

# MMCircuitEval: A Comprehensive Multimodal Circuit-Focused Benchmark for Evaluating LLMs

Chenchen Zhao<sup>\*1,7</sup>, Zhengyuan Shi<sup>\*1,7</sup>, Xiangyu Wen<sup>\*1,7</sup>, Chengjie Liu<sup>2,7</sup>, Yi Liu<sup>1,7</sup>,  
Yunhao Zhou<sup>1,7</sup>, Yuxiang Zhao<sup>3,7</sup>, Hefei Feng<sup>4,7</sup>, Yinan Zhu<sup>7</sup>, Gwok-Waa Wan<sup>7</sup>,  
Xin Cheng<sup>5,7</sup>, Weiyu Chen<sup>2,7</sup>, Yongqi Fu<sup>4,7</sup>, Chujie Chen<sup>6,7</sup>, Chenhao Xue<sup>3,7</sup>,  
Ying Wang<sup>6</sup>, Yibo Lin<sup>3</sup>, Jun Yang<sup>4,7</sup>, Ning Xu<sup>5,7</sup>, Xi Wang<sup>4,7</sup>, and Qiang Xu<sup>+1,7</sup>

<sup>1</sup>Department of Computer Science and Engineering, The Chinese University of Hong Kong

<sup>2</sup>School of Electronic Science and Engineering, Nanjing University

<sup>3</sup>School of Integrated Circuits, Peking University

<sup>4</sup>School of Intergrated Circuits, Southeast University

<sup>5</sup>School of Computer Science and Engineering, Southeast University

<sup>6</sup>Department of Computer Science and Technology, University of Chinese Academy of Sciences

<sup>7</sup>National Center of Technology Innovation for EDA

**Abstract**—The emergence of multimodal large language models (MLLMs) presents promising opportunities for automation and enhancement in Electronic Design Automation (EDA). However, comprehensively evaluating these models in circuit design remains challenging due to the narrow scope of existing benchmarks. To bridge this gap, we introduce MMCircuitEval, the first multimodal benchmark specifically designed to assess MLLM performance comprehensively across diverse EDA tasks. MMCircuitEval comprises 3614 meticulously curated question-answer (QA) pairs spanning digital and analog circuits across critical EDA stages—ranging from general knowledge and specifications to front-end and back-end design. Derived from textbooks, technical question banks, datasheets, and real-world documentation, each QA pair undergoes rigorous expert review for accuracy and relevance. Our benchmark uniquely categorizes questions by design stage, circuit type, tested abilities (knowledge, comprehension, reasoning, computation), and difficulty level, enabling detailed analysis of model capabilities and limitations. Extensive evaluations reveal significant performance gaps among existing LLMs, particularly in back-end design and complex computations, highlighting the critical need for targeted training datasets and modeling approaches. MMCircuitEval provides a foundational resource for advancing MLLMs in EDA, facilitating their integration into real-world circuit design workflows. Our benchmark is available at <https://github.com/cure-lab/MMCircuitEval>.

## I. INTRODUCTION

In recent years, large language models (LLMs) have showcased remarkable capabilities across various domains, effectively automating tasks and enhancing productivity in fields such as natural language processing, software engineering, and data analysis. Recognizing this potential, the semiconductor industry has started exploring the role of LLMs in circuit design. Recent initiatives, such as ChipNemo [1] and SemiKong [2], underscore the promise of LLMs in assisting engineers with circuit analysis and design optimization within Electronic Design Automation (EDA) workflows.

To evaluate the efficiency and effectiveness of LLMs in the circuit design domain, researchers have developed various circuit-focused benchmarks aimed at assessing these models

from diverse perspectives. Some existing benchmarks [3]–[6] concentrate on assessing the quality of Verilog code snippets generated by LLMs. While these benchmarks provide a range of design specifications paired with corresponding implementations or testbenches, the designs are generally far smaller and less complex than those encountered in practical scenarios. Other benchmarks [7], [8] attempt to evaluate LLMs based on their abilities to select EDA tools and generate design flow scripts. However, these narrowly focused tasks limit the scope of tool planning and leave unanswered questions about whether LLMs truly understand the intricate circuit features. Consequently, comprehensively evaluating LLMs for circuit design remains a critical challenge.

To address this gap, we introduce MMCircuitEval: a comprehensive multimodal circuit-focused benchmark specifically designed to evaluate LLM performance across various stages and types of circuit designs. MMCircuitEval comprises 3,614 question-answer (QA) pairs collected from diverse and reliable sources, including open-source *textbooks*, technical *question banks*, and *online resources*. To enhance realism and practicality, we also generate additional questions derived from datasheets, register-transfer level (RTL) codes, and netlists of *real-world products*, covering a wide range of circuit-related challenges. Each QA pair undergoes meticulous manual review by domain experts to ensure it meets high standards of accuracy, relevance, and technical depth.

MMCircuitEval is structured to offer a dual perspective, catering to both circuit designers and LLM developers. On the one hand, MMCircuitEval systematically categorizes QA pairs based on circuit types (digital and analog) and distinct stages of EDA workflows (general knowledge, design specification, front-end design, and back-end design). This organization allows for a detailed evaluation of how effectively LLMs can assist hardware engineers at different stages of circuit design. On the other hand, MMCircuitEval includes detailed metadata for each QA pair, specifying the abilities being tested (domain-specific knowledge, multimodal comprehension, logical reasoning, and numerical computation) and difficulty levels (easy,

<sup>\*</sup>Equal contribution

<sup>+</sup>Corresponding author: Qiang Xu (qxu@cse.cuhk.edu.hk)

medium, and hard). This detailed labeling not only provides granular insights into model performance but also serves as a guide for optimizing LLM design to better support complex circuit design from the perspective of LLM developers.

We conduct extensive experiments to evaluate a variety of common LLMs using our proposed benchmark. Experimental results reveal that most widely-used LLMs fail to achieve satisfactory performance in circuit-focused question answering. Notably, the models struggle the most with *back-end design* and *circuit-related computations*, highlighting the urgent need for relevant training corpora and processing techniques tailored to circuit-focused LLM training. We also discuss potential solutions to address these challenges, including illustrative experiments to validate their impacts.

To the best of our knowledge, MMCircuitEval is the first benchmark designed to assess LLM capabilities across various multimodal circuit-related questions and different EDA stages. We hope that MMCircuitEval will serve as a foundational resource, inspiring further innovation in leveraging LLMs to develop advanced solutions to EDA challenges.

