to East and Inner Asian Buddhism*, pages 361–382.

Tianxing Wu, Cong Gao, Guilin Qi, Lei Zhang, Chuanqi Dong, He Liu, and Du Zhang. 2017. Kg-buddhism: the chinese knowledge graph on buddhism. In *Semantic Technology: 7th Joint International Conference, JIST 2017, Gold Coast, QLD, Australia, November 10-12, 2017, Proceedings 7*, pages 259–267. Springer.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, and 1 others. 2024a. Cmoraleval: A moral evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2408.09819*.

Linhao Yu, Qun Liu, and Deyi Xiong. 2024b. Lfed: A literary fiction evaluation dataset for large language models. *arXiv preprint arXiv:2405.10166*.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2023. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *arXiv preprint arXiv:2312.17080*.

Helen Jiahe Zhao and Jiamou Liu. 2018. Finding answers from the word of god: Domain adaptation for neural networks in biblical question answering. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

Dan Zigmond. 2021. Toward a computational analysis of the pali canon. *Journal of the Oxford Centre for Buddhist Studies*, 20.

A Appendix

A.1 BuddhismEval Question Taxonomy

To design our own question taxonomy, we had to carefully go through some of the standard question taxonomies that exist in the world which directly or indirectly relates to our work such as the Sri Lanka Buddhism education assessment, how questions are being addressed in Buddhism itself, Bloom's Taxonomy and modern evaluation benchmarks.

Sri Lankan Buddhist education comprises standard examinations such as Dhamma school exams, School level Buddhism exams, YMBA exams, university major course level exams, exams for Buddhist monks etc. These exam questions are organized around key traditional domains based on the Pali Canon (Thripitaka) which contains 3 major categories called Sutta Pitaka, Vinaya Pitaka and Abhidhamma Pitaka. "Sutta Pitaka" contains the Buddha's discourses delivered in the form of conventional teaching using conventional terms, depending on the intellect of the audience while "Vinaya Pitaka" contains the code of discipline for the Buddhist monks and nuns. Abhidhamma, which is believed to be the Buddha's ultimate teaching is considered to be at a higher level than the Sutta Pitaka and the Vinaya Pitaka, both of which deal with the practical aspects of Buddhism. It is more technical and analyzes the mental and physical phenomena into ultimate constituents which are classified and tabulated in a uniquely systematic way.

The level of difficulty and the Buddhist knowledge and theory coverage done by these Sri Lankan exams can be different. A typical O/L exam syllabus includes: the Buddha's life which discusses about biological events from birth to Enlightenment, before and beyond), Buddhist history and civilization (Sasana) which includes the Buddhist councils, spread of the religion, significant events happened and so on, Core Dhamma teachings such as doctrines and practices, Pali scriptures and verses and Abhidhamma principles which all these theory types can be fall into the sections of Pali Canon. There are a few more categories related to Buddhism in Sri Lanka such as the evolution of Buddhism in Sri Lanka, Sri Lankan Buddhist literature, significant and historical events happened and how we can apply the teachings of Pali canon into our today's life.

These exams heavily emphasize factual recall and understanding, critical thinking, philosophical interpretations, contextual application of Buddhist ethics, moral virtues and judgement, cultural sensitivity and historical understanding often via written tests with multiple choice questions (MCQ), short answer questions, structured essays and essay ques-

tions.

Buddhism itself recognizes different types of questions and appropriate ways to handle them because Buddhism constantly encourages people to reason before accepting any facts or truths are credible. The Lord Buddha categorizes questions into four types in Anguttara Nikaya which goes as (1) those that deserves a straight forward answer, (2) those that requires an analytical explanation, (3) those that best answered with a counter-question, and (4) those to be set aside as unanswerable. For example, a factual question could be what are the "five precepts or five rules of morality training for Buddhist lay people?" which gets a direct answer, where as a complex metaphysical question such as "is the universe eternal or not?" or "what is the end of the universe?" might be unanswerable and can put aside. Also, the "Four Noble Truths" in Buddhism: (1) suffering, (2) cause of suffering, (3) end of suffering, (4) path to end the suffering are the foundational teachings of Buddhism that provide a framework for understanding the nature of suffering and the path to liberation. This framework is constantly reminded by the Buddhist clergy to the laity whenever we encounter real life events that brings us sadness, anger, sorrow and negative thoughts.

