This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

EEE-Bench: A Comprehensive Multimodal Electrical And Electronics Engineering Benchmark

Ming Li[1*] Jike Zhong[2*] Tianle Chen[3] Yuxiang Lai[4] Konstantinos Psounis[2]
[1]University of Tokyo [2]University of Southern California [3]Boston University [4]Emory University

Abstract

*Recent studies on large language models (LLMs) and large multimodal models (LMMs) have demonstrated promising skills in various domains including science and mathematics. However, their capability in more challenging and real-world related scenarios like engineering has not been systematically studied. To bridge this gap, we propose EEE-Bench, a multimodal benchmark aimed at assessing LMMs' capabilities in solving practical engineering tasks, using electrical and electronics engineering (EEE) as the testbed. Our benchmark consists of **2860** hand-picked and carefully curated problems spanning **10** essential subdomains such as analog circuits, control systems, etc. Compared to other domains, engineering problems are intrinsically 1) more visually complex and versatile and 2) less deterministic in solutions. Successful solutions to these problems often demand more-than-usual rigorous integration of visual and textual information as models need to understand intricate images like abstract circuits and system diagrams while taking professional instructions. Alongside EEE-Bench, we provide extensive quantitative evaluations, fine-grained analysis, and improvement methods using **17** widely-used open- and closed-sourced LLMs and LMMs and **7** popular prompting techniques. Our results reveal notable deficiencies in current foundation models for EEE, including an average performance ranging from 19.48% to 46.78% and a tendency toward "laziness" in overlooking essential visual context. In summary, we believe EEE-Bench not only reveals some noteworthy limitations of LMMs but also provides a valuable resource for advancing research on their application in practical engineering tasks, driving future improvements in their capability to handle complex, real-world scenarios. The EEE-Bench is publicly released on this link.*

1. Introduction

Electrical and electronics engineering (EEE) embodies a profound synthesis of theoretical knowledge and practical application, serving as the backbone of modern digital and power technological advancements [5, 6, 34]. From designing intricate circuitry to advancing power systems, this discipline demands rigorous logical reasoning and a deep understanding of physical laws [7, 70]. Addressing problems in EEE requires that researchers and engineers not only grasp the related theoretical knowledge but also comprehend the associated visual contexts, such as circuit diagrams and signal waveforms. Leveraging machine learning models equipped with robust knowledge and reasoning capabilities, alongside accurate perception of visual contexts, can significantly enhance various aspects of our lives. These models can aid in intricate hardware design processes [10, 50, 87], optimize operations within the electric energy sector [57], drive innovation by proposing novel approaches to traditional scientific challenges [75], and assist in solving complex educational problems.

Recent Large Language Models (LLMs) [2, 8, 18, 35, 78] and Large Multimodal Models (LMMs) [21, 48, 76, 97] have achieved much progress and demonstrated remarkable capabilities in various domains, e.g., mathematical reasoning [90, 91] and general science [54]. To evaluate the abilities of these foundation models, various specialized or versatile benchmarks [56, 92, 99] have been developed. However, the reasoning ability of these foundation models on engineering problems with visual contexts has not been systematically examined. Therefore, to better harness the potential of foundation models, it is crucial to develop a new benchmark specific to engineering problems, and we select EEE as our case study for the reasons mentioned before. This will facilitate the evaluation of LLMs and LMMs in terms of their ability to tackle rigorous engineering reasoning tasks, thereby advancing research progress in this field.

Compared with other subjects such as mathematics, the visual contexts in EEE are often more abstract and heavily logic-oriented. Solving EEE problems demands specialized technical knowledge, as illustrated in Figure 1. Specifically, unlike mathematical problems [56, 81, 98] which might involve clear numerical data and operations, visual representations in EEE, such as circuit diagrams, are intricate and logical, often containing minimal textual information (more

*Equal contribution.

