

Dharmamitra: AI-Assisted Multilingual Research Infrastructure for Buddhist Studies

Dharmamitra is a public, research-oriented digital platform for Buddhist Studies. It integrates machine translation, grammatical analysis, OCR, semantic search, and intertextuality exploration for the major classical languages of Buddhist textual traditions. The project began at UC Berkeley in April 2023 and is currently co-hosted by the Tsadra Foundation, Tohoku University, and the Berkeley AI Research Lab. Its central aim is to bring advanced data-driven methods into the daily workflow of translators, philologists, historians, and scholars of Buddhist textual materials, while keeping scholarly verification and source transparency directly accessible to its users.

Textual oriented Buddhist Studies is structurally multilingual. Important textual traditions are preserved across Sanskrit, Pāli, Tibetan, Chinese, and other Asian languages, while modern scholarship is distributed across English, Japanese, German, French, Chinese, and many other languages. Researchers often need to compare parallel passages, trace citations, identify translations, locate doctrinal vocabulary, and move between canonical corpora, manuscripts, lexicons, and secondary literature. These tasks are intellectually demanding not only because of the difficulty of the languages, but also because the relevant materials are distributed across many databases and technical environments. Even a relatively simple question, such as whether an image or doctrinal formulation in one Sanskrit passage has parallels in Tibetan, Chinese, or Pāli sources, may require repeated keyword searches, language-specific expertise, and substantial familiarity with the structure of several corpora.

Dharmamitra addresses this problem by providing a unified research environment in which classical Buddhist texts can be translated, grammatically analyzed, searched, and contextualized across languages. The platform currently supports Sanskrit, Tibetan, Buddhist Chinese, and Pāli, with English as the primary target language for translation and explanation. Its tools include machine translation, grammatical and lexical explanation, OCR for images and documents, MITRA Explore, MITRA Deep Research, browser extensions, DharmaNexus as an intertextual reference database, and a large Sanskrit-Tibetan dictionary derived from parallel sentence data. The system is publicly accessible at dharmamitra.org and is designed as a non-commercial infrastructure for scholarly and educational use.

One of Dharmamitra's major contributions is its attempt to combine the accessibility of large language models with the power of specialized digital humanities systems designed for philological research. In recent years, AI systems have become increasingly powerful for translation, summarization, and multilingual access. Yet for Buddhist Studies, these tools are risky if used without grounding in primary sources. They can produce fluent but unverifiable

answers, obscure uncertainty, and encourage researchers to rely on generated text rather than source-based evidence. Dharmamitra therefore does not treat AI as a replacement for philological judgment. Instead, it uses AI as part of a retrieval-grounded workflow in which generated explanations are connected to source passages, dictionary entries, grammatical analysis, and links into the underlying textual database.

MITRA Translate provides machine translation and grammatical explanation for Sanskrit, Tibetan, Buddhist Chinese, and Pāli. For Sanskrit, Tibetan, and Chinese, the system provides segmentation, lexical information, morpho-syntactic explanation, and, where possible, dictionary links to external resources such as the Digital Dictionary of Buddhism and Tibetan lexical resources. The goal is not merely to produce a polished English rendering, but to make the structure of the source text more inspectable and references directly accessible. This is especially useful in Buddhist Studies, where researchers often need to work with texts in several languages at different levels of competence. Dharmamitra can offer quick orientation, draft translations, grammatical explanations, and alternative readings, while allowing the user to access additional resources and to evaluate the output critically.

MITRA Explore extends this approach from translation to discovery. Traditional keyword search remains useful, but it is limited when dealing with Sanskrit sandhi, compounding, morphology, and multilingual retrieval tasks. Dharmamitra therefore combines classical search techniques with semantic and multilingual retrieval. Users can search for passages that express a similar meaning in different wording within the same language, or ask a query in one language and retrieve semantically related passages in another. For example, a Sanskrit sentence, an English description, or a Japanese research question can be used to locate relevant passages in Sanskrit, Tibetan, Chinese, or Pāli. This makes it possible to search not only for words, but for concepts, images, doctrinal ideas, and explore textual relationships across language boundaries.