## II. RELATED WORK

### A. LLM for EDA

The rapid advancements of AI and LLMs have expanded their applications to various professional domains, including EDA [9]. Specifically, LLMs have demonstrated significant potential for enhancing EDA workflows in areas such as hardware code generation and verification, EDA tool planning, and interactive question answering systems. For instance, specific LLM-based approaches [10]–[13] focus on automatically generating circuit designs, while others are employed to plan and manage EDA tools [14], [15] (e.g., automating task decomposition, script generation, and task execution within the EDA design flow). Interactive systems such as ChipGPT [16] and ChipNemo [1] further showcase the utility of LLMs for question answering (QA) and knowledge retrieval, enabling engineers to access relevant design information and troubleshoot issues more effectively. While these advancements highlight the versatility of LLMs in EDA, there remains a notable lack of general and comprehensive benchmarks to evaluate LLM performance across a wide range of EDA tasks.

### B. Multimodal Benchmarks for LLM Evaluation

Recent research has developed a variety of multimodal benchmarks [17]–[19] to quantitatively measure the performance of LLMs across diverse use cases. These benchmarks conduct extensive evaluations of domain knowledge capacity, comprehension, and logical inference of LLMs in a wide range of professional vision-language scenarios such as mathematics, medicine, art, and engineering. Model performance is quantitatively represented by the average correctness of answers to multimodal questions of different types (e.g., choice, open-ended, etc), tested capabilities, and downstream scenarios.

In the domain of circuit design and EDA, existing benchmarks for evaluating LLMs remain limited in narrow scopes. For instance, studies [3], [5], [6] focus primarily on assessing

LLM capabilities in generating hardware description code and testbenches. Meanwhile, Wu et al. [7] introduces an open-source dataset centered around the OpenROAD [15] EDA toolchain, comprising question-answer pairs, code snippets, and their corresponding OpenROAD scripts. Similarly, another benchmark [8] offers high-quality question-document-answer triplets that address various queries within the EDA design flow. However, these benchmarks primarily focus on applying EDA tools based on documentation rather than understanding design methodologies or analyzing specific designs. More recently, Xu et al. propose ChatICD-Bench [20], a comprehensive benchmark spanning multiple subfields of chip design. While promising, ChatICD-Bench still lacks comprehensive performance analyses and offers limited question diversity.

## III. THE MMCIRCUITEVAL BENCHMARK

### A. Overview

The MultiModal Circuit-focused Evaluation (MMCircuitEval) benchmark aims to extensively assess the professional capabilities of multimodal LLMs across all stages of the EDA circuit design workflow through closely-related QA samples.

MMCircuitEval consists of 3614 manually curated test questions from diverse sources, covering general EDA knowledge and major circuit design stages for both digital and analog circuits. The questions specifically target different circuit-related abilities of multimodal LLMs.

Building on MMCircuitEval, we conduct extensive experiments to evaluate and horizontally compare the performance of existing LLMs in chip design, providing detailed analyses of their accuracies and errors.

### B. Data Collection and Curation

1) *Data collection and organization*: To ensure the diversity and comprehensive coverage of MMCircuitEval for fair comparisons, we collect and categorize the data according to the following three key aspects:

**Design stages.** Typical circuit design workflows are divided into front-end and back-end designs, based on predefined specifications. Our data collection spans all the three stages.

- *Design specification.* To address the lack of systematic test questions related to design specifications, we gather 42 datasheets from online sources and generate high-quality questions focusing on specific parameters of the corresponding products with human experts review. The questions involve comprehension of specification documents, multi-step inference based on circuit design knowledge, and parameter-related computation.
- *Front-end design.* We collect questions covering basic front-end circuit knowledge and code logic, netlist comprehension and computation, and circuit behavior analysis. Most questions are extracted from textbooks and online materials. Additionally, we manually craft 98 Verilog code snippets to test the models' RTL code comprehension capabilities.
- *Back-end design.* We collect questions related to fundamental layout design knowledge and usages of layout

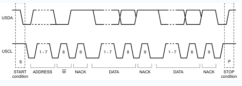
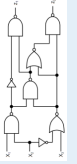
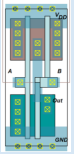
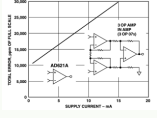
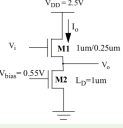
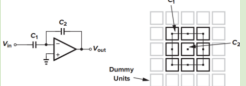
	General Knowledge	Design Specification	Front-End Design	Back-End Design
Digital Circuit	<p>What is the primary function of a place-and-route tool?</p> <p>(A) Logic synthesis (B) HDL simulation (C) Arrangement of standard cells and their interconnections (D) Power analysis</p> <p>Count: 302 (8.4%)</p>	<p>How does the controller handle multiple communications without generating a STOP condition?</p>  <p>Count: 259 (7.2%)</p>	 <p>Compute the arrival time, required time, and slack for every net with output requirement of 8ns. Identify the critical region</p> <p>Count: 697 (19.3%)</p>	 <p>Write the logic function of this standard cell layout</p> <p>Count: 551 (15.3%)</p>
Analog Circuit	<p>When both nMOS and pMOS transistors of CMOS logic gates are ON, the output is:</p> <p>(A) 1 or V<sub>dd</sub> or HIGH state (B) 0 or ground or LOW state (C) Crowbarred or Contention(X) (D) None of the mentioned</p> <p>Count: 302 (8.4%)</p>	<p>If the AD621 is operating at a gain of 10 with a source resistance of 1 k<math>\Omega</math>, what is the expected input voltage noise at 1 kHz?</p>  <p>Count: 231 (6.4%)</p>	 <p>Calculate the width of M2 to provide this level shift</p> <p>Count: 819 (22.6%)</p>	<p>Write the reason of implementing the layout of C1 and C2 like this?</p>  <p>Count: 453 (12.5%)</p>

Fig. 1. The overview of MMCircuitEval and sampled questions. Better viewed with zoom-in.

design tools. The questions are extracted from textbooks and online materials.

- **General knowledge.** 604 test questions related to general EDA knowledge (e.g., fundamentals of digital and analog circuits) are also included in the dataset.

**Circuit types.** We ensure a balanced data proportion between digital and analog circuit types. For *specification* and *back-end design*, we collect both digital and analog datasheets and layouts for test question extraction. For *front-end design*, given that RTL codes are inherently digital, we specifically formulate 185 questions related to analog concepts (e.g, voltage, current, and power calculations) to assess the models' analog computation capabilities and maintain data balance.

**Question types and modalities.** The questions in MMCircuitEval can be categorized into single-answer and multi-answer choice questions, fill-in-the-blank questions, and open-ended questions with respect to question type, and can be categorized into text-only questions and multimodal questions with respect to data modality.

In addition, we paraphrase some of the question texts for augmentation. We also include explanations for each question-solution pair where applicable. This feature enables more objective and logical answer evaluation outlined in III-C.

