

The Myanmar AI Gap

Burmese is inside today's AI models, but it is built elsewhere and it trails its neighbours. This paper measures the gap in data, models, infrastructure and people, and explains why it grows every year it is left alone.

Arckara · September 2026 · arckara.com/research/myanmar-ai-gap

DOI 10.5281/zenodo.22781767

ABSTRACT

About 43 million people speak Burmese, but the language makes up 0.0164 percent of the pages in the latest Common Crawl, the open web archive behind most public training data. Thai has about 23 times as many pages, and Vietnamese about 66 times. The shortage shows inside the models. On the same reading test, Llama 2 70B scored 90.9 in English, 46.2 in Thai and 24.1 in Burmese, below random guessing. The best models have improved fast, and the strongest now scores 72 percent on a 2026 Burmese benchmark. But Myanmar's own capacity to build AI has moved the other way. It recorded the most internet shutdowns of any country in 2024 and 2025, generates power for about half of its demand, and published fewer research articles in 2023 than in 2019, while Viet Nam's output nearly doubled. Its neighbours have launched their own language models since 2023. We argue that the gap compounds, and that it is smaller today than it will be in any year that follows, unless it is closed on purpose.

1 Where Burmese stands

About 43 million people speak Burmese.^[1] The language is no longer missing from AI. Google Translate added it in December 2014.^[2] Google lists Burmese among the languages every Gemini model can read and answer in,^[3] ChatGPT offers a Burmese interface,^[4] and Meta's NLLB translation model covers it.^[5] On BURMESE-SAN, a Burmese benchmark published in 2026, Gemini 2.5 Pro scored 72.35 percent and GPT-5 scored 66.46 percent, up from 51.61 percent for GPT-4o.^[1]

Support is thinner than those headlines suggest. The Gemini app's own list of supported languages leaves out Burmese, while it includes Khmer and Lao.^[6] Meta's Llama 4 officially supports Thai, Vietnamese, Indonesian and Tagalog, and not Burmese.^[7] Open models are the ones local developers can download, adapt and run on their own hardware, so their gaps matter most for the products a country builds for itself.

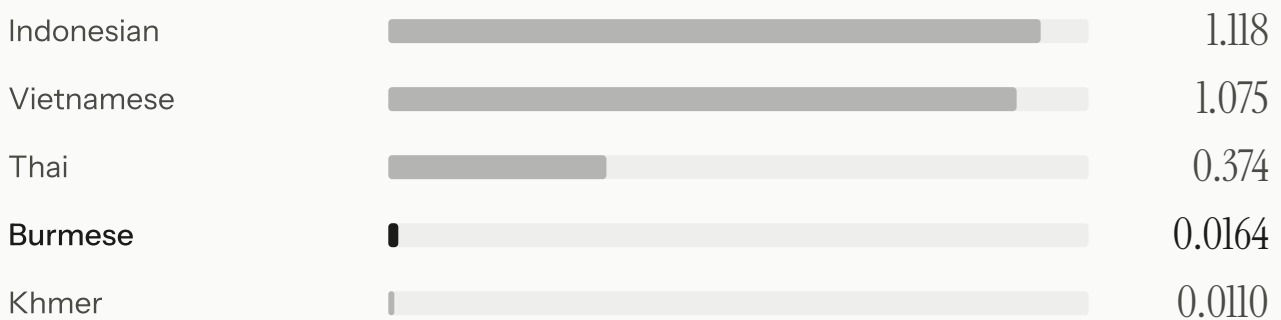
So the useful question is not whether AI can handle Burmese. It is how well it works compared with other languages, who builds it, and whether anyone in Myanmar can shape it. On all three, the evidence points the same way. Burmese has a sliver of the text that models learn from. It scores far below its neighbours on the same models. And Myanmar's capacity to build its own tools has shrunk while its neighbours built theirs.

2 Too little text

Language models learn from text, most of it collected from the public web. Common Crawl, the open archive behind most public training sets, labels each page it saves by language. In its latest crawl, from August 2026, Burmese is 0.0164 percent of pages. Thai is 0.374 percent, about 23 times as much. Vietnamese is 1.075 percent, about 66 times as much, and English is 40.45 percent.^[8]

FIGURE 1

Share of web pages by language in the latest Common Crawl



Common Crawl, crawl CC-MAIN-2026-34, primary language of each page, in percent. English, at 40.45 percent, is left out so the smaller shares can be seen.

