GAOKAO-MM: A Chinese Human-Level Benchmark for Multimodal Models Evaluation

Yi Zong, Xipeng Qiu*
School of Computer Science, Fudan University
Shanghai Collaborative Innovation Center of Intelligent Visual Computing
yzong22@m.fudan.edu.cn, xpqiu@fudan.edu.cn

Abstract

The Large Vision-Language Models (LVLMs) have demonstrated great abilities in image perception and language understanding. However, existing datasets either focus solely on primary perception abilities and commonsense knowledge, or have a low level of text comprehension difficulty, which are insufficient to reflect the comprehensive capabilities of LVLMs, particularly in terms of Chinese language proficiency. We propose GAOKAO-MM, a multimodal benchmark based on the Chinese College Entrance Examination (GAOKAO), comprising of 8 subjects and 12 types of images, such as diagrams, function graphs, maps and photos. GAOKAO-MM derives from native Chinese context and sets human-level requirements for the model's abilities, including perception, understanding, knowledge and reasoning. We evaluate 10 LVLMs and find that the accuracies of all of them are lower than 50%, with GPT-4-Vision (48.1%), Qwen-VL-Plus (41.2%) and Gemini-Pro-Vision (35.1%) ranking in the top three positions. The results of our multi-dimension analysis indicate that LVLMs have moderate distance towards Artificial General Intelligence (AGI) and provide insights facilitating the development of multilingual LVLMs.[1]

1 Introduction

Over the past years, large language models (LLMs) (Brown et al., 2020; OpenAI, 2023; Bubeck et al., 2023; Anil et al., 2023) have achieved great success and have been on par with or surpassed human in certain areas. Inspired by the development of LLMs, large vision-language models (LVLMs) also make much progress on perception, understanding and reasoning, which shows great potential towards Artificial General Intelligence (AGI). However, previous multimodal datasets, including VQA (Goyal et al., 2017), OK-VQA (Marino et al., 2019), TextVQA (Singh et al., 2019), focus on a single capability or image type. Consequently, there is a growing need for a more human-level benchmark to evaluate LVLMs on comprehensive tasks. Lu et al. propose MathVista to evaluate the mathmatical reasoning of foundation models; MME (Fu et al., 2023) measures both perception and cognition abilities of LVLMs; MM-Vet (Yu et al., 2023) defines 6 core VL capabilities and examines the 16 integrations of interest.

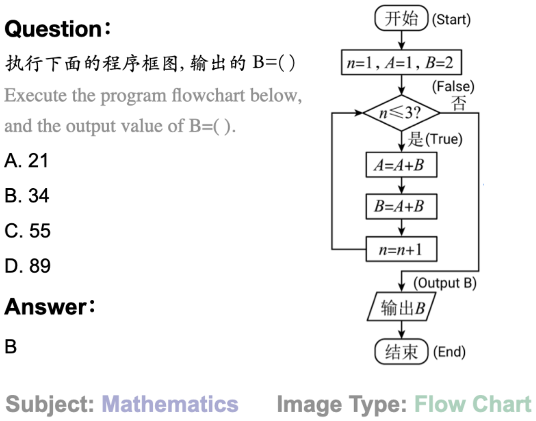
Question:
执行下面的程序框图, 输出的 B=( )
Execute the program flowchart below, and the output value of B=( ).
A. 21
B. 34
C. 55
D. 89
Answer:
B
Subject: Mathematics **Image Type:** Flow Chart

Figure 1: **An example of math question in GAOKAO-MM.** The English translation in the text and images is added by the author for readers' understanding.

In this work, we propose GAOKAO-MM, a benchmark derived from Chinese College Entrance Examination (GAOKAO). It comprises of 8 subjects, including Chinese, mathematics, physics, chemistry, biology, politics, history and geography, and its average question length is 4 times longer than existing benchmarks. GAOKAO-MM covers 12 types of images, such as diagrams, charts, photos and maps. LVLMs need to perceive and understand the images and text, fuse their inner

---

*Corresponding Author.

[1] The dataset and evaluation code are available through: `https://github.com/OpenMOSS/GAOKAO-MM`

Findings of the Association for Computational Linguistics: ACL 2024, pages 8817–8825
August 11-16, 2024 ©2024 Association for Computational Linguistics

knowledge and make inference to get correct answers, which more closely mirrors the way humans perceive the world and make decisions. We evaluate 10 LVLMs, including open-source and closed-source models, and the results demonstrate the existing models still have considerable room for improvement.

