DharmaBench: Evaluating Language Models on Buddhist Texts in Sanskrit and Tibetan

Kai Golan Hashiloni[1,2], Shay Cohen[1,2], Asaf Shina[1,2], Jingyi Yang[3], Orr Meir Zwebner[1,2], Nicola Bajetta[3], Guy Bilitski[1,2], Rebecca Sundén[3], Guy Maduel[1], Ryan Conlon[3], Ari Barzilai[1,2], Daniel Mass[1,2], Shanshan Jia[3], Aviv Naaman[1], Sonam Choden[3], Sonam Jamtsho[3], Yadi Qu[3], Harunaga Isaacson[3], Dorji Wangchuk[3], Shai Fine[1], Orna Almogi[3], Kfir Bar[1,2]

[1]Data Science Institute, Reichman University, Herzliya, Israel
[2]Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
[3]University of Hamburg, Germany

Correspondence: kai.golanhashiloni@post.runi.ac.il

Abstract

We assess the capabilities of large language models on tasks involving Buddhist texts written in Sanskrit and Classical Tibetan—two typologically distinct, low-resource historical languages. To this end, we introduce **DharmaBench**,[1] a benchmark suite comprising 13 classification and detection tasks grounded in Buddhist textual traditions: six in Sanskrit and seven in Tibetan, with four shared across both. The tasks are curated from scratch, tailored to the linguistic and cultural characteristics of each language. We evaluate a range of models, from proprietary systems like GPT-4o to smaller, domain-specific open-weight models, analyzing their performance across tasks and languages. All datasets and code are publicly released, under the CC-BY-4 License and the Apache-2.0 License respectively, to support research on historical language processing and the development of culturally inclusive natural-language-processing systems.

1 Introduction

Large Language Models (LLMs) have made rapid progress in natural language processing (NLP) and understanding (NLU), achieving strong performance across a wide range of tasks in high-resource languages such as English, Chinese, and French (Brown et al., 2020; Rae et al., 2021; OpenAI et al., 2024; Tay et al., 2023; Team et al., 2025). These advancements have been fueled by large-scale datasets, architectural innovations, and the development of general-purpose benchmarks like GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2020), and MMLU (Hendrycks et al., 2021). However, the majority of existing benchmarks focus on contemporary, high-resource languages, overlooking historical and low-resource languages that are vital for scholarly research and cultural preservation. Sanskrit and Classical Tibetan are two such languages. Both are central to the Buddhist textual tradition, covering over a millennium of philosophical, literary, and religious discourse. Despite their cultural and scholarly importance, they remain vastly underrepresented in modern NLP research. These languages pose unique challenges: rich morphology, long and complex sentence structures, and domain-specific terminology that differs significantly from modern language usage. Furthermore, the scarcity of annotated corpora, standardized tools, and benchmarks for these languages has limited progress in developing and evaluating LLMs for historical and domain-specific language understanding.

The 2025 AI Index Report from Stanford's Human-Centered AI (HAI) initiative[2] shows that LLMs are surpassing newly introduced benchmarks at an accelerating rate. Benchmarks released as recently as 2023 have already seen major performance gains, highlighting the need to continually create new benchmarks, particularly in underexplored domains, to effectively evaluate and challenge the growing capabilities of modern AI systems.

Ancient languages pose distinct challenges, including highly heterogeneous corpora spanning centuries, orthographic variation, a lack of standardization, and the absence of native speakers for validation. Benchmarking is especially difficult, since tasks must reflect real philological needs rather than generic NLP setups. Prior work on Latin, Ancient Greek, and Classical Arabic—such as treebanks (Mcgillivray et al., 2009; Bamman

[1]Dharma is a key Indic term with multiple senses, including "teaching", "law", and "truth". In Buddhist contexts, it often refers to the Buddha's teaching or the textual tradition itself.

[2]https://hai.stanford.edu/ai-index/2025-ai-index-report

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2088–2110

and Crane, 2011; Eckhoff et al., 2018; Vierros, 2018), morphological tagging (Dukes and Habash, 2010; Sharaf and Atwell, 2012), and genre classification (Alrabiah et al., 2013; Ahmed et al., 2025)—has demonstrated the value of domain-specific benchmarks. However, such resources are still missing for Sanskrit and Tibetan.

Prior work on low-resource benchmarks has largely focused on contemporary spoken languages. The recent IndicGenBench (Singh et al., 2024) set a first large benchmark for Indic languages, including Sanskrit and Tibetan. Focusing solely on Tibetan, we see the recent TLUE benchmark (Gao et al., 2025). Both leave the classical form of the languages unattained. While modern versions of a language reflect its contemporary usage, classical forms refer to earlier, often literary or religious stages that differ significantly in grammar, vocabulary, and orthography. Additionally, while there has been growing interest and notable progress in developing language models tailored to Tibetan and Sanskrit (Huang et al., 2025; Nehring et al., 2024; Chaudhari et al., 2024; Lv, 2024), the creation of corresponding benchmarks has lagged behind.

To bridge this gap, we introduce **DharmaBench**, the first benchmark specifically designed to evaluate language models on classification and detection tasks involving Buddhist texts written in Sanskrit and Classical Tibetan. The benchmark consists of 13 tasks—six in Sanskrit, seven in Tibetan—with four designed as cross-lingual tasks that span both languages. All tasks are carefully curated from historical Buddhist corpora, tailored to reflect both the linguistic complexity and domain relevance of the source materials.

We evaluate a range of LLMs, including state-of-the-art closed-weight models (e.g., GPT-4o, Claude 3.7 Sonnet, Gemini 2.5) as well as open-weight models (like Llama 4 or Qwen-2.5). Our results highlight key challenges that current models face when applied to ancient, typologically diverse languages, especially in low-resource and domain-specific settings.

Our main contributions can be summarized as follows: **(1)** We release DharmaBench, a multi-task benchmark suite for evaluating Buddhist classification and detection tasks in Sanskrit and Classical Tibetan. The benchmark includes 13 tasks, with four of them implemented similarly in both languages. By providing a rigorous, culturally informed benchmark for ancient Buddhist languages, we aim to advance research on multilingual, historically grounded NLP systems. **(2)** We conduct extensive evaluations across a range of LLMs, identify key limitations, and provide guidance for future work in historical and low-resource NLP. This sets a first baseline and benchmark for future models.

Once they reach a satisfactory level, we expect LLMs for these languages to be primarily used by philologists, digital humanists, and historians, who will leverage capabilities rooted in these tasks.

All datasets and evaluation code are publicly released to foster reproducibility and community involvement.

2 Related Work

2.1 Multilingual Evaluation Benchmarks for Tibetan and Sanskrit

For multilingual evaluation, including Tibetan and Sanskrit, Singh et al. (2024) introduced IndicGenBench, which assesses LLM generation across Indic languages using an auto-translation (machine-translation) and post-correction pipeline. GlotStoryBook (Kargaran et al., 2023) provides a multilingual story collection, including Sanskrit and Tibetan, for narrative understanding and generation. FLORES-200 (NLLB Team, 2022), built on FLORES-101 (Goyal et al., 2022), is a comprehensive multilingual benchmark with sentence-level parallel translations, including Sanskrit and Tibetan. The recent Massive Multilingual Text Embedding Benchmark (MMTEB)(Enevoldsen et al., 2025), which includes a large, multilingual set of evaluation tasks for embedding models, also covers both languages. SansTib (Nehrdich, 2022) is a Sanskrit–Classical Tibetan parallel corpus with a 6,916-pair gold test set, enabling translation evaluation. Nehrdich et al. (2025) introduced a Buddhist Chinese evaluation dataset and Gemma-2-mitra, a multilingual LLM for translation and retrieval across Pāli, Sanskrit, Buddhist Chinese, and Tibetan.

2.2 Tibetan Evaluation Benchmarks

Gao et al. (2025) introduced TLUE, a benchmark for Tibetan language understanding across multiple tasks using auto-translation. Deng et al. (2023) released MiTC, a multilingual text classification dataset with 82,662 samples across five Chinese minority languages, including Tibetan. WCM (Yang et al., 2022) provides a Wikipedia-based classification benchmark for Tibetan and other minority languages. TNCC (Zhang et al., 2022) is a Tibetan news classification dataset with 9,276

samples in 12 classes. TUSA (Zhang et al., 2024) and TNEC (Kong et al., 2024) offer 10,000 and 100,000 samples for Tibetan sentiment analysis, respectively.

TibetanQA (Sun et al., 2021) contains 20,000 QA pairs, while TibetanQA2.0 (Dan and Sun, 2024) refines the dataset with improved quality and 12,054 entries including unanswerable questions. Jin et al. (2024) built Tibetan medical resources for named-entity recognition (NER), and Pan et al. (2025) introduced four Tibetan reasoning benchmarks, a 70GB unannotated corpus on which they trained a Qwen2.5-7B-based LLM.

Most existing benchmarks target modern Tibetan. The Annotated Corpus of Classical Tibetan (ACTib) (Wallman et al., 2017), a 185M-word segmented and POS-tagged version of the Buddhist Digital Resource Center's etexts[3], supports linguistic-level evaluation such as POS tagging.

2.3 Sanskrit Evaluation Benchmarks

Several Sanskrit benchmarks exist, most focusing on text generation, with some covering Sanskrit. Anveshana (Jagadeeshan et al., 2025) targets cross-lingual information retrieval with English queries and Sanskrit documents, bridging modern–classical language evaluation. In ancient Indian philosophy, VedantaNY-10M (Mandikal, 2024) provides a dataset for long-form QA on Advaita Vedanta texts. Chaudhari et al. (2024) introduced a simile element detection task requiring models to identify the four components of Sanskrit similes (*upamā*) in classical poetry, expanding from 400 annotated to 17K examples, along with a Semantic Analogy Prediction task. Jadhav et al. (2025) released a 128-sample multitask dataset on *upamā alaṅkāra* (similes), evaluating LLMs on simile classification and component detection.

2.4 Summary

Despite these advances, low-resource and typologically distinct languages—especially historical ones such as Sanskrit and Classical Tibetan—remain significantly underrepresented. Moreover, the unique challenges associated with processing ancient textual material, including specialized classification and detection tasks, are largely absent from existing general-purpose benchmarks. This gap motivates the need for domain-specific, linguistically grounded evaluation suites like ours.

[3] https://www.bdrc.io/

3 DharmaBench

We design a novel multilingual, multi-task benchmark to assess a model's ability to understand Sanskrit and Classical Tibetan (hereafter referred to as Tibetan). All tasks are domain-specific and probe the model's grasp of core linguistic and conceptual phenomena that are especially prominent in these languages. We define and curate datasets for six tasks in Sanskrit and seven in Tibetan, with four tasks shared across both languages. An example of the input and output of each task is depicted in Table 2 and Table 3 (both in Appendix A). For convenience, each task is assigned a code, with the last character either `S` or `T` to indicate the language. The benchmark consists of classification and detection tasks that require the model to identify and label specific text spans. For some tasks where the curation process is less labor-intensive, we also provide train-test splits to support fine-tuning of small LMs. In all classification tasks, unless stated otherwise, there is an equal number of texts from each label.