The training sets built from that archive inherit the gap. Google's mC4 has 0.9 billion Burmese tokens, against 11 billion for Thai and 2,733 billion for English.^[9] FineWeb2, released in 2025, has 12.35 GB of Burmese text and 278.68 GB of Thai.^[10] When AI Singapore trained its first SEA-LION model for the region's languages, Burmese made up 0.49 percent of the training tokens, and reaching even that meant repeating the same 1.2 billion tokens four times.^[11]

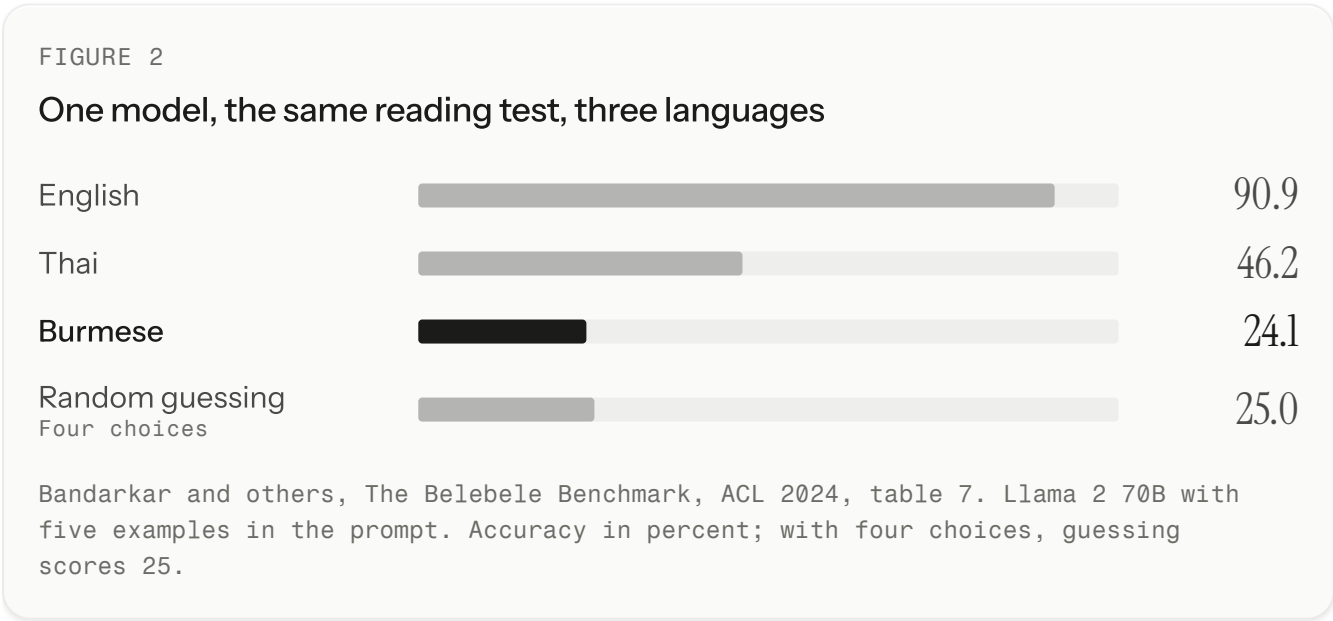
Wikipedia shows how few people are writing. In September 2026 the Burmese Wikipedia had 286 active users. The Thai Wikipedia had 2,727 and the Vietnamese 4,955.^[12]

Burmese also carries a problem of its own. For most of the last decade, most Burmese text was typed in Zawgyi, a font encoding that uses the Unicode code points for Myanmar script in its own, incompatible way. Text written in one looks garbled in the other.^[13] Myanmar officially moved to Unicode on 1 October 2019, when an estimated 90 percent of people still used Zawgyi.^[14] Much of the Burmese text on the web was written before the switch. The team behind Google's MADLAD-400 dataset found that "a large fraction of Myanmar script data on the internet is Zawgyi encoded data" and had to convert it,^[15] and in the CC-100 corpus the Zawgyi file for Burmese is nearly four times the size of the Unicode one.^[16] A dataset that skips this step can lose that text, or learn from scrambled characters.

