Challenging Multimodal LLMs with African Standardized Exams: A Document VQA Evaluation

Victor Olufemi[1] Oreoluwa Babatunde[1] Kausar Moshood[1] Emmanuel Bolarinwa[1]
[1]LyngualLabs
{victor, oreoluwa, kausar, emmanuel}@lynguallabs.org

Abstract

Despite rapid advancements in multimodal large language models (MLLMs), their ability to process low-resource African languages in document-based visual question answering (VQA) tasks remains limited. This paper evaluates three state-of-the-art MLLMs—GPT-4o, Claude-3.5 Haiku, and Gemini-1.5 Pro—on WAEC/NECO standardized exam questions in Yoruba, Igbo, and Hausa. We curate a dataset of multiple-choice questions from exam images and compare model accuracies across two prompting strategies: (1) using English prompts for African language questions, and (2) using native-language prompts. While GPT-4o achieves over 90% accuracy for English, performance drops below 40% for African languages, highlighting severe data imbalance in model training. Notably, native-language prompting improves accuracy for most models, yet no system approaches human-level performance, which reaches over 50% in Yoruba, Igbo, and Hausa. These findings emphasize the need for diverse training data, fine-tuning, and dedicated benchmarks that address the linguistic intricacies of African languages in multimodal tasks, paving the way for more equitable and effective AI systems in education.

1. Introduction

The rapid advancements in artificial intelligence (AI) have led to the emergence of multimodal large language models (MLLMs) capable of processing and understanding both textual and visual information [2, 13]. Notable examples include OpenAI's GPT, Anthropic's Claude and Google's Gemini. These models exhibit impressive capabilities in interpreting combined visual-textual inputs, allowing them to extract text from images and answer questions about that content. However, their ability to accurately process text from images in low-resource languages remains an open question [1]. Despite the progress in multilingual NLP, most state-of-the-art models are primarily trained on high-resource languages, resulting in suboptimal performance for many African languages. Low-resource languages are severely underrepresented in the datasets used to train and evaluate MLLMs [1, 5], and African languages such as Yoruba, Igbo, and Hausa present unique linguistic and orthographic challenges that differ significantly from dominant languages on which these models are typically trained [12]. The scarcity of high-quality training data for these languages exacerbates the performance disparity between high- and low-resource languages [9]. Recent benchmarks confirm that multimodal models perform very well on English but struggle on many African languages due to data limitations [2]. In this work, we present a novel evaluation of state-of-the-art MLLMs using real exam questions from WAEC and NECO standardized assessments from West Africa, spanning visual inputs and low-resource African languages. We assess whether these models can read and reason over multimodal exam content in Yoruba, Igbo, and Hausa without relying on external OCR or preprocessing. Our findings highlight both the potential and current limitations of MLLMs in African educational contexts, offering actionable insights for the development of more inclusive and equitable AI systems.

1.1. Research Objectives

This study aims to systematically evaluate the performance of GPT-4o, Claude-3.5 Haiku and Gemini-1.5 Pro in natural language comprehension for African languages by addressing the following objectives:

- **Objective 1:** Assess the ability of multimodal LLMs to accurately extract and process text from WAEC/NECO examination images.
- **Objective 2:** Compare performance under different prompt languages, analyzing whether using English vs. native-language prompts affects answer accuracy.

2. Related Works

2.1. Multimodal Large Language Models and Their Capabilities

Multimodal large language models integrate multiple data modalities, such as text and images, to enhance comprehension and reasoning [2, 13]. These models build on advances in vision-language pre-training that combine visual encoders with language models [6, 14]. State-of-the-art MLLMs have achieved impressive performance on many text-based visual tasks, including image captioning, document understanding, and visual question answering. In general, these models perform well on tasks in high-resource languages. However, studies have shown that their effectiveness diminishes significantly in low-resource languages such as Yoruba, Igbo, and Hausa [1, 16]. For instance, the IrokoBench evaluation found a substantial drop in GPT-4o's performance on African language understanding compared to English. Similarly, a culturally diverse VQA benchmark [15] demonstrated that even powerful vision-language models fail to generalize across linguistically diverse or culturally unfamiliar inputs. [17] introduced M3Exam, a multilingual, multimodal exam benchmark, and reported major performance discrepancies between high-resource and low-resource languages. While current MLLMs can process Latin-script inputs with high accuracy, they struggle with the complex morphology and orthographic variations present in many African languages [7]. This gap underscores that simply scaling to multimodal inputs is not sufficient for broad multilingual competency.