2 GAOKAO-MM

2.1 Dataset Description

The GAOKAO, Chinese College Entrance Examination, serves as a pivotal assessment tool for evaluating the scholarly aptitudes of high school students. The GAOKAO is the most popular Chinese standardized test and we regard it as a suitable evaluation on human-level tasks.

The GAOKAO-MM comprises of 646 questions and covers 8 subjects, including Chinese, mathematics, physics, chemistry, biology, politics, history and geography. There are 897 images in 12 types, such as diagrams, function graphs, maps, line charts and geometric shapes. The distribution of image types is shown in Appendix A.

The GAOKAO-MM poses a significant challenge for LVLMs: models are supposed to perceive and understand both the image and text inputs, recall their inner knowledge and conduct reasoning to answer the rigorous and comprehensive questions. Thanks to the popularity and authoritative of the GAOKAO, GAOKAO-MM provides a human-level approach to evaluate the capabilities of LVLMs. An example of math question is shown in Figure 1.

2.2 Data Collection

Data Source We collect the GAOKAO papers from 2010 to 2023 in PDF format from the Internet.

Data Collection We manually extract the multi-choice questions with images using OCR, and the mathematical formulas within the questions were converted into LaTeX format.

Data Curation We select the questions that require the integration of images for accurate answering and filter out duplicate questions. And we perform Lanczos resampling on low-resolution images, ensuring that at least one dimension of the image is increased to no less than 512 pixels.

2.3 Comparisons with Existing Benchmarks

We compare GAOKAO-MM with existing benchmarks which are knowledge-based (Lu et al., 2022b; Marino et al., 2019, Kembhavi et al., 2016; Kahou et al., 2018; Lu et al., 2022a; Yue et al., 2023; Liu et al., 2023b). And GAOKAO-MM is different from them from three aspects.

Benchmark	Size	Avg. Q. Leng.	Expl.	Question	Image	Lang.
IconQA	107K	8.30	✖	MC+FB	Icon Image	Eng.
OK-VQA	14K	8.09	✖	Open	Photos	Eng.
Ai2D	5K	9.78	✖	MC	Diagrams	Eng.
FigureQA	>1M	6.07	✖	BC	5 Types	Eng.
ScienceQA	6K	12.11	✔	MC	5 Types	Eng.
MMMU	11.5K	59.33	✔	MC+Open	30 Types	Eng.
MM-Bench-CN	3K	15.48	✖	MC	Not Mentioned	T. Chi.
GAOKAO-MM	0.65K	260.19	✔	MC	12 Types	N. Chi.

Table 1: **The comparison between GAOKAO-MM and other existing benchmarks.** Avg Q. Leng. stands for Average Question Length; FB stands for Fill-in-the-Blank Questions; MC stands for Multiple-Choice Quesions; BC stands for Binary-Choice Questions; T. Chi. stands for translated Chinese; N. Chi. stands for native Chinese.

Firstly, the average question length of GAOKAO-MM is much longer than existing benchmarks, even more than 4 times to the longest one. Previous multimodal benchmarks focus more on the visual perception, while the text input is just an auxiliary to images; GAOKAO-MM puts emphasis both on image perception and text understanding, which simulates the way human receive multimodal information more closely.

Secondly, GAOKAO-MM provides detailed explanations of the correct answers annotated by human. And we provide an example of explanation in Appendix B.2.

Lastly, most Chinese multimodal benchmarks are translated from English benchmarks. However, GAOKAO-MM derives from native Chinese context and contains questions more related to Chinese culture.

3 Experiments

3.1 Methodology

We evaluate 10 LVLMs on GAOKAO-MM, including open-source models and closed-source models. Following GAOKAO-Bench (Zhang et al., 2023), we utilize a zero-shot setting strategy to emulate how human partake in examinations. We design prompts tailored to different subjects and adopt a robust and rule-based approach to extract answers. Models are prompted to output the reasoning process and final answers in an explicit format. The examples of prompts are shown in Appendix B.1 and examples of model outputs are

