JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark

Shota Onohara[1*] **Atsuyuki Miyai**[1*] **Yuki Imajuku**[1*] **Kazuki Egashira**[1*] **Jeonghun Baek**[1*]
Xiang Yue[2] **Graham Neubig**[2] **Kiyoharu Aizawa**[1]
[1]The University of Tokyo [2]Carnegie Mellon University

Abstract

We introduce **JMMMU** (Japanese MMMU), an expert-level benchmark that can truly evaluate the performance of large multimodal models (LMMs) in Japanese. Compared to other existing Japanese multimodal benchmarks, JMMMU requires a deep understanding of Japanese culture and advanced reasoning skills, and it includes more than ten times the number of questions found in similar benchmarks, enabling more reliable quantitative evaluations. We believe our findings inspire the development of high-standard benchmarks in more languages, and pave the way for LMM developments that are more inclusive of non-English languages. Project page is available at `https://mmmu-japanese-benchmark.github.io/JMMMU/`.

1 Introduction

Recent large multimodal models (LMMs) have demonstrated remarkable performance across various tasks, ranging from common sense reasoning to those requiring expert-level, domain-specific knowledge. This highlights the critical role of benchmarks in evaluating the diverse capabilities of LMMs. However, current benchmarks focus primarily on performance in English, with less emphasis on the utility in other languages. Notably, performance evaluations of LMMs in Japanese, despite its unique culture spreading around the world, remain underrepresented. Current Japanese multimodal benchmarks exhibit the following weaknesses:

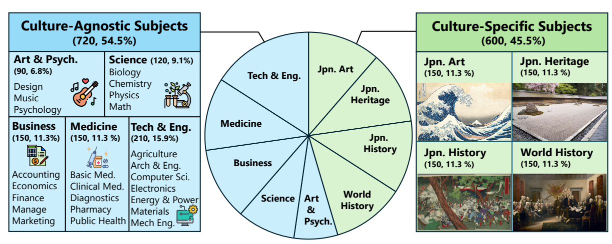

Figure 1: **Overview of our JMMMU dataset.**

- **(W1)** Existing benchmarks [1–7] focus on common sense knowledge, but none of them cover expert-level knowledge.
- **(W2)** Many of them do not account for cultural differences. They are often created by translating existing English benchmarks [1–3], and thus the questions are unfamiliar to Japanese people.
- **(W3)** Recent benchmarks try to consider cultural differences [4–7], but they are all limited in size (only up to 102 questions [4]), raising concerns about whether reliable quantitative evaluation can be achieved.

This work: Creating a Massive, Expert-level, Truly-Japanese Multimodal Benchmark Given the circumstance, we introduce **JMMMU** (*Japanese MMMU*), a multimodal benchmark that can truly evaluate expert-level LMM performance in Japanese. An overview of our JMMMU can be found

*Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: **Overall results.** A grayed column represents the evaluation in English (for the questions we translated). The rest of the results are average and individual subjects' scores on JMMMU. Overall, JMMMU leaves great room for improvement (up to 40.5% for open-source, and 58.6% for GPT-4o).