We proceed to describe each of the tasks comprising the DharmaBench benchmark. Dataset sizes, task types, and average text length in characters are summarized in Table 1.

All tasks were developed in collaboration with a team of scholars, all experts in Sanskrit and Tibetan, who are either PhD candidates or hold doctoral degrees. For more information about the annotation team, see Appendix D.

Task-specific annotation guidelines were developed by at least two annotators and approved by the first author before the annotation process. This team was first asked to collect relevant texts for each task from diverse sources to facilitate a broader, more general evaluation of the tested models. For the full description of the source texts used for each task, see Appendix C. We release all annotation guidelines and source references to facilitate further curation and extension of this benchmark. In some tasks, we used Label Studio (`LS`)[4] as our annotation platform. For example screenshot see Figure 2 (Appendix G). In all tasks, a *sample* is defined as a given sentence, passage, or text, unless stated otherwise explicitly. In other words, a sample is a segment of continuous text.

3.1 Task Definitions

Simile and Metaphor Detection (SMD). This is a detection task targeting figurative language—specifically similes and metaphors—in Sanskrit

[4] https://labelstud.io/

	SMDS	QUDS	RCMS	RCDS	VPCS	MCS	SDT	QUDT	RCMT	VPCT	AACT	SCCT	THCT
Type	D	D	C	D	C	C	D	D	C	C	C	C	C
Labels	2	3	2	1	2	10	1	1	2	2	2	2	13
Train	0	0	510	510	0	0	0	1,000	705	600	600	600	600
Test	410	400	400	400	400	200	328	406	400	400	400	400	400
Avg. Length	107	672	639	884	144	148	231	1,009	1,028	608	1,213	1,287	1,287

Table 1: Task type (*D* stands for detection and *C* for classification), number of labels/span types, split sizes and average text length in characters.

and Tibetan Buddhist texts. Figurative expressions play a central role in conveying abstract philosophical and doctrinal content in both literary traditions. The task evaluates language models' ability to identify such expressions, both when explicitly marked (e.g., by lexical cues) and when implicit or structurally embedded. Each figurative expression is labeled as either a simile (`SIM`) or a metaphor (`MET`). A sample may contain multiple annotated spans.

Simile and Metaphor Detection Sanskrit (SMDS). Detection of *upamā* (similes) and *rūpaka* (metaphors) in classical and philosophical Sanskrit prose and poetry. Similes are often explicitly marked by words such as *iva*, *yathā*, or *sadṛśa*, while metaphors usually appear as compounds or identity substitutions. A total of 144 similes and 142 metaphors are annotated across 182 texts, and 228 texts contain no annotated spans.

Simile Detection Tibetan (SDT). Detection of *dpe rgyan* (similes) in Tibetan, focusing on their role in both indigenous and translated Buddhist literature. Similes are typically signaled by expressions like *'dra ba*, *dang 'dra*, *bzhin*, *ji ltar*, *de ltar*, or *lta bu*. These constructions serve as key vehicles for analogy and conceptual clarification in Tibetan exegesis. The dataset includes 404 annotated similes across 179 texts, with 149 texts containing no annotated spans.

Quotation Detection (QUD). This detection task targets the identification of explicit citations in texts and the extraction of the authors or titles mentioned. Citations are defined here as direct speech attributed to another text, excluding silent borrowings, stock phrases or maxims, reported speech, dialogue, and root texts embedded within commentaries. The task evaluates a model's ability to recognize the discourse structure of citations, including lead-in and concluding phrases, and to associate bibliographic metadata when available.

Quotation Detection Sanskrit (QUDS). Detection of explicit quotations from other works in Sanskrit, typically introduced by citation markers and attributed to an author or text name. We annotate three span types: (1) `QUOTE`, the quoted passage itself, reproducing content from another text; (2) `TITLE`, the name of the cited work when explicitly mentioned; and (3) `AUTHOR`, the name of the cited author when explicitly mentioned. A sample may contain multiple annotated spans, with identifiers linking each quote to its corresponding title or author. Overall, 521 quotations are annotated across 351 texts, with 210 title spans and 124 author spans. Forty-nine texts contain no annotated spans.

Quotation Detection Tibetan (QUDT). Detection of quotations in Tibetan, where citations are often marked by formulaic phrases attributing a quote to a particular author or work, especially in scholastic or polemical writing. Evaluation here focuses only on detecting the `QUOTE` label, defined consistently with QUDS above. Additional annotation labels are listed in Appendix F. The dataset includes 335 annotated quotations across 162 texts, while 244 texts contain no annotated spans.

Root-Text and Commentary Matching (RCM). This task evaluates a model's ability to determine whether a given commentary passage comments on a specific verse or excerpt from a root-text. Such capabilities are particularly significant in both Sanskrit (RCMS) and Tibetan (RCMT) literary traditions, where extensive commentarial writing forms the backbone of textual scholarship and interpretation. Success in this task reflects a model's potential for sophisticated corpus-level reasoning—automatically identifying commentarial relationships and aligning related texts across large collections. Each sample consists of a pair of passages: the first is a root-text segment, and the second is a candidate commentary. Labels are `TRUE` and `FALSE`, indicating whether the two passages form a matching pair.

Root-Text and Commentary Detection Sanskrit (RCDS). A related detection variant, defined only for Sanskrit, requires the model to identify the precise boundaries (span) of a commentarial

passage on a given root-verse or excerpt within a larger text passage. Each sample again consists of a pair of passages—the root-text and a passage potentially containing a commentary segment that must be localized within it.

Verse vs. Prose Classification (VPC). This classification task (VPCS for Sanskrit and VPCT for Tibetan) distinguishes verse from prose at the document level, determining whether an entire text is predominantly composed in verse or prose. The task supports structural understanding of complex works and enables downstream applications such as *root–commentary separation* (see Task RCC), since root texts are typically written in verse and their commentaries in prose. By recognizing formal rather than semantic features, models can help automatically isolate root verses for further analysis.

Verse vs. Prose Classification Tibetan (VPCT). In Tibetan, verse passages (*tshigs bcad*) follow distinctive metrical patterns: (1) lines usually contain an odd number of syllables—most often seven; (2) grammatical particles tend to occur on even-numbered syllables (e.g., 2nd, 4th, 6th), producing a rhythmic cadence; and (3) each verse line ends with a *double shad* (‖), marking completion. Texts often contain both verse and prose (*tshig lhug*); therefore, this task serves as a first step toward segmenting mixed works based on rhythmic and structural cues.

Metre Classification Sanskrit (MCS). This classification task evaluates a model's ability to identify the metrical pattern of Sanskrit verse. The dataset includes verses labeled with one of ten common metres: *anuṣṭubh-pathyā*, *anuṣṭubh-vipulā*, the *upajāti*-family (*indravajrā*, *upendravajrā*, *upajāti*), *śārdūlavikrīḍita*, *vasantatilakā*, *drutavilambita*, *sragdharā*, *śālinī*, *mandākrāntā*, and *āryā*. Beyond metre recognition, this task supports quotation detection and authorship attribution by capturing metrical preferences characteristic of specific authors or periods. Significant challenges include: (1) syllable segmentation and weight detection (*guru* vs. *laghu*); (2) the large variety of metre types and sub-variants; and (3) the presence of partial or embedded verses within prose. Applications include improving e-text quality, aiding philologists in metre identification, and reconstructing root verses from commentaries.

Allochthonous vs. Autochthonous Classification Tibetan (AACT). This classification task distinguishes between Tibetan *allochthonous* (`ALLO`) works—texts translated primarily from Sanskrit—and *autochthonous* (`AUTO`) works—indigenous Tibetan compositions. The task contributes to understanding the dynamics of cultural and linguistic transfer in Buddhist textual production. It also serves as a foundation for other analyses, such as authorship attribution and translation studies, including the classification of works into *Ancient Translations* (mainly mid-8th to mid-9th century) and later translation strata.

Scriptures vs. Non-Scriptures Classification Tibetan (SCCT). This classification task divides Tibetan canonical literature (allochthonous works) into two categories: *scriptures* (`SCR`) and *non-scriptures* (`NSCR`). `SCR` includes texts believed to be the Word of the Buddha or another Awakened Being—such as *sūtra* and *tantra*—while `NSCR` includes works ascribed to Buddhist masters and scholars, covering a broad range of genres such as commentaries, treatises, and manuals, including secular topics like medicine and grammar. Using large language models, this task explores how scriptures evolved linguistically and stylistically, aiming to uncover recurring features associated with canonical authority and acceptance.

Thematic Classification Tibetan (THCT). This task classifies Tibetan canonical literature (allochthonous) by genre and theme, following the subdivisions of the Tibetan Canon. It supports content-based analysis by testing whether thematic coherence can serve as an indicator of religio-philosophical affiliation and intellectual context.

There are **13** distinct Buddhist-specific themes: Vinaya, Tantra, Dhāraṇī, Epistles, Sūtra, Non-tantric eulogies, Tantric treatises, Madhyamaka treatises, Abhidharma treatises, Vinaya treatises, Sanskrit treatises, medical treatises, and treatises on arts and crafts.

Real world relevancy. Beyond their linguistic and cultural value, these tasks have clear real-world relevance for research and digital scholarship. **SMDS/SDT** assist scholars by reducing manual annotation effort and by testing models' ability to capture subtle semantic relations central to reasoning and literary expression. **QUDS/QUDT** facilitate tracing intertextuality and intellectual lineages across vast corpora. **RCMS/RCMT/RCDS** can semi-automate one of the most labor-intensive processes in Buddhist studies, enabling large-scale comparative analyses. **VPCS/VPCT/MCS/AACT** (structural classifica-

tion) support segmentation, digital edition preparation, and the recovery of embedded verse, while revealing stylistic and metrical patterns tied to authors and periods. **AACT/SCCT/THCT** (thematic and source classification) distinguish indigenous from translated works, canonical from non-canonical texts, and classify Buddhist themes—informing research on canon formation, translation practices, and cultural diffusion. Overall, these tasks target key bottlenecks in digital humanities, benefiting philologists, historians of religion and philosophy, and digital archivists, while showcasing how LLMs can enable fine-grained analysis of low-resource, culturally significant corpora.

