

Evaluating Large Language Models in Mongolian

Dorjnyam Tumur-Ochir Fei Cheng Yugo Murawaki Chenhui Chu
Kyoto University

dorjnyam@nlp.ist.i.kyoto-u.ac.jp
{feicheng, murawaki, chu}@i.kyoto-u.ac.jp

Abstract

This paper presents a comprehensive evaluation for assessing large language model (LLM) capabilities in the Mongolian language, addressing a critical gap in multilingual LLM evaluation. We introduce MonMLU, a novel benchmark derived from native-level university entrance exams, alongside Mongolian adaptations of established benchmarks including Vicuna, MT-Bench, MGSM, and XCOPA. Our evaluation of leading commercial and openly available models reveals that while GPT-4o-mini achieves the highest performance (8.86 on Vicuna, 8.10 on MT-Bench), openly available models significantly underperform. These findings highlight future opportunities for improving LLM performance in Mongolian and other low-resource languages.

1 Introduction

The advancements in large language models (LLMs) have been especially significant in high-resource languages such as English and Chinese [1, 2]. These models have demonstrated exceptional capabilities in text generation, classification, understanding, and reasoning tasks [3]. While many state-of-the-art models are multilingual, their effectiveness in low-resource languages often lags due to limited training data and benchmarks. Multilingual benchmarks such as MMLU and MGSM have been developed to address these gaps by extending the original English benchmarks to various languages [4, 5].

The Mongolian language, with over 6 million speakers across Mongolia and Inner Mongolia (China), represents a significant yet underserved language [6]. Digital resources in Mongolian comprise less than 0.02% of Common Crawl data, highlighting the substantial disparity in representation compared to high-resource languages. This underrepresentation necessitates dedicated evaluation frameworks to

assess and advance LLM capabilities in Mongolian.

To address the resource gap in Mongolian, we develop and release MonMLU,¹⁾ a multi-subject multiple-choice language understanding benchmark based on native entrance examination materials in Mongolian. We adapt the widely used English benchmarks such as Vicuna and MT-Bench [7] to evaluate Mongolian rapidly. We show that current models can generate intelligible responses and follow simple instructions in Mongolian. In addition, to measure reasoning and understanding abilities, MGSM [5] and XCOPA[8] are adapted.

The main contributions of our work include:

- MonMLU, a comprehensive benchmark based on native Mongolian university entrance exams, testing cultural and linguistic knowledge across eight subjects.
- Mongolian-specific versions of established benchmarks (Vicuna, MT-Bench, MGSM, XCOPA) through translation and localization.
- We evaluate leading commercial and open-source LLMs, providing a quantitative assessment of their Mongolian language capabilities.

2 Related Work

2.1 Monolingual benchmarks

Monolingual benchmarks play a crucial role in assessing the performance of large language models (LLMs) within a specific language. Benchmarks such as MMLU, MT-Bench, and GSM8k are widely used to evaluate LLMs for reasoning, conversational ability, and mathematical problem-solving tasks. These benchmarks are designed to test models on diverse, challenging tasks and have significantly advanced the capabilities of LLMs for high-resource languages like English.

1) <https://huggingface.co/ku-nlp/monmlu>

For instance, MMLU assesses knowledge across 57 subjects, including humanities, sciences, and general knowledge. MT-Bench evaluates models' multi-turn conversation abilities, ensuring context and coherence across interactions are maintained. Similarly, GSM8k focuses on grade school math word problems, testing models' mathematical reasoning and problem-solving skills.

However, such comprehensive benchmarks have not been developed for low-resource languages like Mongolian. The lack of Mongolian benchmarks limits the ability to evaluate models comprehensively and hinders progress in improving LLMs for the language. Developing tailored benchmarks for the Mongolian language is critical for understanding and enhancing model performance in this context.

2.2 Multilingual benchmarks

Efforts to extend evaluations beyond high-resource languages have led to the development of multilingual benchmarks. These benchmarks aim to assess LLM performance across a wide range of languages, providing insights into multilingual and cross-lingual capabilities. Key examples include:

- **MMMLU (Multilingual MMLU):** An extension of MMLU, MMMLU provides evaluations in 14 languages by translating the original English dataset.
- **XCOPA (Cross-lingual Choice of Plausible Alternatives):** Derived from the original English COPA benchmark, XCOPA evaluates commonsense reasoning in multiple languages by asking models to select the most plausible outcome or cause of a given scenario.
- **MGSM (Multilingual Grade School Math):** This is a multilingual adaptation of the GSM8k dataset, containing grade school-level math problems translated into various languages. MGSM assesses the ability to solve basic mathematical problems within a multilingual context.