Researchers have a name for languages in this position. In a 2020 study that sorted the world's languages by the resources available for them, Joshi and colleagues placed Burmese in class 1 on a scale from 0 to 5, the "Scraping-Bys": languages with some unlabelled text and almost no labelled data. Thai is in class 3 and Vietnamese in class 4. The study singled out Burmese as a language with millions of speakers and almost no research attention. ^[17]

3 The gap inside the models

The clearest test gives one model the same questions in different languages. Belebele does this with 900 reading comprehension questions, each with four possible answers, in 122 language variants, so random guessing scores 25. ^[18] In the Belebele paper, Llama 2 70B scored 90.9 in English, 46.2 in Thai and 24.1 in Burmese. GPT-3.5 Turbo, answering without examples, scored 87.7, 55.7 and 30.3. ^[18]



Those are 2023 models, and the best systems have moved on. BURMESE-SAN measured the change: GPT-4o scored 51.61 percent, GPT-5 66.46 percent, and Gemini 2.5 Pro led at 72.35 percent. Open models trail. The best open model in the study, ERNIE 4.5, scored 54.68 percent, and Llama 3.3 70B scored 23.07 percent. ^[1]

The size of the gap now depends on who built the model. On AI Singapore's SEA-HELM leaderboard, in its August 2026 version, Google's Gemma 4 31B scores 69.13 in Burmese and 71.09 in Thai. Qwen 3.6 27B scores 47.92 in Burmese and 68.85 in Thai, and NVIDIA's Nemotron 3 Super scores 6.17 in Burmese and 59.28 in Thai. ^[19] The tasks differ between languages, so these are not exact comparisons, but the pattern is plain: a few builders treated Burmese as a priority, and others did not.

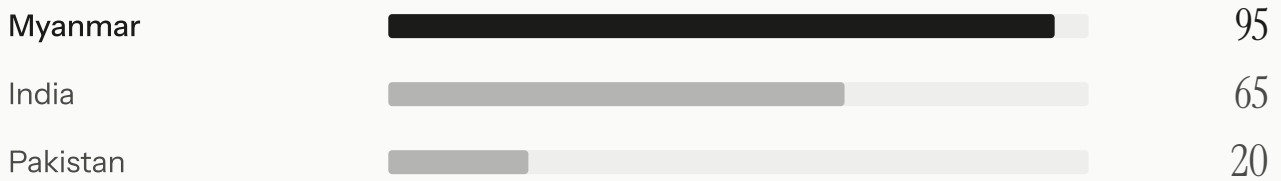
Burmese also costs more to run. Models read text as tokens, and a tokenizer built mostly from English breaks Burmese into many small pieces. Petrov and colleagues found that the tokenizer used by GPT-4 needs 11.7 times as many tokens for a Burmese sentence as for the same sentence in English, one of the highest ratios of any language they measured. Tokenizers designed for many languages, such as mT5's, need about 1.6 times. ^[20] Every extra token costs money and fills the model's limited context, so the same number of tokens carries "less than a tenth of the content" in Burmese. ^[20] The fix exists. It depends on whoever designs the tokenizer choosing to use it.

4 Why the text does not grow

More Burmese text would exist if more people in Myanmar could get online, stay online and publish. Since 2021, each of those has become harder.

Myanmar recorded more internet shutdowns than any other country in 2024, with 85, ^[21] and again in 2025, with 95. Since 2021, shutdowns have reached all 14 of its states and regions. ^[22] Freedom House reports that mobile networks must block every website not on an approved list, and that a block on VPNs in 2024 also disrupted services from Google and Amazon. ^[23] Text that is never typed never reaches a training set.

FIGURE 3

Internet shutdowns in 2025, the three countries with the most

Access Now, internet shutdowns in 2025, March 2026. Incidents documented during the year.

