Advancing the Evaluation of Traditional Chinese Language Models: Towards a Comprehensive Benchmark Suite

Chan-Jan Hsu*, Chang-Le Liu*, Feng-Ting Liao*, Po-Chun Hsu*, Yi-Chang Chen*, Da-shan Shiu
MediaTek Research

September 2023

Abstract

The evaluation of large language models is an essential task in the field of language understanding and generation. As language models continue to advance, the need for effective benchmarks to assess their performance has become imperative. In the context of Traditional Chinese, there is a scarcity of comprehensive and diverse benchmarks to evaluate the capabilities of language models, despite the existence of certain benchmarks such as DRCD, TTQA, CMDQA, and FGC dataset. To address this gap, we propose a novel set of benchmarks that leverage existing English datasets and are tailored to evaluate language models in Traditional Chinese. These benchmarks encompass a wide range of tasks, including contextual question-answering, summarization, classification, and table understanding. The proposed benchmarks offer a comprehensive evaluation framework, enabling the assessment of language models' capabilities across different tasks. In this paper, we evaluate the performance of GPT-3.5, Taiwan-LLaMa-v1.0, and Model 7-C-Chat, our proprietary model, on these benchmarks. The evaluation results highlight that Model 7-C-Chat achieves performance comparable to GPT-3.5 with respect to a part of the evaluated capabilities. In an effort to advance the evaluation of language models in Traditional Chinese and stimulate further research in this field, we have open-sourced our benchmark and opened the model for trial.

1 Introduction

The evaluation of large language models (LLMs) has long been a crucial task. With the advancement of technology, LLMs have become more sophisticated, providing higher-quality responses akin to human responses to open-ended questions. However, evaluating these models is challenging, and there is a need for well-designed benchmarks to assess their performance comprehensively and consistently. Existing English benchmarks such as MMLU [Hendrycks et al., 2021], IMDB [Maas et al., 2011], and XSum [Narayan et al., 2018] cover measurements of models' capabilities in question answering, sentiment classification, and summarization, respectively. In Traditional Chinese, while there exist some benchmarks such as Delta Reading Comprehension Dataset (DRCD) [Shao et al., 2019], Taiwanese Trivia Question Answering (TTQA) [Ennen et al., 2023], and Formosa Grand Challenge (FGC) dataset [STPI, 2020], there is limited availability of comprehensive and diverse benchmarks for evaluating language models' capabilities.

In this paper, to address the need for a comprehensive suite of evaluations in Traditional Chinese, we propose a set of new benchmarks. The benchmarks are built upon available Traditional Chinese and English datasets to test the capabilities of language models in Traditional Chinese. Our proposed benchmarks assess the capabilities of tasks related to contextual question answering, world knowledge, summarization, classification, table understanding. In terms of evaluating world knowledge, we further propose a new dataset - Taiwan Massive Multitask Language Understanding (TMMLU) - encompassing exams from high school entrance exams to vocational exams across 55 subjects in total.

We evaluate the performance of proprietary and open-source models, namely GPT-3.5, Taiwan-LLaMa-v1.0 [Lin and Chen, 2023a], and Model 7-C-Chat (our fine-tuned model for chatting capability), using our proposed Traditional Chinese benchmarks. Notably, our proposed benchmarks provide a comprehensive set of evaluation tasks for language models, allowing us to assess their performance on various tasks. For some of the evaluated capabilities, the evaluation outcomes demonstrate that Model 7-C-Chat matches the performance of the state-of-the-art GPT-3.5 model in Traditional Chinese.

To promote more research on advancing state-of-the-art language models in Traditional Chinese, we have open-sourced our benchmark code and relevant datasets and opened for trial of our proprietary model, Model 7-C-Chat, for comparison[1].

2 Related work

There exists a wealth of English benchmarks for evaluating different capabilities of language models. EluetherAI's Language Model Evaluation Harness [Gao et al., 2021] is a unified framework to test generative language models on a large number of different evaluation tasks. Holistic Evaluation of Language Models (HELM) [Liang et al., 2022] is an evaluation framework that consists of evaluations in 42 scenarios.