These multilingual benchmarks have significantly broadened the scope of LLM evaluations. Nevertheless, Mongolian is often underrepresented, or even absent, in many existing multilingual benchmarks. This underscores the need for targeted evaluation datasets that reflect the linguistic and cultural specifics of the language.

2.3 Low-resource language benchmarks

Low-resource languages are frequently incorporated into massively multilingual benchmarks such as SIB-200 [9] and Taxi-200 [10]. However, massively multilingual benchmarks often prioritize breadth of language coverage over data quality, particularly for low-resource languages. Furthermore, they are often limited to classification tasks, neglecting other important and diverse tasks such as reasoning and generation.

3 Mongolian Language

Mongolian, a member of the Mongolic language family, presents a significant yet underrepresented language in natural language processing. It is primarily spoken in not just Mongolia, but also in some parts of China (Inner Mongolia) and Russia, with approximately 5-6 million speakers. The language employs diverse writing systems, including the traditional Mongolian and Cyrillic scripts. A Romanized transliteration of the Cyrillic script is also common in informal online communication. This study focuses on Mongolian written in Cyrillic, the most prevalent form. Mongolian presents unique challenges due to its complex linguistic features such as agglutinative morphology and vowel harmony. These linguistic characteristics, coupled with limited digital resources, make Mongolian a compelling case study for enhancing LLM performance in low-resource language settings.

4 Benchmarks

To overcome the scarcity of evaluation resources for Mongolian, we employ a two-pronged approach: adapting existing multilingual benchmarks and creating novel benchmarks tailored specifically for the language. These benchmarks comprehensively evaluate Natural Language Understanding (NLU) and Natural Language Generation (NLG) tasks. This section introduces and discusses the following benchmarks: Mongolian Vicuna and MT-Bench, MonGSM and MonCOPA, and MonMLU. See Appendix A.2 for samples.

4.1 Adapting existing benchmarks

To adapt Vicuna, MT-Bench, MGSM, and XCOPA for Mongolian, we followed a machine translation, post-editing, and localization approach:

Table 1: Summary of benchmarks.

Benchmark	Description	Task type	Size	Construction method
Mongolian Vicuna	Single-turn dialogue	NLG	80	Translation
Mongolian MT-Bench	Multi-turn dialogue	NLG	80	Translation
MonGSM	Math word problems	NLU (Reasoning)	250	Translation
MonCOPA	Commonsense reasoning	NLU (Reasoning)	500	Translation
MonMLU	General knowledge	NLU (Knowledge)	400	Entrance Examination Collection

- Machine translation:** We used GPT-4o to translate the original English benchmarks into Mongolian.
- Manual post-editing:** A native Mongolian speaker manually edited the translations to ensure linguistic accuracy and cultural appropriateness.
- Localization:** Nouns and named entities such as persons and places were replaced by Mongolian names to make their context more native. For example, robe to ᠳᠠᠨᠢᠭᠡᠨᠠᠯ , Hawaii to ᠬᠡᠪᠴᠢᠭᠠᠯ .

Mongolian Vicuna and MT-Bench To quickly evaluate models in the Mongolian language, we adapted the widely used Vicuna and MT-Bench, which are widely used benchmarks for evaluating conversational abilities in LLMs. These benchmarks assess the capacity to generate coherent, contextually relevant, and logically consistent responses in single-turn and multi-turn dialogues.

MonGSM and MonCOPA MonGSM is a Mongolian adaptation of MGSM dataset, which consists of grade school-level math word problems. These problems test basic arithmetic, logical reasoning, and problem-solving abilities. XCOPA is a benchmark for evaluating models’ ability to transfer commonsense reasoning across languages.

These benchmarks highlight the reasoning capabilities of models in Mongolian and offer a comparative perspective on their performance relative to other languages.

4.2 Constructing native MonMLU

MonMLU is a benchmark specifically designed to evaluate general knowledge and reasoning abilities in the Mongolian language. We obtained test questions from the Mongolian General Entrance Examination and received permission from the relevant institution,²⁾ making it a reliable source for testing models on diverse subjects. It covers eight subjects: Mongolian language, Mongolian history, Biology, Chemistry, Physics, and Social Science. Each subject contains 50 multiple-choice questions with

five choices.

We obtained exam questions and answers, publicly available on 2), in PDF format. We manually extract the text from the exams, excluding images and tables to focus on unimodal models.

MonMLU evaluates models on the accuracy of responses, the depth of knowledge across various subjects, and the ability to understand complex queries and provide contextually correct answers.