Models	Overall	Chinese	Math.	Phys.	Chem.	Biol.	Poli.	Hist.	Geog.
VisCPM-Chat	16.8%	0.0%	18.8%	3.7%	22.4%	23.8%	3.0%	32.4%	24.9%
VisualGLM-6B*	20.7%	**21.9%**	16.3%	12.1%	25.4%	28.6%	21.2%	38.2%	24.0%
CogVLM-Chat*	13.2%	3.1%	17.5%	0.0%	16.4%	9.5%	12.1%	20.6%	21.3%
LLaVA-1.5-7B*	17.9%	6.2%	12.5%	7.2%	23.9%	14.3%	21.2%	32.4%	24.9%
LLaVA-1.5-13B*	18.0%	3.1%	16.3%	10.6%	19.4%	33.3%	12.1%	23.5%	23.5%
mPLUG-Owl2	24.1%	6.2%	22.5%	12.4%	22.4%	28.6%	48.5%	32.4%	30.3%
Qwen-VL-Chat	14.2%	0.0%	17.5%	3.2%	16.4%	14.3%	15.2%	26.5%	19.9%
GPT-4V	**48.1%**	18.8%	**45.0%**	**26.7%**	41.8%	38.1%	**75.8%**	79.4%	**62.0%**
Qwen-VL-Plus	41.2%	18.8%	33.8%	14.4%	**44.8%**	**42.9%**	63.6%	**82.4%**	55.7%
Gemini-Pro-Vision	35.1%	18.8%	23.7%	17.2%	38.8%	**42.9%**	54.5%	55.9%	46.6%

Table 2: **Accuracy of LVLMs on GAOKAO-MM.**The models above the line are open-source models, models below the line are closed-source models. *: Model does not support the simultaneous input of multiple images.

shown in Appendix B.3.

Models We evaluate 7 open-source and 3 closed-source LVLMs on GAOKAO-MM, and deploy the open-source models through their checkpoints from Huggingface and access closed-source models via their API calls.

Open-source models include: (1) *VisCPM-Chat* (Hu et al., 2024); (2) *VisualGLM-6B* (Du et al., 2022); (3) *CogVLM-Chat* (Wang et al., 2024); (4) *LLaVA-1.5* (Liu et al., 2023a) We test on its 7B and 13B versions; (5) *mPLUG-Owl2* (Ye et al., 2023); (6) *Qwen-VL-Chat* (Bai et al., 2023).

Closed-source models are: (1) *GPT-4V* (OpenAI, 2023); (2) *Qwen-VL-Plus* (Bai et al., 2023); (3) *Gemini-Pro-Vision* (Team et al., 2023).

3.2 Results

Overall Performance The accuracies of all of the LVLMs are lower than 50%, and GPT-4V ranks the first with an accuracy of 48.1%. This reflects GAOKAO-MM poses a considerable challenge to the multimodal models due to its high demands on human-level capabilities, including perception, understanding, knowledge and reasoning.

Performance of Open-Sourced and Closed-Source Models All of the open-source models reach accuracy rates lower than 25%, and the gap between the open-source and closed-source models is greater than 11%. This indicates there is still huge room for improvement for the open-source community.

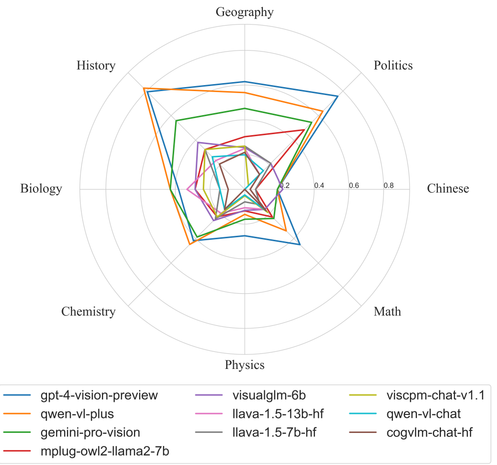

Figure 2: Different Performance in Subjects.

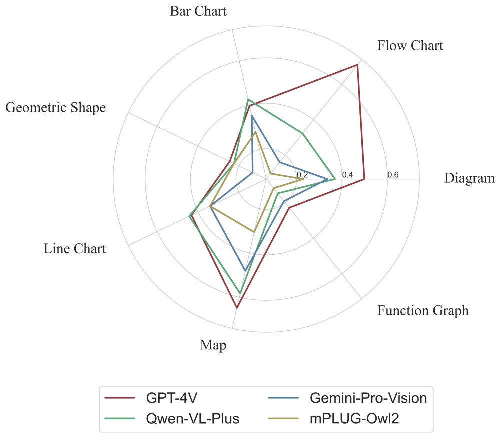

Figure 3: Different Performance in Image Types.