*These authors contributed equally to this work and are arranged in alphabetical order

[1] `https://github.com/mtkresearch/MR-Models`

BIG-bench [BIG-bench authors, 2023] is a collaborative benchmark designed to examine LLMs across diverse task topics ranging from linguistics and childhood development to software development and social bias. AGIEval [Zhong et al., 2023] is a benchmark tailored to assess models on human cognition and problem-solving, derived from 20 prominent admission and qualification exams including the Gaokao, SAT, law school tests, and civil service exams. These English benchmarks and the evaluations therein are commonly evaluated at the release of the models such as BLOOM [Scao et al., 2022], Pythia [Biderman et al., 2023], Falcon [Penedo et al., 2023], Llama (1 [Touvron et al., 2023b] and 2 [Touvron et al., 2023a]), and their fine-tuned variants.

As to the notable open benchmarks in Traditional Chinese, at the time of this writing (mid-August, 2023), we summarize them below. DRCD, a reading comprehension peer-reviewed dataset, contains 30k question-answer pairs based on Wikipedia articles. TTQA [Ennen et al., 2023], a trivia question-answering not-peer-reviewed dataset, consists of 64 expert-selected paragraphs from Wikipedia for testing a model's knowledge on Taiwanese-specific topics. Chinese Movie Dialogue Question Answering (CMDQA) [Luo et al., 2022], a dialogue-based information-seeking question-answering dataset, contains 10k QA dialogues (40k turns in total) about movie information parsed from Wikipedia. Formosa Grand Challenge (FGC) dataset is a passage question answering dataset of 750 samples created from Taiwanese news articles and government announcements.

Language models have been shown to provide responses akin to human responses to open-ended questions. The open-ended types of questions however cannot easily be mapped 1-on-1 to a single answer. At the time of this writing, notable evaluation benchmarks with GPT-4 as judge have been wildly adopted by the community, albeit its tendency to favour longer text and texts generated by LLMs [Lin and Chen, 2023a, Liu et al., 2023]. Vicuna [Chiang et al., 2023] consists of 80 questions spanning across 8 tasks. Similar to Vicuna, WizardLM [Xu et al., 2023] constructed a test set of 218 open-ended questions covering 29 areas such as writing, role-play, and philosophy. As for Traditional Chinese open-ended questions, a translated version of Vicuna benchmark is used to test Taiwan-LLaMa [Lin and Chen, 2023a,b].

3 Benchmark

Here we give a succinct introduction to each benchmark we will use in this study. We categorize the proposed set of benchmarks into capabilities. Table 1 lists the evaluation benchmarks used in this study and Appendix A shows some examples. As source datasets in Traditional Chinese are limited, we translate the listed English datasets to Traditional Chinese for the evaluation.

Capabilities	Evaluation Dataset	Source Language
Contextual QA	DRCD [Shao et al., 2019] FGC [STPI, 2020]	Traditional Chinese Traditional Chinese
World Knowledge	TTQA [Ennen et al., 2023] TMMLU (ours)	Traditional Chinese Traditional Chinese
Summarization	XSum-TC [Narayan et al., 2018]	English
Classification	IMDB-TC [Maas et al., 2011]	English
Table Understanding	Penguins-in-a-Table-TC [BIG-bench authors, 2023]	English

Table 1: The datasets and their respective nature for benchmarking capabilities in this study. We translate English datasets to Traditional Chinese for the evaluation, which is indicated by the "-TC" suffix.

3.1 Capabilities

Below are summaries of the benchmarked capabilities as listed in Table 1 and the corresponding datasets used in evaluating the respective capabilities in this study.