Even counting who is online has become difficult. The official internet use figures that the World Bank publishes from the International Telecommunication Union stop at 45 percent for Myanmar in 2020, with nothing since. The same series puts Thailand at 91 percent and Viet Nam at 84 percent in 2024.^[24] The World Bank's own estimate is that 44 percent of Myanmar's people were online in January 2025.^[25]

Power is the next limit. Myanmar generated an average of 3,034 MW between January and September 2025, against an installed capacity of 6,880 MW.^[25] In June 2026 the World Bank described generation of roughly 3,000 MW a day, covering only half of demand, and firms surveyed in October 2025 reported a median outage of four hours.^[26] Training and serving models needs steady power, and so does every office, school and phone that would produce Burmese text.

The economy that would pay for all of this has shrunk. Real GDP fell 9.0 percent in fiscal year 2020/21 and 12.0 percent in 2021/22.^[25] The World Bank expects it to stay about 11 percent below its level before the pandemic,^[27] and estimates that 29.9 percent of people lived in poverty in 2025.^[26] Myanmar has been on the Financial Action Task Force's list of high risk jurisdictions since October 2022,^[28] a listing the World Bank says raises the cost of trade finance and capital flows, and the authorities limit payments abroad that are not for trade.^[25, 26]

5 The people who would build it

AI is built by people, and the supply of people has been cut at every stage. Myanmar's schools were fully closed for 532 days between February 2020 and February 2022, the longest closure in East Asia and the Pacific. In 2021, about 30 percent of teachers were dismissed. The share of young people aged 6 to 22 in education fell from 69.2 percent in 2017 to 56.8 percent in 2023. ^[29] ISP-Myanmar, a research group, reports that candidates for the matriculation exam fell from over 900,000 in 2020 to roughly 250,000 in 2026. ^[30]

Research output shows the same break. Myanmar published 274 scientific and technical journal articles in 2019 and 234 in 2023. Over the same years Viet Nam went from 5,684 to 10,644, Thailand from 12,304 to 16,656, and Cambodia from 137 to 201. ^[31] In 2019 Viet Nam published about 21 articles for every one from Myanmar. By 2023 it published about 45.