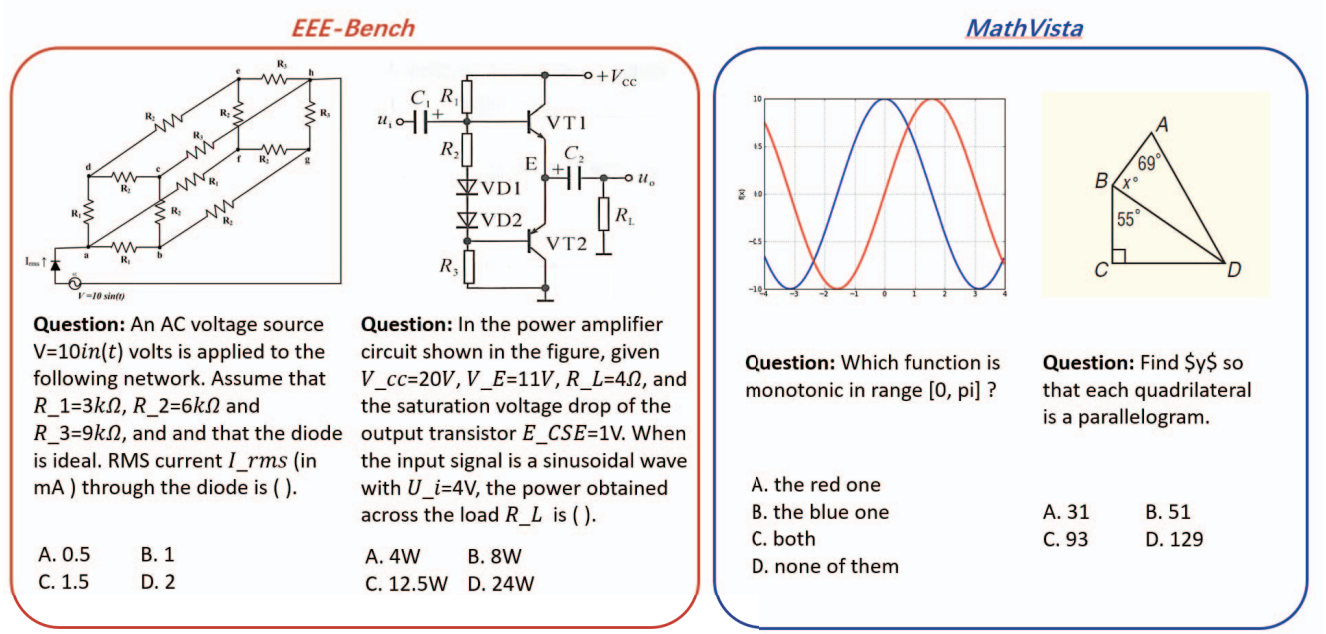


Figure 1. Comparison between EEE-Bench and MathVista problems: EEE problems feature significantly more complex and intricate visual information compared to mathematical problems, requiring specific technical knowledge to interpret diagrams effectively.

example images can be found in Appendix F). Hence effective problem-solving in this domain requires a deep understanding of the logical relationships between different components of the diagram and the application of specific technical knowledge to interpret these visuals.

In view of the above, the proficiency of existing LMMs in mathematical domains does not guarantee their effectiveness in EEE tasks. Owing to the same cause, existing benchmarks in other domains [22, 55, 56, 81, 92, 94, 98] also lack some necessary rigor to accurately faithfully inspect LMMs' problem-solving capability in challenging real-world scenarios where the visual context is crucial but involves serious complexity. Consequently, there is a pressing need for a specialized benchmark focused on EEE to thoroughly assess the multimodal reasoning abilities of LMMs within this specific field as well as to provide a broader outlook of their problem-solving capabilities in these practical engineering problems in the real world.

To this end, we present EEE-Bench, a pioneering multimodal EEE reasoning benchmark in visual contexts. To ensure that EEE-Bench encompasses a thorough range of knowledge in EEE, we initially identified ten pivotal topics for problem collection: (1) *Digital Logic Circuits and Microprocessor Design*, (2) *Circuit Theory and Network Analysis*, (3) *Analog Circuits*, (4) *Power Electronics and Power Systems*, (5) *Signals and Systems*, (6) *Communication Systems*, (7) *Control Systems*, (8) *Electronic Devices*, (9) *Electrical Machines*, and (10) *Electromagnetics*. These topics comprehensively cover both theoretical foundations and practical applications within EEE. When collecting problems for EEE-Bench, we include a broad range of visual contexts such as electric and digital circuits, system diagrams, abstract scenes, electronic component images, and assorted tables, charts, and plots. Overall, EEE-Bench comprises 2,860 entirely new problems.

We conduct extensive experiments using EEE-Bench to assess the reasoning abilities of 17 leading foundation models, including both open-source [16, 45, 48] and closed-source [62, 76]. Figure 2 illustrates the overall experimental results. We discover that most existing LMMs struggle to understand EEE diagrams and perform poorly on EEE-Bench. Our results show that GPT-4o [62] achieves the best overall performance across different topics with 46.78% accuracy. Notably, closed-source models generally outperform open-source models. When evaluating performance across various subjects, we observed that many closed-source models excel in subjects with straightforward visual contexts but require significant computational resources, such as Electronic Devices. However, these models tend to struggle with subjects that involve complex visual diagrams, such as Circuit Theory and Network Analysis. These findings underscore the proficiency of closed-source LMMs in numerical computation, but also reveal their limitations in processing intricate visual information. Through the fine-grained error analysis, we found that current LMMs struggle to understand EEE diagrams and reason effectively. Taking GPT-4o as an example, we found that over 50% of the errors stem from reasoning issues, and 26.5% are due to errors in image perception. These findings suggest that better EEE visual understanding capabilities and more robust

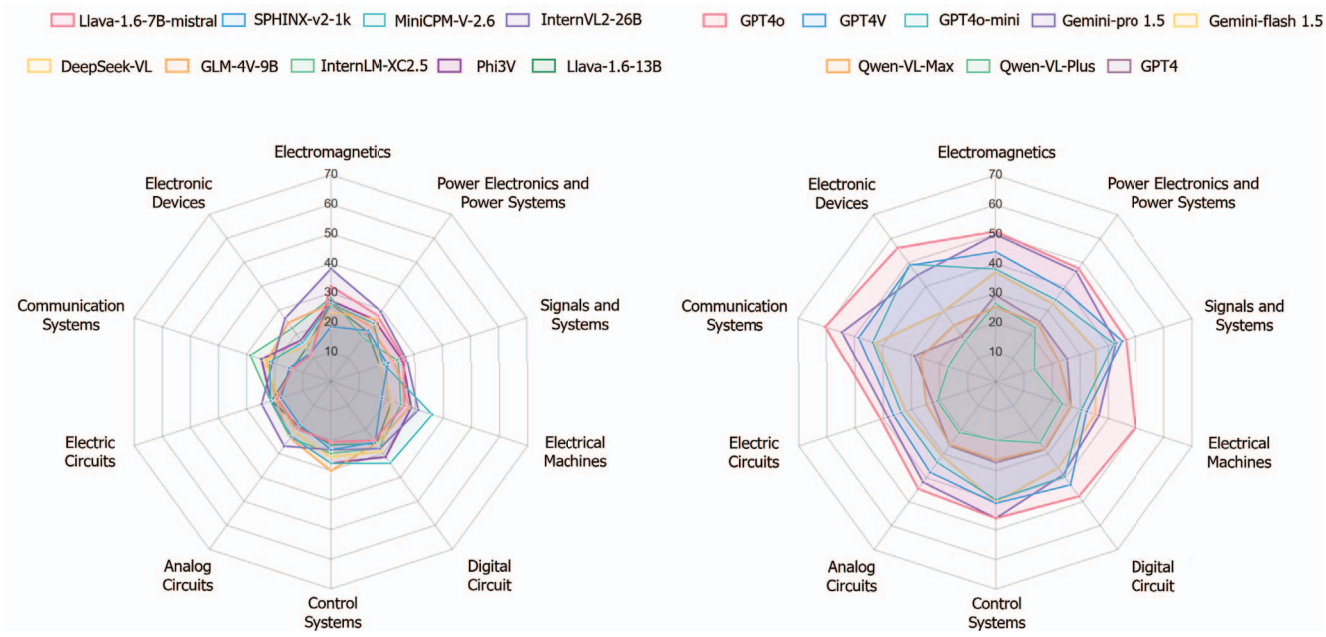


Figure 2. Overview of performance of **left: open source** models and **right: close source models**.