3.2 Data Collection

All our data is collected from public domain sources and/or under permissive licenses; details in Table 6 and Table 7 (Appendix C) and in Table 11 (Appendix J). Our Tibetan data derives from e-texts like ACIP[5] and repositories such as Esukhia[6], as well as transcripts from modern printed editions of Buddhist works. Our data initially consisted of several different transliteration systems, EWTS,[7] ACIP,[8] which in turn use different styles for marking punctuation, pagination, and catalog numbers (for example: `@66b`, or `[66b.1]` for pagination and `{D120}` for catalog numbers). The data was slightly preprocessed to remove some noise, such as the removal of various numerical references, and then converted into Unicode for consistency.[9] For Sanskrit, we collect texts from transcripts of printed editions—primarily but not exclusively scholarly Buddhist works—produced either by human transcribers or through OCR corrected by human reviewers. A large proportion of the data is fetched from GRETIL.[10] Data is provided in IAST[11] with a mix of punctuation styles: sometimes using the more modern conventions of periods, commas, and question marks, and sometimes with the more traditional *daṇḍa* (represented by the vertical bar, |). Sandhi—phonological transformations that occur at word boundaries—is generally applied in the texts, but not always, which is representative of how Sanskrit is generally written—editors generally apply or do not apply sandhi according to what feels natural to them. Spaces between words often follow the standard IAST style; however, some texts also follow the spacing conventions typical of printed Devanagari texts.

3.3 Data Extraction and Annotation

LLMs have a strict context length, which is the amount of tokens they can process at once (Vaswani et al., 2023). Due to their limited ability to encode Sanskrit and Tibetan correctly and efficiently (Chaudhari et al., 2024; Nehrdich et al., 2024; Meelen et al., 2024; Zhuang and Sun, 2025; Yang et al., 2022), we set a 2,000-character maximum on all of our samples, across all tasks. Texts are split based on this length limitation.

To ensure reliability and consistency during annotation, each task was annotated by at least one expert scholar. The annotator's task is to validate the data and to assign a label or mark spans, etc., depending on the task. Another, different scholar was tasked to validate the annotated results by selecting at least 20-30% of the samples in each task. We encountered occasional issues, which typically stemmed from input errors in the underlying data or from ambiguities in the initial task definitions. Input errors were straightforward to correct, while annotation guidelines were continuously refined through discussion with other experts as new edge cases emerged. Specifically, we iterated to refine the definition and guidelines during the annotation of the first 10-20% of the data, and since then, it has remained fixed. For complex or subjective cases, disagreements were resolved collectively; when no clear resolution was possible, the samples were excluded to preserve label quality. Overall, validations reported very few systematic problems, and the iterative process of refining guidelines and adjudicating disagreements gave us confidence in the reliability of the final gold labels.

For the Tibetan tasks AACT, SCCT, and THCT, we collect texts that are collectively known to be of a given label. For example, the Derge Kangyur is known to be *allochthonous* and *scripture*, for the AACT and SCCT tasks in Tibetan, respectively. We then apply a heuristically designed cutting mechanism, based on spaces and punctuation (like the *shad* །) to cut chunks of several sentences

[5] Data fetched from https://asianlegacylibrary.org/library/

[6] https://github.com/Esukhia/derge-kangyur.

[7] The Extended Wylie Transliteration Scheme is a reversible, machine-readable variant of Wylie, which converts Tibetan script into Latin characters using only ASCII.

[8] The ACIP (Asian Classics Input Project) Tibetan transliteration system is a specific scheme developed for inputting Tibetan texts into a computer.

[9] https://github.com/OpenPecha/pyewts

[10] https://textgridrep.org/project/TGPR-2ba9cb1b-9602-202d-71ce-67e63a29de55

[11] The International Alphabet of Sanskrit Transliteration (IAST) is a transliteration scheme that allows the lossless romanization of Indic scripts as employed by Sanskrit and related Indic languages.

at a time. To allow for diversity in the lengths of the output texts, the chunks contain a varying number of sentences, randomly selected during the cutting process. The final label is set based on the majority of the given text.

For SMDS and SDT, some potentially relevant texts were identified as containing or not containing similes and/or metaphors and were annotated using the LS platform. The annotator is asked to mark the minimal span containing the metaphor or simile.

In QUDT, the annotation team first identified relevant source texts and then cut the longer texts in the same way as described above (AACT). The texts are validated and discarded if they don't meet the quality standard during the annotation process, using the LS platform. QUDS is done similarly, but manually cut.

In the VPCS task, a scholar selects candidate texts known to be either verse or prose, validating or discarding them based on criteria such as excessive length. The verses are then manually assigned to the appropriate metre type based on metrical patterns. Verses that fall into metre types outside the scope of our study are discarded from both the MCS and VPCS datasets. In some cases, preliminary metre predictions are obtained using existing tools[12] and subsequently corrected by hand.

To prepare pairs for the RCMS and RCMT, either a complete root text or a substantial section (e.g., a chapter) was selected first. In Sanskrit, all root texts are composed in verse, whereas in Tibetan, some are also in prose. In both cases, the texts are segmented into individual verse/prose. One or more traditional commentaries on the root text were then chosen to showcase different styles of writing. The experts read the commentary and manually identified and segmented the portions corresponding to each root verse. During annotation, we mostly avoided trivial pairs in which the commentary quotes the root text or a significant portion of it. To create the same amount of negative pairs, we simply randomly couple root texts and commentaries that are not aligned.

For the Sanskrit detection variant of the RCDS task, to create a more realistic and challenging setting, a random number of surrounding commentary sentences (ranging from 0 to roughly 10) were inserted before and/or after the target span, drawn from the natural sequential flow of the same text. The pairs—including the negative ones—are identical to those used in the RCMS task, and the validation process follows the same procedure.

4 Experiments

4.1 Models

We set a zero-shot baseline following recent work on LLMs for non-generative information extraction tasks (Liu et al., 2023; Sun et al., 2023; Smădu et al., 2024) and evaluate diverse models; checkpoint details and licenses are in Table 10 (Appendix I).

4.1.1 Generative LLMs

We prompt generative LLMs with task instructions and inputs, asking them to respond in a structured JSON format. In classification tasks, the format includes a *label* field containing the label as a string. In detection tasks, the output includes a *prediction* key containing a list of dictionaries in the format `{"LABEL": LABEL "SPAN": SPAN}`, where `SPAN` denotes the extracted substring. Prompting is implemented via the LangChain[13] framework, utilizing the respective model provider (see Table 11). All models and tasks are evaluated using a straightforward zero-shot approach, setting the temperature to 0.3 for stability. Each model is initialized with a system prompt that defines the task and specifies the required output format; an example prompt is provided in Appendix H, and all prompts are released in the paper's repository.[14] Prompts are refined through an iterative trial-and-error process, aiming to maximize overall performance. Given the high cost and manual effort involved in prompt engineering, we use the Gemini 2.0 Flash model during development and transfer the optimized prompt to all other models. Before running the prompt with all models, we refined it with models that exhibited a high rate of hallucinations. All prompts are given in English, and relevant phrases are given in the source language. If a model fails to produce a response in the expected format, we treat the output as a hallucination and replace it with a negative label or an empty prediction list. We allow some flexibility in the output format, as preliminary results indicated that certain models struggled with rigid formatting requirements, despite producing correct predictions.

[12] https://sanskritmetres.appspot.com/identify
https://www.skrutable.info/

[13] https://www.langchain.com/

[14] https://github.com/Intellexus-DSI/DharmaBench

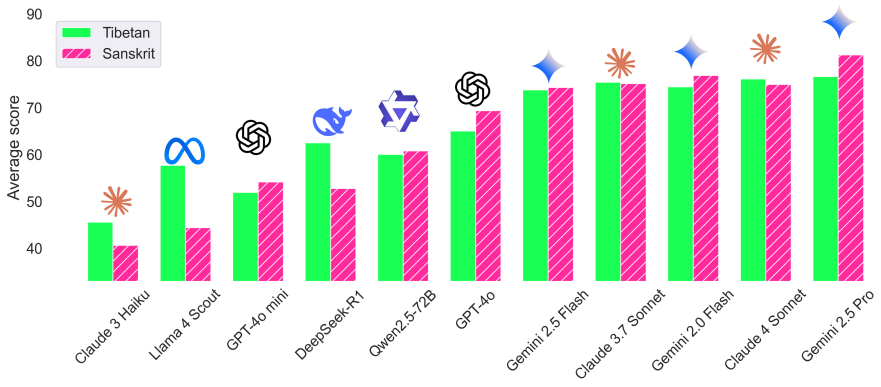

Figure 1: Average performance (F1-score) across all tasks, per language, for each model.

In addition to Gemini 2.0 Flash, we evaluate Gemini 2.5 Flash and Pro (Comanici et al., 2025). We include OpenAI's GPT-4o[15] as a strong closed-weight model, along with its smaller and more cost-efficient variant, GPT-4o mini. Similarly, we evaluate Anthropic's Claude 3.7 Sonnet, Claude 4 Sonnet, and Claude 3 Haiku.[16] For comparison with open-weight models, we evaluate Llama 4 Scout and Qwen2.5-72B (Qwen et al., 2025), and DeepSeek-R1, which is designed to enhance reasoning capabilities (DeepSeek-AI et al., 2025). A few domain-specific models exist and demonstrate language understanding and domain knowledge when prompted with open-ended questions. However, our preliminary experiments indicate that in zero-shot *formatted response* settings, they can struggle to produce consistent outputs, partly due to formatting-related errors. Notable examples include T-LLaMA (Lv, 2024), Gemma-2-mitra (Nehrdich et al., 2025), and ByT5-Sanskrit (Nehrdich et al., 2024). We intend to explore fine-tuning these models for such tasks in future work.

Several recently released models were unavailable at the time of writing because their resources were not yet available. For instance, Huang et al. (2025) introduced Sun-Shine, an LLM trained on Classical Tibetan, while SansGPT (Chaudhari et al., 2024) serves as another example.

4.1.2 Encoder-based LMs

To support a wider comparative analysis, we also test encoder-based language models, for tasks where a training set is available. Fine-tuning for RCDS and QUDT, both of which are detection tasks, is omitted from the experiments. This is due to the models' fixed 512-token context window, while the average text lengths are 884.67 and 1009.36 characters, respectively. Typically, this issue can be addressed by splitting the text into chunks; however, in our case, many annotated spans exceed 512 tokens on their own. On average, `QUOT` spans in QUDT are 420.77 characters long, while `COMM` spans in RCDS average 542.51. The problem is exacerbated in non-language-specific models, where tokenization can result in more tokens than characters. In the classification tasks, we truncate the input text to 512 tokens.

The hyperparameters for fine-tuning are given in Table 9 (Appendix E). **Multilingual models** include mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), both of which are multilingual encoder models that have proven to be strong baselines across a variety of cross-lingual tasks (Ruder et al., 2021). In **Tibetan** we evaluate Tibetan-RoBERTa,[17] and the Chinese minority languages model CINO (Yang et al., 2022). For **Sanskrit** we evaluate `bert-base-buddhist-sanskrit` (Lugli et al., 2022), which focuses on Buddhist Sanskrit, and IndicBERTv2 (Doddapaneni et al., 2023).