	MMMU val (720)	Overall (1,320)	Culture Specific (600)	Culture Agnostic (720)	Jpn. Art (150)	Jpn. Heritage (150)	Jpn. History (150)	World History (150)	Art & Psychology (90)	Business (150)	Science (120)	Health & Medicine (150)	Tech & Eng. (210)
Random	24.6	24.8	25.0	24.6	25.0	25.0	25.0	25.0	25.4	25.0	22.8	25.6	24.3
Large Multimodal Models: Text + Image as Input													
LLaVA-ov-05b [11]	29.4	26.0	23.3	28.2	22.7	22.7	24.0	24.0	26.7	27.3	24.2	30.7	30.0
xGen-MM [12]	35.7	28.6	28.2	28.9	30.0	20.7	22.7	39.3	32.2	21.3	22.5	36.7	31.0
Phi-3v [13]	37.6	29.5	26.5	31.9	31.3	18.7	29.3	26.7	26.7	28.7	25.8	37.3	36.2
LLaVA1.6-13b [14]	29.9	31.1	33.7	29.0	32.0	24.0	32.0	46.7	25.6	28.7	30.0	34.0	26.7
Phi-3.5v [13]	39.2	32.4	34.3	30.8	37.3	27.3	35.3	37.3	27.8	31.3	30.0	36.7	28.1
LLaVA CALM2 [15]	29.9	34.9	41.5	29.4	42.7	36.7	40.0	46.7	27.8	26.0	26.7	34.0	31.0
EvoVLM JP v2 [16][17]	33.9	38.1	45.2	32.2	44.0	40.0	42.0	54.7	32.2	28.7	28.3	38.7	32.4
Internvl2-8b [18][19]	43.3	38.3	42.5	34.7	41.3	38.0	35.3	55.3	40.0	36.0	34.2	34.0	32.4
LLaVA1.6-34b [14]	45.7	39.8	43.2	37.1	42.0	36.0	40.7	54.0	42.2	41.3	25.0	36.7	39.0
LLaVA-ov-7b [11]	45.1	40.5	43.0	38.5	36.0	30.7	37.3	68.0	41.1	36.7	31.7	38.7	42.4
GPT-4o [20]	52.1	58.6	66.7	51.8	60.7	70.7	58.7	76.7	53.3	55.3	45.8	61.3	45.2
Large Language Models: Only Text as Input													
GPT-4o text	44.9	38.1	35.5	40.3	32.7	32.0	35.3	42.0	38.9	36.0	41.7	45.3	39.5

in Figure 1. To address **(W1)**, we created a benchmark based on the validation set of MMMU [8] consisting of 900 samples, which is widely used to evaluate LMMs' expert-level reasoning with domain-specific knowledge. For **(W2)**, we first carefully analyzed the existing MMMU benchmark for its cultural dependencies. Then, for questions in culture-agnostic subjects, we employed native Japanese speakers who are experts in each subject, and asked them to translate both the texts and images (e.g. the title of a graph) into Japanese. Further, we replaced culture-dependent subjects with new subjects that are conceptually similar, but better aligned with Japanese culture. For example, the original MMMU contains a subject called *History*, which we divided into *Japanese History* and *World History*. For each subject, we sourced images from the web that had no licensing issues and created questions based on content typically found in Japanese textbooks. Finally, in response to **(W3)**, JMMMU consists of 720 translation-based (culture-agnostic) and 600 brand-new (culture-specific) questions, for a total of 1,320 questions, updating the size of the existing culture-aware Japanese benchmark by >10x.

Implication of Our Benchmark Creation Our JMMMU benchmark for the first time enables the community to reliably evaluate LMM's expert-level reasoning capabilities in Japanese. Our observations suggest that focusing solely on performance evaluation in English could risk a biased development competition that overlooks the utility in non-English languages. Conversely, a benchmark for a specific language can stimulate interest among model developers to improve its accuracy, as is currently observed with Chinese [9, 10]. We hope that our benchmark will not only trigger the community's interest in Japanese language performance, but also serve as a catalyst for benchmark creation in other languages, leading to the development of LMMs that are more inclusive of non-English languages.

2 Experiments and Findings

In Table 1, we provide the evaluation results on our JMMMU benchmark. In our experiment, the performance is up to 40.5% for open-source, and 58.6% for closed-source models, leaving great room for improvement. In this section, we summarize our key observations on the culture-agnostic (CA) and culture-specific (CS) splits.

CA Split: The Effect of Translation The score on the CA split is lower than its English counterpart (MMMU CA) for most of the models (except for LLaVA CALM2 [15], a Japanese LMM). This suggests that, even for the same questions, many models perform worse when asked in Japanese.