EEE reasoning for LMMs could be the potential for future advancement in addressing multi-modal EEE problems.

Lastly, we analyze which modality—vision or text—LMMs predominantly rely on. Our investigation revealed a "laziness" phenomenon within existing LMMs. In contrast to previous studies that concentrated on hallucinations of nonexistent objects with natural images [46, 49, 66], our focus is on the robustness of the understanding and reasoning process in LMMs dealing with technical images. Specifically, we found that when presented with spurious captions that contradict the accompanying images, LMMs tend to disregard essential visual information, relying instead on the textual content and reasoning, even if it is misleading. This tendency leads to flawed reasoning processes within the models. Consequently, augmenting original question text with such spurious captions results in significant accuracy declines—7.79% for GPT-4o and 6.78% for Gemini-Pro 1.5 on the EEE-Bench. These results reveal that existing LMMs mainly rely on text information and may overlook visual information even if it is explicitly provided in text.

Overall, our contributions can be summarized as follow:

- We introduce EEE-Bench, a novel multimodal benchmark designed for assessing the reasoning abilities of LMMs in EEE problems. EEE-Bench consists of 2860 samples spanning 10 essential subjects in EEE, featuring a diverse range of visual contexts such as electric and digital circuits. This benchmark is specifically crafted to evaluate how well LMMs can handle complex visual and logical challenges within the EEE field.
- We conduct extensive experiments using EEE-Bench to assess the reasoning abilities of 17 leading open-source and closed-source models. We found that most existing LMMs struggle to deal with EEE problems, especially in subjects with complex visual contexts, yielding an average performance ranging from 19.48% to 46.78%. In turn, we studied and found unique challenges of technical images and strived to resolve these issues using 7 popular prompting techniques for multimodal.
- We further explore which modality—text or vision—current LMMs primarily rely on. Our research identifies a "laziness" phenomenon in these models: LMMs will not try to read images if relevant visual information for reasoning already appears in the text, even if this extra-textual information is spurious.

Together, these findings provide valuable insights for future development of improving LMMs in the context of challenging, real-world technical images.

2. The EEE Benchmark

In this section, we introduce **EEE-Bench**. First, we introduce our data collection process in subsection 2.1. Next, we provide a summary of EEE-Bench in subsection 2.2. Lastly, we provide a high-level analysis in Appendix F.

2.1. Data Collection

Guidelines. As our general principle is to establish a challenging real-world-oriented benchmark capable of assessing both visual understanding and logical reasoning with integrated multimodal knowledge, we adhere to the following guidelines for initial data collection: 1) questions with

high-resolution and context-rich images (diverse complex patterns, domain-specific markings, etc); 2) problems that require logical reasoning steps (calculations, cross-domain information integration) to solve in addition to visual comprehension; 3) problems with deterministic answers for robust evaluation (multiple choices questions and free-form questions requiring answer no longer than one word as did in [56]); 4) balanced collection source, sub-domain representation, and problem type.

Collection and curation. We collect all our questions from official EEE exams as well as verified online sources in a multimodal setting Note that as our work is first in this direction, no existing dataset or benchmark is available, hence all questions are newly collected. For data curation, we employ a two-stage filtering process. To ensure the solution quality, we first filter for questions appearing in at least two distinct sources having identical solutions. To ensure the significance of visual input, we subsequently adopt methods similar to [93] where two trials are run using text-only GPT-4 [62] in which questions with two correct answers are eliminated. These processes result in 2860 high-quality examples across ten sub-domains. Two final quality checks for typos, image resolution, and prompt completeness were also conducted by graduate students.

2.2. Benchmark Summary

Our benchmark consists of a total of 2860 examples divided into 10 diverse but essential sub-domains. We first summarize general statistics. Questions in our benchmark are categorized into two types similar to [56]: multiple choice and free-form, counting 2059 and 801, respectively. Regarding answer type, four forms are present: text (single word) represents the largest group of 2059 questions, 595 are float, 204 are integer, and 2 are list. The average token length for question text is 306. For clarity, we provide a more detailed summary of EEE-Bench in Appendix F. In addition, we provide more information regarding the wide coverage and significance of vision in our benchmark in Appendix F.

3. Experiments

In this section, we conduct a comprehensive evaluation of 17 existing LLMs and LMMs on EEE-Bench. subsection 3.1 introduces the experimental setup. Subsequently, subsection 3.2 presents the quantitative results. Finally, section 3.3 provides a detailed, fine-grained error analysis.