MITRA Deep Research combines the strengths of MITRA Explore and MITRA Translate in a unified workflow. A user can enter a passage, a question, or a research idea, and the system retrieves potentially relevant materials from DharmaNexus across several languages. It then provides translations and contextual explanations while linking back into the source database. These deep links are crucial: they allow users to inspect the passage in context, compare related materials, and verify whether the AI-generated synthesis is justified by the underlying sources. In this sense, Dharmamitra tries to make AI useful while also making it more transparent. The system is designed around the principle that generated answers must remain accountable to texts.

The central evidence layer behind these tools is DharmaNexus, an intertextuality exploration database that provides monolingual intertextual links for Sanskrit, Pāli, Tibetan, and Chinese, and multilingual sentence-level parallels between Sanskrit and Tibetan, Sanskrit and Chinese, and Chinese and Tibetan. The database currently contains more than two million multilingual parallel sentences and serves as the main textual reference system for Dharmamitra. It allows users to move from a translation or search result into a broader network of textual relations, including parallel passages, translations, citations, and related formulations.

A further resource developed within the project is the MITRA Sanskrit-Tibetan Dictionary. This dictionary contains approximately four million entries automatically extracted from sentence-level parallel data. It is designed for expert use and provides extensive evidence for entries, often including hundreds or thousands of reference sentences. Although automatically extracted entries require critical use and are not always perfectly segmented, the dictionary offers a new kind of evidence-rich lexical resource for scholars working on the junction between Sanskrit and Tibetan Buddhist materials. An online version is in preparation, and the current version is already available for use with GoldenDict.

Dharmamitra contributes to Buddhist Studies by lowering barriers of access. The project is motivated by the recognition that language knowledge remains central to philological work, but that the distribution of linguistic training is uneven. Researchers may have strong domain knowledge but lack access to one of the relevant primary or secondary languages. Younger scholars may face increasing pressure to work across multiple languages while having limited time and institutional support for long-term training. Dharmamitra cannot and should not replace language learning, but it can make multilingual research more navigable. Used responsibly, it allows scholars to locate materials, test hypotheses, compare passages, and engage with texts that would otherwise remain difficult to access.

The project has already reached a substantial scholarly and public user base. Current usage includes roughly 900 to 1,100 core translation users who send more than one query per day, approximately 8,000 to 11,000 daily translation requests, and more than 1,000 daily requests for MITRA Explore and MITRA Deep Research. Feedback has come from users across the international Buddhist Studies community, and usage patterns show broad global interest. This suggests that Dharmamitra is addressing a real infrastructural need: scholars, translators, students, and practitioners want tools that help them work across Buddhist textual traditions with direct access to the relevant source evidence.

The project continues to expand. Recent updates include the addition of Tibetan data such as the Rin chen gter mdzod and rNying ma bka' ma in collaboration with the Tsadra Foundation, a strategic data partnership with BDRC, continuous expansion of Sanskrit electronic texts, the addition of premodern Chinese materials via Kanripo, preparation for Japanese material, and ongoing development of OCR models for Sanskrit and Tibetan. Future work will focus on improving retrieval quality, expanding corpora, integrating more reliable OCR/HTR workflows, and strengthening the connection between AI-generated assistance and verifiable textual evidence.

Dharmamitra fits closely with the spirit of the Aming Tu Prize. Professor Aming Tu's work helped make Buddhist textual heritage available in the digital realm through foundational projects such as CBETA and related digital Buddhist resources. Dharmamitra seeks to extend this legacy into the age of multilingual AI and computational philology. Its contribution is not simply the application of new technology to old texts, but the design of a research infrastructure in which access, verification, intertextual discovery, and scholarly responsibility are integrated. By combining public accessibility with advanced multilingual retrieval, machine translation, grammatical analysis, OCR, and DharmaNexus as an evidence layer, Dharmamitra offers a new

model for Digital Buddhist Studies: one in which AI helps researchers discover and understand Buddhist textual materials, while the authority of interpretation remains grounded in sources, context, and human judgment.