4.2 Evaluation

For all classification tasks, we report the micro-averaged F1. For the span-based detection tasks, we use the MUC-5 evaluation metrics[18] (Chinchor and Sundheim, 1993), focusing on *mode: type*, which permits partial overlap between predicted and gold spans while requiring matching span tags.

We estimate about 180,000 API calls across all experiments, costing around $660. To balance robustness and cost, we run the efficient models—like Gemini 2.0 Flash, GPT-4o mini, Llama 4

[15] `https://openai.com/`
[16] `https://www.anthropic.com/`
[17] `https://huggingface.co/sangjeedondrub/tibetan-roberta-base`
[18] `https://github.com/cyk1337/eval4ner`

Scout—with three seeds and report with mean and standard deviation. The more expensive ones, like GPT-4o, are tested with one seed only.

5 Results and Discussion

Figure 1 shows the average performance of each model across all tasks by language—referred to as the DharmaBench Score. For the full results, including the encoder-based models, see Tables 4 and 5 (both in Appendix B).

Overall, larger and more recent models perform better, as expected. We observe a clear advantage for the Gemini family, with all Gemini variants achieving strong results across both languages. Notably, Gemini 2.5 Pro demonstrates the best overall performance, reaching an F1-score of ~83 in Sanskrit and ~76 in Tibetan.

Beyond general trends, several factors could explain the observed performance patterns. **Model-wise**, larger and newer models benefit from higher capacity, more diverse pretraining (including low-resource data), and improved architectures and training methods. Differences in (non-public) pretraining corpora also play a role. For instance, DeepSeek, despite its strong reasoning design, underperforms on several tasks, likely because its reasoning chains stall before producing concrete answers (e.g., in MCS). **Task-wise**, detection tasks are generally more complex than classification ones. Claude models, for example, struggle with detection but perform well in classification. Tasks requiring domain knowledge or nuanced interpretation—such as SMDS and QUDS—show higher variability. Conversely, tasks like RCMS, RCMT, and VPCS can sometimes be solved using formal cues (e.g., repeated phrasing or verse formatting), resulting in higher scores. Multi-class setups (e.g., THCT with 13 themes) and subtle origin-based distinctions (AACT) also add difficulty. **Methodologically**, stronger prompting—such as few-shot, self-consistency, or chain-of-thought setups—could further enhance results. Beyond prompting, domain-specific models could potentially show better results. Preliminary results are promising, and we plan to present them in future work.

As outlined in 4.1.2, we evaluate encoder-based language models on six Tibetan tasks and one Sanskrit task, constrained by the limited training data available for the others. Overall, language-specific models generally outperform broad multilingual ones, though both show considerable variation across tasks. For certain tasks, such as RCMT, all models perform poorly, likely due to input truncation that removes critical context. The small size of the training data is another contributing factor, limiting the models' capacity to generalize effectively. A few tasks reach very high scores (above 95% micro-F1), which we now state explicitly. Interestingly, they remain challenging for human annotators, often requiring expert philological knowledge and discussion. In some cases—such as AACT—humans can only classify correctly by relying on prior knowledge about the source or provenance of a text, rather than on the textual surface itself. When restricted to our 2,000-character evaluation chunks, such distinctions are frequently opaque. Moreover, no single model performs well across all tasks: models that excel in one often underperform in another, suggesting that each task isolates different capabilities rather than measuring trivial features.

Interestingly, Sanskrit tasks favor large closed-weight models (e.g., Gemini, GPT), while Tibetan tasks show comparatively stronger performance by smaller or open-weight models. This asymmetry underscores current gaps in multilingual representation and highlights the need for continued research into smaller, accessible models for philological applications.

6 Conclusion

In this work, we assess the capabilities of modern language models on NLP tasks involving ancient textual material, consisting of Buddhist texts written in Sanskrit and Classical Tibetan. To enable this evaluation, we introduced **DharmaBench**, a benchmark comprising diverse, domain-specific tasks drawn from canonical and commentarial sources. We evaluated a range of state-of-the-art models, including instruction-tuned, multilingual, and general-purpose LLMs. As expected, larger and more recent models generally perform better, with the Gemini family leading overall and Gemini 2.5 Pro achieving the highest scores across both languages. This suggests strong adaptability for tasks involving complex, low-resource classical texts. By providing a dedicated evaluation suite for Sanskrit and Tibetan, DharmaBench fills an important gap and supports the NLP and digital humanities communities in developing culturally and linguistically inclusive tools for processing historical texts.

Limitations

Our study shed light on LLMs' capabilities for understanding Sanskrit and Tibetan. However, it also has several limitations. First, the tasks we curate in **DharmaBench** have a relatively small test set, a couple of hundred samples each. Some have a relatively small training split. Our novel benchmark focuses on text understanding capabilities and contains no generative tasks. All of our data is gold-standard and is labeled by human experts. However, it is noteworthy that in some datasets, not all the samples are reviewed by a second annotator. Within the scope of this paper, we do not report encoder-based results for all tasks, as outlined in Section 4.1.2. For the classification tasks, texts are truncated to 512 tokens, and more advanced approaches are left for future work.

In future work, we plan to expand the datasets and provide training splits to other datasets as well. Due to budget constraints, we restricted our experiments to a representative subset of configurations: we used three random seeds for the more cost-efficient models and only a single run for the more expensive ones.

A potential limitation when working with LLMs is risk of contamination. Since all data seem to be retrieved from public sources, it could be that some models have seen at least the raw texts during training. We thus included several LLMs, trained by different groups and with different cut-off dataes, as a partial coping strategy.

Another limitation is that we manually tuned our prompts on a subset of models and reused the same configurations across all models. In other words, we did not perform model-specific prompt optimization, which could potentially improve performance but would be prohibitively cost-intensive. While this approach allows for consistent comparisons, it may not yield optimal performance for each model. Exploring automatic or model-specific prompt tuning could further improve results.

These limitations point to several promising directions for future work, including larger datasets and new, potentially generative, tasks, as well as investigating prompt optimization strategies and model-specific tuning to enhance performance and generalization.

Ethics Statement

We use publicly available data sources, models, and other artifacts under their licenses and their intended usage; details are listed in Table 10 (Appendix I) and Table 11 (Appendix J). No personally identifiable information or offensive data is processed and annotators make sure no information as such exist in the released texts. Our work is intended for research purposes only. We see no potential risk in our work.

Acknowledgments

This study is supported in part by the European Research Council (Intellexus, Project No. 101118558). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authorities can be held responsible for them.

References

Munef Abdullah Ahmed, Raed Abdulkareem Hasan, Mostafa Abdulghafoor Mohammed, Peter Mwangi, and Tirus Muya. 2025. Classification arabic language (classical arabic poetry, al-hur arabic poetry and prose) using machine learning. *EDRAAK*, 2025:94–104.

Maha Alrabiah, Abdulmalik Alsalman, and Eric Atwell. 2013. The design and construction of the 50 million words ksucca king saud university corpus of classical arabic.

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Rhugved Pankaj Chaudhari, Bhakti Jadhav, Pushpak Bhattacharyya, and Malhar Kulkarni. 2024. SansGPT: Advancing generative pre-training in Sanskrit. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 432–441, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Proceedings of the 5th Conference on Message Understanding*, MUC5 '93, page 69–78, USA. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3284 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Zhengcuo Dan and Yuan Sun. 2024. TibetanQA2.0: Dataset with unanswerable questions for tibetan machine reading comprehension. *Data Intelligence*, 6.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023. MiLMo:minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 329–334.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Csaba Dezső, Dominic Goodall, and Harunaga Isaacson. 2024. *The Lineage of the Raghus*, page 93. Number 38 in Murty Classical Library of India. Harvard University Press, Cambridge, Massachusetts and London.

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.

Kais Dukes and Nizar Habash. 2010. Morphological annotation of Quranic Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Hanne Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Haugen, and Marius Jøhndal. 2018. The proiel treebank family: a standard for early attestations of indo-european languages. *Language Resources and Evaluation*, 52.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. MMTEB: Massive multilingual text embedding benchmark. *Preprint*, arXiv:2502.13595.

Fan Gao, Cheng Huang, Nyima Tashi, Xiangxiang Wang, Thupten Tsering, Ban Ma-bao, Renzeg Duojie, Gadeng Luosang, Rinchen Dongrub, Dorje Tashi, Hao Wang Xiao Feng, and Yongbin Yu. 2025. TLUE: A Tibetan language understanding evaluation benchmark. *Preprint*, arXiv:2503.12051.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Cheng Huang, Fan Gao, Yutong Liu, Nyima Tashi, Xiangxiang Wang, Thupten Tsering, Ban Ma-bao, Renzeg Duojie, Gadeng Luosang, Rinchen Dongrub, Dorje Tashi, Xiao Feng, Hao Wang, and Yongbin Yu. 2025. Sun-Shine: A foundation large language model for Tibetan culture and heritage. *Preprint*, arXiv:2503.18288.

Bhakti Jadhav, Himanshu Dutta, Shruti Kanitkar, Malhar Kulkarni, and Pushpak Bhattacharyya. 2025. An introduction to computational identification and classification of upamā alaṅkāra. In *Computational Sanskrit and Digital Humanities - World Sanskrit Conference 2025*, pages 1–14, Kathmandu, Nepal. Association for Computational Lingustics.

Manoj Balaji Jagadeeshan, Prince Raj, and Pawan Goyal. 2025. Anveshana: A new benchmark dataset for cross-lingual information retrieval on English queries and Sanskrit documents. *Preprint*, arXiv:2505.19494.

Zhang Jin, Zhang Ziyue, Yeshi Lobsang, Tashi Dorje, Wang Xiangshi, Cai Yuqing, Yu Yongbin, Wang Xiangxiang, Tashi Nyima, and Luosang Gadeng. 2024. Tibetan medical named entity recognition based on syllable-word-sentence embedding transformer. ResearchGate preprint.

S. D. Joshi and J. A. F. Roodbergen. 1986. *Patañjali's Vyākaraṇa-mahābhāṣya: Paspaśāhnika*, pages 27–29. University of Poona, Poona.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 6155–6218. Association for Computational Linguistics.

Chunwei Kong, Xueqiang Lv, Le Zhang, Haixing Zhao, Zangtai Cai, and Yuzhong Chen. 2024. RPEPL: Tibetan sentiment analysis based on relative position encoding and prompt learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(12).

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

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, Marseille, France. European Language Resources Association.

Pu C. Duo-L. Li Y. Zhou Q. Shen J Lv, H. 2024. T-LLaMA: a Tibetan large language model based on LLaMA2.

Priyanka Mandikal. 2024. Ancient wisdom, modern tools: Exploring retrieval-augmented LLMs for Ancient Indian philosophy. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 224–250, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Barbara Mcgillivray, Marco Passarotti, and Paolo Ruffolo. 2009. The index thomisticus treebank project: Annotation, parsing and valency lexicon. *TAL*, 50:103–127.