3.1. Experiment Setup

Model selection. We evaluate a total of 1 LLM and 16 LMMs against EEE-Bench. The abundant model selection captures the diversity of LMMs today and ensures a thorough and comparable assessment. In detail, the LMM set includes **(a) 9 open-source** LMMs: LLaVA1.6-7B, LLaVA-1.6-13B [47, 48], DeepSeek-VL-7B [52], Phi3-Vision-4.2B [1], MiniCPM-Llama3-V 2.6 [88], SPHINX-v2-1k [45], InternVL2-26B [16, 17], GLM-4V-9B [26], and InternLM-XC2.5 [96], **(b) 7 closed-source** LMMs: GPT-4o [62], GPT-4o-mini [63], GPT-4V [61], Gemini 1.5 Pro, Gemini 1.5 Flash [69, 76], Qwen-VL-Max and Qwen-VL-Plus [4]. To further ground the impact of vision in LMMs, we provide two additional input-only evaluation baselines using GPT-4 [2]: 1) a text-only baseline where the visual inputs are entirely hidden and 2) an improved baseline of text-only input but with image captions generated by GPT-4o. The prompt for caption generation and the details about all models can be found in Appendix C.

Evaluation. We utilize an evaluation pipeline similar to [56]. Raw responses are first generated from LMMs and then processed by GPT-4o-mini to extract the concise answers given an exemplary system prompt. Finally, accuracy is obtained through calculation against the ground truth.

Implementation details. We evaluate all models in a zero-shot manner for better generalization examination. We run each evaluation setting twice for each LMM and record their average scores for the final report. Hyperparameters and model setups are provided in Appendix E.

3.2. Experiment Results

In this section, we compare the performance of 17 open-source and closed-source models. We report the overall accuracy and the accuracy among each subject in Table 1.

Finding 1: Existing LMMs struggle to address EEE problems. The results presented in Table 1 highlight that existing LMMs can not deal well with EEE problems in visual contexts. A representative failure case is shown in Appendix D. For example, GPT-4o achieves the highest average accuracy of just 46.78% (less than half), and the average accuracies of proprietary and open-source models are only 35.28% and 23.19%, respectively. These results indicate that current LMMs are inadequate in providing satisfactory assistance for real-world engineering tasks with their existing performance levels.

Finding 2: Proprietary models exhibit superior generalization compared to open-source models. Similar to prior works on benchmarking the performance of LLMs and LMMs in mathematics and sciences [56, 92], we found it evident that a significant performance disparity exists between closed-source and open-source models, as shown in Table 1. The top two best-performing closed-source LMMs, GPT-4o and Gemini-Pro 1.5, significantly outperform their open-source counterparts. For instance, GPT-4o, the best-performing closed-source LMM, has an average accuracy of 46.78%, while the highest-scoring open-source LMM, InternVL2-26B, achieves only 26.89%—a gap of nearly 20

Model	ALL	EMag	PEPS	SS	EMac	DLC	CorS	AC	CTNA	ComS	ED
Random Chance	17.45	28.43	19.65	15.22	18.01	18.55	15.90	18.18	16.89	15.84	10.50
LLMs (text input)											
GPT-4	25.52	29.41	25.33	25.36	26.71	28.23	27.20	26.59	22.47	28.71	19.18
GPT-4 w/ caption	28.22	28.43	28.82	26.81	31.06	31.25	25.10	25.21	27.20	34.65	30.59
Open-source LMMs											
Llava-1.6-7B	21.36	32.35	27.51	26.09	26.09	24.60	20.50	19.55	19.09	13.86	11.42
Llava-1.6-13B	20.73	26.47	20.96	17.39	21.74	26.01	21.34	18.70	21.11	13.86	14.16
SPHINX-v2-1k	19.48	18.63	21.40	20.29	18.01	25.40	23.01	18.01	17.74	14.85	11.87
MiniCPM-V-2.6	25.21	26.47	24.45	18.84	36.02	34.07	27.62	22.98	21.45	21.78	16.44
InternVL2-26B	26.89	38.24	29.26	26.81	31.06	28.02	23.01	26.93	24.66	20.79	26.48
DeepSeek-VL	22.31	23.53	26.20	17.39	21.74	29.23	25.52	21.10	18.58	23.76	14.61
GLM-4V-9B	24.06	26.47	23.14	22.46	27.95	26.01	30.13	23.50	19.76	22.77	24.66
InternLM-XC2.5	23.85	28.43	18.34	23.91	24.84	27.82	24.27	23.50	21.45	28.71	22.37
Phi3V	24.90	27.45	25.76	25.36	28.57	31.45	27.62	22.81	21.28	24.75	17.35
Closed-source LMMs											
GPT-4o	46.78	50.98	47.60	46.38	49.69	47.78	46.03	44.60	40.88	60.40	56.16
GPT-4V	40.21	44.12	38.86	44.93	32.30	42.94	41.00	37.74	36.32	48.51	48.86
GPT-4o-mini	36.99	38.24	34.50	42.03	30.43	39.52	39.75	33.62	32.77	43.56	49.32
Gemini-Pro 1.5	41.99	50.00	46.29	42.75	36.65	38.71	46.03	41.85	38.34	54.46	44.75
Gemini-Flash 1.5	33.95	37.25	32.75	35.51	35.40	35.89	40.59	30.87	31.93	42.57	29.68
Qwen-VL-Max	25.73	25.49	24.45	22.46	26.71	28.02	26.36	26.24	24.66	26.73	23.74
Qwen-VL-Plus	21.33	26.47	22.71	13.77	23.60	25.40	19.67	21.10	20.78	16.83	17.35

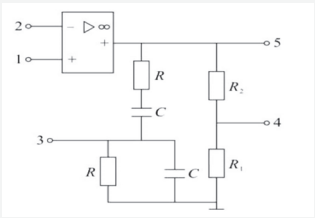
Table 1. Comparison of model performances across various EEE subjects. Subjects: EMag: Electromagnetics, PEPS: Power Electronics and Power Systems, SS: Signals and Systems, EMac: Electrical Machines, DLC: Digital Logic Circuits and Microprocessor Design, CorS: Control Systems, AC: Analog Circuits, CTNA: Circuit Theory and Network Analysis, ComS: Communication Systems, ED: Electronic Devices. The highest accuracy for **closed-source** and **open-source** LMMs is marked in **red** and **blue** respectively.

percentage points. Moreover, even the lightweight closed-source models, GPT-4o-mini and Gemini-Flash 1.5, surpass all open-source LMMs in performance. This discrepancy not only underscores the superior generalization capability of state-of-the-art (SOTA) closed-source models but also underscores the pressing need for the development of open-source LMMs that are proficient in EEE tasks.