CS Split: Capturing Deep Understanding of Japanese Culture Even when models perform similarly on the CA split, their performance on the CS split can vary significantly. For instance, (i) Phi-3v [13] (no multilingual support), (ii) Phi-3.5v [13] (a multilingual model with Japanese support), and (iii) EvoVLM JP v2 (a Japanese LMM) [17] show similar results on the CA split ($31.5 \pm 0.7\%$). However, their CS scores differ markedly: (i) Phi-3 scores worse (-5.4%), (ii) Phi-3.5 scores slightly better ($+3.5\%$), and (iii) EvoVLM excels ($+13.0\%$). This highlights how Japanese-focused training can significantly impact performance in Japanese-specific contexts, and JMMMU is capable of capturing these differences.

Limitation

While JMMMU can assess the latest LMMs' expert-level skills, it cannot evaluate model performance on subjects outside of those currently covered. As models gain more knowledge and improve their reasoning abilities, it will be necessary to expand the range of subjects and include more challenging questions. Moreover, since JMMMU only covers the Japanese language, evaluating model performance in other languages and cultural contexts remains an important area for future work. We reiterate here that we hope these efforts will help mitigate the underrepresentation of diverse cultures and languages.

Acknowledgments and Disclosure of Funding

This work was partially supported by JST BOOST, Japan Grant Number JPMJBS2418, JSPS KAKENHI Grant Number 24K23882, and JST JPMJCR22U4.

References

[1] Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *COLING*, 2018.

[2] Turing. Llava-bench-ja. https://github.com/turingmotors/heron/tree/main/playground/data/llava-bench-ja, 2024.

[3] Turing. Llava-bench-in-the-wild (japanese). https://github.com/turingmotors/heron/tree/main/playground/data/llava-bench-in-the-wild, 2024.

[4] Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. Heron-bench: A benchmark for evaluating vision language models in japanese. In *CVPR workshop*, 2024.

[5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023.

[6] Sakana AI. Ja-vlm-bench-in-the-wild. https://huggingface.co/datasets/SakanaAI/JA-VLM-Bench-In-the-Wild, 2024.

[7] Sakana AI. Ja-multi-image-vqa. https://huggingface.co/datasets/SakanaAI/JA-Multi-Image-VQA, 2024.

[8] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

[9] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*, 2024.

[10] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[11] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[12] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models, 2024. URL https://arxiv.org/abs/2408.08872.

[13] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

[15] Aozora Inagaki. llava-calm2-siglip. https://huggingface.co/cyberagent/llava-calm2-siglip, 2024.

[16] Inoue Yuichi, Akiba Takuya, and Makoto Shing. Llama-3-evovlm-jp-v2. URL https://huggingface.co/SakanaAI/Llama-3-EvoVLM-JP-v2.

[17] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *arXiv preprint arXiv:2403.13187*, 2024.

[18] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[19] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[20] OpenAI. Gpt-4o, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in abstract are justified in the experimental results in Section 2

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitation is stated in Section 2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: While we understand the importance of reproducibility, it is difficult to include all the detailed parameters in the two-page tiny paper. Instead, we will release the dataset and the codebase which are sufficient to ensure reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset created and used for the experiments will be released on Hugging Face, and the evaluation code will be made available on our GitHub repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: While we understand the importance of reproducibility, it is difficult to include all the detailed parameters and its justification in the two-page tiny paper. Instead, we will release the dataset and the codebase which are sufficient to ensure reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experiments were conducted with the temperature parameter set to zero, ensuring that the results were theoretically deterministic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We will release the dataset and the codebase which are sufficient to ensure reproducibility instead of including all the detailed parameters in the tiny paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics and ensured compliance.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: While the positive impacts of our work are stated throughout our paper, we believe our work does not contain any significant negative impact on society, and thus omitted.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the original MMMU dataset and have properly cited it. We will publish our dataset along with the information required by the license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We created a new dataset. We discussed the process of creating it, as well as the absence of any licensing issues in Section 1.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing have been conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No experiments with human subjects have been conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.