Table 1 shows the summary of all benchmarks. These benchmarks collectively provide a comprehensive evaluation suite for the Mongolian language, covering conversational skills, reasoning, and general knowledge.

5 Evaluation

5.1 Settings

Models For our evaluation, we selected the leading commercial and openly available models. For commercial models, we use a family of OpenAI models, namely GPT-3.5-Turbo, GPT-4-Turbo, and GPT-4o-mini.³⁾ Openly available models are Gemma-2-9B and 27B,⁴⁾ Llama-3.1 8B,⁵⁾ Mistral 7B,⁶⁾ EMMA-500,⁷⁾ and Qwen2.5 7B.⁸⁾ We also include versions of Gemma-2-9B and Llama-3.1-8B, Gemma-2-9B-Alpaca-MN and Llama-3.1-8B-Alpaca-MN respectively, fine-tuned on the original Alpaca instruction dataset that is machine-translated to Mongolian.⁹⁾

Evaluation method We implemented a consistent evaluation protocol across all benchmarks. For reasoning

3) <https://platform.openai.com/docs/models>

4) <https://huggingface.co/google/gemma-2-9b-it>

5) <https://huggingface.co/meta-llama/Llama-3.1-8B>

6) <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

7) <https://huggingface.co/MaLA-LM/emma-500-llama2-7b>

8) <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

9) <https://huggingface.co/datasets/sailab/alpaca-mongolian-cleaned>

2) <https://eec.mn/>

Table 2: Performance of various models across NLG and NLU tasks.

	NLG		NLU		
	Vicuna	MT-Bench	MonGSM	MonMLU	MonCOPA
GPT-4o-mini	8.86	8.10	68.0	52.3	61.0
GPT-4-Turbo	8.49	7.89	74.4	51.1	59.8
GPT-3.5-Turbo	2.95	3.22	17.6	18.0	15.3
Gemma-2-27B-IT	7.01	6.50	59.2	43.7	49.6
Gemma-2-9B-IT	4.22	3.96	46.4	32.2	38.1
Gemma-2-9B-Alpaca-MN	5.95	4.76	14.0	21.6	23.4
Llama-3.1-8B-Instruct	1.98	2.40	34.0	21.1	23.5
Llama-3.1-8B-Alpaca-MN	5.47	3.99	4.0	17.5	16.7
Mistral-7B-Instruct-v0.3	1.38	1.10	4.0	0.0	10.3
EMMA-500-llama2-7b	3.96	2.54	0.0	4.0	0.0
Qwen-2.5-7B-Instruct	2.50	2.64	12.4	23.1	15.6

tasks such as MonGSM, we employed a zero-shot chain-of-thought setting to assess natural problem-solving capabilities. Conversational tasks, Vicuna and MT-Bench, were evaluated using GPT-4o as an independent judge as proposed by [11]. Multiple-choice questions were scored using exact match criteria.

5.2 Results

As shown in Table 2, our evaluation revealed significant performance variations across models and tasks. Among commercial models, GPT-4o-mini demonstrated superior performance in both NLG and NLU tasks, achieving scores of 8.86 on Vicuna and 52.3 on MonMLU. GPT-4-Turbo followed closely with strong results of 8.49 on Vicuna and 51.1 on MonMLU. However, GPT-3.5-Turbo showed notably lower performance, scoring only 2.95 on Vicuna and 18.0 on MonMLU.

In the openly available category, Gemma-2-27B-IT achieved the strongest results among non-commercial models, scoring 7.01 on Vicuna and 6.50 on MT-Bench. The 9B variant also shows reasonable performance. Smaller models demonstrated limited effectiveness, as evidenced by Mistral-7B and Llama-3.1-8B-Instruct.

As illustrated by Figures 1 and 2, fine-tuned models on Alpaca, most notably Llama-3.1-8B, show considerable performance improvement in generation tasks in categories such as writing and roleplay. At the same time, they lose their performance in categories such as extraction, coding, and math. However, we can see a huge performance drop in NLU tasks. This may be due to a lack of task-specific training data as Alpaca contains mostly generic instructions.

6 Conclusion

In this work, we introduced a set of benchmarks for the Mongolian language. Four benchmarks were adapted from widely used English and multilingual benchmarks including MT-Bench and MGSM. MonMLU, a multi-subject multiple-choice benchmark based on native examination material, was constructed. We evaluated various types of models on these newly constructed benchmarks. Openly available models lag behind commercial models considerably. Also, simply fine-tuning models with additional Mongolian instruction data, while improving text coherence in generated output, does not enhance understanding and reasoning in the Mongolian language. In future work, we plan to explore methods aimed at simultaneously improving both generative and understanding capabilities.