Bloom's Taxonomy is another framework for categorizing educational goals, developed by a committee of educators chaired by Benjamin Bloom in 1956. It was first introduced in the publication Taxonomy of Educational Objectives: The Classification of Educational Goals. This taxonomy defines a hierarchy of cognitive objectives from lower to higher-order. In Bloom's original taxonomy, questions may test simple knowledge, comprehension, application, analysis, synthesis and evaluation. The revised Bloom's taxonomy similarly spans from remember and understand up to analyze, evaluate and create. Each level corresponds to different question forms, for example "Who brought Buddhism to Sri Lanka?" is a factual recall question where as "if Sri Lanka hadn't had Buddhism, what kind of belief system Sri Lankans might have followed?" considers as an analytical level question with a hypothetical scenario.

Modern AI evaluation benchmarks similarly categorize questions by the type of reasoning or knowledge required. LLMs are tested on their factual accuracy, counterfactual reasoning, philosophical or open-ended reasoning, ethical reasoning, practical problem solving and analytical reasoning. Recent research findings highlight the importance of evaluating moral and ethical reasoning in LLMs because they need to handle not just facts but value-laden decisions. Also, ethical reasoning and moral values can be culturally dependent. Studies show multilingual, multi-cultural prompts lead to different moral judgements. Given that there are considerable number of Buddhist chatbots offering Buddhist guidance and knowledge, a robust evaluation should include ethical questions grounded in specific religious and philosophical traditions to see if the models if the models can align with these value systems.

A.2 Sample Bilingual MCQs by Question Category

1) Factual Understanding (Literal Meaning)

> *According to Buddhist teaching, what follows a person who speaks or acts with an impure mind?*
> 1. Suffering follows, like the wheel follows the foot of the ox.
> 2. Enlightenment
> 3. Joy and prosperity
> **Correct:**Suffering follows, like the wheel follows the foot of the ox.

2) Interpretative Questions (Philosophical Analysis)

> *From the perspective of Buddhist philosophy as presented in The Dhammapada, how are the karmic consequences of harmful speech described?*
> 1. As a means to achieve social influence and rewards
> 2. As ethically neutral, lacking any significant personal or spiritual effect
> 3. It harms oneself and others, creating negative karma
> **Correct:** It harms oneself and others, creating negative karma.

3) Ethical and Moral Reasoning

> *You witness a friend excel in a task, which stirs jealousy (Mára) within you, tempting you to undermine their success. According to the*

Dhammapada, what is the most ethical way to overcome this temptation?
1. Avoid contact with your friend to suppress the jealousy.
2. Subdue your mind through mindfulness and discipline.
3. Spread rumors to damage your friend's reputation.
Correct: Subdue your mind through mindfulness and discipline.

4) Practical Applications for Life Advice

How can Buddhist teachings guide someone struggling with guilt after doing something wrong?
1. Blame others and justify the action
2. Reflect on the harm caused and learn from it
3. Ignore the guilt and move on
Correct: Reflect on the harm caused and learn from it.

5) Comparative and Analytical Questions

What does the Dhammapada say about protection from harm compared to modern security methods?
1. Protection comes from ethical living and mindfulness.
2. Security depends mostly on wealth and technology.
3. Isolation is the best form of protection.
Correct: Protection comes from ethical living and mindfulness.

1) Factual Understanding (Literal Meaning)

"බෞද්ධ ජීවිතයක ප්‍රකට වියයුතු මූලික සද්ගුණ බඹ විහරණ නමින් හැඳින්වෙයි. එම සද්ගුණ හඳුන්වන තවත් ධර්මානුකූල ව්‍යවහාරයක් නම්,
1. චතුඅප්පමඤ්ඤා යන්නයි.
2. චතුපටිසම්භිදා යන්නයි.
3. චතුරාරක්ඛා යන්නයි.
නිවැරදි පිළිතුර: චතුඅප්පමඤ්ඤා යන්නයි.