FIGURE 4

Scientific and technical journal articles, 2023

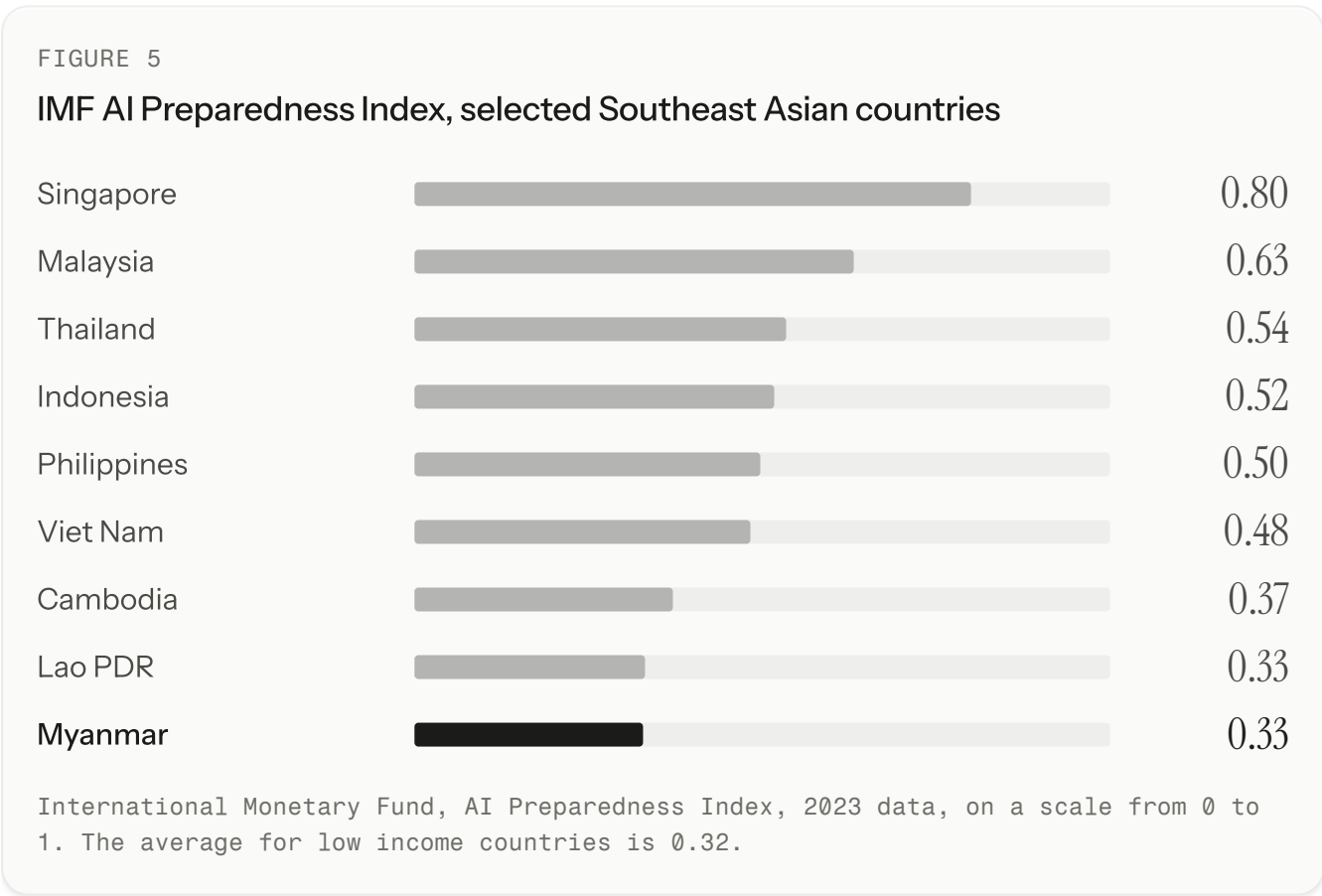


World Bank, World Development Indicators, from the US National Science Foundation. Fractional counts, rounded. Each row notes the 2019 count.

The people trained before 2021 are leaving. In a World Bank survey of 2,400 employed university graduates in early 2024, 35 percent said they would take a similar job abroad for the same pay. ^[32] The World Bank's firm survey found resignations linked to migration rising to 16 percent in March 2026, and 22 percent in Yangon, with 53

percent of those who resigned going abroad. ^[26] After a conscription law took effect in February 2024, entries from Myanmar into Thailand in March and April were more than three times those of a year before. ^[33]

International indices bring these pieces together. Oxford Insights ranks Myanmar 173rd of 195 countries in its 2025 Government AI Readiness Index, below every Southeast Asian country except Timor-Leste. ^[34] The IMF's AI Preparedness Index gives Myanmar 0.33, the lowest score in Southeast Asia and about the average for low income countries. ^[35]



6 Neighbours are building their own

While this happened, Myanmar's neighbours began building AI for their own languages. Singapore announced a S\$70 million national programme for multimodal language models in December 2023, ^[36] and AI Singapore's SEA-LION models cover eleven languages used in the region, Burmese among them. ^[11] In Thailand, SCB 10X

released the Typhoon models from late 2023, ^[37] and a government backed model, ThaiLLM, opened to the public in 2026. ^[38] In Viet Nam, VinAI released PhoGPT, trained from scratch on 102 billion Vietnamese tokens, in 2023. ^[39] Indonesia's Sahabat-AI launched in November 2024, ^[40] and Malaysia's prime minister launched ILMU in August 2025. ^[41]

We found no comparable programme for Burmese. That does not prove there is none, but the Burmese work we could find is small and led by universities and individuals. The Asian Language Treebank was built with the University of Computer Studies, Yangon, ^[42] which also produced a Myanmar and English corpus of about 200,000 sentence pairs. ^[43] Ye Kyaw Thu released open tools for splitting Burmese text into words, ^[44] and developers published community models such as MyanmarGPT and Burmese-GPT on Hugging Face in 2023 and 2024. ^[45, 46]

The money is moving the other way. Private investment in AI reached 344.7 billion US dollars in 2025, and 285.9 billion of it went to companies in the United States. Singapore, with 1.82 billion, was the only Southeast Asian country among the top 15. ^[47] The World Bank reports that high income countries account for 87 percent of notable AI models and 91 percent of AI venture capital, with 17 percent of the world's people. ^[48]

7 Why the gap compounds

None of this would be urgent if the gap stayed the same size. It does not, because AI improves where it already has users, data and builders, and each of those feeds the others. A language with more text gets better models. Better models attract more users. More users produce more text, and more reasons to build the next model. A language that starts at the bottom of that loop falls further behind each year it stays there.