Contextual Question Answering is the task in which a model is given a contextual input and is asked to respond to a given question related to the input. This task is most similar to standard benchmarks in closed QA or common sense reasoning. DRCD is a Traditional Chinese machine reading comprehension dataset containing 10,014 paragraphs from 2,108 Wikipedia articles and over 30,000 questions. FGC dataset is a passage question answering dataset of 750 samples created from Taiwanese news articles and government announcements.

World Knowledge task requires a model to have a certain level of knowledge about the real world. TTQA is for assessing language models' common sense abilities on Taiwanese terms, comprising 64 passages from Wikipedia about diverse Taiwanese cultural topics, necessitating model comprehension and reasoning. Taiwan Massive Multitask Language Understanding (TMMLU) is curated from examinations in Taiwan, consisting of 55 subjects spanning across multiple disciplines, from vocational to academic fields, and covering elementary to professional proficiency levels. It is designed to identify a model's knowledge and problem-solving blind spots similar to human evaluations. See Appendix B for the list of subjects.

Summarization task requires a model to summarize a given passage in an abstract manner. Extreme Summarization (XSum) dataset evaluates abstractive summarization with 226,711 BBC news articles across diverse domains, aiming for one-sentence summaries.

Classification task is defined as requesting a model to determine the category of given input text, such as sentiment analysis and natural language inference. IMDB dataset offers binary sentiment classification with 25,000 polar movie reviews each for training and testing sentiment classifiers.

Table Understanding task evaluates a model's capacity to construct an accurate depiction of the data presented to it in both tabular and natural language formats, and its ability to identify and retrieve the pertinent details required to address a straightforward query. The "penguins in a table" task contained in BIG-bench asks a language model to answer questions about the animals contained in a table, or multiple tables, described in the context.

To assess the capability of the models, we adopt metrics from academic benchmarks like HELM. For evaluations in areas like Contextual QA, World Knowledge, Classification, and Table Understanding, we provide the prefix exact match (EM) scores. For exception, TTQA is formatted as multiple choice questions, so we report the accuracy. In terms of Summarization, ROUGE-2 is reported.

3.2 Capability of Answering Open-Ended Questions

To assess language models' ability to provide helpful answers to open-ended questions, we use TAIDE-14 [TAIDE, 2023]. TAIDE-14 consists of 14 different text generation tasks covering 50 topics and includes a total of 140 prompts specifically designed to evaluate Traditional Chinese LLMs. These prompts were created by GPT-4 using the provided task, domain, and keywords, and were further validated by human experts.
Inspired by [Fu et al., 2023, Lin and Chen, 2023a, Liu et al., 2023], we evaluate model responses using GPT-4 as the evaluator. We adopt an evaluation strategy similar to [Fu et al., 2023] where for each task we require GPT-4 to provide scores adhering to criterions of various aspects. The aspects for different types of tasks are as follows:

- **Tasks with the golden answer**: accuracy;
- **Extraction tasks**: consistency, conciseness;
- **Summarization tasks**: semantic coverage, consistency, conciseness[2];
- **General tasks**: relevance, depth, diversity, organization[3].

Details of the criteria of aspects for evaluating different tasks are presented in Appendix C. We observe that several samples of the summarization and extraction tasks in TAIDE-14 dataset come with blank input, and thus we re-categorize them as general tasks for evaluation.

4 Results

4.1 Models Compared

In this study, we analyze the performances of three models: GPT-3.5, Taiwan-LLaMa-v1.0 and Model 7-C-Chat. The version of GPT-3.5 utilized for this comparison is a snapshot titled GPT-3.5-Turbo-0613, dated June 13, 2023. Taiwan-LLaMa-v1.0, on the other hand, is a refined version of the Llama 2 model, configured for Traditional Chinese. It has been pre-trained on a dataset encompassing over 5 billion tokens and further fine-tuned using a rich set of more than 490,000 instruction-response samples.