2) Interpretative Questions (Philosophical Analysis)

"බුදුදහම අනුව සසර දුක තුන් ආකාර ය. එනම් දුක්ඛ දුක්ඛතා, විපරිණාම දුක්ඛතා හා සංඛාර දුක්ඛතා යනුවෙනි. මෙහි දුක්ඛ දුක්ඛතා
1. ජාති ජරා ව්‍යාධි මරණාදී දුක් ය.
2. මානසිකව විඳීමට සිදු වන දුක් පීඩාවන්
3. වෙනස්වීම පිළිබඳ දුක ය.
නිවැරදි පිළිතුර: ජාති ජරා ව්‍යාධි මරණාදී දුක් ය.

3) Ethical and Moral Reasoning

"මංගල සූත්‍රයෙහි "අනවජ්ජානි කම්මානි"" යන පාඨයෙන් කියැවෙන නිවැරදි කර්මාන්ත ගණයට අයත් නොවන්නේ,
1. කෘෂි කර්මාන්තයයි.
2. සත්ත්ව වෙළඳාමයි.
3. ගව පාලනයයි.
නිවැරදි පිළිතුර: සත්ත්ව වෙළඳාමයි.

4) Practical Applications for Life Advice

"රාජ්‍ය පාලනය පිළිබඳ බෞද්ධ ඉගැන්වීම් අතර හමු වන සප්ත අපරිහානීය ධර්මවල සඳහන් පරිදි, පාලකයා විසින් ප්‍රජාවගේ යහපත උදෙසා ගතයුතු ක්‍රියාමාර්ගයක් වන්නේ,
1. ධනෝපායන මාර්ග විවර කර දීම ය.
2. පූජනීය ස්ථානවලට ගරුසත්කාර කිරීම ය.
3. යුද භටයන්ට රක්ඛාවරණ ගුප්තිය සැලැස්වීම ය.
නිවැරදි පිළිතුර: පූජනීය ස්ථානවලට ගරුසත්කාර කිරීම ය.

5) Comparative and Analytical Questions

"ලොව පහළ වූ ආගමික දර්ශනවාද අතර බුදුසමය ප්‍රකට කරන සුවිශේෂ ලක්ෂණයක් ලෙස හැඳින්විය හැක්කේ,
1. නවීන විද්‍යාව සමග කොහෙත්ම එකඟ නොවීම ය.
2. චින්තනයට හා වීමංසනයට දෙන නිදහස සීමා රහිත වීම ය.
3. බටහිර දර්ශනය හා හැමවිටම පරස්පර නොවීම ය.
නිවැරදි පිළිතුර: බටහිර දර්ශනය හා හැමවිටම පරස්පර නොවීම ය.

6) Cultural and Historical Context

"බුදුරජාණන් වහන්සේ දෙවිලොව සිට සංකස්ස පුරයට වැඩම කිරීම චිත්‍රයට නගා ඇති පොළොන්නරු යුගයට අයත් පූජනීය ස්ථානය කුමක් ද?
1. හිඳගල විහාරය
2. තිවංක පිළිම ගෙය
3. මැදවල විහාරය
නිවැරදි පිළිතුර: තිවංක පිළිම ගෙය.

1

Figure 2: Sample of Sinhala MCQs by Question Category

A.3 Category Distribution of Each Evaluation Dataset

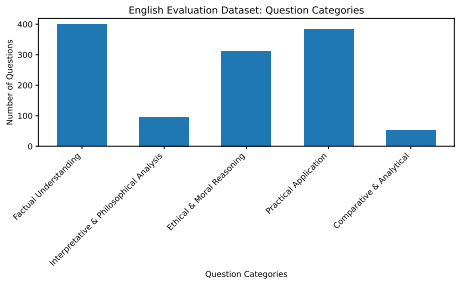

Figure 3: Number of English MCQs per question category.