International institutions describe the same risk. In January 2024 the head of the IMF wrote that many countries "don't have the infrastructure or skilled workforces to harness the benefits of AI, raising the risk that over time the technology could worsen inequality among nations." ^[49] A 2025 UNDP report names Myanmar among the

countries that "risk missing early dividends" from AI because of gaps in skills, investment, power and connectivity.^[50] Oxford Insights now separates the model makers, the few countries with the resources and talent to build frontier AI, from the model takers.^[34]

Part of the distance can already be measured as a trend. Between 2019 and 2023, Viet Nam's research articles for each one from Myanmar went from about 21 to about 45.^[31] Myanmar's recorded shutdowns went from 37 in 2023, second in the world, to 95 in 2025, first.^[22,51] Between 2023 and 2026, five neighbours put language models of their own into public use. If these trends hold, the gap in 2030 will not be today's gap with a few years added. It will be the gap between countries that shape AI in their own language and a country that waits for others to do it.

The loop can also run the other way, and the evidence shows the gap is not fixed. Tokenizers designed for many languages bring the cost of Burmese down from 11.7 times English to about 1.6.^[20] Google's Gemma 4 31B already scores within two points of Thai in Burmese.^[19] A good open Burmese model makes Burmese products cheaper to build, and those products bring more Burmese speakers online in their own language, which produces the text the next model needs.

8 What would change the curve

The work that would change the curve is concrete, and much of it does not need a data centre.

- **Burmese text, in Unicode and open.** Convert legacy Zawgyi text, clean it, record where it came from, and publish it under open licences, so every model builder starts from the same better data.
- **Public Burmese evaluations.** Test sets such as Belebele and BURMESE-SAN already exist.^[1,18] Running them on every new model, with native speakers reviewing the answers, turns progress into something measured instead of claimed.

- **Models that fit the conditions.** Where power and connections fail, a smaller open model that runs on an ordinary computer without a connection is worth more than a larger one in a data centre abroad.
- **Burmese speakers as builders.** Writing data, judging answers and fixing mistakes needs people who read Burmese, inside Myanmar and abroad, more than it needs hardware.

This is the work Arckara was started to do. Our first model, Arc 1.0, will be open source and is expected in December 2026, and we will publish how it is evaluated, including where it falls short.

We are already far behind. If we do nothing, today is the closest we will ever be.

References

- [1] Thura Aung, Jann Railey Montalan, Jian Gang Ngui and Peerat Limkonchotiwat. BURMESE-SAN: Burmese NLP Benchmark for Evaluating Large Language Models. LREC 2026. arxiv.org/abs/2602.18788
- [2] Catherine Shu. Google Translate Adds 10 More Languages, Including Burmese. TechCrunch, 11 December 2014. techcrunch.com/2014/12/11/google-translate-adds-10-more-languages
- [3] Google Cloud. Google models, supported languages. Vertex AI documentation, updated 3 September 2026. cloud.google.com/vertex-ai/generative-ai/docs/models
- [4] OpenAI. How to change your language setting in ChatGPT. OpenAI Help Center, retrieved 14 September 2026. help.openai.com/en/articles/8357869-how-to-change-your-language-setting-in-chatgpt
- [5] NLLB Team. No Language Left Behind: Scaling Human-Centered Machine Translation. Meta AI, 2022. arxiv.org/abs/2207.04672