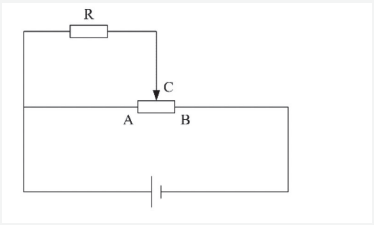
Finding 3: Diverse subdomains present different challenges. Problems across various subjects demand specific capabilities for effective resolution (refer to Appendix F for more details). This section delves into the performance of existing LMMs across diverse subjects and discusses their strengths and weaknesses. Most closed-source LMMs excel in handling problems related to Communication Systems (ComS) and Electronic Devices (ED), yet they perform significantly poorer on problems from Analog Circuits (AC) and Circuit Theory and Network Analysis (CTNA). For instance, the best-performing LMM GPT-4o achieves average accuracies of 60.40% and 56.16% in ComS and ED respectively, which are substantially higher than the overall average accuracy of 46.78%. In contrast, its performance drops to 44.60% and 40.88% in AC and CTNA, respectively, both of which fall below the overall average.

Problems within ComS and ED typically present straightforward visual information, yet solving them demands considerable computational effort. Conversely, problems from AC and CTNA often feature complex visual contexts, such as intricate circuit diagrams, necessitating the use of advanced mathematical and computational tools. Thus, tackling issues in these domains requires not just strong numerical capabilities but also sophisticated vision recognition and reasoning skills. This performance disparity highlights that while many existing closed-source LMMs are adept at managing computationally intensive tasks due to their robust numerical calculation capabilities, they struggle with EEE problems that require more nuanced vision recognition and complex reasoning abilities. This suggests a crucial area for further development in enhancing the vision and reasoning capabilities of LMMs to better handle the full spectrum of EEE challenges.

Conversely, open-source LMMs face challenges in handling tasks related to Coms and ED, despite the simplicity



Question: In the circuit shown in figure 10, if the parameters are reasonable and the circuit is to form an oscillator that can oscillate, the correct connections are ( )
(A). Connect 1-5, Connect 2-3
(B). Connect 1-4, Connect 2-3
(C). Connect 1-3, Connect 2-5
(D). Connect 1-3, Connect 2-4
Response from GPT-4o: To form an oscillator, the circuit needs to have the correct feedback and phase shift. In this case, the connections that will allow the circuit to oscillate are:
B. Connect 1-4, Connect 2-3.



Question: In the circuit shown in Figure, it is known that the sliding contact C of the variable resistor is at the midpoint of AB. The voltage across the resistor R in the circuit is ( ).
(A). $\frac{1}{2}U_{AB}$ (B). Less than $\frac{1}{2}U_{AB}$
(C). Greater than $\frac{1}{2}U_{AB}$ (D). U_{AB}
Response from GPT-4o: In the given circuit, the sliding contact C is at the midpoint of AB, which means the variable resistor is divided into two equal parts. Therefore, the voltage across each part of the resistor is half of the total voltage across AB.Thus, the voltage across the resistor R is: $\frac{1}{2}U_{AB}$.
The correct option is: A. $\frac{1}{2}U_{AB}$.

Figure 3. Error examples of response from GPT4o. Left: Image Perception Error. Right: Reasoning Error. Error is marked with red color.

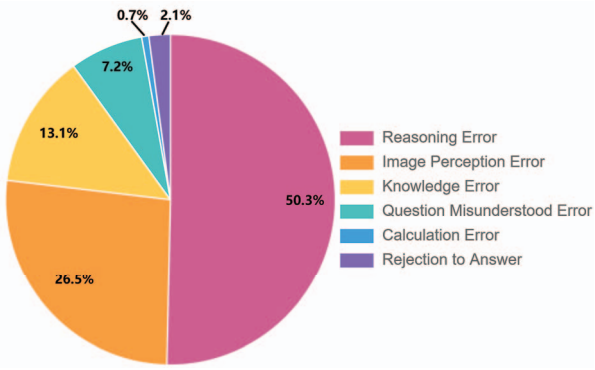


Figure 4. Error distribution of the wrong results from GPT-4o.

of the associated visual contexts. This performance discrepancy between open-source and closed-source LMMs underscores a significant shortfall in the computational capabilities of the open-source models. Urgent development is needed to enhance the computational efficiency and numerical problem-solving abilities of open-source LMMs.

3.3. Error Analysis

In this section, we provide an in-depth analysis of the common errors observed in evaluations on EEE-Bench using GPT-4o as an example. Figure 4 shows the breakdown by error measured by 6 types: reasoning error, image perception error, question misunderstood error, knowledge error, calculation error, and rejection to answer. We focus our analysis on the two largest error groups, **reasoning error** and **perception error**, constituting **50.3%** and **26.5%** of total error respectively.

Lack of reasoning. These happen when LMMs can read the image but cannot reason correctly to arrive at the answer. For instance, on the right of Figure 3, the LMM correctly reads the location of contact C but fails to consider the location of resistor R in the overall circuit, thus leading to the wrong answer.

Lack of visual understanding. These happen when LMM fails to answer correctly due to misunderstandings about the composition of the image. Concretely, the circuit diagram on the left of Figure 4 shows a filter circuit using an operational amplifier with multiple feedback loops involving resistors and capacitors. In such cases even if the explanation is fine the verdict is wrong as the diagram is more complex than current LMMs can handle.