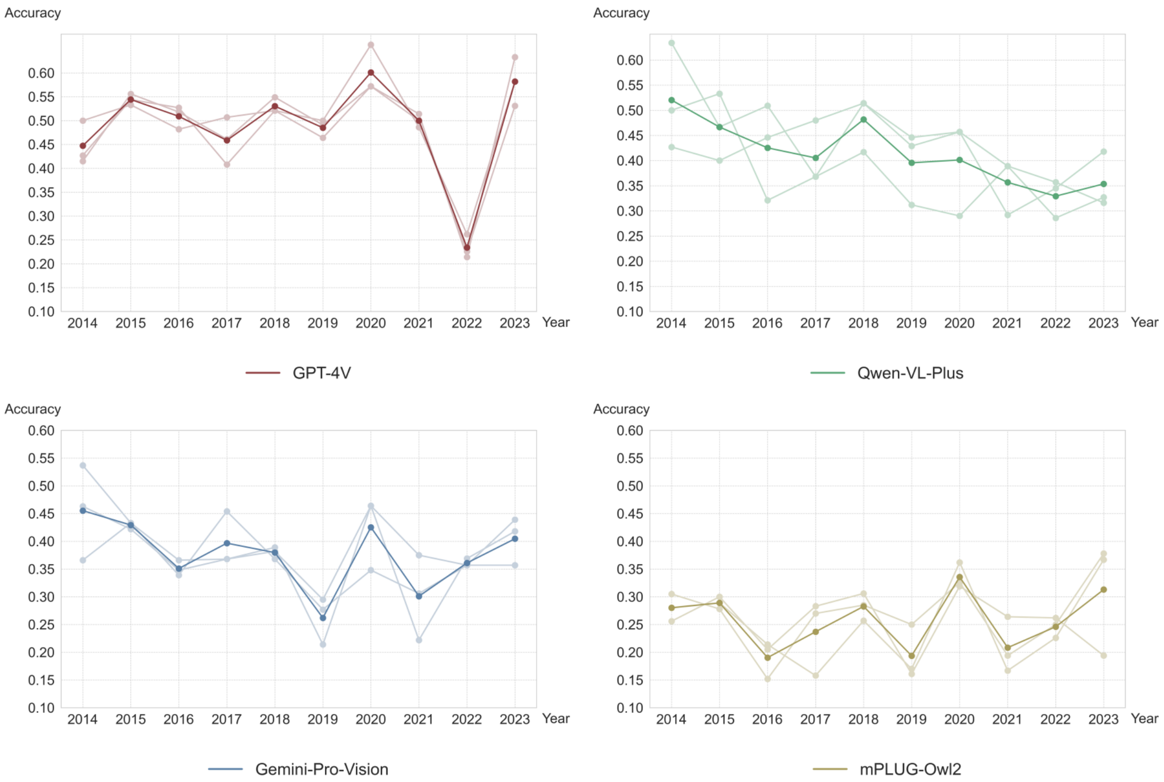

Figure 4: **Difference in Annual Trends.** The light-colored lines represent the accuracy obtained from three tests, while the dark-colored line represents the average accuracy.

4 Analysis

4.1 Difference in Subjects

Models perform differently among the 8 subjects as is shown in Figure 2. Most models perform worse on mathematics and physics, which can be attributed to the the weak reasoning and calculation abilities. Notably, the accuracy of Chinese is the lowest for most models due to the reason that the average length of Chinese questions is the longest and models have to identify details in the text that correspond to the information in the images. It reveals that fine-grained understanding of long texts and images remains a significant challenge for LVLMs.

4.2 Difference in Image Types

We present the accuracy rates of the top 4 models on the primary image types in Figure 3. All of them have accuracy rates exceeding their average accuracies on map-type images, while none of them excel in function graphs and geometric shapes. This is related to the imbalanced distribution of image types in image datasets and the varying demands of different images on specific domain knowledge. However, GPT-4V remarkably outperforms other models on flow charts, which reflects its superior logical reasoning abilities.

4.3 Difference in Annual Trends

We observe differences in the performance of models on questions from various years in Figure 4. The top four models do not exhibit consistent trends in performance over the past decade's questions. And the accuracy of the same model can vary by up to 20% across different years. Considering the consistent difficulty and variety of questions in the GAOKAO each year, we believe that this fluctuation is related to the robustness of large-scale models. Their output quality is sensitive to the input questions and images. This might be one of the focal points for future research.