- [6] Google. Gemini Apps Help, supported languages and countries. Retrieved 14 September 2026.
support.google.com/gemini/answer/13575153
- [7] Meta. Llama 4 model card. 5 April 2025.
github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md
- [8] Common Crawl. Distribution of languages, crawl CC-MAIN-2026-34. Common Crawl statistics, retrieved 14 September 2026.
commoncrawl.github.io/cc-crawl-statistics/plots/languages
- [9] Linting Xue and others. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. NAACL 2021.
arxiv.org/abs/2010.11934
- [10] Guilherme Penedo and others. FineWeb2: One Pipeline to Scale Them All. 2025.
arxiv.org/abs/2506.20920
- [11] AI Singapore. SEA-LION v1 7B model card. Hugging Face, 2023.
huggingface.co/aisingapore/SEA-LION-v1-7B
- [12] Wikimedia. List of Wikipedias, active users by language. Retrieved 14 September 2026.
meta.wikimedia.org/wiki/List_of_Wikipedias
- [13] Nick LaGrow and Miri Pruzan. Integrating autoconversion: Facebook's path from Zawgyi to Unicode. Engineering at Meta, 26 September 2019.
engineering.fb.com/2019/09/26/android/unicode-font-converter
- [14] AFP. Myanmar switches to international Unicode on October 1. Published by The Thaiger, 27 September 2019.
thethaiger.com/news/regional/myanmar/myanmar-switches-to-international-unicode-on-october-1
- [15] Sneha Kudugunta and others. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. 2023.
arxiv.org/abs/2309.04662
- [16] Alexis Conneau and others. Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020. The CC-100 corpus files.
data.statmt.org/cc-100
- [17] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali and Monojit Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. ACL 2020.
arxiv.org/abs/2004.09095
- [18] Lucas Bandarkar and others. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. ACL 2024.
arxiv.org/abs/2308.16884
- [19] AI Singapore. SEA-HELM leaderboard, version of 5 August 2026. Retrieved 14 September 2026.
leaderboard.sea-lion.ai

- [20] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr and Adam Bibi. Language Model Tokenizers Introduce Unfairness Between Languages. NeurIPS 2023.
arxiv.org/abs/2305.15425
- [21] Access Now. Emboldened offenders, endangered communities: internet shutdowns in 2024. February 2025.
www.accessnow.org/wp-content/uploads/2025/02/KeepItOn-2024-Internet-Shutdowns-Annual-Report.pdf
- [22] Access Now. Rising repression meets global resistance: internet shutdowns in 2025. March 2026.
www.accessnow.org/wp-content/uploads/2026/03/KeepItOn-Internet-Shutdowns-2025-Annual-Report.pdf
- [23] Freedom House. Myanmar: Freedom on the Net 2024.
freedomhouse.org/country/myanmar/freedom-net/2024
- [24] World Bank. World Development Indicators: individuals using the Internet, percent of population, from the International Telecommunication Union. Updated 13 July 2026.
data.worldbank.org/indicator/IT.NET.USER.ZS?locations=MM-TH-VN
- [25] World Bank. Myanmar Economic Monitor, December 2025.
documents1.worldbank.org/curated/en/099120625204042781/pdf/P507203-7c4662b6-c1d8-4c3c-9b0c-d4835f2763cb.pdf
- [26] World Bank. Myanmar Economic Monitor, June 2026.
documents1.worldbank.org/curated/en/099061526075033335/pdf/P507203-a497789a-01dd-4d79-bb47-faaf34d61123.pdf
- [27] World Bank. Myanmar's economy shows tentative stabilization but fuel shock intensifies pressures. Press release, 16 June 2026.
www.worldbank.org/en/news/press-release/2026/06/16/myanmar-s-economy-shows-tentative-stabilization-but-fuel-shock-intensifies-pressures
- [28] Financial Action Task Force. High-Risk Jurisdictions subject to a Call for Action, June 2026.
www.fatf-gafi.org/en/publications/High-risk-and-other-monitored-jurisdictions/call-for-action-june-2026.html
- [29] World Bank. State of Education in Myanmar. July 2023.
thedocs.worldbank.org/en/doc/716418bac40878ce262f57dfbd4eca05-0070012023/original/State-of-Education-in-Myanmar-July-2023.pdf
- [30] ISP-Myanmar. Six Million: Half of Myanmar's Students Are Shut Out of School. 24 June 2026.
ispmyanmar.com/sb2026-01
- [31] World Bank. World Development Indicators: scientific and technical journal articles, from the National Science Foundation. Retrieved 14 September 2026.
data.worldbank.org/indicator/IP.JRN.ARTC.SC?locations=MM-VN-TH-KH
- [32] Ghorpade, Imtiaz and Han. High-Skilled Migration from Myanmar. World Bank Policy Research Working Paper 10878, August 2024.
documents1.worldbank.org/curated/en/099807008212435597/pdf/IDU-c61abeb6-7dba-4323-9f19-3f7b21feed4d.pdf