Marieke Meelen, Sebastian Nehrdich, and Kurt Keutzer. 2024. Breakthroughs in Tibetan NLP digital humanities.

Sebastian Nehrdich. 2022. SansTib, a Sanskrit - Tibetan parallel corpus and bilingual sentence embedding model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6728–6734, Marseille, France. European Language Resources Association.

Sebastian Nehrdich, Avery Chen, Marcus Bingenheimer, Lu Huang, Rouying Tang, Xiang Wei, Leijie Zhu, and Kurt Keutzer. 2025. MITRA-zh-eval: Using a buddhist Chinese language evaluation dataset to assess machine translation and evaluation metrics. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 129–137, Albuquerque, USA. Association for Computational Linguistics.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751, Miami, Florida, USA. Association for Computational Linguistics.

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. 2024. Large language models are echo chambers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10117–10123, Torino, Italia. ELRA and ICCL.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Leiyu Pan, Bojian Xiong, Lei Yang, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, Jianxiang Peng, Juesi Xiao, Tianyu Dong, Zhuowen Han, Zhuo Chen, Sangjee Dondrub, Caizang Tai, Haixing Zhao, Huaque Cairang, and 21 others. 2025. Banzhida: Advancing large language models for Tibetan with curated data and continual pre-training. *Preprint*, arXiv:2507.09205.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin

Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abdul-Baquee Sharaf and Eric Atwell. 2012. QurAna: Corpus of the Quran annotated with pronominal anaphora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 130–137, Istanbul, Turkey. European Language Resources Association (ELRA).

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating large language models for complex word identification in multilingual and multidomain setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16764–16800, Miami, Florida, USA. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Yuan Sun, Sisi Liu, Chaofan Chen, Zhengcuo Dan, and Xiaobing Zhao. 2021. (construction of high-quality Tibetan dataset for machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218, Huhhot, China. Chinese Information Processing Society of China.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms. *Preprint*, arXiv:2205.05131.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Marja Vierros. 2018. *Linguistic Annotation of the Digital Papyrological Corpus: Sematia: Case Studies on the Digital Edition of Ancient Greek Papyri*, pages 105–118.

Jeff Wallman, Zach Rowinski, Ngawang Trinley, Chris Tomlinson, and Kurt Keutzer. 2017. Collection of Tibetan etexts compiled by the Buddhist digital resource center.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Preprint*, arXiv:1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jiangyan Zhang, Deji Kazhuo, Luosang Gadeng, Nyima Trashi, and Nuo Qun. 2022. Research and application of Tibetan pre-training language model based on BERT. In *Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics*, ICCIR '22, page 519–524, New York, NY, USA. Association for Computing Machinery.

Jing Zhang, Hang Ren, Jin Yang, Nuo Qun, and Shengqiao Ni. 2024. Sentiment analysis of Tibetan short texts based on graph neural network with keyphrases integration. pages 474–480.

Wenhao Zhuang and Yuan Sun. 2025. CUTE: A multilingual dataset for enhancing cross-lingual knowledge transfer in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10037–10046, Abu Dhabi, UAE. Association for Computational Linguistics.

A Examples

See Table 2 for examples of input and output of each task in Sanskrit and Table 3 for Tibetan.

B Full Results

Table 5 summarizes the results for all Tibetan tasks, and Table 4 shows the Sanskrit results. The first row represents the chance level: for classification tasks, it corresponds to random guessing, while for detection tasks, it is based on an empty prediction with no spans detected. Interestingly, DeepSeek-R1 got a solid 0 in the task of MCS. A qualitative analysis revealed that in most API calls, the model wasn't able to finish the reasoning part, not reaching the final answer stage.

C Datasets Sources Distribution

In Table 6, we detail the source from which we collect and extract our texts for the datasets in Sanskrit. In Table 7 for Tibetan, where for tasks VPCT, AACT, SCCT and THCT we provide the sources of the initial texts that were later on cut into chunks as detailed in Section 3.3.

D Annotators Demographics

All annotators are expert scholars in Sanskrit and/or Tibetan, either current PhD candidates or holders of doctoral degrees, and are co-authors of this paper. Consequently, no monetary compensation was provided, as the annotation work was conducted as part of their academic research. In total, 11 annotators participated in the project (5 male, 6 female).

E Encoder-based Models Fine-tuning Hyperparameters

See Table 9 for the list of hyperparameters we use during fine-tuning, found after a greedy optimization process. The rest of the hyperparameters are the default values of the `Trainer`[19] class from the `transformers` package. We conduct our fine-tuning experiments on a NVIDIA GeForce RTX 3090 machine with 24GB of memory.

Overall, all runs and tests took approximately 30 hours. The exact code and package versions required are published in the project's repository.

F Quotation Detection Tibetan (QUDT) - Full Dataset

See the full released dataset's label distribution and statistics in Table 8.

G Label Studio Example

We provide an example of the Label Studio annotation platform we use for some of the annotations in Figure 2.

H Prompt Example

We provide an example prompt, from the Similes Detection task in Tibetan, see Figure 3.

I Model Checkpoints

In Table 10, we present the checkpoints used in this work and the models' sizes and license.

J Artifacts

We detail artifacts we use and their respective usage and licenses in Table 11. Working with LangChain provides a modular and reproducible framework for LLM evaluation, particularly when working with multiple providers, as it wraps their APIs in a unified layer.

K AI assistants

We used AI assistants (e.g., ChatGPT) to support code formatting, phrasing suggestions, and LaTeX styling during writing. All outputs were reviewed and edited by the authors. No content was directly generated or used without human verification.

[19] https://huggingface.co/docs/transformers/en/main_classes/trainer

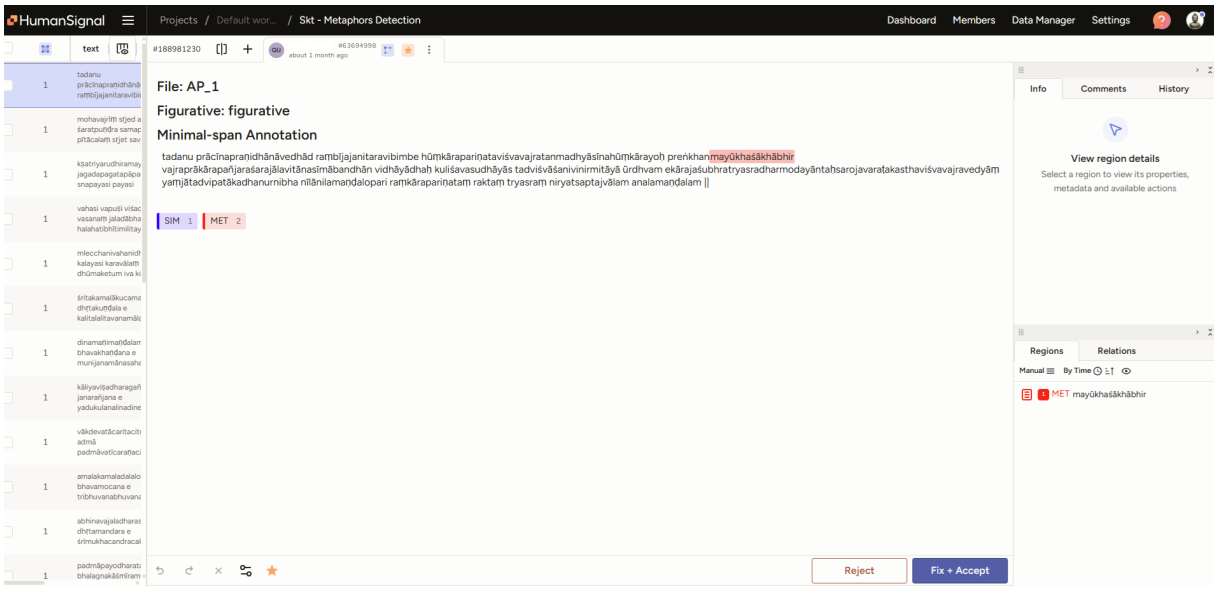

Figure 2: Label Studio annotation example.

Prompt for SDT

```
You are a computational linguist and philologist specializing in
identifying similes expressions in Classical Tibetan texts.
Your task is to analyze a given sentence or verse written in Tibetan and
extract simile expressions it contains.
Definitions:
- A simile: an explicit comparison between two entities, typically marked
by comparison words such as "ཚུ་བཏུ", "བཞིན", "ཚུ", "འདྲ", etc.

Annotation Guideline:
- Identify all spans containing a simile.
- Label each identified span as "SIM" (simile).
- Do not annotate literal or descriptive phrases that do not rely on
identification or comparison.
- Mark the absolute minimal span that contains the simile, even if it is
part of a larger phrase.

Return your output as a JSON with "prediction" key. The value is a list
of dictionaries, each with:
- LABEL: "SIM"
- SPAN: the exact text span that contains the simile (minimal required
span)
Example for an item: "LABEL": "SIM", "SPAN": "དེ་བཞིན་དུ"

If no simile is found, return an empty list under the "prediction" key.
Only respond with the JSON output, do not include any additional text or
explanations.

Text: ཤེའི་ཐུགས་གང་གི་ཚུ་གཏེར་འདྲ
```

Figure 3: Prompt example for Similes Detection task in Tibetan.