2) **Data curation:** To ensure the correctness and the quality of the collected questions, answers and explanations, our team, with specialized EDA engineers and Ph.D. students, manually review each question-solution pair in the dataset according to the following systematic set of rules:

- **Relevance:** whether the question is strictly circuit-focused;
- **Alignment:** whether the question fits within the designated category;
- **Quality:** whether the question is of low quality;
- **Content correctness:** whether the solution and explanation (if applicable) are correct;
- **Attribute correctness:** whether the question type, circuit type, and difficulty level of the question are correct and objective;

- **Format:** whether the question adheres to the pre-defined model-friendly format.

We filter out irrelevant and low-quality questions that:

- Are over-simple;
- Lack essential information (e.g., figures of document pages in *design specification*);
- Need to be answered in figure (e.g., layout sketching in *back-end design*);
- Are beyond the capacity of existing large vision-language models or cannot be fairly evaluated (e.g., those with extremely high complexity or subjectivity).

After filtering, we correct false information in the remaining questions to ensure accuracy, relevance and technical depth.

In addition, we utilize GPT-4o to assign a *tested ability* label to each question, which reflects the specific ability that the question assesses in the models. The possible labels include:

- **Knowledge-related**, indicating whether the models possess the necessary circuit-related knowledge stored in their memory.
- **Comprehension-related**, indicating whether the models can extract relevant information from the provided material in the question.
- **Reasoning-related**, indicating whether the models can perform logical inference based on the given context.
- **Computation-related**, indicating whether the models can apply correct formulas to derive correct numerical results.

In our benchmark preparation, all contextual outputs produced by LLMs (e.g. the label assignment process) are followed by manual verification and correction to ensure accuracy.

The detailed data composition is presented in Table I. Note that a very small proportion of the data cannot be categorized as either digital or analog. In addition, we do not categorize simple multimodal questions into *general knowledge*. Instead, we consider them as simple and straightforward cases within their respective circuit design stages (e.g., gate circuits as simple gate-level netlists in *front-end design*, analog circuits as basic cases in analog *front-end design*, etc). This results in no multimodal question of category *general knowledge*.

TABLE I  
DATA STATISTICS OF MMCIRCUITEVAL.

Data compositions	Statistics
Categories (stages)	4
Total	3614
General knowledge	604 (16.7%)
Specification	490 (13.6%)
Front-end	1516 (41.9%)
Back-end	1004 (27.8%)
Total (digital : analog)	50.0% : 50.0%
General knowledge (digital : analog)	50.0% : 50.0%
Specification (digital : analog)	52.8% : 47.2%
Front-end (digital : analog)	46.0% : 54.0%
Back-end (digital : analog)	54.9% : 45.1%
Total (text-only : multimodal)	58.4% : 41.6%
General knowledge (text-only : multimodal)	100.0% : 0.0%
Specification (text-only : multimodal)	50.0% : 50.0%
Front-end (text-only : multimodal)	50.0% : 50.0%
Back-end (text-only : multimodal)	50.0% : 50.0%
Single-answer choice	738 (20.4%)
Multi-answer choice	86 (2.4%)
Fill-in-the-blank	396 (11.0%)
Open-ended	2394 (66.2%)
Knowledge-related	1446 (40.0%)
Comprehension-related	410 (11.3%)
Reasoning-related	832 (23.0%)
Computation-related	926 (25.6%)
With solution explanation	2271 (62.8%)
Easy : Medium : Hard	15.2% : 58.7% : 26.1%

### C. Benchmark Construction

We select a broad series of text-only and multimodal LLMs, and test their performance on the curated dataset. We comprehensively assess their average accuracies across different circuit types, design stages, tested abilities, data modalities, question types, and difficulty levels.

The evaluation of each question is based on the similarity between the provided solution and the model output. Since they are both text-only, we employ four prevalent text-centered metrics to measure their similarity:

- Bilingual evaluation understudy (BLEU) score, which estimates the ratio of the output phrases existing in the solution. In MMCircuitEval, BLEU score is used to evaluate whether the model produces misleading information or hallucination. We specifically adopt the 4-gram BLEU score in our settings.
- Recall-oriented understudy for gisting evaluation (ROUGE) score, which estimates the ratio of the solution phrases existing in the model output. In MMCircuitEval, ROUGE score is used to evaluate whether the model output contains all the critical information needed to answer the question. We utilize the average of the 1-gram, 2-gram and longest common subsequence ROUGE scores in our settings.
- Embedding cosine similarity, which estimates the semantic consistency between the two answers. In MMCir-

cuitEval, embedding cosine similarity is used to evaluate whether the model output follows a reasoning process logically similar to the solution. We employ the Text-Embedding-3-Large model provided by OpenAI in our settings.

- GPT preference. In MMCircuitEval, we leverage GPT-4-turbo [21], proficient in text processing, for overall correctness rating from a well-trained expert’s perspective.

We assign a weight of 2 for GPT preference and a weight of 1 for others in the MMCircuitEval evaluation. Additionally, we require the tested models to provide an explanation alongside each answer to facilitate better correctness judgment. As validated in section IV-B, an integration of the metrics is qualified for answer similarity evaluation in circuit-specific scenarios, even though GPT may not always perform well in providing correct answers.

### D. Highlights of MMCircuitEval

A statistical comparison between MMCircuitEval and existing circuit-focused benchmarks is presented in Table II. We highlight the key advantage of MMCircuitEval over existing circuit-focused benchmarks:

- **Large data volume and broad data spectrum.** According to our knowledge, MMCircuitEval is the first benchmark that encompasses different stages of typical circuit design workflows. It also covers diverse circuit types, question types, and data modalities.
- **Comprehensive evaluation.** Fine-grained data categorization and tested ability assignment enable MMCircuitEval’s multi-dimensional model performance evaluation, and also allow for a more nuanced understanding of LLM capabilities.
- **High scalability.** Usages of GPT with manual data curation achieves high efficiency and scalability of data collection without compromising data quality. Furthermore, GPT-generated questions exhibit minimal overlap with the training corpora of existing foundation models, ensuring fair horizontal comparisons.

## IV. EXPERIMENTS

### A. Baseline Models

We evaluate various model families that are widely-applied in the field of multimodal QA. For many model families with multiple variants and parameter scales, we select at least two variants from each to test their circuit-focused upscaling capabilities. These models can be categorized based on their image processing techniques:

- Text-only, which lack the ability to process visual data. For these models, we leverage a BLIP [22] captioning model to generate textual descriptions as substitutes for visual information.
- Image-to-string, which process visual information through hard-coded image-to-string conversions.
- Image encoding, which incorporate embedded visual encoders to directly process images.