Gap between open and proprietary models. One plausible cause for this gap is the lack of domain knowledge of open-source models. Hence, to ensure a fair comparison, we test open models with the captions and domain knowledge generated by proprietary models. We first prompt GPT-4o to generate two sets of domain priors: a) generic image caption and b) question-targeted knowledge. The results for LLaVA-1.6-7B/InternVL2-26B with this background knowledge as context is for a): 21.10/27.02, b): 21.24/27.40, and a)+b): 21.30/27.45(-0.06/+0.56). These suggest **reasoning capability** is indeed the likely cause.

Model		ALL	EMag	PEPS	SS	EMac	DLC	CorS	AC	CTNA	ComS	ED
GPT-4o	w/ caption	45.35	48.04	41.48	44.93	44.72	43.75	44.35	46.31	42.23	48.51	57.99
	Δ	-1.43	-2.94	-6.12	-1.45	-4.97	-4.03	-1.68	+1.71	+1.35	-11.89	+1.83
	w/ spurious caption	38.99	50.00	38.43	44.93	36.02	40.12	38.91	38.42	33.11	45.54	44.75
	Δ	-7.79	-0.98	-9.17	-1.45	-13.67	-7.66	-7.12	-6.18	-7.77	-14.86	-11.41
Gemini-Pro 1.5	w/ spurious caption	35.21	46.08	41.92	38.41	37.89	36.9	38.49	31.39	29.39	38.61	36.07
	Δ	-6.78	-3.92	-4.37	-4.34	+1.24	-1.81	-7.54	-10.46	-8.95	-15.85	-8.68

Table 2. Performance comparison of GPT-4o and Gemini-Pro 1.5 with and without (spurious) **captions**. Positive changes are in blue and negative changes in red.

Model	0-shot COT [37]	1-Shot CoT [86]	LTM [103]	PROGRAM [19]	Retrieval [25]	Self-Correction [65]	Self-Consistency [83]	DDCot [101]	Cantor [24]
GPT-4o	46.82 (+0.04)	46.59 (-0.19)	46.96 (+0.18)	46.26 (-0.52)	47.22 (+0.44)	48.26 (+1.48)	50.32 (+3.54)	49.12 (+2.34)	49.94 (+3.16)
Gemini-Pro 1.5	41.67 (-0.34)	41.67 (-0.32)	42.02 (+0.03)	40.64 (-1.35)	42.32 (+0.33)	43.23 (+1.24)	45.74 (+3.73)	44.89 (+2.9)	44.98 (+3.99)

Table 3. Performance comparison of GPT-4o and Gemini-Pro 1.5 using different **prompting strategies**. Positive changes are in blue and negative changes in red.

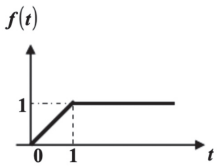
The examples above clearly underline the challenge of engineering problems to LMMs. We provide more of such instances in Appendix I. Compared to other domains such as mathematics, these results demonstrate that engineering problems are intrinsically more **complex** and **dynamic** as components in EEE domain could be arranged in very different but meaningful ways to form more complicated diagrams whereas their counterparts in, for example, mathematics are mostly stagnated and fixed. On the other hand, it highlights that LMMs struggle to understand electrical engineering diagrams and are particularly challenged by increased visual complexity and detailed components such as resistors and symbols. Overall, it indicates that solving these electric engineering problems requires more delicate incorporations of both visual and textual inputs.

4. Discussion

The unique challenge of EEE problems. subsection 3.2 demonstrates that even SOTA LMMs struggle to perform on EEE problems, and in subsection 3.3, we found that this may be due to the difficulty in visual understanding. To further validate our hypothesis, we conduct a simple experiment that appends an **image caption** extracted using GPT-4o to the original question prompt, aiming to provide direct visual assistance. The detailed steps can be found in Appendix C. Surprisingly, as shown in Table 3, the improvements on the EEE-Bench with such caption are **negative** across 7 out of 10 subdomains, with an average performance change of -1.43% using GPT-4o. We note that a similar procedure yields minor improvement on benchmarks such as MathVista [56]. In the case of EEE-Bench, the results demonstrate that simply adding captions cannot help, and more importantly, the extraction model failed to accurately extract information from the technical images and such inaccurate context has been used by the model to solve problems. The fact that even a captioner as strong as GPT-4o cannot extract accurate and helpful captions implies the unique challenge of **technical images**: high abstraction, versatility, and rich domain-specific details. Intriguingly, these observations lead to two additional questions which we investigate in the next paragraphs: ***1) are there other ways to improve the performance?*** and ***2) do LMMs rely more on vision or text?***

Improving performance on EEE-Bench. Here we seek to address question (1) in the previous paragraph. We experiment with GPT-4o [63] and Gemini-Pro 1.5 [76] with 9 popular and reliable prompting methods for LLM: 0-Shot CoT [37], 1-Shot CoT [86], LTM [103], PROGRAM [19], Retrieval [25], Self-Correction [65], Self-Consistency [83], DDCot [101], Cantor [24]. We leave these methods' detailed description and implementation to Appendix E. As shown in Table 3, only Self-Consistency, Cantor, and DDCot provide meaningful improvements. Interestingly, we found that while many of these methods are useful for language-only tasks, they are not always helpful for MLLMs, especially when challenging visuals are present.