4.2 Capabilities Benchmark Results

Table 2 illustrates the comparative performance of various models on designated datasets. We carry out all evaluation zero-shot and use greedy decoding for a fair comparison. It is evident that GPT-3.5 predominantly surpasses other models in benchmark tests spanning all assessed capabilities. Both Taiwan-LLaMa-v1.0 and Model 7-C-Chat manage to approximate GPT-3.5's performance in limited instances, exhibiting less than a 5% discrepancy in certain benchmarks. Specifically, Taiwan-LLaMa-v1.0 showcases parallel performance in the IMDB-TC benchmark, whereas Model 7-C-Chat is comparable in the DRCD, TTQA and XSum-TC benchmarks.

Notwithstanding, the table understanding task reveals a discernible deficiency in both Taiwan-LLaMa-v1.0 and Model 7-C-Chat, with frequent hallucinations evident in numerous samples. Moreover, summarization tasks delineated suboptimal results; even though the models were instructed to condense the context into a single concise sentence, they demonstrated low Rouge-2 scores universally. This underperformance was manifested as over-extended summaries in GPT-3.5 and Model 7-C-Chat, and occasional lack of summaries in Taiwan-LLaMa-v1.0. We assessed the XSum dataset and found the presence of summaries incorporating elements not delineated in the original documents, potentially a causal factor in the diminished performance metrics observed in the tasks.

[2] We use "conciseness" to evaluate whether a response is shorter than the original text and meets all word and sentence count conditions while remaining to-the-point.

[3] We use "organization" to evaluate whether a response is presented in a clear and easy-to-understand manner.

[4] We sub-sampled 5000 samples from the test set of the original XSum dataset for evaluation.

[5] Although some of the questions in this dataset have their options listed out, and can be considered as multiple choice, the model's output does not always fall into one of the choices, leading to a score lower than random guessing. We include "the ability to identify this task as a multiple choice" as part of the task objective and do not fix this "bug".

Capability tested	Dataset (metric)	Models		
		GPT-3.5	Taiwan-LLaMa	Model 7-C-Chat
Contextual QA	DRCD (EM)	0.771	0.719	0.765
	FGC (EM)	0.48	0.33	0.43
World Knowledge	TTQA (Accuracy)	1.00	0.81	1.00
	TMMLU (EM)	0.515	0.307	0.404
Summarization	XSum-TC (Rouge-2)[4]	0.032	0.001	0.039
Classification	IMDB-TC (EM)	0.941	0.929	0.916
Table Understanding	Penguins-in-a-Table-TC (EM)[5]	0.32	0.00	0.08

Table 2: The benchmark result of models.

4.3 Open-Ended Benchmark Results

We present the win-rate chart to demonstrate the capability of answering TAIDE-14 tasks, judged by GPT-4. Our proprietary model fine-tuned for chatting capability, Model7-C-Chat, matches GPT-3.5 on 98, out of all the 140, test samples. Though, Taiwan-LLaMa-v1.0 shows slightly better capability than Model 7-C-Chat on TAIDE-14. See Figure 1 for reference.