5 Conclusion

In this work, we propose GAOKAO-MM, a novel Chinese multimodal benchmark focusing on the human-level capabilities of LVLMs, including perception, understanding, knowledge and reasoning. Deriving from Chinese College Entrance Examination(GAOKAO) and comprising of diverse subjects and images, it is well-suited for accessing comprehensive abilities in Chinese context. We evaluate 10 prominent LVLMs such as GPT-4V, Gemini-Pro-Vision and Qwen-VL-Plus, and the accuracies of them are lower than 50%, demonstrating considerable room for improvement. Our multi-dimension analysis reveal that the models' mathematical reasoning ability and comprehension of long texts and images still need to be enhanced, and the models' robustness to different texts and images is one of the future research directions. We believe GAOKAO-MM can propel the development of LVLMs towards AGI and provide insights into their applications in areas such as education.

Limitations

While we make contributions to Chinese multimodal benchmarks, there are still some limitations in this work. Firstly, due to the cost of collecting high-quality data, our benchmark covers the GAOKAO questions from 2010 to 2023 and has room for improvement in terms of the scalability and balanced distribution of data. Secondly, the evaluation lacks detailed analysis of the inference process of LVLMs. We plan to maintain GAOKAO-MM as an ongoing project, supplementing it annually with new questions from the GAOKAO.

Ethics Statement

In the process of developing GAOKAO-MM, we strictly adhere to ethical and legal standards. In China, the copyright of GAOKAO questions is generally considered to be in the public domain, allowing us to use them as an open evaluation dataset. We ensure that no private or sensitive information is included. We believe that open access to this dataset will contribute to the advancement of research in the field of multimodal models while maintaining high ethical standards.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (No.2022ZD0160102). The computations in this research were performed using the CFFF platform of Fudan University. We extend our gratitude to all reviewers for their valuable feedback, and we also wish to thank Botian Jiang, Bo Wang, Qingyuan Cheng and Jiachun Li for their insightful discussions.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, and Dmitry Lepikhin et al. 2023. Palm 2 technical report.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Large multilingual models pivot zero-shot multimodal learning across languages.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player?

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2022b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read.

Gemini Team, Rohan Anil, and Sebastian Borgeaud et al. 2023. Gemini: A family of highly capable multimodal models.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogvlm: Visual expert for pretrained language models.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark.

A Key Statistics of GAOKAO-MM

In this appendix, we show the distribution of image types in GAOKAO-MM in Figure 3. The statistics of 8 subjects are shown in Table3.

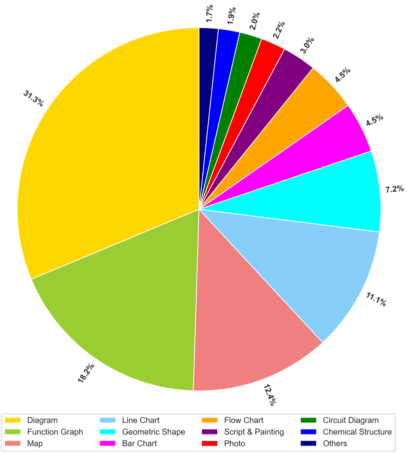

Figure 5: Distribution of Image Types in GAOKAO-MM

Question Type	Type	Number	Avg. Length
Geography MCQs	single choice	221	179.8
Physics MCQs	multi choice	174	299.1
Math MCQs	single choice	80	157.8
Chemistry MCQs	single choice	67	399.1
History MCQs	single choice	34	115.1
Political Science MCQs	single choice	33	183.1
Biology MCQs	single choice	21	257.0
Chinese Practical Lit	multi choice	16	1349.3

Table 3: Statistics of Question Types.

B Examples

In this appendix, we present examples of prompts used in the zero-shot setting experiment, questions and corresponding explanations and the outputs of GPT-4V, Qwen-VL-Plus, Gemini-Pro-Vision and mPLUG-Owl2. The English translation is added by the author for readers' understanding.

B.1 Examples of Prompts

We design prompts to standardize the output format of the models and leverage their reasoning abilities. And we present an example of a single-choice question prompt and an example of a multiple-choice question prompt below.