TABLE II  
STATISTICAL COMPARISON BETWEEN MMCircuitEval AND OTHER CIRCUIT-FOCUSED BENCHMARKS. SC, MC, B, O OF “QUESTION TYPES” REFER TO SINGLE-ANSWER CHOICE, MULTI-ANSWER CHOICE, FILL-IN-THE-BLANK, AND OPEN-ENDED.

Benchmarks	Size	Categories	Modalities	Question types	Sources
<b>MMCircuitEval</b>	3614	4	Text-only, text&image	SC, MC, B, O	Textbook, Internet, Handcraft, Synthesis
EDA Corpus [7]	1533	2	Text-only	O	Handcraft
ORD-QA [8]	90	4	Text-only	O	Synthesis
ChatICD-Bench [20]	622	7	Text-only	O	Textbooks, Internet, Handcraft

The list of models evaluated is presented in Table III. According to our knowledge, ChipExpert [20] is currently the only publicly available circuit-focused LLM in this domain.

### B. Correctness Validation of the Proposed Evaluator

We first conduct experiments to validate the effectiveness of our proposed comprehensive answer evaluation metric stated in Section III-C. We randomly select 100 questions from each design stage across multiple models and obtain multiple groups of answers. Our team then manually check whether the overall scores calculated with the proposed metric are accurate enough to reflect their correctness. Results show that the testers are generally positive about the quantitative results, indicating the effectiveness of the proposed evaluator.

### C. Evaluation Results and Discussions

The overall evaluation results of the models are detailed in Table III, with the average results of each model category shown in Table IV. The baseline models are tested and horizontally compared across the following aspects:

- **Global baseline performance (Overall)**, comprehensively reflecting circuit-related model performance;
- **Circuit design stages (G / S / F / B)**, targeting the models’ knowledge base and logic capabilities in specific circuit design scenarios. In the table, G, S, F, and B respectively refer to general knowledge, design specification, front-end design, and back-end design stages;
- **Tested abilities (K / Cph / R / Cpt)**, indicating whether the models’ original capabilities on general-purpose benchmarks maintain effectiveness in the EDA field. In the table, K, Cph, R, and Cpt respectively represent that the questions test the abilities of knowledge, comprehension, reasoning, and computation;
- **Data modalities (T / M)**, indicating whether the models can effectively read and process circuit-related visual materials. In the table, T and M respectively refer to text-only and multimodal questions.

**Global baseline performance (Overall).** From this table, we observe that most of LLMs struggle to reach satisfactory levels of accuracy on the overall problem set. For instance, several LLMs, such as InstructBLIP [26] and BLIP2 [27] only perform less than 20% accuracy in our MMCircuitEval. The primary reason for this shortfall is the lack of sufficient circuit-related training materials in existing vision-language corpora. In addition, due to the scarcity of circuit-specific data in LLM training, the circuit-focused horizontal performance between them may not align well with general-purpose benchmarks.

Nonetheless, most models exhibit certain scalability as previously validated by other benchmarks.

Among the evaluated models, GPT-4v [21] achieves the highest overall performance with 69.4% problem solved out, while ChipExpert [20] ranks highest among open-source models by answering 67.1% questions correctly. Models such as LLaMa3.2-Vision-Instruct-90B [28] (58.5%) and Gemini1.5-Pro [45] (62.2%) also show relatively better performance than most other LLMs. Notably, these models all have significantly large parameter scales except ChipExpert, indicating their high scalability, and that we can improve their performance in circuit design simply by upscaling. As a circuit-focused model, the text-only ChipExpert, based on the LLaMa3-8B [28] architecture, outperforms most other models by a large margin, particularly in knowledge mastery and back-end processing, further highlighting the significance of high-quality training data in circuit-related LLM applications.

**Performance across different circuit design stages (G / S / F / B).** We further investigate the results according to the different design stages. The results are evident that back-end design generally exhibits the lowest accuracy across models. For example, GPT-4v [21] achieves a relatively high accuracy of 69.4% on overall questions, but its performance on back-end design questions is considerably lower, with only 48.2% accuracy. This trend is mostly consistent across other models, with ChipExpert [20] answering 61.1% of the back-end design questions correctly, while models like LLaMa3.2-Vision-Instruct-90B [28] and Gemini1.5-Pro [45] only manage 36.6% and 42.6%, respectively. On average, different categories of LLMs have a 12.0%-21.8% performance decline on back-end problems compared with those of other circuit design stages.

The performance disparity can be attributed to several key factors. First, questions related to general knowledge (e.g., basic digital and analog circuits) and front-end design (e.g., comprehension of code snippets and simple circuit diagrams) are more prevalent in the existing training corpora, making it easier for LLMs to handle these tasks. In contrast, back-end design questions often involve highly specific layouts and details that require more specialized data, which is currently scarce in most circuit-related corpora used to train LLMs. Second, back-end design is more context-dependent and multimodal, with many questions requiring a deep understanding of placement, routing, and other intricate layout-specific challenges. These complex visual and spatial relationships pose a significant difficulty for LLMs that have been trained primarily on textual data with limited circuit-specific examples.

**Performance across different tested abilities (K / Cph /**

TABLE III  
COMPREHENSIVE EVALUATION RESULTS OF THE SELECTED TEXT-ONLY AND MULTIMODAL LLMs. AMONG THE MODEL CATEGORIES, THE HIGHEST CORRECTNESS IS **BOLD**, AND THE SECOND-HIGHEST IS UNDERLINED. THE HIGHER VALUE ( $\uparrow$ ) INDICATES BETTER PERFORMANCE.