2.2. Challenges in Multilingual NLP for Low-Resource African Languages

The lack of training data remains a fundamental challenge in multilingual NLP research, particularly for African languages [1]. Unlike English or other widely spoken languages, Yoruba, Igbo, and Hausa have relatively limited corpora and annotated datasets available for training or fine-tuning large models. This data scarcity negatively impacts model performance on both text-only and multimodal tasks [16]. Even large multilingual language models like XLM-R [3] or BLOOM struggle on African languages that were underrepresented in their training data. In addition, many African languages have unique linguistic properties – for example, tonal phonology and extensive use of diacritics in Yoruba, or complex noun classes in some Bantu languages – which prove difficult for pre-trained LLMs to handle. These orthographic and grammatical nuances are often lost or misinterpreted by models not specifically adapted to them [12]. Recent studies such as [9] highlight that vision-language models exhibit poor understanding of culturally or linguistically specific content, reinforcing the importance of developing benchmarks that reflect real-world linguistic diversity. There have been efforts to bolster NLP for African languages – for example, the Masakhane project's participatory approach to machine translation [10] and the creation of language-specific models like AfriBERTa [11], but these are text-only initiatives. Until similar resources and benchmarks are created on the multimodal front, AI models will continue to exhibit biases favoring high-resource languages over under-represented ones [2, 8]. Our work addresses this gap by providing a focused evaluation on Yoruba, Igbo, and Hausa, thereby pushing towards more inclusive multimodal model development.

2.3. Standardized Exam Benchmarks in AI Research

Standardized exams have become a widely adopted benchmark for evaluating AI models. The structured format of exam questions—where each item follows a consistent style and has a known correct answer offers a controlled environment for assessing an AI's reading comprehension, reasoning, and problem-solving abilities. Several recent studies have used exam-based benchmarks to evaluate large language models. For example, M3Exam [17] compiles real multilingual exam questions and shows that GPT-4o and similar models perform well on high-resource languages but struggle on under-represented languages. Similarly, the MEGAVERSE benchmark [2] evaluated LLMs across 83 languages and highlighted substantial performance gaps in low-resource linguistic settings. Our study follows a similar methodology of exam-driven evaluation but narrows the focus specifically to structured educational content in popular Nigerian languages. By concentrating on WAEC/NECO multiple-choice questions in Yoruba, Igbo, and Hausa, we provide an in-depth look at model capabilities in a context that had not been examined in prior multilingual benchmarks. This approach also complements efforts like [4]. A MMLU which included a broad range of subjects and some languages: we add the dimension of image-based text understanding in an educational assessment scenario.

3. Methodology

3.1. Dataset Curation

The dataset for this study was curated from past WAEC and NECO examination questions in Yoruba, Igbo, Hausa, and English. We targeted multiple-choice questions (MCQs) from recent years to ensure a representative sample of modern usage. The curation process involved several steps:

3.1.1 Data Collection

We obtained past examination papers from students and bookshops that sell educational materials. However, ac-

quiring exam questions for language subjects (Hausa, Igbo, Yoruba) online proved extremely challenging, if not nearly impossible, due to their limited availability compared to more widely documented subjects. To ensure a sizable dataset in each target language, we focused on examination papers from the years 2008–2024. Table 1 below summarizes the provisional composition of the dataset.

3.1.2 Question Segmentation

Each question‘ was manually cropped from scanned examination sheets to isolate it as an individual image. This ensured that each image contained exactly one question for the model to answer, standardizing the input format. Only multiple-choice questions were included to maintain a uniform evaluation style.

3.1.3 Answer Key Verification

Many exams came with official answer keys, which we treated as gold-standard answers. For questions lacking official keys (or in cases where only the exam paper was available), we consulted linguistic and subject matter experts fluent in Yoruba, Igbo, or Hausa to determine the correct answer. These expert-verified answers were cross-checked to ensure accuracy.