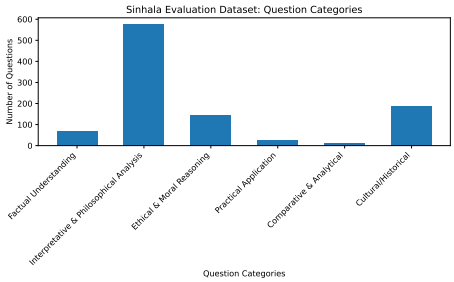

Figure 4: Number of Sinhala MCQs per question category.

A.4 System Prompts

Below are the exact prompt templates used in our evaluation, presented as plain text.

English Short Prompt

> [INST] Choose the correct answer based on the following Theravada Buddhist teachings and explain your reasoning:
>
> Question: *[question here]*
>
> Options: 1. *[option 1]* 2. *[option 2]* 3. *[option 3]*
>
> Answer by using only one of the options (1, 2, or 3): [/INST]

English Long Prompt

> [INST] You are a Theravada Buddhist scholar with deep knowledge of the Dhammapada. Carefully analyze the following question based on Lord Buddha's teachings. Think step-by-step internally to find the right answer, but only output the final answer (1, 2, or 3).
>
> Question: *[question here]*
>
> Options: 1. *[option 1]* 2. *[option 2]* 3. *[option 3]*
>
> Your final answer (1, 2, or 3): [/INST]

Sinhala Short Prompt

පහත බෞද්ධ ප්‍රශ්නයට පිළිතුරු දෙන්න. පිළිතුර අංකයෙන් පමණක් දක්වන්න (1, 2, හෝ 3).

ප්‍රශ්නය: *[question here]*

අභ්‍යන්තර: 1. *[option 1]* 2. *[option 2]* 3. *[option 3]*

පිළිතුර:*[correct option]*

Sinhala Long Prompt

ථෙරවාදී බෞද්ධාගම, බෞද්ධ දර්ශනය පිළිබඳ දැනුම භාවිතා කරමින් පහත බෞද්ධ ප්‍රශ්නයට පිළිතුරු දෙන්න.

පිළිතුරට පෙර තර්කන පියවර භාවිතා කරන්න.

ප්‍රශ්නය: *[question here]*

අභ්‍යන්තර: 1. *[option 1]* 2. *[option 2]* 3. *[option 3]*

ඔබගේ අවසාන පිළිතුර (1, 2, හෝ 3):

Sinhala Bilingual Prompt

ථෙරවාද බෞද්ධ විද්වතෙකු ලෙස සවිස්තරාත්මක විශ්ලේෂණය කරන්න. (As a Theravada Buddhist scholar, conduct a detailed analysis)

පිළිතුරට පෙර තර්කන පියවර භාවිතා කරන්න. (Use reasoning steps before answering)

ප්‍රශ්නය: *[question here]*

අභ්‍යන්තර: 1. *[option 1]* 2. *[option 2]* 3. *[option 3]*

ඔබගේ අවසාන පිළිතුර (1, 2, හෝ 3): (Your final answer (1, 2, or 3):