- **Single-Choice Questions**

请你做一道数学选择题。

请你结合文字和图片一步一步思考,并将思考过程写在【解析】和<eoe>之间。你将从A，B，C，D中选出正确的答案，并写在【答案】和<eoa>之间。
例如：【答案】: A <eoa>
完整的题目回答的格式如下：
【解析】...<eoe>
【答案】...<eoa>
请你严格按照上述格式作答。
题目如下：

Please solve the following mathematical multiple-choice question:
Please integrate textual explanations with accompanying diagrams to guide step-by-step problem-solving. Record your thought process between 【Analysis】and <eoe>. Select the correct answer from options A, B, C, and D, and indicate it between 【Answer】and <eoa>.
For instance: 【Answer】A<eoa>
The format of the complete question answer is as follows:
【Analysis】...<eoe>
【Answer】...<eoa>
Ensure adherence to the prescribed format for comprehensive question-answering.
The question is as follows:

- **Multiple-Choice Questions**

请你做一道语文阅读理解题。
请你结合文字和图片一步一步思考,并将思考过程写在【解析】和<eoe>之间。你将从A，B，C，D，E中选出所有符合题意的答案，并写在【答案】和<eoa>之间。
例如：【答案】AB <eoa>
完整的题目回答的格式如下：
【解析】... <eoe>
【答案】... <eoa>
请你严格按照上述格式作答。
题目如下：

Please complete a Chinese reading comprehension question. Please integrate textual explanations with accompanying diagrams to guide step-by-step problem-solving. Record your thought process between 【Analysis】and <eoe>. Select all suitable answers from A, B, C, D, and E, and indicate it between 【Answer】and <eoa>.
For instance: 【Answer】AB <eoa>
The format of the complete question answer is as follows:
【Analysis】...<eoe>
【Answer】...<eoa>
Ensure adherence to the prescribed format for comprehensive question-answering.
The question is as follows:

B.2 Examples of Questions and Explanations

In this appendix, we present an example of history questions and its explanations annotated by human. Each question in GAOKAO-MM is equipped with detailed explanation to analysis the reasoning process of LVLMs.

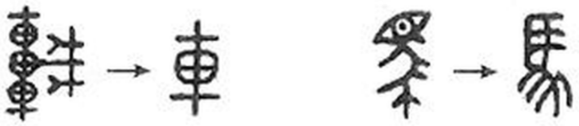

Figure 6: Image of An Example of History Questions.

- **Question**

1. 如图是西周与战国两个时期相同文字的不同写法, 反映出字形发生了变化, 促成这一变化的主要因素是( )

A. 文字的频繁使用
B. 书写材料的不同
C. 各国变法的实施
D. “书同文”的推行

The picture shows the different writing methods of the same characters in the Western Zhou Dynasty and the Warring States Period, reflecting changes in the shape of the characters. The main factors that contributed to this change are ( )
A. Frequent use of writing
B. Differences in writing materials
C. Implementation of reforms in various countries
D. Promotion of "shu tongwen"

- **Answer**

A

- **Explanation**

【解答】
A. 中国古代汉字发展的总体趋势是由繁到简, 之所以出现这样的变化是由于中国的现实, 西周到战国时期社会政治经济市场化不断的发展, 文字在社会发展中的表述功能越来越强, 被频繁地使用, 所以必须由繁到简便于书写。
B. 从西周到战国时期, 我国出现了简读, 都是写在竹和木片上, 而且都用笔和墨。
C. 各国变法从春秋时期开始，但是各国变法中并没有涉及到文字的。
D. 书同文是秦始皇推行的措施, 和材料时代不符。
故选: A。

【Analysis】
A. The overall trend of the development of ancient Chinese characters is from complexity to simplicity. The reason for such changes is due to the continuous development of society, politics, economy, and marketization from the Western Zhou Dynasty to the Warring States Period in China. As writing became increasingly important in social development, it was used frequently, hence the need for simplification for ease of writing.
B. From the Western Zhou Dynasty to the Warring States Period, simplification emerged in China, with writing done on bamboo and wooden slips, using pens and ink.
C. Reforms in various countries began in the Spring and Autumn Period, but they did not involve changes to writing systems.
D. "Shu tongwen" was a measure implemented by Emperor Qin Shi Huang, which is not consistent with the materials used during that period.
The answer is A.

B.3 Examples of Model Outputs

In this appendix, we show the corresponding outputs of the history questions mentioned above generated by GPT-4V, Qwen-VL-Plus, Gemini-Pro-Vision and mPLUG-Owl2.