Your LMM might be lazy. Here we investigate question (2) mentioned before. Previous works [46, 49, 66] have explored hallucination problems of recognition of nonexisting items in natural images. In contrast, our research delves into this issue with technical images, emphasizing the comprehension of image information and reasoning rather than object recognition. We design a straightforward experiment that prepends, instead of the accurate caption X_c, a spurious caption X_c' which contains a spurious caption of the image augmented from the original caption by GPT-4 (see details in Appendix C). We intentionally make such argumentation small in token but major in semantic effect, for example, changing "*increasing*" into "*decreasing*" when describing a function like shown in Figure 5. Note that such caption merely serves as extra context as we explicitly prompt LMMs to answer based on the *figure* or *image*. We display results on EEE-Bench in Table 3 and on MathVista [56] in Appendix D. Interestingly, across the board, the results



Question: Given that the waveform of $\boldsymbol{f(t)}$ is as shown in the figure, the expression for $\boldsymbol{f(t)}$ is ( ).
A. $\boldsymbol{f(t) = t\varepsilon(t) - t\varepsilon(t-1)}$ B. $\boldsymbol{f(t) = (1-t)\varepsilon(t) - (1-t)\varepsilon(t-1)}$
C. $\boldsymbol{f(t) = t\varepsilon(t) - (t-1)\varepsilon(t-1)}$ D. $\boldsymbol{f(t) = (1+t)\varepsilon(t) - (t+1)\varepsilon(t+1)}$

Spurious context: Image caption context:
The diagram is a plot of a function $f(t)$ versus time t. Here is a detailed description:
- Axes: - The horizontal axis represents time t. - The vertical axis represents the function $f(t)$.
- Function Behavior: - The function starts at the origin, $(0, 0)$. - From $t = 0$ to $t = 1$, the function decreases linearly from 1 to 0. This indicates a slope of -1. - At $t = 1$, the function reaches a value of 0. - For $t > 1$, the function remains constant at $f(t) = 0$.
- Key Points: - $f(0) = 1$ - $f(1) = 0$ - The slope from $t = 0$ to $t = 1$ is -1. - The function is constant for $t > 1$.
This plot represents a unit ramp function that transitions into a unit step function at $t = 1$.

Correct Response w/o spurious caption: The waveform $f(t)$ is a ramp function starting at $t = 0$ and ending at $t = 1$, where it becomes constant. This can be expressed using unit step functions $\varepsilon(t)$.
The correct expression is: $f(t) = t\varepsilon(t) - (t-1)\varepsilon(t-1)$ This matches option C.

Wrong Response w/ spurious caption: The function $f(t)$ starts at 1 and decreases linearly to 0 from $t = 0$ to $t = 1$, then remains constant at 0 for $t > 1$. This behavior can be represented using the unit step function $\varepsilon(t)$.
The correct expression for $f(t)$ is: $f(t) = (1-t)\varepsilon(t) - (1-t)\varepsilon(t-1)$ This matches the description of the function, where it decreases linearly from 1 to 0 and then stays at 0. The correct option is: **B**

Figure 5. Examples of LMM "laziness" in EEE-Bench. The spurious text in the caption and error in response is marked with red color. The LMM used here is GPT-4o.

show a significant negative impact, measuring a 7.79% and 8.4% drop on EEE-Bench and MathVista respectively using GPT-4o, where the same can be found with Gemini-Pro 1.5. Figure 5 shows a failure case on EEE-Bench where LMMs fail to reason correctly with a minor change in the caption of "*increasing*" into "*decreasing*", even if the function in the image is clearly increasing. More examples are displayed in Appendix J. Given the explicit instruction to follow the image, the results reveal a key shortcoming of LMMs: the tendency to rely more on text information when both visual and textual inputs are given. We note that a recent work [72] indeed finds that LLMs are easily disturbed by irrelevant content, we thus believe such an issue for LMMs is noteworthy for further study.

5. Related Works

Recently, Large Multimodal Models (LMMs) have been constructed and developed by integrating Large Language Models (LLMs) [8, 35, 78, 85] with large vision encoders [9, 15, 64, 68]. Notably, leading LMMs [13, 23, 40, 41, 47, 48, 77, 89, 97] such as GPT-4o [62] and Gemini-Pro 1.5 [76] have achieved exceptional visual reasoning capabilities across various domains, including Visual Question Answering [39] and Mathematics [56]. Recent efforts have also focused on leveraging large models for engineering tasks, such as chip design [5, 6, 11, 34] and optimizing operations in the energy sector [57].

To evaluate the strengths and limitations of these foundation models, thereby facilitating their ongoing development and refinement, various text-only benchmarks [29, 30, 36, 95, 102] and general-purpose Visual Question Answering (VQA) tasks [3, 27, 28, 33, 58, 59, 59, 74] have been proposed. To provide a multi-faceted evaluation for LMMs, several benchmarks have been introduced, including specialized ones [42, 53, 56, 67, 80, 84, 99] and versatile benchmarks [14, 39, 44, 51, 71, 82, 92].

However, the reasoning capabilities of existing LMMs in visual contexts have not yet been thoroughly examined for EEE problems. Therefore, this paper proposes an EEE-Bench to comprehensively evaluate the EEE multimodal reasoning abilities of current LMMs. To the best of our knowledge, we respectfully believe the EEE-Bench, which comprises 2860 meticulously curated problems from ten pivotal EEE subjects, is the first in the direction to offer a robust and comprehensive assessment of LMMs' reasoning abilities in complex visual contexts within EEE.

6. Conclusion

This study focuses on benchmarking modern LMMs' performance in challenging real-world scenarios such as in engineering where technical images are involved. We use EEE as the testbed and start by creating a pioneering benchmark termed EEE-Bench which consists of **2860** hand-picked and carefully curated questions spanning **10** subdomains. We evaluate **17** open and closed-source LLMs (1) and LMMs (16) on EEE-Bench and explore **7** popular strategies to improve on this task. Our results and analysis highlight the unique challenges of EEE-Bench, and more broadly, technical images in multimodal reasoning and understanding, raising the need for further investigation and more curated designs of LMMs while providing a valuable resource for advancing research in this direction.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 4

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 8, 14

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 4

[5] Bimal K Bose. Energy, environment, and advances in power electronics. In *ISIE'2000. Proceedings of the 2000 IEEE International Symposium on Industrial Electronics (Cat. No. 00TH8543)*, pages TU1–T14. IEEE, 2000. 1, 8, 14