Task	Input	Output
SMDS	uddharaty andhatamasād viśvam ānandavarṣiṇī | paripūrṇā jayaty ekā devī **ciccandracandrikā ‖** (The Goddess is supreme. She lifts the world from profound darkness and showers it with bliss. She is the one and only, full and complete, **her consciousness the moonlight of the moon**.)	[['MET', 'ciccandracandrikā']]
QUDS	ayam arthaḥ—rāgādir evāntaraṃ viṣam. tad uktaṃ **bhagavatā—rāgo dveṣaś ca mohaś ca ete loke trayo viṣāḥ ‖** iti. (The meaning is as follows: precisely passion and the rest are the inner poisons. That is taught by the **Blessed One**: **"Passion, hatred, ignorance—these are the three poisons in this world."**)	[['AUTHOR', 'bhagavatā'], ['QUOTE', 'rāgo dveṣaś ca mohaś ca ete loke trayo viṣāḥ ‖']]
RCMS	**Root-text:** baddhaś cec cittamātaṅgaḥ smṛtirajjvā samantataḥ | (If the elephant that is the mind is tied all around with the rope of mindfulness …) **Commentary:** tasyāyattīkaraṇe guṇam āha. yadi baddhaḥ kathaṃcid bhavet. smṛtir vakṣyamāṇalakṣaṇā. saiva rajjur bandhanopāyatvāt. samantataḥ sarvathā, asatpakṣe pracāranirodhāt. (He states the virtues of bringing [the mind] under control. If it were bound in some way. The characteristics of mindfulness will be explained later. Just that [mindfulness] is a rope insofar as it serves as the means of binding. [It ties] all around, or in every way, because it prevents wandering towards the unwholesome,…)	TRUE
RCDS	**Root-text:** baddhaś cec cittamātaṅgaḥ smṛtirajjvā samantataḥ | (for translation see RCMS) **Passage:** paraloke avīcyādau yāṃ karoti svacchandatayāvasthitaṃ cittam eva mataṅgaja eva, tathāgatājñāṅkuśena kathaṃcid vaśīkriyamāṇatvāt. **tasyāyattīkaraṇe guṇam āha. yadi baddhaḥ kathaṃcid bhavet. smṛtir vakṣyamāṇalakṣaṇā. saiva rajjur bandhanopāyatvāt. samantataḥ sarvathā, asatpakṣe pracāranirodhāt.** (… which the mind does while residing in the next world in Avīci hell and so forth, acting according to its own will—this mind being itself an elephant insofar as it is somehow brought under control by the goad of the Tathāgata's command. **He states the virtues of bringing [the mind] under control. If it were … [see RCMS for the remainder of the translation]**)	[['COMM', 'tasyāyattīkaraṇe guṇam āha. yadi baddhaḥ kathaṃcid bhavet. smṛtir vakṣyamāṇalakṣaṇā. saiva rajjur bandhanopāyatvāt. samantataḥ sarvathā, asatpakṣe pracāranirodhāt.']]
VPCS	kāni punaḥ śabdānuśāsanasya prayojanāni? rakṣohāgamalaghvasandehāḥ prayojanam. rakṣārtham vedānām adhyeyam vyākaraṇam. lopāgamavarṇavikārajñaḥ hi samyak vedān paripālayiṣyati. ūhaḥ khalu api. (But what are the uses of instruction in words? *Rakṣā* 'preservation [of the Vedic texts]', *ūha* '[suitable] adaptation [of a *mantra* according to the requirements of a particular ritual]', *āgama* '[complying with a] vedic injunction', *laghu* 'simplicity/economy [in acquiring knowledge of the correct forms of language]', and *asaṃdeha* 'removal of doubt' are the use. One should study grammar for the preservation of the Vedas. Because one who is acquainted with [the techniques of] *lopa* 'deletion', *āgama* 'augment' and *varṇavikāra* 'sound-substitution' will [be able to] preserve the Vedas correctly. Certainly, the [suitable] adaptation [of a *mantra* according to the requirements of a particular ritual is] also [a use of grammar].) (Joshi and Roodbergen, 1986)	PROSE
MCS	tam īśaḥ kāmarūpāṇām atyākhaṇḍalavikramam | bheje bhinnakaṭair nāgair anyān uparurodha yaiḥ ‖ (The same elephants, with temples splitting in their musth, that the lord of Kamarupa had once used to lay siege to others he now made over to Raghu, whose valor exceeded Indra's.) (Dezső et al., 2024)	śālinī

Table 2: Task inputs and expected outputs for Sanskrit tasks. Translation is given in brackets for convenience. In extraction tasks, we highlight the ground truth spans that should be identified by the models.

Task	Input	Output
SDT	འཆགས་དང་འཇིག་བཅས་སེམས་ཅན་བཅས་པ་ཡི། འཇིག་རྟེན་ཁམས་ཀུན་**སྒྱུ་མ་ལྟ་བུར་**མཐོང་། དབང་ལྡན་དེ་ལྟར་ཇི་ལྟར་འདོད་བཞིན་དེ་དག་དང་། རྣམ་བཅས་སྣ་ཚོགས་དག་ཀྱང་ཡང་དག་སྟོན།། (All worldly realms, with formation and disintegration, together with sentient beings, are seen **as an illusion**. The Powerful One displays them with different forms as he pleases.)	[["SIM","སྒྱུ་མ་ལྟ་བུར་"]]]
QUDT	དཀོན་མཆོག་སྤྲིན་ལས་ཇི་སྐད་དུ། **སྨན་པའི་དཔེས་བདག་བསྔོད་པ་ཉེས་པ་མེད་པར་བསྟན་པ་ལྟ་བུའོ།། ཡང་བསོད་ནམས་བསྡུང་བར་འདོད་པས། རྙེད་དང་བཀུར་སྤྱིས་འཇིགས་ཐུ་ཞིང་།། ཁེངས་པ་རྣམ་དུ་སྤང་བར་བྱ།།** (As it says in the Ratnamegha: **"It is like what is shown with the example of a doctor—that self-praise is without fault. Furthermore, those who wish to protect merit should fear gain and honor, and always abandon pride."**)	[["QUOTE","སྨན་པའི་དཔེས་བདག་བསྔོད་པ་ཉེས་པ་མེད་པར་བསྟན་པ་ལྟ་བུའོ།། ཡང་བསོད་ནམས་བསྡུང་བར་འདོད་པས། རྙེད་དང་བཀུར་སྤྱིས་འཇིགས་ཐུ་ཞིང་།། ཁེངས་པ་རྣམ་དུ་སྤང་བར་བྱ།།"]]
RCMT	**Root-text:** སྨྲའོ་པ་ཡོ་ཡིས་ངེས་པར་བཅང་བ་ཡི།། རྡོ་རྗེ་དེ་དག་གང་ཡིན་བཤད་པར་བྱ།། སོར་བསྐུ་བསྐུ་གཉིས་བསྐུ་དྲུག་བཅོ་བརྒྱད་དང་།། མཆོག་གི་ཚད་ནི་སོར་ཡང་ཉེ་ཤུ་བའོ།། (Now I shall explain what those *vajra*s are that a practitioner must surely wield. They are ten, twelve, sixteen, eighteen fingers [in length]; and also twenty fingers, which is the length of the greatest [*vajra*].) **Commentary:** སོར་བསྐུ་ཞེས་བྱ་བ་ལ་སོགས་པས་ནི་རྡོ་རྗེ་རྣམས་ཀྱི་ཚད་བསྟན་ཏོ།། (By "ten fingers" and so forth, the length of the *vajra*s is taught.)	TRUE
VPCT	མཁོན་ཏོ་ཚངས་སྤྱོད་གནས་སྐད་དང་།། དབྱིངས་ཀྱང་ཤེན་ཏུ་མཛེས་སྐད་དུ།། ཁྱོད་ཀྱིས་ཐ་ན་ཤེད་ལགས་ཤིང་།། ཐ་བར་བགྱི་བ་མིན་པའང་ལགས།། ཕྱགས་རྗེས་ཁེ་ཚོམ་ཅན་རྣམས་ལ།། ཁྱོད་ཀྱིས་མཚན་མ་གཞན་ཚུལ་ཐྱེས།། (… because of your pure conduct and because of their extraordinarily beautiful form, although [these body parts] are shameful, Lord, to you [read: *khyod kyi*] they are not shameful. Out of compassion [you have shown] them to doubters as a small fraction of your auspicious marks.)	VERSE
AACT	པའི་དོན་གསུངས་སོ།། ལེའུའི་མཚན་གྱི་མཇུག་བསྡུ་བ་ནི་དེ་བཞིན་གཤེགས་པ་ཐམས་ཅད་ཀྱི་སྙིང་པོ་བསྒྲུབ་བ་ཞེས་པ་དོན་དངོས་སྒྲུབ་ཀྱི་དོན་ཅི་ཡང་མ་ཐོབ་པ་ཐོབ་པར་བྱེད་ཅིང་། ཐོབ་པའི་ཟུགས་པ་བརྟན་ཞིང་ཕྱིར་མི་ལྡོག་པར་ཡུར་པས་ན་གནས་སྐབས་དང་མཐར་ཐུག་གི་འབྲས་བུ་ཕྱུང་པར་ཅན་ཐོབ་པའི་དོན་ཏོ།། ལེའུ་བཅུ་པའི་རྣམ་པར་བཤད་པའོ།། (The meaning of … is taught. To summarize the title of the chapter, the meaning of [read: *zhes pa'i don*] "Invoking the Essence of All the Tathāgatas" is that, since it causes to obtain the unobtained benefit of *siddhi*s and since the obtained power proves to be stable and irreversible, one obtains the supreme temporary and ultimate results. This is the commentary on the tenth chapter.)	AUTO
SCCT	ཆྲཱུ་ཆྲཱུ་བ་ས་ཀྵི་རྟན་ཀྲ་ཤྲ་ལ་ཨ་བ་ལཱི་ནེ་ཀྱི་ཀྱི་ཕཊ། ཨོ་སྱ་ར་ཆྲིླྀ་ལ་ཀྱི་ཀྱི་ཕཊ། ཨོ་ཡིྃ་ཐྲིྃ་ས་པ་ཀ་ཤྲ་ཀྲ་ལ་ག་ཀ་ཆུ་ཇཾ་ག་སརྦ་ཤ། ཁརཊ་ལ་ཁརཊ་ལ་ཀྱི་ཀྱི་ཕཊ། ཨོ་ཛྲ་མ་དེ་ཝི་ལ་ཀྱི་ཀྱི་ཕཊ། ཨོ་ཨྃ་གཱུཾ་ཨྃ་གཱུཾ་ཀྱི་ཀྱི་ཕཊ། ཨོ་སྱ་ (Mantra.])	NSCR
THCT	ཚོགས་ཀྱི་ནང་ནས་ཁོ་ལ་བ་ནི་སོགས་བརྒྱད་པོ་རྣམས་འཇིག་པར་མི་འགྱུར་ཏེ། གོ་བ་བ་ནུ། ཅི་ག་བྱེ། ། བཅོ་བྱེ། ། བརྒྱད་ (Technical grammatical explanation.])	Treatises on Sanskrit

Table 3: Task inputs and expected outputs for Tibetan tasks. Translation is given in brackets for convenience. In detection tasks, we highlight the ground truth spans that should be identified by the models.

Model	SMDS	QUDS	RCMS	RCDS	VPCS	MCS	Avg
Chance (see the text)	55.60	12.25	50.00	48.75	50.00	10.00	37.77
GPT-4o mini	35.22±0.56	29.46±0.55	84.25±0.90	82.08±0.60	80.92±0.38	13.17±0.76	54.18
GPT-4o	56.25	45.58	98.00	91.71	99.25	25.50	69.38
Claude 3 Haiku	23.16±0.36	30.13±0.47	51.75	41.20±1.35	77.42±0.14	20.17±0.29	40.64
Claude 3.7 Sonnet	40.27	55.00	98.75	91.71	98.75	66.50	75.16
Claude 4 Sonnet	53.85	66.92	**99.25**	**95.46**	98.75	35.50	74.96
Gemini 2.0 Flash	**66.10±0.35**	71.66±0.77	97.33±0.14	92.22±0.24	99.00	34.83±0.76	76.86
Gemini 2.5 Flash	59.22±0.96	53.30±1.25	96.83±0.72	88.72±1.62	88.83±0.80	58.83±2.36	74.29
Gemini 2.5 Pro	45.18	**72.63**	97.00	91.23	**99.50**	**82.00**	**81.26**
Llama 4 Scout	24.73±0.97	24.76±10.84	82.58±0.52	61.61±1.50	60.33±49.44	12.50	44.42
Qwen2.5-72B	32.74±0.49	34.68±0.78	97.83±0.29	91.60±0.18	96.42±1.46	11.50	60.8
DeepSeek-R1	50.49	20.38	94.50	59.92	91.25	0.00	52.76
mBERT	—	—	98.65±0.52	—	—	—	—
XLM-RoBERTa	—	—	67.80±26.13	—	—	—	—
IndicBERTv2	—	—	50.75±3.02	—	—	—	—
bert-base-buddhist-sanskrit-v2	—	—	88.75±20.83	—	—	—	—

Table 4: Micro-F1 scores on the DharmaBench benchmark Sanskrit tasks and the average (Avg.) score of each model. The highest score per column is emphasized.