1

Figure 5: Sinhala short prompt, Long prompt, Bilingual Prompt

A.5 Comparison of the Chosen LLMs

Model	Training Type	OS/P	Reasoning	Size	Lang.	Scores (MMLU / GSM8K)
BERT	Pre-trained + fine-tuned	OS	Basic	<10 B	MultiL	Low (60)
Command R	Instruction-tuned	P	Inter-mediate	>30 B	MultiL	Moderate (70)
Mistral 7B Instruct v0.3	Instruction-tuned	OS	Inter-mediate	<10 B	MultiL	High (75)
Qwen3 32B	Instruction-tuned	OS	Advanced	>30 B	MultiL	High (80)
Grok-3 Mini (beta)	Instr. + RL-tuned	P	Inter-mediate	<10 B	MultiL	Moderate (70)
Command R7B	Instruction-tuned	P	Advanced	10–30 B	MultiL	High (75)
RoBERTa	Pre-trained + fine-tuned	OS	Basic	<10 B	MultiL	Low (65)
GPT-3.5 Turbo	Instr. + RL-tuned	P	Advanced	<30 B	MultiL	MMLU 70; GSM8K 60
Claude 3 Haiku	Instr. + RL-tuned	P	Advanced	<30 B	MultiL	MMLU 75; GSM8K 65
Gemini Flash 1.5	Instr. + RL-tuned	P	Advanced	<30 B	MultiL	MMLU 80; GSM8K 70
Phi 4 Reasoning Plus	Instruction-tuned	OS	Inter-mediate	<10 B	MultiL	Moderate (70)
Gemini 2.5 Flash Preview	Instr. + RL-tuned	P	Advanced	>30 B	MultiL	MMLU 80; GSM8K 75
DeepSeek R1	Instruction-tuned	OS	Inter-mediate	10–30 B	MultiL	Moderate (70)

Table 3: Comparison of the chosen LLM models. Instr: instruction-tuned; RL-tuned: reinforcement-learning tuned; OS: open-source; P: proprietary; MultiL: multilingual.

A.6 Category-wise Performance Heatmaps

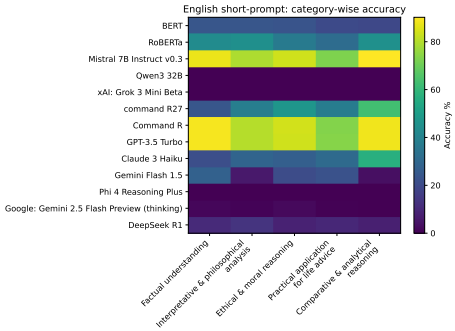

Figure 6: English short-prompt category-wise accuracy heatmap with corrected model names (e.g., "command R 27" → Command R7B, "xAI: Grok 3 Mini Beta" → Grok-3 Mini (beta)). Rows=LLMs; columns=question categories.

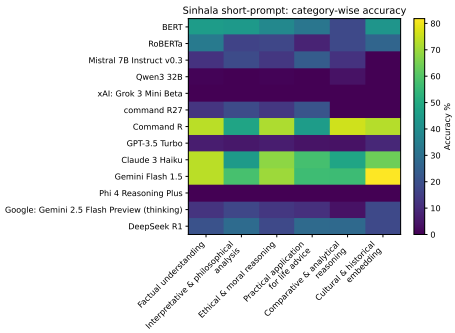

Figure 7: Sinhala short-prompt category-wise accuracy heatmap with corrected model names (e.g., "command R 27" → Command R7B, "xAI: Grok 3 Mini Beta" → Grok-3 Mini (beta)). Rows=LLMs; columns=question categories.

A.7 Source-wise Performance Heatmap (Sinhala)