- [33] International Organization for Migration, Thailand. Myanmar migrants in Thailand, situation brief. January 2025.
thailand.iom.int/sites/g/files/tmzbd11371/files/documents/2025-03/myanmar_migrants_thailand_jan25_final-1.pdf
- [34] Oxford Insights. Government AI Readiness Index 2025. January 2026 edition.
oxfordinsights.com/wp-content/uploads/2026/01/2025-Government-AI-Readiness-Index-Report_01_26.pdf
- [35] International Monetary Fund. AI Preparedness Index, 2023 data. IMF DataMapper, retrieved 14 September 2026.
www.imf.org/external/datamapper/AI_PI@AIPI/ADVEC/EME/LIC
- [36] GovInsider. Report on Singapore's S\$70 million National Multimodal LLM Programme. 4 December 2023.
govinsider.asia/intl-en/article/singapore-to-channel-usdollar52-million-into-building-capacities-for-seas-first-regional-llm
- [37] Kunat Pipatanakul and others. Typhoon: Thai Large Language Models. 2023.
arxiv.org/abs/2312.13951
- [38] National Science and Technology Development Agency, Thailand. ThaiLLM announcement. April 2026.
www.nstda.or.th/en/news/news-years-2026/thai11m.html
- [39] Dat Quoc Nguyen and others. PhoGPT: Generative Pre-training for Vietnamese. 2023.
arxiv.org/abs/2311.02945
- [40] GoTo. Indosat Ooredoo Hutchison and GoTo launch Sahabat-AI, Indonesia's open source LLM. Press release, 14 November 2024.
www.gotocompany.com/en/news/press/indosat-ooredoo-hutchison-and-goto-launch-sahabat-ai-indonesias-open-source-llm-for-empowering-digital-sovereignty
- [41] Bernama. Report on the launch of ILMU by the Prime Minister of Malaysia. 12 August 2025.
www.bernama.com/en/news.php?id=2455958
- [42] National Institute of Information and Communications Technology, Japan. Asian Language Treebank (ALT) Project.
www2.nict.go.jp/astrec-att/member/mutiyama/ALT
- [43] Toshiaki Nakazawa and others. Overview of the 5th Workshop on Asian Translation. WAT 2018.
aclanthology.org/Y18-3001.pdf
- [44] Ye Kyaw Thu. myWord: syllable, word and phrase segmenter for Burmese. GitHub, 2021.
github.com/ye-kyaw-thu/myWord
- [45] Min Si Thu. MyanmarGPT. Hugging Face, December 2023.
huggingface.co/jojo-ai-mst/MyanmarGPT
- [46] Wai Yan Nyein Naing. Burmese-GPT. Hugging Face, January 2024.
huggingface.co/WYNN747/Burmese-GPT

- [47] Stanford Institute for Human-Centered Artificial Intelligence. AI Index Report 2026, chapter 4: Economy.
hai.stanford.edu/assets/files/ai_index_report_2026_chapter_4_economy.pdf
- [48] World Bank. Strengthening AI Foundations: Emerging Opportunities for Developing Countries. Factsheet, 21 November 2025.
www.worldbank.org/en/news/factsheet/2025/11/21/strengthening-ai-foundations-emerging-opportunities-for-developing-countries
- [49] Kristalina Georgieva. AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity. IMF Blog, 14 January 2024.
www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity
- [50] United Nations Development Programme. The Next Great Divergence: Why AI May Widen Inequality Between Countries. December 2025.
www.undp.org/sites/g/files/zskgke326/files/2025-12/undp-rbap-the-next-great-divergence.pdf
- [51] Access Now. Shrinking democracy, growing violence: internet shutdowns in 2023. May 2024.
www.accessnow.org/wp-content/uploads/2024/05/2023-KIO-Report.pdf