Sebastian Nehrdich

Berkeley AI Research Lab (BAIR) EECS, University of California, Berkeley 2121 Berkeley Way, CA 94704

Mobile: +1 510 717 9377

Email: nehrdich@berkeley.edu

Current position

Associate Specialist, Berkeley AI Research Lab, University of California, Berkeley

Areas of specialisation

Natural Language Processing, Large Language Models (LLMs), deep learning and database construction with focus on traditional Buddhist languages such as Pāli, Sanskrit, Chinese and Tibetan as well as modern research languages such as English, Japanese, French, and German.

Research Experience

Apr 2023-present — Associate Specialist, CTO of the Dharmamitra project, under the guidance of Prof. Kurt Keutzer, Berkeley AI Research Lab, University of California, Berkeley

Feb 2021-Mar 2023 — Research Assistant, Heinrich-Heine-Universität Düsseldorf. Development and implementation of deep learning models and managing the technical infrastructure for the project "ChronBMM: Dating text corpora using Bayesian Mixture Models" led by Dr. Oliver Hellwig.

Aug 2019-Mar 2023 — Contract work at the Khyentse Center for Tibetan Buddhist Textual Scholarship, Universität Hamburg: Design and implementation of the BuddhaNexus database for research on parallel passages in Ancient Buddhist corpora.

May-Nov 2020 — Research Assistant at the Khyentse Center for Tibetan Buddhist Textual Scholarship, Universität Hamburg for the BuddhaNexus project.

Education

Mar 2023-present — Doctoral candidate at the Institute for Language and Information, Heinrich-Heine Universität Düsseldorf. Working title: "Development of a Framework for the Intertextual Analysis of Indian Abhidharma- and Yogācāra-Literature". Expected completion: June 2025

Oct 2017-May 2020 — Master of Arts in Buddhist Studies, Universität Hamburg. Title: "The Fourth Chapter of the *Madhyāntavibhāgaṭīkā*: Critical Edition of the Sanskrit/Tibetan and Partial English Translation". *Passed with distinction*

Oct 2018-Mar 2020 — Exchange student at the University of Kyōto, Japan. All classes held in Japanese.

Oct 2012-Sep 2017 — Bachelor of Arts in Indology and Sinology, Universität Hamburg. Title: "Two Commentaries on Vasubandhu's *Triṃśika*: A comparative analysis of Sthiramati's and Xuanzang's exegesis".

Sep 2016-Jun 2017 — Exchange student at the Dharma Drum Institute for Liberal Arts, Republic of China, Taiwan. All classes held in Chinese.

Grants, honours & awards

2024 — Google Cloud Research Credits Grant for use of Vertex AI for Cultural Preservation Tasks at the Dharmamitra project.

2020 — Khyentse Foundation Award for Excellence in Buddhist Studies awarded by the Khyentse Foundation and the Department of Indian and Tibetan Studies, Universität Hamburg, for recognition and distinction in the field of Indo-Tibetan Buddhist Studies for the year 2020.

2015-2020 — Scholarship for bachelor and master degree awarded by the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

Publications & Talks

Publications

2025 — Sebastian Nehrdich, Avery Chen, Marcus Bingenheimer, Lu Huang, Rouying Tang, Xiang Wei, Leijie Zhu, 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*. Albuquerque, USA. 129–137.

2024 — Sebastian Nehrdich, Oliver Hellwig, Kurt Keutzer (2024): One Model is All You Need: ByT5-Sanskrit, a Unified Model for Sanskrit NLP Tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Findings)*.

2024 — Marieke Meelen, Sebastian Nehrdich, Kurt Keutzer (2024): Breakthroughs in Tibetan NLP & Digital Humanities. *Revue d'Etudes Tibétaines*, 72, 5-25.

2023 — Oliver Hellwig, Sebastian Nehrdich, Sven Sellmer (2023): Data-driven dependency parsing of

Vedic Sanskrit. *Language Resources and Evaluation*, 57, 1173-1206.

2023 Sebastian Nehrdich, Marcus Bingenheimer, Justin Brody, Kurt Keutzer (2023): MITRA-zh: An efficient, open machine translation solution for Buddhist Chinese. In *NLP4DH*.

2023 Sebastian Nehrdich (2023): Observations on the Intertextuality of Selected Abhidharma Texts Preserved in Chinese Translation. *Religions*, 14(7), 911.

2022 Sebastian Nehrdich, Oliver Hellwig (2022): Accurate Dependency Parsing and Tagging of Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, LT4HALA 2022.* Marseille, France. 20-25

2022 Sebastian Nehrdich (2022): SansTib, a Sanskrit - Tibetan Parallel Corpus and Bilingual Sentence Embedding Model. In *Proceedings of the 13th Conference on Language Resources and Evaluation, LREC 2022.* Marseille, France. 6728-6734

2021 Oliver Hellwig, Sven Sellmer, Sebastian Nehrdich (2021): Obtaining More Expressive Corpus Distributions for Standardized Ancient Languages. In *Proceedings of the Conference on Computational Humanities Research, CHR2021.* Amsterdam, The Netherlands. 92–107

2020 Sebastian Nehrdich (2020), A Method for the Calculation of Parallel Passages for Buddhist Chinese Sources Based on Million-scale Nearest Neighbor Search. In *Journal of the Japanese Association of the Digital Humanities - Marcus Bingenheimer, Christian Wittern, Jinhua Chen (Guest Editors) Special Issue: Buddhism and Technology*. Vol. 5. ISSN: 2188-7276.