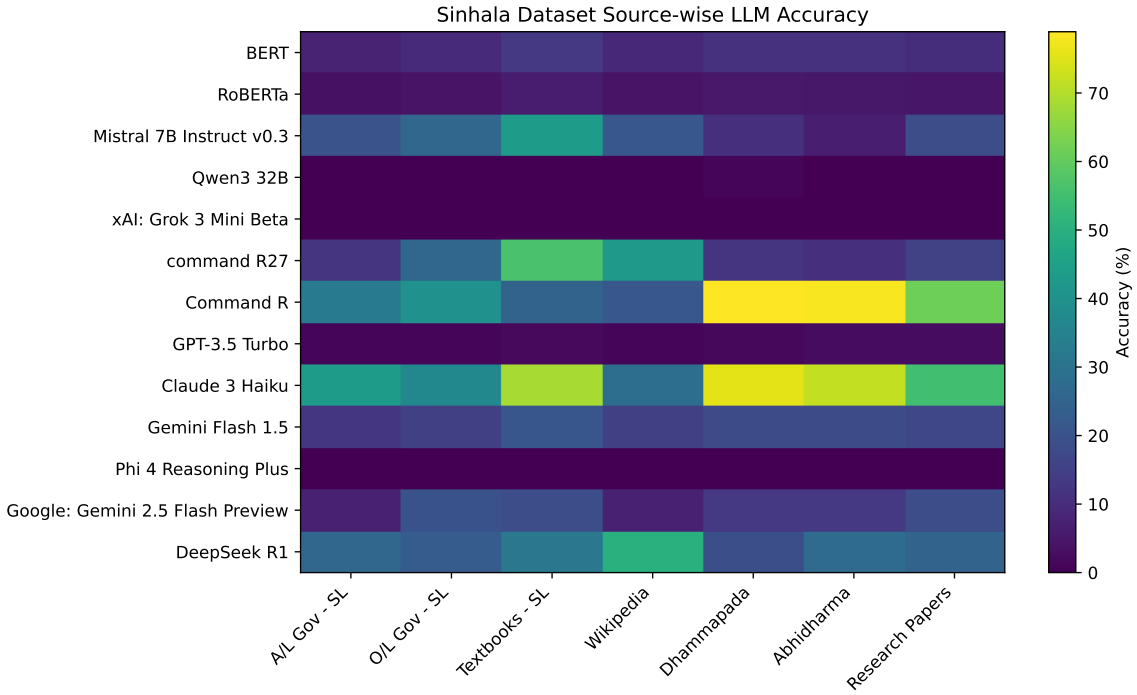

Figure 8: Source-wise accuracy heatmap for the Sinhala dataset, showing each model's performance across different source categories (A/L Gov, O/L Gov, Textbooks, Wikipedia, Dhammapada, Abhidharma, Research Papers) with corrected model names (e.g., "command R 27" → Command R7B, "xAI: Grok 3 Mini Beta" → Grok-3 Mini (beta)).

A.8 Source-Location Performance

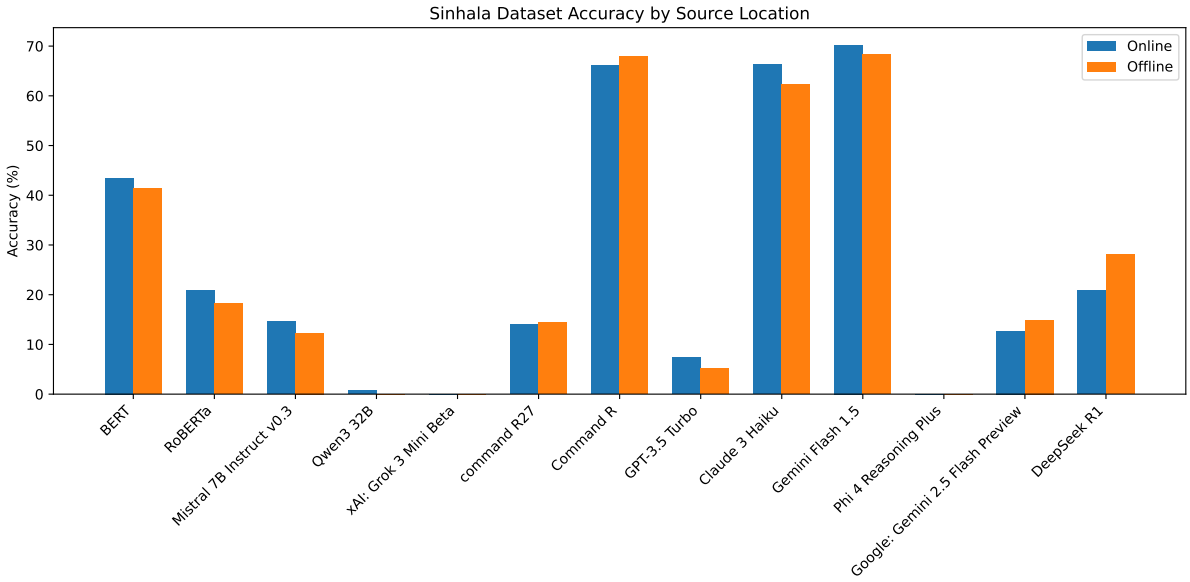

Figure 9: Sinhala dataset accuracy by source location (online vs. offline), with corrected model names (e.g., "command R 27" → Command R7B, "xAI: Grok 3 Mini Beta" → Grok-3 Mini (beta)).