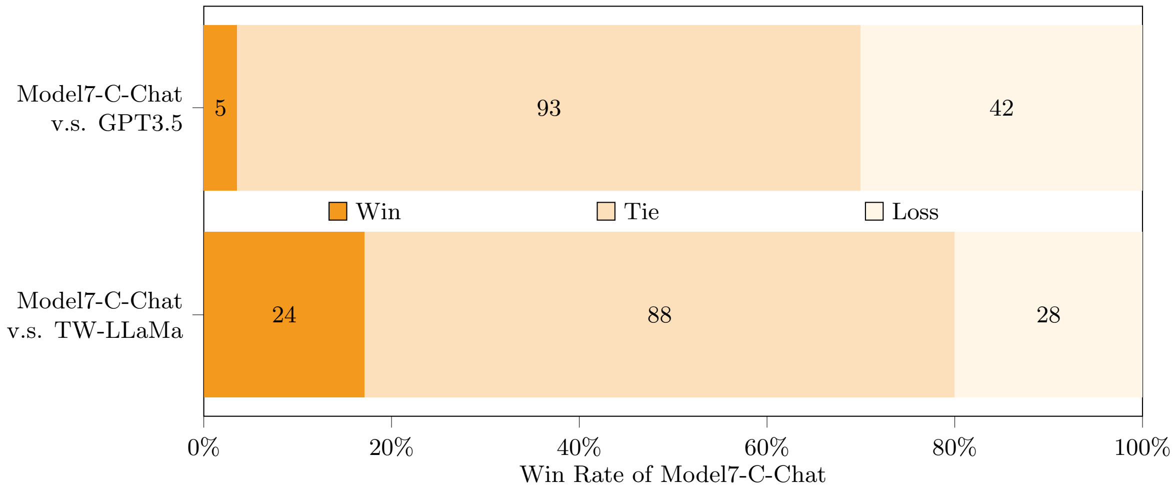

Figure 1: Win-rate between models on the TAIDE-14 benchmark. The win-rate chart shows comparisons between GPT-3.5, Taiwan-LLaMa-v1.0 and Model 7-C-Chat.

5 Conclusion

In conclusion, the evaluation of large language models, particularly in the context of Traditional Chinese, is a critical and challenging task. This study proposes a comprehensive set of benchmarks, built upon existing Traditional Chinese and English datasets, to assess the capabilities of these models across various tasks. The evaluation of models such as GPT-3.5, Taiwan-LLaMa-v1.0, and our proprietary model, Model 7-C-Chat, demonstrated the effectiveness of these benchmarks. Notably, Model 7-C-Chat showed comparable performance to the state-of-the-art GPT-3.5 model regarding certain of evaluated Traditional Chinese tasks.

The introduction of these benchmarks is a significant step towards advancing the evaluation of language models in Traditional Chinese. By making our benchmark code and relevant datasets open-source, and releasing our base model, Model 7-C-Chat, for trial, we aim to stimulate further research in this field. We believe that these resources will provide a valuable foundation for future studies aiming to improve the capabilities of language models in Traditional Chinese.

References

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

BIG-bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=uyTL5Bvosj`.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Philipp Ennen, Po-Chun Hsu, Chan-Jan Hsu, Chang-Le Liu, Yen-Chen Wu, Yin-Hsiang Liao, Chin-Tung Lin, Da-Shan Shiu, and Wei-Yun Ma. Extending the pre-training of bloom for improved support of Traditional Chinese: Models, methods and results, 2023.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021. URL https://doi.org/10.5281/zenodo.5371628.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2022.

Yen-Ting Lin and Yun-Nung Chen. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada, July 2023a. Association for Computational Linguistics. URL https://aclanthology.org/2023.nlp4convai-1.5.

Yen-Ting Lin and Yun-Nung Chen. Taiwanese-aligned language models based on meta-llama2, 2023b. URL https://github.com/adamlin120/Taiwan-LLaMa. Code and models available at https://github.com/adamlin120/Taiwan-LLaMa.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

Shang-Bao Luo, Cheng-Chung Fan, Kuan-Yu Chen, Yu Tsao, Hsin-Min Wang, and Keh-Yih Su. Chinese movie dialogue question answering dataset. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 7–14, Taipei, Taiwan, November 2022. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). URL https://aclanthology.org/2022.rocling-1.2.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only, 2023.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. DRCD: A Chinese machine reading comprehension dataset, 2019.

STPI. 2020 「科技大擂台與AI對話」訓練資料集, 2020. URL https://scidm.nchc.org.tw/dataset/grandchallenge2020.