2018 Oliver Hellwig, Sebastian Nehrdich (2018): Sanskrit Word Segmentation Using Character-level Recurrent and Convolutional Neural Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium. 2754–2763.

Invited Talks, Workshops, and Seminars

Mar 2025 Talk, CEAL (Council on East Asian Libraries) Technology Forum, Columbus, Ohio: "MITRASearch: Building Information Retrieval Systems for Classical Asian Languages in the Age of AI"

Mar 2025 Workshop, Annual Conference of the Association of Asian Studies, Columbus, Ohio: "Machine Translation for Asian Studies"

Dec 2024 Invited Talk, International Symposium "Buddhist Studies and Digital Humanities: 100 Years of the Taishō Tripiṭaka and 30 Years of SAT", Ito International Research Center, University of Tokyo, Tokyo, Japan: "MITRA Search: Exploring Buddhist Literature Preserved in Classical Asian Languages with Multilingual Approximate Search"

Nov 2024 Invited for Workshop "Case studies from current research projects - Conversations on Digital Scholarly Editing", Śivadharma Project Headquarters, University of Naples L'Orientale, Naples, Italy: "Dharmamitra: Developing a Toolkit for Philological Work on Premodern Asian Low-

Resource Languages"

Nov 2024 Invited Talk, Manuscriptology and Digital Humanities Online Lecture Series, Heidelberg University and Otani University: "Dharmamitra: New Tools for Sanskrit Translation, Grammatical Analysis, Search, and Digital Philology"

Oct 2024 Invited seminar talk, Center for Language and Speech Processing,Johns Hopkins University, Baltimore, USA: "MITRA: Beyond Just Machine Translation for Premodern Asian Low Resource Languages"

Oct 2024 Invited Talk, Conference "AI and the Future of Buddhist Studies Conference", Numata Center for Buddhist Studies, University of California Berkeley, Berkeley, USA: "Dharmamitra Search: Leveraging Multilingual Language Models for Search and Detection of Textual Reuse across Diverse Text Collections"

Aug 2024 PNC Annual Conference and Joint Meetings, Seoul, South Korea: "MITRA: Developing Language Models for Machine Translation and Search in Buddhist Source Languages"

Apr 2024 Invited Talk, Conference "International Conference on Retrospect and Prospect of Digitalization of Buddhist Resources", National Taiwan University, Taipei, Taiwan: "Massive Multilingual Machine Translation and Search for Buddhist Languages: The MITRA Project"

Mar 2024 Invited Keynote talk, "Interdisciplinary Symposium on Machine Translation and Digital Humanities", National University of Singapore, Singapore: "Dharmamitra: Enabling Massive Multilingual Machine Translation for Ancient Languages of the Buddhist Tradition"

Feb 2024 Talk, International Sanskrit Computational Linguistics Conference, Auroville, India: "Machine Translation and LLM-Powered Grammatical Explanation for Sanskrit"

Jun 2023 Invited Workshop "Machine Translation and Algorithmic Study of Classical Buddhist Texts", Dongguk University, Seoul, South Korea: "Developing Machine Translation for Ancient Buddhist Texts in Canonical Languages"

Jun 2023 Invited Talk, International symposium "Maitreya Faith/Philosophy and Buddhist Community Service", Chinese University of Hong Kong, Hong Kong: "MITRA: Developing Natural Language Processing Tools for the Languages of Buddhist Literature"

Apr 2023 Invited Talk,Symposium "Advanced Computational Methods for Studying Buddhist Texts",Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences, Vienna, Austria: "Creating a Shared Semantic Vector Space for Buddhist Languages"

Jan 2023 Talk, International symposium "Perspectives of Digital Humanities in the Field of Buddhist Studies", Universität Hamburg, Hamburg, Germany: "ChronBMM – Dating Text Corpora Using Bayesian Mixture Models"

Jan 2023 Talk, International symposium "Perspectives of Digital Humanities in the Field of Buddhist Studies", Universität Hamburg, Hamburg, Germany: "Multilingual Semantic Mining for Text Align-

ment and Parallel Corpus Building for Buddhist Languages"

Sep 2022 Talk, Conference "Deutscher Orientalistentag (DOT) 2022", Freie Universität Berlin, Berlin, Germany: "ChronBMM: Bayesian Mixture Models für die Datierung von Textkorpora"