A.9 Total time taken by each LLM

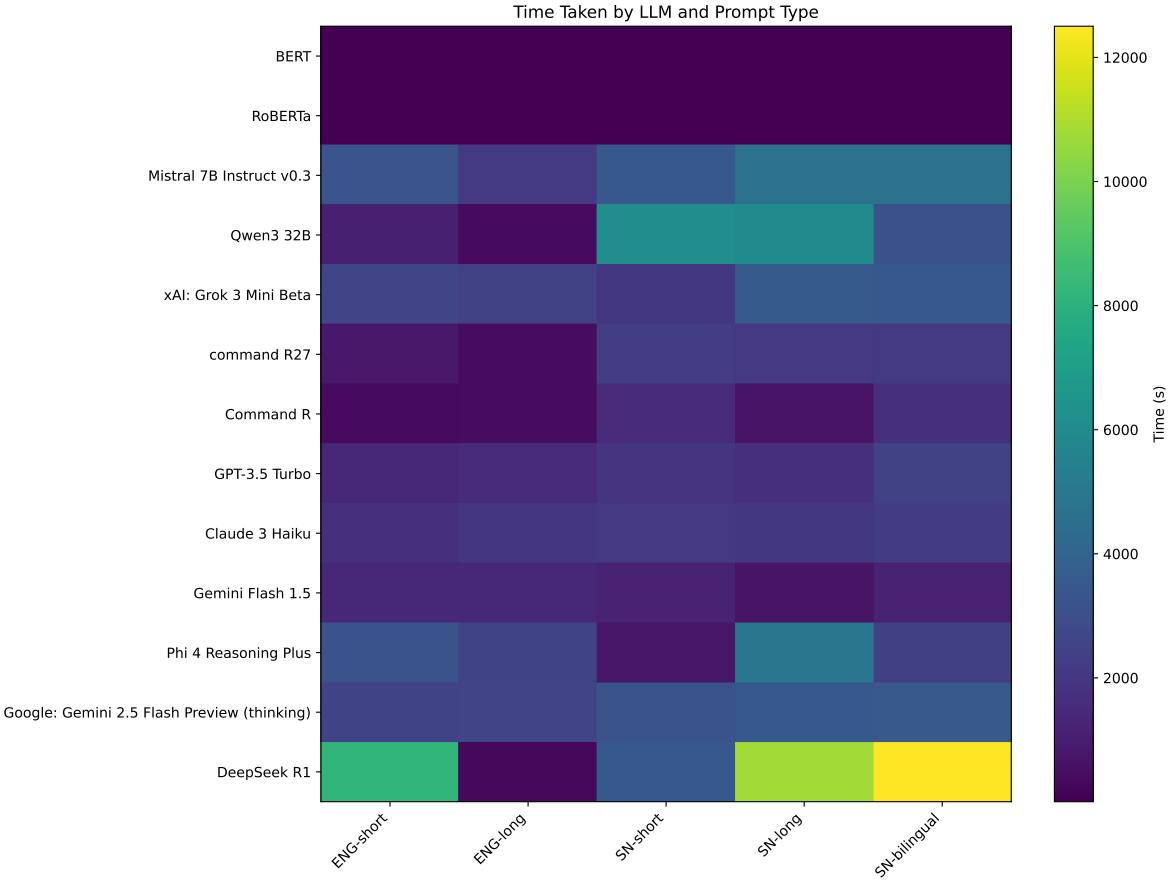

Figure 10: Heatmap of inference time (in seconds) for each LLM Across Prompt Types, with corrected model names (e.g., "command R 27" → Command R7B, "xAI: Grok 3 Mini Beta" → Grok-3 Mini (beta)). Darker cells indicate shorter runtime.