TAIDE. Taide-14-tasks, 2023. URL https://huggingface.co/datasets/taide/TAIDE-14-tasks.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng

Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions, 2023.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

A Benchmark Question Examples

- DRCD

要探討從梨俱吠陀到波你尼時代梵語的發展，可以考察印度教其它文本，如娑摩吠陀、夜柔吠陀、阿闥婆吠陀、梵書和奧義書。在此期間，這門語言的威望、它的神聖用途及其正確發音的重要性，形成了一股強大的保守力量，防止梵語像普通語言一樣隨時間而演變。現存最古老的梵語文法是波你尼的《八篇書》，大約於公元前四世紀成形。它本質上是規範性文法，就是說它定義了正確梵語的用法，儘管它包含了描述成分，但大多是處理在波你尼時代已經廢棄了的某些吠陀形式。這裡所說的「梵語」不作脱離於其他語言的特殊語言看待，而是視作講話的高雅純正或完美方式。通過梵語文法家如波你尼的精密分析，梵語的知識在古印度是社會等級層次高和教育程度高的標誌，並主要教授給高等世襲階級的成員。梵語作為古印度的學術語言，與俗語同時共存，而俗語演化成了中古印度-雅利安語方言，並最終演化成了當代的各種印度-雅利安語言。

Q: 夜柔吠陀與阿闥婆吠陀均可以最為研究哪一門語言的參考？
A: 梵語

- TTQA

它是位於亞熱帶的台灣內唯一一種溫帶性魚類，也是只產於台灣的特有櫻鮭亞種，為冰河孑遺生物。由於其相當稀有且瀕臨絕種，加上它的生活習性迥異於其他魚類，遂得「國寶魚」之美譽。

Q: 該動物的名稱是：
A: 櫻花鉤吻鮭

- FGC

海倫·凱勒於1880年6月27日出生在美國阿拉巴馬州的塔斯坎比亞。海倫·凱勒原為健康的嬰兒，但在19個月大的時候患了急性腦充血病，失去了聽覺和視覺。長大後運用自創的手語與家庭成員溝通。隨著年歲的增長，簡單的交流不能滿足她，脾氣變得暴躁。6歲時，她的父母在家庭醫生的協助下，邀請柏金斯啓明學校的安妮·蘇利文老師作為海倫·凱勒的啓蒙導師。在1887年，藉著她的導師安妮·蘇利文對她耐心的教導和關愛，並找到專家使她學會發音，讓她學會流暢的表達，才開始與其他人溝通並接受教育。海倫·凱勒不但學會閱讀和説話，還以驚人的毅力完成了哈佛大學的學業並於1904年畢業，成為有史以來第一個獲得文學學士學位的盲聾人士。成年後，她繼續廣泛閱讀刻苦學習，掌握了英語、法語、德語、拉丁語和希臘語，成為盲聾的作家和教育家。她致力於殘疾人事業，四處募捐以改善殘疾人的生活環境和受教育水平。她的事跡使她入選美國《時代周刊》「人類十大偶像之一」，被授予「總統自由獎章」。

Q: 海倫凱勒出生於哪一個城市？
A: 塔斯坎比亞

- TMMLU

Q: 「臺灣原住民的布只有形制屬傳統或較現代的分別，像圓領的剪裁、鈕扣和棉布的使用等，都是受漢人的影響而來。泰雅族的貝珠鈴衣，是貝珠串底下加銅鈴裝飾，銅鈴也是和漢人交易而來。日治時代的原住民服裝，還出現以漢人棉布做底、日本布做袖口、原住民圖案做主要裝飾的混搭法。」這段文字的主旨最可能是下列何者？
(A)不同文化的碰撞，可融合並產生新的火花
(B)外來文化的入侵，讓在地的傳統文化日漸消失
(C)臺灣原住民的文化，影響了漢人與日本人的穿著
(D)觀察不同族群的服飾，就能了解不同文化的差異

A: (A)

- XSum-TC

埃文斯一開始就以一記漂亮的弧線球從20 碼處射入底角，幫助矮腳雞隊取得領先。蝦米隊發起反擊，瑞恩·倫納德(Ryan Leonard) 的一記猛烈遠射迫使布拉德福德門將本·威廉姆斯(Ben Williams) 做出精彩撲救。當蒂龍·巴內特的凌空抽射擊中橫梁時，主隊幾乎扳平比分，但布拉德福德堅持了下來。