Year	English	Yoruba	Igbo	Hausa
2008	0	20	0	0
2009	0	16	0	0
2010	0	19	0	0
2011	0	20	0	0
2012	0	17	0	0
2013	0	20	0	0
2014	0	20	0	0
2015	0	19	0	0
2016	0	20	0	0
2017	0	0	0	0
2018	29	19	0	0
2019	30	19	0	0
2020	30	19	0	0
2021	60	38	24	0
2022	60	40	45	20
2023	60	36	45	20
2024	0	36	0	36
Total	**269**	**378**	**114**	**76**

Table 1. Dataset composition by year and language. WAEC and NECO Combined

3.2. Model Selection and Evaluation Criteria

We selected three state-of-the-art multimodal LLMs for benchmarking: GPT-4o (OpenAI), Claude-3.5 Haiku (Anthropic), and Gemini-1.5 pro (Google DeepMind). These models although uneven in sizes, were chosen due to their cutting-edge performance and diverse origins (industry leaders in AI). We accessed GPT-4o, Claude-3.5 Haiku, and Gemini-1.5 Pro via their official API endpoints, While other emerging models (such as Mistral) could be considered, we limited our testing to these three due to time and resource constraints. Our evaluation was based on two primary criteria:

- **Answer Accuracy:** The percentage of questions for which the model's answer matched the expert-verified correct answer. This is a direct measure of performance on the multiple-choice questions.
- **Language-wise Performance:** We compare accuracy across the four languages (English, Yoruba, Igbo, Hausa) to identify any performance disparities.

3.3. Experimental Setup

We designed a uniform evaluation pipeline and prompting strategy to ensure a fair comparison between models. Key aspects of the experimental setup are outlined below:

3.3.1 Prompting Strategy

We employed two query strategies for each question image:

1. An English-prompted query.
2. A native-language-prompted query.

In the English prompt condition, the model was instructed in English (*e.g.* "Analyze the image and answer the question) while being given an image containing a Yoruba/Igbo/Hausa question. In the native prompt condition, we translated the instruction into the question's language (Yoruba, Igbo, or Hausa) so that the model received the prompt in the same language as the question. This allows us to test whether prompting in the local language improves understanding or not. Each model thus answers every question twice: once with an English prompt and once with a native-language prompt.

3.3.2 Prompt Template

We crafted a consistent system message for all models, emphasizing the task and format. Below is a simplified example of the prompt content used (shown here in English for brevity):

System Prompt:

"You are a knowledgeable assistant for answering exam questions. Carefully read the question in the image and evaluate each of the four choices. Provide the answer by indicating the option (A, B, C, D, or E) with the highest probability of being correct, along with probability scores for each option in JSON format."

User Prompt:

"Analyze the following question image and determine the correct answer (A, B, C, D, or E). Respond in JSON with your probabilities for each option."

For native-language trials, the prompts were translated appropriately (*e.g.* to Yoruba). All models were thus given a very similar cue and format requirement, to the extent their API allowed system instructions.

3.4. Evaluation Metric

We used a strict accuracy metric for each model's responses. A model receives a score of 1 for a question if its highest-probability choice matches the correct answer, and 0 otherwise. We then compute overall accuracy as well as per-language accuracy.

The above methodology enables a controlled and fair evaluation of each model's ability to interpret exam images and answer questions in multiple languages. All model outputs and metadata are logged for analysis.

4. Results

The evaluation results provide insights into the performance of GPT-4o, Gemini-1.5 Pro, and Claude-3.5 Haiku on multiple-choice exam questions in Yoruba, Hausa, Igbo, and English. We analyze accuracy under two prompting conditions:

1. Prompting in English.
2. Prompting in the respective African language.

We also compare the models' performance to human baseline scores.

4.1. Model Performance Across Languages

The Table 2 below presents the accuracy scores for each model across different languages and prompt conditions:

Prompt	GPT-4o Accuracy	Gemini-1.5 Pro Accuracy	Claude-3.5 Haiku Accuracy
Yoruba Exam Questions			
Yoruba Prompt	32.80% (124/378)	29.63% (112/378)	26.72% (101/378)
English Prompt	31.74% (121/378)	33.86% (128/378)	25.92% (98/378)
Hausa Exam Questions			
English Prompt	39.47% (30/76)	36.84% (28/76)	28.95% (22/76)
Hausa Prompt	43.42% (33/76)	44.74% (34/76)	23.68% (18/76)
English Exam Questions			
English Prompt	90.33% (243/269)	73.61% (198/269)	55.39% (149/269)
Yoruba Prompt	79.55% (214/269)	72.49% (195/269)	39.03% (105/269)
Hausa Prompt	80.30% (216/269)	72.86% (196/269)	40.89% (110/269)
Igbo Prompt	81.04% (218/269)	72.12% (194/269)	36.43% (98/269)
Igbo Exam Questions			
English Prompt	27.19% (31/114)	31.58% (36/114)	18.42% (21/114)
Igbo Prompt	28.95% (33/114)	35.96% (41/114)	23.68% (27/114)

Table 2. Accuracy scores for GPT-4o, Gemini, and Claude across different languages and prompt conditions.