The analysis provided by GPT-4V is the most comprehensive, but it fails to grasp the key aspect of the options; Qwen-VL-Plus and Gemini-Pro-Vision each analyzes only one option, with Qwen-VL-Plus overlooking crucial information in the question, while Gemini-Pro-Vision's reasoning and analysis are correct; mPLUG-Owl2, on the other hand, lacks a meaningful analytical process altogether.

- **Qwen-VL-Plus Output**

【解析】本题考查的是秦朝巩固统一的措施。秦始皇统一六国后，实行“书同文”，把七国文字统一成小篆，故选D。
【答案】D

This question examines the measures taken by the Qin Dynasty to consolidate its unification. After the unification of the six states by Emperor Qin Shi Huang, the policy of "shu tongwen" was implemented, unifying the scripts of the seven states into Xiaozhuan (small seal script). Therefore, option D is chosen.
Answer：D

- **mPLUG-Owl2 Output**

【解析】: 在这个问题中，我们需要找到西周和战国时期相同的文字，但是它们的写法不同。这表明字形发生了变化，因此我们需要找到这种变化的主要原因。
【答案】: B. 书写材料的不同

In this question, we need to identify characters that are the same between the Western Zhou Dynasty and the Warring States Period but are written differently. This indicates changes in character shapes, so we need to identify the main reasons for these changes.
Answer：B. Differences in writing materials

- **GPT-4V Output**

【解析】此题考查的是中国古代文字的发展变化。西周时期的文字主要是金

文，而战国时期的文字则逐渐演变为秦篆。金文的书写多用于铸造在青铜器上，而秦篆的书写则更多地出现在竹简和帛书上。从西周到战国，书写材料的变化对文字的演变起到了重要作用。竹简和帛书的使用使得书写更加便捷，书写工具从刻刀变为毛笔，这促使了文字形态由繁复向简化发展。因此，书写材料的不同是促成字形变化的主要因素。选项A“文字的频繁使用”虽然也对文字的演变有一定影响，但不是主要因素；选项C“各国变法的实施”与文字演变的直接关系不大；选项D“‘书同文’的推行”是秦朝统一六国后的事情，与战国时期文字的变化无直接关系。因此，最合适的答案是B。
【答案】: B

This question examines the development and changes in ancient Chinese characters. During the Western Zhou Dynasty, the predominant script was the Jinwen (inscriptions on bronze vessels), while in the Warring States Period, the script gradually evolved into Qin Zhuan (seal script). Jinwen was mainly used for casting on bronze vessels, whereas Qin Zhuan appeared more frequently on bamboo slips and silk books. The change in writing materials from the Western Zhou Dynasty to the Warring States Period played a significant role in the evolution of characters. The use of bamboo slips and silk books made writing more convenient, and the transition from engraving tools to brushes facilitated the simplification of characters from complexity to simplicity. Therefore, the difference in writing materials is the primary factor contributing to changes in character shapes. Option A, "Frequent use of writing," although influencing the evolution of characters to some extent, is not the main factor. Option C, "Implementation of reforms in various countries," has little direct relevance to the evolution of characters. Option D, "Promotion of ’shu tongwen’," occurred after the unification of the six states by the Qin Dynasty and is not directly related to the changes in characters during the Warring States Period. Therefore, the most appropriate answer is B.
Answer：B

- **Gemini-Pro-Vision Output**

【解析】从图中可以看出，西周时期的文字较为繁杂，而战国时期的文字较为简单。这一变化的主要原因是文字的频繁使用。随着社会的发展，人们对文字的需求量越来越大，为了提高书写的效率，就需要简化字形。此外，书写材料的不同也会导致字形的变化。西周时期，人们主要使用甲骨和金文来书写，而战国时期，人们开始使用竹简和纸张来书写。这些书写材料的特性不同，也对字形的变化产生了一定的影响。
【答案】A

From the image, we can see that the characters during the Western Zhou Dynasty were more complex, while those during the Warring States Period were simpler. The main reason for this change is the frequent use of writing. With the development of society, there was an increasing demand for writing, leading to the need for simplifying character forms to enhance writing efficiency. Additionally, differences in writing materials can also lead to changes in character shapes. During the Western Zhou Dynasty, people primarily used oracle bones and Jinwen (inscriptions on bronze vessels) for writing, while during the Warring States Period, they began using bamboo slips and paper. The characteristics of these writing materials also had an impact on the changes in character shapes.
Answer：A