Figure 1: Category-wise scores of various models on Mongolian Vicuna.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23K28144.

References

- [1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- [2] 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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- [4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2020.
- [5] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022.
- [6] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics.
- [7] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- [8] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics.
- [9] David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects, 2024.
- [10] Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schütze. Taxi1500: A multilingual dataset for text classification in 1500 languages, 2024.
- [11] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

A Appendix

A.1 MT-Bench

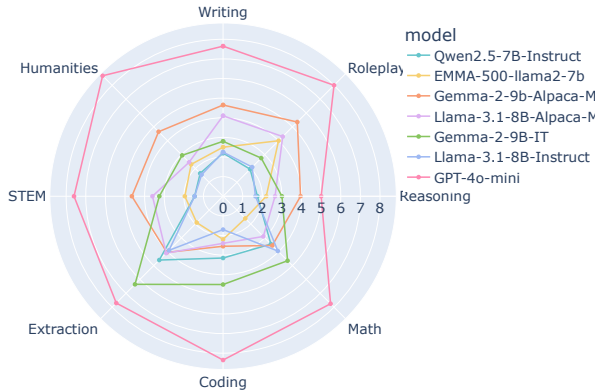


Figure 2: Category-wise scores of various models on Mongolian MT-Bench.

A.2 Samples

Vicuna

Question:
Цаг төлөвлөлтөө хэрхэн сайжруулах вэ?
Answer:
Цаг төлөвлөлтөө сайжруулах нь таны ажлын үр productivity, амьдралын чанарыг дээшлүүлэхэд ихээхэн тусална. Энд хэд хэдэн үр дүнтэй аргууд байна:

- Зорилго тодорхойлох:** Эхлээд та урт хугацааны болон богино хугацааны зорилгуудаа тодорхойлж, тэдгээрийг бичиж аваарай. Энэ нь таны цагийг ямар зүйлийн төлөө зарцуулж байгааг ойлгоход тусална.
- Prioritization:** Ажлуудаа чухал байдлаар нь ангилаарай. Эдгээрийг "Eisenhower Matrix" ашиглан ангилах нь үр дүнтэй. Энэ арга нь ажлуудыг "чухал", "яаралтай", "чухал биш", "яаралтай биш" гэсэн дөрвөн ангилалд хуваахад тусална.
- Цагийн хуваарь гаргах:** Өдөр тутмын, долоо хоногийн, сарын хуваарийг гаргаж, түүгээрээ ажиллах. Хуваарийг гаргахдаа чухал ажлуудаа анхнаасаа оруулж, бусад ажлуудаа түүнд тохируулан зохион байгуулаарай.
- Технологи ашиглах:** Тохиромжтой програм, аппликейшн ашиглаарай. Жишээлбэл, Google Calendar, Todoist, Trello гэх мэт. Эдгээр нь цаг төлөвлөлт, хуваарь гаргахад ихээхэн тусалдаг.

Figure 3: A sample instruction and response pair of GPT-4o-mini on Mongolian Vicuna.

MT-Bench

Turn 1:
Саяхан Хөвсгөлд хийсэн аяллынхаа талаар соёлын өвүүд болон заавал үзэх ёстой газруудыг онцлон харуулсан сонирхолтой аялын блог нийтлэл бич. (Compose an engaging travel blog post about a recent trip to **Khuvsgul**, highlighting cultural experiences and must-see attractions.)

Turn 2:
Өмнөх хариултаа дахин бичнэ үү. Бүх өгүүлбэрийг **X** үсгээр эхлүүлнэ үү. (Rewrite your previous response. Start every sentence with the letter **X**.)

Figure 4: A sample from MT-Bench. Words in red are localized words.

MonGSM

Нэг **дээл** хийхэд хөх өнгийн 2 боодол болон түүнээс хоёр дахин бага цагаан өнгийн даавуу ордог. Нэг дээл хийхэд нийт хэдэн боодол даавуу ордог вэ? 3. (A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take? 3.)

Figure 5: A sample from MonGSM. Words in red are localized words.

MonMLU

Эртний улсуудын мал сүргийн бүтэц өнөөгийн монголчуудтай адил бөгөөд хамгийн цөөн тоотой мал аль нь вэ? (The livestock structure of ancient countries is similar to that of modern Mongolians. Which livestock has the smallest number?)

A) Хонь (Sheep)
B) Тэмээ (Camel)
 C) Үхэр (Cattle)
 D) Адуу (Horse)
 E) Ямаа (Goat)

Figure 6: A sample from MonMLU. The correct answer is in bold.