Models	Overall	G	S	F	B	K	Cph	R	Cpt	T	M
<b>MLLMs (Image encoding)</b>											
QWen-VL-Chat [23]	32.2	41.6	41.0	32.7	21.4	37.3	46.4	25.6	16.9	34.7	28.6
InternLM-XComposer-VL-7B [24]	19.7	41.2	11.9	22.9	5.8	32.5	17.0	15.2	12.5	25.5	11.6
InternVL2-8B [25]	38.8	<u>48.3</u>	34.9	45.5	24.8	51.8	47.1	39.5	29.0	50.3	22.5
InternVL2-40B [25]	41.7	<u>48.3</u>	35.4	51.8	<u>25.7</u>	<u>54.4</u>	51.8	<u>45.7</u>	<u>30.3</u>	<u>54.3</u>	24.2
InstructBLIP-Flan-T5-XL [26]	10.8	11.9	18.9	8.6	9.6	8.2	16.5	6.8	7.1	8.1	14.6
InstructBLIP-Flan-T5-XXL [26]	10.7	17.7	11.1	6.7	12.2	11.6	13.2	9.5	3.3	9.8	11.9
BLIP2-Flan-T5-XL [27]	14.3	14.4	20.3	13.0	13.2	12.4	24.2	12.4	7.7	13.8	15.0
BLIP2-Flan-T5-XXL [27]	14.0	14.1	17.3	12.9	13.9	14.8	18.8	15.0	9.8	15.6	11.7
LlaMa3.2-Vision-Instruct-11B [28]	<u>43.2</u>	42.1	<u>58.0</u>	<u>53.5</u>	21.1	38.9	<u>54.5</u>	42.7	29.9	39.7	<u>48.1</u>
LlaMa3.2-Vision-Instruct-90B [28]	<b>58.5</b>	<b>64.2</b>	<b>63.6</b>	<b>69.1</b>	<b>36.6</b>	<b>62.2</b>	<b>71.4</b>	<b>57.0</b>	<b>42.7</b>	<b>61.2</b>	<b>54.7</b>
MiniCPM-V [29]	21.1	23.0	19.2	25.9	13.6	24.5	28.7	22.5	11.3	24.5	16.3
MiniCPM-V2 [29]	11.5	17.7	13.3	11.8	6.6	14.2	15.5	7.8	0.7	10.5	13.1
MiniCPM-LlaMa3-V2.5 [29]	43.0	46.7	47.4	52.5	24.3	44.5	53.5	44.5	25.6	43.7	41.9
Yi-VL-6B [30]	17.3	30.0	12.9	20.2	7.4	23.8	19.3	15.4	6.4	19.5	14.1
Yi-VL-34B [30]	32.5	34.4	40.6	38.9	17.6	34.1	45.8	28.9	16.8	32.9	31.8
Kosmos2 [31]	13.6	10.9	15.3	13.7	14.1	13.7	15.1	13.3	9.4	12.9	14.5
<b>MLLMs (Image-to-string)</b>											
GPT-4 [21]	63.7	64.7	63.8	74.0	47.6	66.7	69.5	61.3	45.2	62.0	66.1
GPT-4-Turbo [21]	67.4	67.9	68.9	<u>77.9</u>	<b>50.6</b>	<b>69.4</b>	74.0	<u>66.7</u>	54.0	<b>67.2</b>	67.8
GPT-4v [21]	<b>69.4</b>	<b>69.9</b>	<u>80.3</u>	<b>79.8</b>	48.2	<u>67.9</u>	<u>75.5</u>	<b>67.3</b>	<b>59.2</b>	<u>66.5</u>	<b>73.6</b>
GPT-4o [32]	<u>68.0</u>	<u>69.4</u>	<b>80.7</b>	76.2	<u>48.6</u>	66.1	<b>75.9</b>	66.3	<u>56.9</u>	65.5	<u>71.4</u>
Reka-Flash [33]	55.7	61.8	54.7	68.3	33.5	63.0	66.5	61.3	39.3	63.4	44.8
Reka-Edge [33]	36.7	43.4	35.4	45.8	19.5	42.5	46.2	41.9	19.4	42.0	29.3
<b>Text-only LLMs</b>											
ChipExpert* [20]	<b>67.1</b>	<b>75.5</b>	<b>61.9</b>	69.5	<b>61.1</b>	<b>77.4</b>	<b>77.0</b>	<b>69.1</b>	<b>63.7</b>	<b>77.7</b>	<b>52.4</b>
GPT-3.5-Turbo [34]	54.9	51.7	58.6	66.7	37.2	59.6	66.4	58.4	42.1	60.8	46.7
DeepSeek-MoE-16B-Chat [35]	34.0	28.0	34.6	36.7	33.3	33.2	38.8	34.8	34.5	35.6	31.8
DeepSeek-LLM-7B-Chat [36]	35.1	27.4	39.4	37.1	34.7	33.5	42.8	36.4	34.3	36.2	33.5
DeepSeek-Math-7B-RL [37]	39.9	42.3	42.8	41.6	34.3	42.6	44.9	39.8	42.7	44.1	33.9
DeepSeek-Math-7B-Instruct [37]	40.5	53.3	40.1	41.0	32.3	46.2	42.8	38.3	46.0	47.4	30.8
DeepSeek-V2-Lite-Chat [38]	39.7	35.0	40.7	43.1	36.8	40.4	44.3	41.2	40.1	42.6	35.5
QWen-2-Instruct-0.5B [39]	13.8	13.6	23.1	14.4	8.3	13.9	25.8	11.1	6.9	15.3	11.6
QWen-2-Instruct-7B [39]	48.2	54.3	51.2	59.5	26.0	55.2	58.2	51.3	30.4	53.8	40.3
QWen-2-Instruct-72B [39]	50.2	66.4	37.6	59.8	32.3	64.0	45.2	53.1	40.7	58.3	39.0
QWen-2.5-Instruct-7B [39]	53.0	57.0	57.4	63.3	32.9	58.9	65.9	54.0	38.9	59.1	44.4
QWen-2.5-Instruct-72B [39]	60.9	64.2	<b>61.9</b>	<u>73.3</u>	39.7	66.0	<u>72.3</u>	64.3	<u>49.7</u>	67.5	<u>51.6</u>
InternLM-Chat-20B [40]	33.3	35.9	38.9	39.0	20.5	37.7	42.6	34.6	20.2	37.2	27.8
InternLM2-Chat-7B [40]	45.0	53.0	49.7	52.7	26.3	50.3	55.5	48.5	32.3	51.1	36.4
InternLM2.5-Chat-7B [40]	46.7	57.5	51.0	55.3	25.0	54.0	60.5	45.1	33.6	53.4	37.3
LlaMa2-Chat-HF-7B [41]	29.7	37.7	31.3	35.8	14.8	35.5	40.3	29.9	12.5	33.6	24.2
LlaMa2-Chat-HF-13B [41]	33.6	42.9	38.8	38.0	18.9	39.5	48.1	33.4	11.2	37.3	28.5
LlaMa3-Instruct-8B [28]	46.1	51.7	47.2	55.6	27.8	53.1	59.2	47.5	31.1	53.0	36.3
LlaMa3.1-Instruct-8B [28]	47.6	53.0	50.6	57.8	27.5	54.8	60.7	50.3	35.9	56.2	35.6
MiniCPM-SFT-1B [42]	20.5	31.0	22.5	20.7	12.9	26.6	27.5	15.1	6.1	21.9	18.6
MiniCPM-SFT-2B [42]	24.8	29.1	23.6	26.9	19.6	29.2	31.2	22.8	10.9	26.1	23.0
MiniCPM3-4B [42]	48.6	55.1	49.3	58.0	30.3	53.7	61.6	50.9	31.6	53.7	41.4
Yi-Chat-6B [30]	31.7	41.8	35.9	35.6	17.6	37.4	45.7	31.8	11.4	35.7	26.1
ChatGLM3-6B [43]	32.2	39.6	37.4	37.8	16.7	36.9	46.6	32.7	12.3	35.9	27.0
Gemini1.0-Pro <sup>+</sup> [44]	18.8	41.1	1.9	9.0	28.6	34.7	8.5	15.4	16.4	28.6	5.1
Gemini1.5-Pro <sup>+</sup> [45]	<u>62.2</u>	<u>72.2</u>	50.2	<b>75.2</b>	<u>42.6</u>	<u>75.1</u>	66.0	<u>65.3</u>	45.2	70.6	50.4
Claude3.5-Sonnet <sup>+</sup> [46]	53.5	60.6	35.9	65.4	39.9	74.4	62.5	54.4	40.0	<u>70.8</u>	29.3

\* A circuit-focused LLM based on the LlaMa3-8B architecture, trained with circuit-related curated data.