Summary: 憑藉李·埃文斯的早早進球，布拉德福德城擊敗紹森德聯隊，確保獲得聯賽附加賽席位。

- IMDB-TC

我不在乎是否有人認為這部電影不好。如果你想知道真相，這是一部非常好的電影！它具有電影應有的一切。你真的應該買這個。

Sentiment: 正面

B Taiwan Massive Multitask Language Understanding (TMMLU)

Subject	Model 7-C-Chat	Taiwan-LlaMa-v1.0	GPT-3.5
企業管理	28/50	23/50	37/50
國際關係與近代外交史	9/20	4/20	10/20
基礎醫學	40/76	25/76	59/76
分科測驗物理	1/9	2/9	0/9
中式麵食加工	57/119	37/119	71/119
普通物理	9/30	8/30	13/30
分科測驗地理	9/25	6/25	13/25
油壓	15/47	23/47	26/47
分科測驗歷史	21/38	15/38	23/38
國際法	12/21	4/21	14/21
中醫臨床醫學	95/325	77/325	133/325
中餐烹調—葷食	23/60	22/60	32/60
中餐烹調—素食	29/59	22/59	35/59
普通生物學	25/48	16/48	36/48
會考國文	24/63	24/63	33/63
冷凍空調	16/56	17/56	25/56
分科測驗公民與社會	13/27	9/27	14/27
分科測驗化學	3/7	1/7	1/7
機電整合	24/44	17/44	23/44
分科測驗數學甲	0/5	1/5	0/5
法學知識	16/30	13/30	25/30
營養學	38/86	23/86	39/86
分科測驗生物	6/21	5/21	8/21
商業道德	25/80	26/80	23/80
用電設備檢驗	21/57	14/57	31/57
平版印刷	24/60	26/60	28/60
建築塗裝	31/60	24/60	38/60
普通化學	22/52	17/52	22/47
行銷管理	19/51	14/51	28/51
會考數學	4/7	2/7	2/7
會計學經濟學	9/41	11/41	12/41
資訊安全	35/75	23/75	48/75
農田灌溉排水—灌溉水質管理及檢驗項	22/55	21/55	29/55
牙醫學	168/454	135/454	214/454
社會工作大意	24/50	19/50	32/50
配電電纜裝修	21/56	19/56	26/56
綜合法政知識	26/57	18/57	37/57
網路架設	33/58	19/58	44/58
飛機修護	30/60	17/60	41/60
總體經濟	12/45	10/45	19/45
臨床心理學	124/259	86/259	155/259
臨床血清免疫學與臨床病毒學	34/78	22/78	52/78
裝潢木工	23/58	15/58	32/58
製鞋—製配底	23/57	16/57	34/57
製鞋—製面	22/59	13/59	26/59
觀光資源概要	27/48	22/48	31/48
計算機數學	2/10	1/10	0/10
車輛塗裝	24/59	21/59	34/59
通信技術	24/54	16/54	24/54
邏輯推理	5/15	4/15	9/15
醫學	16/27	13/27	19/27
金銀珠寶飾品加工	20/60	25/60	33/60
電路電子學	3/13	3/13	10/13
食品檢驗分析	28/60	24/60	40/60

C Open-Ended Evaluation

For all aspects, there are four levels of scoring: 0, 1, 2, and 3. GPT-4 is required to assign a score based on the predefined grading criteria for each aspect. We show one of the eight criteria below and the rest can be found in the repository[6].

Criterion of aspect: **Accuracy**
3: The submission is entirely correct and aligns with known sources or facts. The response demonstrates a strong understanding of the subject matter and is consistent with the golden answer.
2: The submission is mostly correct and aligns with known sources or facts. However, there may be minor inaccuracies or omissions that could be improved.

[6] https://github.com/mtkresearch/MR-Models

1: The submission is partially correct. Some aspects of the information align with known sources or facts, but there are significant inaccuracies or omissions.
0: The submission is completely incorrect or fabricated. It does not align with any known sources or facts.