4.2. Key Observations

- **Higher Accuracy in English:** As expected, models performed significantly better on English-only questions, with GPT-4o achieving the highest accuracy (90.33%), followed by Gemini-1.5 Pro (73.61%) and Claude-3.5 Haiku (55.39%). This confirms that the models handle high-resource languages much better than low-resource ones.
- **Effect of Prompting English Questions in African Languages:** Interestingly, when English questions were prompted in Yoruba, Hausa, and Igbo, accuracy dropped compared to using English prompts. GPT-4o's accuracy dropped from 90.33% (English prompt) to 79.55% (Yoruba prompt), 80.30% (Hausa prompt), and 81.04% (Igbo prompt). Gemini-1.5 Pro and Claude-3.5 Haiku showed similar trends, highlighting how translation and linguistic context impact comprehension.
- **Native Language Prompts Improve Accuracy:** For Yoruba, Hausa, and Igbo, prompting the model in the native language generally resulted in higher accuracy than when the prompt was in English. The effect was particularly noticeable in Hausa (*e.g.* GPT-4o: 43.42% Hausa-prompted vs. 39.47% English-prompted).

4.3. Comparison with Human Performance

We also compared model results with human performance, where participants from an independent NLP community answered the same exam questions. The results are presented in Table 3 below:

Language	Human Accuracy
Hausa	68.0%
Igbo	52.3%
Yoruba	56.0%

Table 3. Comparison of human accuracy on multiple-choice exam questions across three African languages.

Human accuracy was significantly higher than all model performances across the three African languages, reinforcing that even non-expert humans outperform state-of-the-art

AI models on structured educational tasks in Yoruba, Igbo, and Hausa.

These results provide strong evidence of the performance gap between AI models and human linguistic abilities, particularly in low-resource African languages.

5. Future Work

This study evaluates proprietary models from OpenAI, Anthropic, and Google due to their state-of-the-art performance and reliable API access, which allows for a standardized evaluation pipeline under limited computational resources. However, we acknowledge that this focus limits generalizability, especially for communities relying on open-source systems. Future work will incorporate multilingual and open-source MLLMs—such as LLaVA, XLM-V, and Mistral-based vision-language systems—to enable broader analysis and support reproducibility in low-resource settings.

Future research can also build upon this work by expanding and improving multimodal datasets for African languages, ensuring high-quality resources that help bridge performance gaps. Fine-tuning LLMs on domain-specific or culturally relevant data may further enhance reasoning capabilities in these languages. Another crucial direction is the development of standardized evaluation benchmarks for African multimodal NLP, enabling more consistent cross-model comparisons. Investigating OCR accuracy for African scripts is equally important, as many languages have distinctive orthographies and diacritics that present unique challenges for vision-language systems. Finally, extending evaluation beyond Yoruba, Igbo, and Hausa to cover more African languages would offer a richer understanding of multilingual and multicultural challenges in NLP.

References

[1] David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. Irokobench: A new benchmark for african languages in the age of large language models, 2025. 1, 2

[2] Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico, June 2024. Association for Computational Linguistics. 1, 2

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. 2

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. 2

[5] 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*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. 1

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[8] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences, 2024. 2

[9] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. 1, 2

[10] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine

translation: A case study in African languages. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, Nov. 2020. Association for Computational Linguistics. 2

[11] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[12] Iroro Orife, David I. Adelani, Timi Fasubaa, Victor Williamson, Wuraola Fisayo Oyewusi, Olamilekan Wahab, and Kola Tubosun. Improving yorùbá diacritic restoration, 2020. 1, 2

[13] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. 1, 2

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[15] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024. 2

[16] Florian Schneider and Sunayana Sitaram. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks, 2024. 2

[17] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, 2023. 2