+ Multimodal models that only support text-only massive testing.

TABLE IV  
AVERAGE EVALUATION RESULTS OF THE SELECTED TEXT-ONLY AND MULTIMODAL LLMs OF EACH MODEL CATEGORY.

Model Categories	Overall	G	S	F	B	K	Cph	R	Cpt	T	M
MLLMs (Image encoding)	26.4	31.7	28.9	30.0	16.6	29.9	33.7	25.1	16.2	28.6	23.4
MLLMs (Image-to-string)	<b>60.0</b>	<b>62.9</b>	<b>65.2</b>	<b>69.7</b>	<b>41.1</b>	<b>62.4</b>	<b>67.8</b>	<b>60.6</b>	<b>45.5</b>	<b>60.9</b>	<b>58.7</b>
Text-only LLMs	41.0	47.1	40.7	46.8	28.7	47.3	49.5	41.6	30.2	46.6	33.1

**R / Cpt).** The evaluation results highlight distinct performance patterns of LLMs across different abilities. Specifically, most models perform strongly on general knowledge retrieval (K) and basic comprehension (Cph) tasks. For instance, GPT-4o [32] achieves an impressive accuracy of 75.9% on comprehension (Cph) tasks, while GPT-4-Turbo [21] successfully answers 69.4% of knowledge retrieval (K) questions. However, there is a noticeable decline in performance when models are tasked with more complex reasoning and computation problems. For example, GPT-4o [32] solves only 56.9% of computation-related (Cpt) questions, and several other LLMs, including InstructBLIP [26], MiniCPM [42], and InternLM [40], answer less than 50% of reasoning and computation questions correctly. On average, different categories of LLMs have a 1.8%-5.7% performance decline on reasoning compared with knowledge retrieval and comprehension, and an 8.9%-15.1% further performance decline on computation.

The underlying reasons for this performance disparity are twofold. First, the LLMs are primarily trained on large vision-language corpora designed to enhance general comprehension and knowledge retrieval. Second, while LLMs excel in general reasoning tasks and numerical computations [18], reasoning and computing circuit-based problems involve specific electronic rules and design methodology. This domain-specific knowledge gap leads to lower accuracy in reasoning and computation tasks within the circuit design context.

**Performance across different data modalities (T / M).** Evaluation results also highlight notable performance gaps when LLMs are tasked with processing circuit-related images. Although multimodal LLMs incorporate image encoders to encode or summarize non-textual information, most models experience performance degradation on multimodal (M) problems compared to purely text-only (T) problems. For example, LLaMa3.2-Vision-Instruct-90B [28] answers 61.2% text-only questions but answers 54.7% with multimodal statements. On average, the tested LLMs have a 2.2%-13.5% performance decline on multimodal problems compared with text-only ones. However, the GPT model family achieve better accuracy on multimodal questions than text-only ones. The image-to-string conversion does not result in severe information loss, which can better cooperate with the well-trained GPT backbones. More importantly, images in string format can be regarded as text and processed in a similar way to the question texts, thus avoiding the accumulation of errors by the visual encoders.

Interestingly, we also find that most models with visual encoders generally perform even worse than text-only models in multimodal QA, given that well-trained visual encoders are supposed to extract more accurate and comprehensive

information than short image captions. For instance, apart from the LLaMa3-backed models (i.e., LLaMa3.2-Vision-Instruct-11B [28], LLaMa3.2-Vision-Instruct-90B [28], and MiniCPM-LLaMa3-V2.5 [29]), the models integrated with visual encoders have the highest multimodal QA accuracy 31.8% (by Yi-VL-34B [30]), lower than 55.6% of the text-only models tested. One possible major reason is that the embedded image encoders are not specifically trained with circuit-related data, and may produce false visual embeddings that mislead the backbone LLMs and negatively impact their final outputs.

#### D. Possible Model Improvements for Circuit-Related Tasks

**How should we process circuit-related images?** As stated above, visual encoders that are not specifically trained with circuit-related data may lead to even worse performance than leveraging image captions. Meanwhile, there is currently no circuit-focused visual encoder that matches the performance of general-purpose ones while adequately covering all circuit design stages. However, embeddings generated by well-trained visual encoders have much more information than short captions. Therefore, with sufficient time and computation resources, visual encoders are apparently better than captioning. On the other hand, encoding images into strings is also a satisfactory approach, as it does not require extra visual processors and does not suffer from severe information loss. However, there is a high proportion of redundant and irrelevant tokens in the encoded strings, which may severely increase the burden on the subsequent LLM. However, this can also be solved by a powerful LLM backbone, which is proved by the high performance of the GPT-4 model family.

**What are the effectiveness and costs of model re-training?** Most tested models process general questions with high accuracies, as validated by other general-purpose benchmarks. With the same parameter scales and the problems restricted within the circuit-related field, it is evident that the models have the potential capabilities to reach equally high or even better performance. This is empirically validated by the performance of ChipExpert [20], which demonstrates that even a small-sized LLM can achieve relatively high circuit-related performance with the help of targeted training.

However, in the circuit-related field, the major challenge of large model training is not the choice of backbone models, but the collection of relevant high-quality data. Most large-scale general-purpose training corpora [47], [48] are established based on open-sourced online materials with clear in-context answers. However, high-quality, open-sourced, circuit-related materials are extremely scarce, reflected by the fact that the scales of existing circuit-related benchmarks are generally only

10% of the general-purpose ones. This poses a great challenge for LLM training for related downstream circuit-related tasks. However, with sufficient high-quality data, it is highly possible for existing LLMs to achieve much higher performance than that reported in the paper.

The time and computation resources required for model training and fine-tuning are the same as training on general-purpose data. From the multimodal perspective, image-to-string models require re-training or fine-tuning the whole backbone, while models with visual encoders can adopt specific techniques to achieve fine-tuning only on the visual encoders. Ultimately, the choice of image processing strategies determines the required resources and difficulty in model training and fine-tuning.