Model	SDT	QUDT	RCMT	VPCT	AACT	SCCT	THCT	Avg.
Chance (see the text)	45.42	60.09	50.00	50.00	50.00	50.00	7.69	44.74
GPT-4o mini	65.86±0.27	55.82±0.49	59.25±1.30	53.34±13.96	39.92±1.38	66.92±0.29	22.33±0.74	51.92
GPT-4o	65.58	52.86	97.75	77.25	62.25	63.25	36.23	65.02
Claude 3 Haiku	38.15±0.05	22.16±1.01	50.17±0.14	65.11±4.51	44.92±0.14	61.33±0.58	37.14±0.14	45.57
Claude 3.7 Sonnet	71.19	57.63	**97.75**	86.75	86.75	73.75	54.09	75.42
Claude 4 Sonnet	74.47	55.24	97.75	87.50	82.00	79.25	56.58	76.11
Gemini 2.0 Flash	72.94±0.34	**76.23±0.47**	95.92±0.14	85.29±0.37	71.25±0.25	72.67±0.14	46.73±3.30	74.43
Gemini 2.5 Flash	**80.13±0.13**	61.81±0.93	95.58±0.63	87.50±0.43	67.08±0.38	72.33±0.63	52.11±1.08	73.79
Gemini 2.5 Pro	77.55	57.22	97.50	88.25	76.25	81.75	**57.82**	**76.62**
Llama 4 Scout	44.73±1.38	59.61±0.25	82.92±1.38	66.96±3.07	58.58±0.80	55.75±0.43	35.24±0.43	57.68
Qwen2.5-72B	56.94±0.92	57.32±1.00	91.92±0.14	79.67±0.14	50.75±0.90	55.75±0.66	27.79±0.50	60.02
DeepSeek-R1	60.56	49.30	87.25	79.00	57.50	64.50	39.45	62.51
mBERT	—	—	57.70±17.20	85.17±1.61	50.00±0.00	53.45±7.71	7.89±0.44	–
XLM-RoBERTa	—	—	50.10±0.22	71.90±10.15	52.40±5.37	73.85±13.35	7.94±0.55	–
tibetan-roberta-base	—	—	52.60±7.71	91.10±0.68	**89.75±1.25**	93.25±0.88	56.08±0.98	–
CINO-V2-base	—	—	59.00±19.30	**91.15±1.22**	82.85±19.52	**97.80±0.41**	19.11±14.51	–

Table 5: Micro-F1 scores on the DharmaBench benchmark Tibetan tasks and the average (Avg.) score of each model. The highest score per column is emphasized.

Task	Sources	Sources
SMDS	*Kāvyādarśa/Kāvyalakṣaṇa* by Daṇḍin	*Gītagovinda* by Jayadeva
	Rāmacarita by Abhinanda	*Raghuvaṃśa* by Kālidāsa
	Buddhacarita by Aśvaghoṣa	*Kapphiṇābhyudaya* by Śivasvāmin
	Kirātārjunīya by Bhāravi	*Kuṭṭanīmata* by Dāmodaragupta
	Mahābhārata attributed to Vyāsa	*Sūktimuktāvalī* comp. by Jalhaṇa
	Śatakatraya by Bhartṛhari	*Śrīkaṇṭhacarita* by Maṅkha
	Surathotsava by Someśvaradeva	*Rāmāyaṇa* by Vālmīki
	Caṇḍamahāroṣaṇatantra with commentary *Padmāvatī* by Mahāsukhavajra	*Vyaktabhāvānugatatattvasiddhi* by Yoginī Cintā
	Abhayapaddhati by Abhayākaragupta	*Hevajratantra*
	Īśvarapratyabhijñāvivṛti by Utpaladeva	*Netratantra*
	Śivastotrāvalī by Utpaladeva	*Tantrāloka* by Abhinavagupta
	Mīmāṃsāślokavārttika by Kumārila Bhaṭṭa with *Kāśikā* by Sucaritamiśra	*Mṛtyuvañcanopadeśa* by Vāgīśvarakīrti
	Pramāṇavārttikavṛttiṭīkā by Karṇakagomin	*Pramāṇavārttikālaṅkāra* by Prajñākaragupta
	Viṣṇupurāṇa attributed to Vyāsa	*Kaṭhopaniṣad*
QUDS	*Abhayapaddhati* by Abhayākaragupta	*Munimatālaṅkāra* by Abhayākaragupta
	Sūtaka (*Caryāmelāpakapradīpa*) by Āryadeva	*Madhyamakāvatārabhāṣya* by Candrakīrti
	"Śāstrārambha" section of the *Nyāyamañjarī* by Jayanta	*Tarkabhāṣā* by Mokṣākaragupta
	Bodhicaryāvatārapañjikā on ch. 5 and ch. 9 vv. 1–20 by Prajñākaramati	*Padminī* by Ratnarakṣita
	Guṇabharaṇī by Raviśrījñāna	*Sāramañjarī* by Samantabhadra
	Sekanirdeśapañjikā by Rāmapāla	*Tattvaratnāvaloka* by Vāgīśvarakīrti
	Vivaraṇa by Vāgīśvarakīrti	*Prasāda* on the *Prakriyākaumudī* by Viṭṭhalācārya
RCDS	*Hevajratantra* ch. 5 w/ *Ratnāvalī Hevajrapañjikā* by Kamalanātha	*Hevajratantra* ch. 5 w/ *Muktāvalī Hevajrapañjikā* by Ratnākaraśānti
	Tattvaratnāvaloka w/ *Vivaraṇa*, both by Vāgīśvarakīrti	*Bodhicaryāvatāra* ch. 5 by Śāntideva w/ commentary by Prajñākaramati
	Caṇḍamahāroṣaṇatantra ch. 1–4 w/ commentary by Mahāsukhavajra	*Gītagovinda* ch. 1–4 by Jayadeva w/ commentary by Mānāṅka
	Mīmāṃsāślokavārttika by Kumārila Bhaṭṭa w/ *Kāśikā* by Sucaritamiśra	*Śivastotrāvalī* by Utpaladeva w/ commentary by Kṣemarāja
	Sekanirdeśa by Advayavajra w/ *Pañjikā* by Rāmapāla	*Nāmasaṃgīti* vv. 1–41 w/ commentary by Vilāsavajra
	Meghadūta vv. 1–25 w/ commentary by Mallināth a	
MCS	*Mahābhāṣya* by Patañjali	*Abhidharmakośa* by Vasubandhu
	Abhidharmakośabhāṣyavyākhyā by Yaśomitra	*Raghuvaṃśa* by Kālidāsa
	Commentary on the *Raghuvaṃśa* by Mallināth a	*Bodhicāryāvatāra* by Śāntideva
	Anargharāghava by Murāri	*Yājñavalkyasmṛti* by Yājñavalkya
	Mitākṣarā by Vijñāneśvara	*Kirātārjunīya* by Bhāravī
	Suvṛttatilaka by Kṣemendra	*Śivalīlārṇava* by Nīlakaṇṭha Dīkṣita
	Pratijñāyaugandharāyaṇa by Bhāṣa	*Campūrāmāyaṇa* by Bhoja
	Śiśupālavadhā by Māgha	*Mūlamadhyamakakārikā* by Nāgārjuna
	Prasannapadā by Candrakīrti	*Āgamaprāmāṇya* by Yāmunācārya
	Prakriyākaumudī by Rāmacandraśeṣa	*Prakriyākaumudīprasāda* by Viṭṭhala
	Manusmṛti	*Abhijñānaśakuntala* by Kālidāsa
	Amaruśataka by Amaru	The Epigrams attributed to Bhartṛhari
	Muktāvalī by Ratnākaraśānti	*Subhāṣitaratnakośa* compiled by Vidyākra
	Sūryaśataka by Mayūrabhaṭṭa	*Meghadūta* by Kālidāsa
	Prajñāpāramitopadeśa by Ratnākaraśānti	*Śiṣyalekha* by Candragomin
	Avadānakalpalatā by Kṣemandra	*Gītagovinda* by Jayadeva
	Vṛttamālāstuti by Jñānaśrīmitra	The Commentary on the *Vasantatilakā* by Vanaratna
	Subhāṣitāvali compiled by Vallabhadeva	*Jātakamālā* by Āryasūra
	Kapphiṇābhyudaya by Śivasvāmin	*Kāvyaprakāśa* by Mammaṭa
	Triṃśikā Vijñaptimātrakārikā by Vasubandhu	*Buddhacarita* by Aśvaghoṣa
	Pramāṇavārttika by Dharmakīrti	*Laṅkāvatārasūtra*
	Caurapañcāśikā by Bihaṇa	*Uttararāmacarita* by Bhavabhūti
	Kāvyālaṅkāra by Bhāmaha	*Vigrahavyāvartanī* by Nāgārjuna
	Āryākośa of Ravigupta	*Nyāyamañjarī* by Bhaṭṭa Jayanta
	Āgamadaṃbara by Bhaṭṭa Jayanta	*Bodhisattvāvadānakalpalatā* by Kṣemendra
	Subhāṣitasaṃgraha compiled by Puruṣottama Mayārāma Paṇḍyā	*Aṣṭāṅgahṛdaya* by Vāgbhaṭa
VPCS	*Abhidharmakośabhāṣya* by Vasubandhu	*Abhidharmakośabhāṣyavyākhyā* by Yaśomitra
	Mitākṣarā by Vijñāneśvara	*Prakriyākaumudīprasāda* by Viṭṭhala
	Commentary on the *Raghuvaṃśa* by Mallināth a	Commentary on the *Raghuvaṃśa* by Hemādri
	Vigrahavyāvartanī by Nāgārjuna	*Prasannapadā* by Candrakīrti
	Mahābhāṣya by Patañjali	Commentary on the *Meghadūta* by Vallabhadeva
	Nyāyamañjarī by Jayantabhaṭṭa	*Kāśikāvṛtti* by Jayāditya and Vāmana
	Padamañjarī by Haradatta	

Table 6: Task source distributions in Sanskrit. SMDS, QUDS, VPCS, RCMS, and RCDS sources are all e-texts; MCS includes various text formats.