Sep 2022 Talk, PNC Annual Conference and Joint Meetings, University of Arizona, Tucson, USA: "The Uncertain Future of Buddhist (Machine) Translation"

Jul 2022 Talk, 16th Seminar of the International Association for Tibetan Studies, Prague, Czech Republic: "Current State of the BuddhaNexus, a Powerful Database for Research on Buddhist Corpora"

Jun 2022 Talk,Digital Orientalist Online Conference "Infrastructure": "Machine Translation and Buddhist Studies – First Results"

Apr 2022 Talk, Khyentse Lecture Series, Universität Hamburg, Hamburg, Germany: "Observations on the Intertextual Relationship of the Sarvāstivāda Abhidharma Literature Preserved in Chinese Translation"

Oct 2021 Talk, Workshop "Buddhism and Language: A Twofold Perspective: The Role of Language in Buddhist Teachings and the Role of Buddhist Sources in Linguistic Research", Ludwig-Maximilians-Universität München, Munich, Germany: "Observations on the Intertextual Relationship and Possible Authorship of the Abhidharma and Yogācāra Works Attributed to Vasubandhu"

Sep 2019 Invited Talk, Young scholars workshop, University of Tsukuba, Tsukuba, Japan: "Towards a New Edition of the Fourth Chapter of the *Madhyāntavibhāgaṭīkā* by Sthiramati"

Jun 2019 Invited Talk, Exzellenzcluster "Asien und Europa im globalen Kontext", Universität Heidelberg, Heidelberg, Germany: "Perspectives on Building a Multilingual Parallel Corpus of Ancient Buddhist Scriptures Based on Machine Learning"

Major Research Projects

2023-present **MITRA Project** – Chief Technology Officer
Berkeley AI Research Lab (BAIR), UC Berkeley
Principal Investigator: Prof. Kurt Keutzer
Leading the development of state-of-the-art machine translation models and semantic search functionality for Classical Asian languages
Coordinating technical infrastructure and research directions for a large team of employees, students and interns
Supervised and implemented pretraining and finetuning of large language models (LLMs) and semantic similarity search models

2018-present **BuddhaNexus Platform** – Chief Developer
Khyentse Center for Tibetan Buddhist Textual Scholarship, Universität Hamburg
Designed and implemented comprehensive database system for detecting intertextual links in

Buddhist sources
Developed NLP technology for processing multilingual Buddhist texts
Created web-based research platform integrating advanced search and analysis tools

2021-2023 **ChronBMM Project** – Technical Lead
Heinrich-Heine-Universität Düsseldorf
Principal Investigator: Dr. Oliver Hellwig
Managed machine learning infrastructure for text dating using Bayesian Mixture Models
Implemented deep learning models for chronological analysis of historical texts
Coordinated computational resources and technical workflows

Additional relevant skills and languages

Spoken languages

German (Native)
English (Fluent in speaking and writing)
Japanese (conversational in speaking, advanced in reading)
Chinese (conversational in speaking, competent in reading)

Research languages

Sanskrit (professional in reading)
Pāli (professional in reading)
Classical Chinese (professional in reading)
Classical Tibetan (competent in reading)

Knowledge in computer science

General programming and system administration skills

- basic knowledge of C/C++, Perl, LISP/Clojure, LaTEX and Bash programming since high school; regular programming with Python since 2013
- Familiarity with Linux/Unix-type environments; administration of the web-servers of the BuddhaNexus project using docker and NGINX (`buddhanexus.net`)
- Managing the training of statistical models and Deep Learning models on the high performance infrastructure of the University of Duesseldorf for the ChronBMM project (Oliver Hellwig, 2021-2023); also managing the calculation of various NLP algorithms for the BuddhaNexus project on the high performance infrastructure of the University of Hamburg since 2019
- Web development: frontend and backend programming in Python (fastapi, ArangoDB) and Javascript (lit elements) for the BuddhaNexus project since 2019

Machine learning experience

- Transformer-based word segmentation tool for Sanskrit in Python with Tensor2Tensor (2018) (https://github.com/OliverHellwig/sanskrit/tree/master/papers/2018emnlp)
- Design and implementation of a word-vector-based system for the detection of textual reuse using nearest neighbor search (HNSW) and local alignment (Smith-Waterman) for the BuddhaNexus project based on NumPy, fastText and Faiss (2020) (https://github.com/BuddhaNexus)
- Dependency Parsers for Vedic Sanskrit (2021) and Latin (2022) in PyTorch (https://github.com/sebastian-nehrdich/latin-parser)
- Training of transformer-based language models for Sanskrit and Tibetan (SansTib 2022) (https://github.com/sebastian-nehrdich/sanstib)
- Since 2022, training and finetuning of language models for machine translation (Chinese to English, Tibetan to English)
- Since 2023, Training and deployment of Large Language Models (LLMs) via pytorch and deepspeed for the MITRA project with up to 10B parameters in size (https://dharmamitra.org; https://huggingface.co/buddhist-nlp)