#### How to improve the models' test-time performance?

There are multiple test-time techniques to improve the performance of LLMs under limited resources. In this paper, we adopt the Chain-of-Thought (CoT) reasoning [49], a widely-adopted test-time inference technique as the core improvement approach. CoT encourages the model to break down complex reasoning tasks into smaller, more manageable steps, effectively improving the model's ability to handle intricate problems. With intermediate reasoning steps, CoT improves the model's decision-making process and overall accuracy.

We select three representative LLMs, BLIP2-Flan-T5-XL [27] (image encoding), GPT-4o [32] (image-to-string), and DeepSeek-LLM-7B-Chat [36] (text-only) to comprehensively evaluate the impact of CoT. We randomly select 100 questions from each circuit design stage, and add detailed CoT instructions to the query prompts. The instructions include the following sequential steps:

- 1) Understand and clarify the core of the question;
- 2) Locate relevant information presented in the material;
- 3) Identify and extract key data crucial to correct solutions;
- 4) Apply related knowledge and logic to solve the question;
- 5) Self-check answer correctness and consistency;
- 6) Summarize the generated content and output the final answer with explanations.

With CoT, the three models have achieved different degrees of performance improvements, with detailed numerical results presented in Table V. BLIP2-Flan-T5-XL [27], GPT-4o [32], and DeepSeek-LLM-7B-Chat [36] have 3.0%, 0.6%, and 1.3% overall performance increases, respectively. In terms of circuit design stages, the models are most significantly improved on front-end QA, with BLIP2-Flan-T5-XL [27] achieving relatively the highest increase in correctness by 9.2%. In terms of tested abilities, the models' reasoning abilities are most significantly improved. GPT-4o [32] achieves relatively the highest increase in reasoning correctness by 4.8%.

Questions of different circuit design stages and tested abilities result in different impacts of CoT on the tested models. Questions related to general knowledge retrieval require sufficient and high-quality knowledge bases recorded in the models' parameters. Therefore, CoT instructions on inference may not be remarkably effective. Questions of specifications and back-end design suffer from the same issue, resulted

TABLE V  
PERFORMANCE COMPARISON OF THE SELECTED MODELS BEFORE AND AFTER CoT INTEGRATION. THE HIGHER CORRECTNESS IS **BOLD**.

Models		Overall	G	S	F	B
BLIP2 [27] w/ CoT		<b>17.3</b>	<b>14.9</b>	20.2	<b>22.2</b>	12.0
BLIP2 [27] w/o CoT		14.3	14.4	<b>20.3</b>	13.0	<b>13.2</b>
GPT-4o [32] w/ CoT		<b>68.6</b>	67.7	80.2	<b>79.9</b>	<b>49.5</b>
GPT-4o [32] w/o CoT		68.0	<b>69.4</b>	<b>80.7</b>	76.2	48.6
DeepSeek [36] w/ CoT		<b>36.4</b>	<b>29.5</b>	38.9	<b>40.3</b>	<b>37.0</b>
DeepSeek [36] w/o CoT		35.1	27.4	<b>39.4</b>	37.1	34.7
Models	T	M	K	Cph	R	Cpt
BLIP2 [27] w/ CoT	<b>20.9</b>	<b>15.5</b>	<b>15.7</b>	<b>25.7</b>	<b>16.9</b>	<b>9.1</b>
BLIP2 [27] w/o CoT	13.8	15.0	12.4	24.2	12.4	7.7
GPT-4o [32] w/ CoT	65.0	<b>73.4</b>	64.4	75.9	<b>71.1</b>	<b>59.8</b>
GPT-4o [32] w/o CoT	<b>65.5</b>	71.4	<b>66.1</b>	75.9	66.3	56.9
DeepSeek [36] w/ CoT	<b>37.3</b>	<b>35.6</b>	32.9	42.4	<b>37.1</b>	<b>34.8</b>
DeepSeek [36] w/o CoT	36.2	33.5	<b>33.5</b>	<b>42.8</b>	36.4	34.3

by the scarce relevant data in LLM training corpora. In contrast, front-end design materials are relatively the easiest to obtain from open sources, while many of the corresponding questions require certain inference abilities. Therefore, CoT has relatively the highest impact on front-end QA. In terms of tested abilities, CoT also has similar impact patterns. The requirements on logic inference abilities grow higher from knowledge retrieval to numerical computation. However, without knowledge of the correct formulas and algorithms to apply, it is still a great challenge to improve the models' computation abilities, even with the help of CoT. Therefore, CoT is more suitable for solving reasoning-related questions. On the other hand, low-performance models may not benefit much from CoT, as their accuracy is bottlenecked by the insufficient data and training processes.

In summary, from the conducted experiments, CoT is an effective approach of test-time model performance improvement on circuit-related QA, and is specifically effective for front-end questions with relatively high occurrence in the training data, but requiring high logic inference capabilities.

## V. CONCLUSION

In this paper, we present MMCircuitEval, a comprehensive circuit-focused benchmark for evaluating multimodal LLMs. With a wide spectrum of test questions covering scenarios such as general circuit knowledge, design specification, front-end design, and back-end design, MMCircuitEval is capable of extensively evaluating general MLLMs on their circuit-related capabilities. In addition to horizontal comparisons, MMCircuitEval highlights that existing efforts on large foundation models are generally underdeveloped in the circuit and EDA fields. Finally, we explore potential solutions and propose directions for future advancements. We believe MMCircuitEval could foster collaboration between the AI and hardware communities and stimulate progress at the intersection of LLMs and circuit design.



## ACKNOWLEDGMENTS

This work was supported in part by the Hong Kong Research Grants Council (RGC) under Grant No. 14212422, 14202824, and C6003-24Y, in part by Huawei Technologies Co. Ltd. under grant No. N2-2c-TH2420350, and in part by National Technology Innovation Center for EDA, China.