Task	Source	
SDT	*Mañjuśrīvikrīḍita* (D96, Esukhia) *Mahāyānottaratantraśāstravyākhyā* by Asaṅga (D4025, ACIP) *Yon tan gzhir gyur ma* by Tsong-kha-pa Blo-bzang-grags-pa (Lotsawa House) *Sugatapañcatriṃśadratnanāmamālāstotra* by Mātṛceṭa (D1142, ACIP) *Namastāraikaviṃśatistotra* (D438, ACIP)	*Sūtrālaṃkārabhāṣya* by Vasubandhu (D4026, ACIP) *Nges don phyag rgya chen po'i smon lam* by Rang-'byung rDo-rje (Lotsawa house) *sGra dbyangs lha mo dbyangs can ma'i bstod pa* by Tsong-kha-pa Blo-bzang-grags-pa (Lotsawa House) *Mahāmudropadeśa* by Tilopa (D2303, ACIP) *dGongs gter sgrol ma'i brgyud 'debs utpa la'i phreng ba* by 'Jam-dbyangs mKhyen-brtse'i dBang-po (Lotsawa House)
QUDT	*mDo kun las btus pa* by Nāgārjuna (D3934, rKTs) *mDo kun las btus pa chen po* by *Adhīśa Dīpaṃkaraśrījñāna (D3961, rKTs) *Subāhuparipṛcchānāmatantrapiṇḍārtha* by Buddhagupta (D2671, ACIP) *Vyaktapadāsuhṛllekhaṭīkā* by Mahāmati (D4190, ACIP) *Dam chos yid bzhin nor bu thar pa rin po che'i rgyan* by sGam-po-pa bSod-nams-rin-chen (BDRC: MW3CN2232) *Grub mtha' shel gyi me long* by Thu'u-bkwan Blo-zang-chos-kyi-Nyi-ma (BDRC MW2124) *sNyan ngag me long dang bod mkhas pa'i 'grel pa* by Daṇḍin & Mi-pham dGe-legs-rnam-rgyal (BDRC: MW1PD137863)	*bSlab pa kun las btus pa* by Śāntideva (D3940, rKTs) *Subāhuparipṛcchātantrapadārthaṭippaṇī*, author unknown (D2672, ACIP) *Subāhuparipṛcchānāmatantrapiṇḍārthavṛtti*, author unknown (D2673, ACIP) *Deb ther sngon po* by 'Gos-lo-tsā-ba gZhon-nu-dpal (BDRC: MW3CN3380) *g.Yung drung bon gyi zhal 'don phyogs bsgrigs dgos 'dod kun 'byung* compiled by Rin-chen Tshe-ring & bSod-nams bKra-shis (BDRC: MW3CN6348) *rDzogs pa chen po klong chen snying thig gi sngon 'gro'i khrid yig kun bzang bla ma'i zhal lung* by rDza-dpal-sprul O-rgyan-'jigs-med-chos-kyi-dbang-po (BDRC: MW3CN4690) *Mi la ras pa'i rnam thar* by gTsang-smyon He-ru-ka-rus-pa'i-rgyan-can & 'Jam-mgon-kong-sprul Blo-gros-mtha'-yas (BDRC: MW1KG3714)
RCMT	*Viṃśikā* by Vasubandhu w/ *Vṛtti* (D4056 & D4057, edition by Jonathan A. Silk) *Subāhuparipṛcchātantra* w/ *Ṭippaṇī*, author unknown (D805 & D2672, Esukhia and ACIP) *Suhṛllekha* by Nāgārjuna w/ *Ṭīkā* by Mahāmati (D4182 & D4190, ACIP) *Ratnāvalī* by Nāgārjuna w/ *Ṭīkā* by Ajitamitra (D4158 & D4159, edition by Michael Hahn and ACIP) *Bodhicittabhāvanā* by Mañjuśrīmitra w/ auto-commentary (D2591 & D2578, ACIP)	*Sekanirdeśa* by Advayavajra w/ *Pañjikā* by Rāmapāla (D2252 & D2253, ACIP) *Subāhuparipṛcchānāmatantrapiṇḍārtha* by Buddhagupta w/ *Piṇḍārthavṛtti*, author unknown (D2671 & D2673, ACIP) *Mūlamadhyamakakārikā* by Nāgārjuna w/ *Vṛtti* by Buddhapālita (D3824 & D3482, ACIP) *Bodhicaryāvatāra* by Śāntideva w/ *Pañjikā* by Prajñākaramati (D3871 & D3872, ACIP) *Vigrahavyāvartanī* w/ *Vṛtti* by Nāgārjuna (D3828 & D3832, ACIP)
VPCT	Autochthonous works from online sites: https://rywikitexts.tsadra.org/index.php/Main_Page https://www.gyalyongsachen.com/	ACIP and rKTs Derge Kangyur and Tengyur
AACT	ACIP and Esukhia Derge Kangyur (allochthonous)* *rNying ma rgyud 'bum*, Chengdu edition (para-canonical) *rNgog slob brgyud dang bcas pa* (BDRC: W1KG16666) (autochthonous)	ACIP Derge Tengyur (allochthonous)* ACIP dGe-lugs collected writings (autochthonous) Rong-zom-pa's collected writings, Chengdu edition (autochthonous)
SCCT	ACIP and Esukhia Derge Kangyur (scripture)	ACIP Derge Tengyur (non-scripture)
THCT	ACIP and Esukhia Derge Kangyur	ACIP Derge Tengyur

Table 7: Task source distributions in Tibetan. *Works whose Indic origin is unclear are excluded. rKTs, ACIP means based on e-texts from those sources. SDT, AACT, SCCT, QUDT, and RCMT sources are all e-texts (except where editions are specified).

Label	Definition	Train	Test
Empty texts	—	369	244
Non-empty texts	—	231	162
QUOTE	The text being cited	467	335
OP	Opening particle marking the start of a citation	334	227
CP	Closing particle marking the end of a citation	352	238
GEN_SRC	No title of the text is provided, only a generic source reference, e.g., "in the same text", "in another", "in the sūtra"	44	32
TITLE	The title of the text being cited	246	160

Table 8: Label statistics for the QUDT task.

Parameter	Value
Training epochs	6
Batch size	16
Learning rate	5e-5

Table 9: Encoders fine-tuning hyperparameters.

Model	Checkpoint	License	Parameters
GPT-4o mini[a]	gpt-4o-mini-2024-07-18	Proprietary	>200B∗
GPT-4o[b]	gpt-4o-2024-08-06	Proprietary	~8B∗
Claude Haiku 3[c]	Used until July 2025	Proprietary	~20B∗
Claude 3.7 Sonnet[d]	Used until July 2025	Proprietary	~175B∗
Claude 4 Sonnet[e]	Used until July 2025	Proprietary	~150-250B∗
Gemini 2.0 Flash[f]	gemini-2.0-flash-001	Proprietary	N/A
Gemini 2.5 Flash[g]	gemini-2.5-flash	Proprietary	N/A
Gemini 2.5 Pro[h]	gemini-2.5-pro	Proprietary	N/A
Llama 4 Scout[i]	Llama-4-Scout-17B-16E-Instruct	LLAMA 4 COMMUNITY LICENSE AGREEMENT	17B active 109B overall
Qwen2.5-72B [j]	Qwen2.5-72B-Instruct-Turbo	Qwen LICENSE	72B
DeepSeek-R1[k]	deepseek-ai/DeepSeek-R1-0528	MIT License	671B
mBERT[l]	google-bert/bert-base-multilingual-cased	Apache-2.0	~110M
XLM-RoBERTa[m]	FacebookAI/xlm-roberta-base	MIT License	~270M
Tibetan-RoBERTa[n]	sangjeedondrub/tibetan-roberta-base	MIT License	~270M
CINO[o]	hfl/cino-base-v2	Apache-2.0	~270M
IndicBERTv2[p]	ai4bharat/IndicBERTv2-MLM-Back-TLM	MIT License	~270M
bert-base-buddhist-sanskrit-v2[q]	Matej/bert-base-buddhist-sanskrit-v2	MIT License	~110M

[a] https://platform.openai.com/docs/models/gpt-4o-mini
[b] https://platform.openai.com/docs/models/gpt-4o
[c] https://www.anthropic.com/news/claude-3-family
[d] https://www.anthropic.com/news/claude-3-7-sonnet
[e] https://www.anthropic.com/news/claude-4
[f] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash
[g] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash
[h] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro
[i] https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E
[j] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[k] https://huggingface.co/deepseek-ai/DeepSeek-R1-0528
[l] google-bert/bert-base-multilingual-cased
[m] https://huggingface.co/FacebookAI/xlm-roberta-base
[n] https://huggingface.co/sangjeedondrub/tibetan-roberta-base
[o] https://huggingface.co/hfl/cino-base-v2
[p] https://huggingface.co/ai4bharat/IndicBERTv2-MLM-Back-TLM
[q] https://huggingface.co/Matej/bert-base-buddhist-sanskrit-v2

Table 10: Checkpoints used during experiments and their License, and their number of parameters. ∗ = non-official estimation, as this information is not public. N/A means not disclosed and an estimation can't be found.

Artifact	Type	License	Usage
LangChain[a]	Framework	MIT License	Prompting
Together AI[b]	Provider	Proprietary	API access
eval4ner AI[c]	Package	Apache-2.0	Metrics calculation
Label Studio AI[d]	Platform	Apache-2.0	Annotations
pyewts[e]	Package	Apache-2.0	Wylie to Tibetan conversion
sanskritmetres[f]	Tool	GPL-2.0 license	Metres pre-annotation
skrutable[g]	Tool	N/A (GitHubg, free acess)	Metres pre-annotation
Esukhia derge-kangyur[h]	Website	Public Domain	Tibetan data
ACIP [i]	Website	Public Domain	Tibetan data
tsadra [j]	Website	Public Domain	Tibetan data
gyalyongsachen [k]	Website	Public Domain	Tibetan data
GRETIL [l]	Website	Public Domain	Sanskrit data

[a] https://www.langchain.com/
[b] https://www.together.ai/
[c] https://github.com/cyk1337/eval4ner
[d] https://labelstud.io/
[e] https://github.com/OpenPecha/pyewts
[f] https://sanskritmetres.appspot.com/identify
[g] https://www.skrutable.info/
[h] https://github.com/Esukhia/derge-kangyur
[i] https://asianlegacylibrary.org/library/
[j] https://rywikitexts.tsadra.org/index.php/Main_Page
[k] https://www.gyalyongsachen.com/
[l] https://textgridrep.org/project/TGPR-2ba9cb1b-9602-202d-71ce-67e63a29de55

Table 11: Packages and artifacts used during experiments, along with their license and usage explanation.