[6] Bimal K Bose. Power electronics and motor drives: advances and trends. 2020. 1, 8, 14

[7] Jennifer AN Brophy and Christopher A Voigt. Principles of genetic circuit design. *Nature methods*, 11(5):508–520, 2014. 1

[8] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1, 8, 14

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 8, 14

[10] Kaiyan Chang, Ying Wang, Haimeng Ren, Mengdi Wang, Shengwen Liang, Yinhe Han, Huawei Li, and Xiaowei Li. Chipgpt: How far are we from natural language hardware design. *arXiv preprint arXiv:2305.14019*, 2023. 1

[11] Kaiyan Chang, Kun Wang, Nan Yang, Ying Wang, Dantong Jin, Wenlong Zhu, Zhirong Chen, Cangyuan Li, Hao Yan, Yunhao Zhou, et al. Data is all you need: Finetuning llms for chip design via an automated design-data augmentation framework. *arXiv preprint arXiv:2403.11202*, 2024. 8, 14

[12] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024. 14

[13] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 8, 14

[14] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, 2023. 8, 14

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 8, 14

[16] 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. 2, 4

[17] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 4, 21

[18] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 1

[19] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 7, 19

[20] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024. 14

[21] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[22] Anna C. Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation, 2024. 2, 14

[23] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 8, 14

[24] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, and Rongrong Ji. Cantor: Inspiring multimodal chain-of-thought of mllm, 2024. 7, 19

[25] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. 7, 19

[26] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 4

[27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 8, 14

[28] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 8, 14

[29] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 8, 14

[30] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 8, 14

[31] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 14

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 14

[33] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 8, 14

[34] J David Irwin and R Mark Nelms. *Basic engineering circuit analysis*. John Wiley & Sons, 2020. 1, 8, 14

[35] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 1, 8, 14

[36] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023. 8, 14

[37] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. 7, 19

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. 19

[39] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 8, 14

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 8, 14

[41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 8, 14

[42] Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, and Jing Ma. Mmcode: Evaluating multi-modal code large language models with visually rich programming problems. *arXiv preprint arXiv:2404.09486*, 2024. 8, 14

[43] Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. *arXiv preprint arXiv:2409.16718*, 2024. 14

[44] Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*, 2024. 8, 14

[45] Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data

and parameters for a family of multi-modal large language models. In *Forty-first International Conference on Machine Learning*, 2024. 2, 4

[46] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 3, 7

[47] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 4, 8, 14

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 4, 8, 14

[49] Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*, 2024. 3, 7

[50] Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, et al. Chipnemo: Domain-adapted llms for chip design. *arXiv preprint arXiv:2311.00176*, 2023. 1

[51] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 8, 14

[52] 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. 4

[53] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 8, 14

[54] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521, 2022. 1

[55] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. 2

[56] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1, 2, 4, 7, 8, 14, 17, 20

[57] Subir Majumder, Lin Dong, Fatemeh Doudi, Yuting Cai, Chao Tian, Dileep Kalathil, Kevin Ding, Anupam A Thatte, Na Li, and Le Xie. Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule*, 8(6):1544–1549, 2024. 1, 8, 14

[58] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 8, 14

[59] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 8, 14

[60] Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*, 2024. 14

[61] OpenAI. Gpt-4v (vision) system card, 2023. Citekey: gptvision. 4

[62] OpenAI. Hello gpt-4o. https://www.openai.com/, 2024. 2, 4, 8, 14

[63] OpenAI. GPT-4o Mini: Advancing cost-efficient intelligence, 2024. 4, 7, 20

[64] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8, 14

[65] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies, 2023. 7, 19

[66] Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. *arXiv preprint arXiv:2402.13220*, 2024. 3, 7

[67] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024. 8, 14

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8, 14

[69] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 4

[70] Hadi Saadat et al. *Power system analysis*. McGraw-hill, 1999. 1

[71] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 8, 14

[72] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context, 2023. 8, 19

[73] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 14

[74] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 8, 14

[75] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. 1

[76] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2, 4, 7, 8, 14

[77] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. 8, 14, 21

[78] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 8, 14

[79] Haixin Wang, Xinlong Yang, Jianlong Chang, Dian Jin, Jinan Sun, Shikun Zhang, Xiao Luo, and Qi Tian. Parameter-efficient tuning of large-scale multimodal foundation model. *Advances in Neural Information Processing Systems*, 36:15752–15774, 2023. 14

[80] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024. 8, 14

[81] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. 1, 2

[82] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023. 8, 14

[83] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. 7, 19

[84] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024. 8, 14

[85] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 8, 14

[86] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 7, 19

[87] Haoyuan Wu, Zhuolun He, Xinyun Zhang, Xufeng Yao, Su Zheng, Haisheng Zheng, and Bei Yu. Chateda: A large language model powered autonomous agent for eda. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024. 1

[88] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 4

[89] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 8, 14

[90] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023. 1

[91] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023. 1

[92] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1, 2, 4, 8, 14

[93] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2024. 4, 21

[94] Sukmin Yun, Haokun Lin, Rusiru Thushara, Mohammad Qazim Bhat, Yongxin Wang, Zutao Jiang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, Haonan Li, Preslav Nakov, Timothy Baldwin, Zhengzhong Liu, Eric P. Xing, Xiaodan Liang, and Zhiqiang Shen. Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal llms, 2024. 2

[95] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish

your sentence? *arXiv preprint arXiv:1905.07830*, 2019. 8, 14

[96] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 4

[97] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 1, 8, 14

[98] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. 1, 2

[99] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. 1, 8, 14

[100] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024. 14

[101] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models, 2023. 7, 19

[102] 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. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 8, 14

[103] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. 7, 19

[104] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 14

[105] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 14

[106] Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *arXiv preprint arXiv:2408.08640*, 2024. 14