## REFERENCES

- [1] M. Liu, T.-D. Ene, R. Kirby *et al.*, “Chipnemo: Domain-adapted llms for chip design,” *arXiv:2311.00176*, 2023.
- [2] C. Nguyen, W. Nguyen, A. Suzuki *et al.*, “Semikong: Curating, training, and evaluating a semiconductor industry-specific large language model,” *arXiv:2411.13802*, 2024.
- [3] S. Liu, Y. Lu, W. Fang, M. Li, and Z. Xie, “Openllm-rtl: Open dataset and benchmark for llm-aided design rtl generation,” in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024, pp. 1–9.
- [4] Y. Lu, S. Liu, Q. Zhang, and Z. Xie, “Rtllm: An open-source benchmark for design rtl generation with large language model,” in *Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2024.
- [5] R. Qiu, G. L. Zhang, R. Drechsler *et al.*, “Autobench: Automatic testbench generation and evaluation using llms for hdl design,” in *International Symposium on Machine Learning for CAD*, 2024.
- [6] M. Liu, N. Pinckney, B. Khailany *et al.*, “Verilogval: Evaluating large language models for verilog code generation,” in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023.
- [7] B.-Y. Wu, U. Sharma, S. R. D. Kankipati, A. Yadav, B. K. George, S. R. Guntupalli, A. Rovinski, and V. A. Chhabria, “Eda corpus: A large language model dataset for enhanced interaction with openroad,” *arXiv:2405.06676*, 2024.
- [8] Y. Pu, Z. He, T. Qiu, H. Wu, and B. Yu, “Customized retrieval augmented generation and benchmarking for eda tool documentation qa,” *arXiv:2407.15353*, 2024.
- [9] L. Chen, Y. Chen, Z. Chu, W. Fang, T.-Y. Ho, R. Huang, Y. Huang, S. Khan, M. Li, X. Li *et al.*, “Large circuit models: opportunities and challenges,” *Science China Information Sciences*, vol. 67, no. 10, p. 200402, 2024.
- [10] S. Thakur, J. Blocklove, H. Pearce, B. Tan, S. Garg, and R. Karri, “Autochip: Automating hdl generation using llm feedback,” *arXiv:2311.04887*, 2023.
- [11] S. Liu, W. Fang, Y. Lu, J. Wang, Q. Zhang, H. Zhang, and Z. Xie, “Rtl-coder: Fully open-source and efficient llm-assisted rtl code generation technique,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.
- [12] S. Thakur, B. Ahmad, H. Pearce *et al.*, “Verigen: A large language model for verilog code generation,” *ACM Transactions on Design Automation of Electronic Systems*, vol. 29, no. 3, pp. 1–31, 2024.
- [13] C. Liu, W. Chen, A. Peng, Y. Du, L. Du, and J. Yang, “Ampagent: An llm-based multi-agent system for multi-stage amplifier schematic design from literature for process and performance porting,” *arXiv:2409.14739*, 2024.
- [14] H. Wu, Z. He, X. Zhang, X. Yao, S. Zheng, H. Zheng, and B. Yu, “Chateda: A large language model powered autonomous agent for eda,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.
- [15] T. Ajayi and D. Blaauw, “Openroad: Toward a self-driving, open-source digital layout implementation tool chain,” in *Proceedings of Government Microcircuit Applications and Critical Technology Conference*, 2019.
- [16] K. Chang, Y. Wang, H. Ren, M. Wang, S. Liang, Y. Han, H. Li, and X. Li, “Chipppt: How far are we from natural language hardware design,” *arXiv:2305.14019*, 2023.
- [17] B. Li, R. Wang, G. Wang *et al.*, “Seed-bench: Benchmarking multimodal llms with generative comprehension,” *arXiv:2307.16125*, 2023.
- [18] X. Yue, Y. Ni, K. Zhang *et al.*, “Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567.
- [19] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao *et al.*, “Mm-bench: Is your multi-modal model an all-around player?” in *European Conference on Computer Vision*. Springer, 2025, pp. 216–233.
- [20] N. Xu, Z. Zhang, L. Qi *et al.*, “Chipexpert: The open-source integrated-circuit-design-specific large language model,” *arXiv:2408.00804*, 2024.
- [21] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv:2303.08774*, 2023.
- [22] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International conference on machine learning*, 2022.
- [23] J. Bai, S. Bai, S. Yang *et al.*, “Qwen-vl: A frontier large vision-language model with versatile abilities,” *arXiv:2308.12966*, 2023.
- [24] P. Zhang, X. Dong, B. Wang *et al.*, “Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition,” *arXiv:2309.15112*, 2023.
- [25] Z. Chen, J. Wu, W. Wang *et al.*, “Internvl: Scaling up vision foundation models and aligning for generic vision-linguistic tasks,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.
- [26] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” 2023.
- [27] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*, 2023.
- [28] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv:2407.21783*, 2024.
- [29] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, “Minicpm-v: A gpt-4v level mllm on your phone,” *arXiv:2408.01800*, 2024.
- [30] A. Young, B. Chen *et al.*, “Yi: Open foundation models by 01.ai,” 2024.
- [31] Z. Peng, W. Wang, L. Dong *et al.*, “Kosmos-2: Grounding multimodal large language models to the world,” *arXiv:2306.14824*, 2023.
- [32] OpenAI, “Gpt-4o system card,” *arXiv:2410.21276*, 2024.
- [33] R. Team, A. Ormazabal, C. Zheng *et al.*, “Reka core, flash, and edge: A series of powerful multimodal language models,” *arXiv:2404.12387*, 2024.
- [34] L. Ouyang, J. Wu, X. Jiang *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.
- [35] D. Dai, C. Deng, C. Zhao *et al.*, “Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models,” *arXiv:2401.06066*, 2024.
- [36] X. Bi, D. Chen, G. Chen *et al.*, “Deepseek llm: Scaling open-source language models with longtermism,” *arXiv:2401.02954*, 2024.
- [37] Z. Shao, P. Wang, Q. Zhu *et al.*, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv:2402.03300*, 2024.
- [38] A. Liu, B. Feng, B. Wang *et al.*, “Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model,” *arXiv:2405.04434*, 2024.
- [39] A. Yang, B. Yang, B. Hui *et al.*, “Qwen2 technical report,” *arXiv:2407.10671*, 2024.
- [40] I. Team, “Internlm: A multilingual language model with progressively enhanced capabilities,” 2023.
- [41] H. Touvron, L. Martin, K. Stone *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv:2307.09288*, 2023.
- [42] S. Hu, Y. Tu, X. Han *et al.*, “Minicpm: Unveiling the potential of small language models with scalable training strategies,” *arXiv:2404.06395*, 2024.
- [43] T. GLM, A. Zeng, B. Xu *et al.*, “Chatglm: A family of large language models from glm-130b to glm-4 all tools,” *arXiv:2406.12793*, 2024.
- [44] G. Team, R. Anil, S. Borgeaud *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv:2312.11805*, 2023.
- [45] G. Team, P. Georgiev, V. I. Lei *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *arXiv:2403.05530*, 2024.
- [46] C. Team, “Introducing the next generation of claude,” 2024. [Online]. Available: <https://www.anthropic.com/news/claude-3-family>
- [47] Z. Yang, P. Qi, S. Zhang *et al.*, “Hotpotqa: A dataset for diverse, explainable multi-hop question answering,” *arXiv:1809.09600*, 2018.
- [48] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” *arXiv:1606.05250*, 2016.
- [49] T. Brown, B. Mann, N. Ryder *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.