WE-MATH: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning?

Runqi Qiao[1*†], Qiuna Tan[1*†], Guanting Dong[1], Minhui Wu[2], Chong Sun[2], Xiaoshuai Song[1], Jiapeng Wang[3†], Zhuoma GongQue[1], Shanglin Lei[4], Yifan Zhang[1], Zhe Wei[1], Miaoxuan Zhang[1], Runfeng Qiao[5], Xiao Zong[1], Yida Xu[1], Peiqing Yang[1], Zhimin Bao[2], Muxi Diao[1], Chen Li[2‡], Honggang Zhang[1‡]

[1]BUPT [2]WeChat Vision, Tencent Inc. [3]SCUT [4]HUST [5]BIT

{qrq, qiunatan, dongguanting, zhhg}@bupt.edu.cn chaselli@tencent.com

https://we-math.github.io

Abstract

Visual mathematical reasoning, as a fundamental visual reasoning ability, has received widespread attention from the Large Multimodal Models (LMMs) community. Existing benchmarks mainly focus more on the end-to-end performance, but neglect the underlying principles of knowledge acquisition and generalization. Instead, we introduce WE-MATH, the first benchmark specifically designed to explore the problem-solving principles. We meticulously collect 6.5K visual math problems and decompose them into 10.9K step-level questions for evaluation, spanning 5 layers of knowledge granularity and 67 hierarchical knowledge concepts. Specifically, we decompose composite problems into sub-problems according to the required knowledge concepts and introduce a novel four-dimensional metric to hierarchically assess inherent issues in LMMs' reasoning process. With WE-MATH, we conduct a thorough evaluation of existing LMMs in visual mathematical reasoning and provide comprehensive analysis and insight for future development. We anticipate that WE-MATH will open new pathways for advancements in visual mathematical reasoning for LMMs. Data and code are available at https://github.com/We-Math/We-Math.

1 Introduction

"I think, therefore I am." — René Descartes

Human cognitive and reasoning patterns have profoundly shaped the progress of deep learning (LeCun et al., 2015). Recently, Large Language Models (LLMs) (Ouyang et al., 2022; Achiam et al., 2023; Touvron et al., 2023; Anil et al., 2023) and Large Multimodal Models (LMMs) (Liu et al., 2024b; Dai et al., 2024; Li et al., 2023; Zhang et al., 2023; Gao et al., 2023b; Bai et al., 2023; Su et al., 2023; Ye et al., 2023; Zhu et al., 2023) showcases strong reasoning abilities that parallel human performance across a wide range of tasks and provide a glimpse into the early outlines of Artificial General Intelligence (AGI).

Mathematical reasoning is a critical capability of foundational models. With the rapid advancements of LMMs, researchers progressively utilize the LMMs for solving visual mathematical problems (Yang et al., 2023; Gao et al., 2023a). To systematically evaluate visual mathematical reasoning capabilities, previous efforts (Lu et al., 2021; Seo et al., 2015; Chen et al., 2021, 2022) have focused on challenging geometric problems. Recently, several benchmarks (Lu et al., 2023; Zhang et al., 2024c) expand the scope to include a wider range of disciplines. However, these benchmarks rely solely on end-to-end results for assessment, which fails to identify inherent issues within the LMMs' reasoning process. While noticing that humans solve complex math problems through gradually mastering and generalizing the *knowledge concepts* (Fitzpatrick, 2008), we claim a fair evaluation of a model's reasoning process should be based on knowledge concepts. Therefore, we pose two questions about mathematical reasoning evaluation:

***Q1**: Does the correct answer truly reflect LMM's ability to reason through such problems accurately?*

***Q2**: Does an incorrect answer suggest a lack of foundational knowledge in LMM's reasoning process?*

As the response, we present **WE-MATH**, as shown in Figure 1, a pioneering benchmark for conducting an in-depth analysis of the underlying principles of LMMs in visual mathematical reasoning. WE-MATH consists of over 6.5K meticulously selected visual math problems, which can be categorized into 5 layers of knowledge granularity across 67 knowledge concepts for ensuring

*Equal contribution

†Work done as intern at WeChat, Tencent Inc.

‡Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20023–20070
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

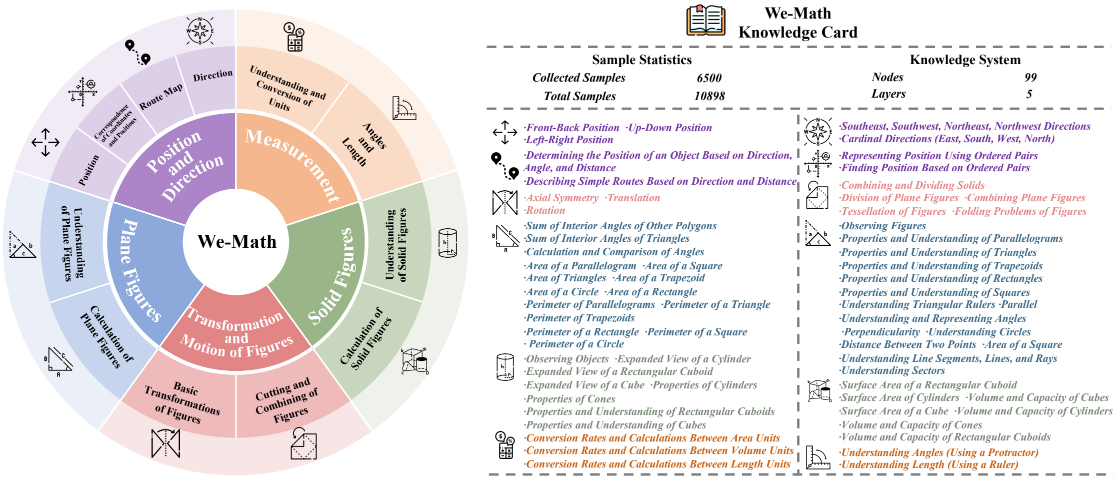


Figure 1: Overview diagram and the statistics of WE-MATH. The left and right side shows the first two layers of WE-MATH's categories and information of different samples and terminal nodes.

comprehensive coverage. We have also observed that real-world math problems typically encompass multiple foundational knowledge concepts, and their difficulty is directly related to the number of concepts involved. Upon this, we decouple the model's ability to solve composite problems with knowledge concepts into two stages:

(1) LMMs can solve individual sub-problems corresponding its knowledge concept;

(2) LMMs reason out the final answer by integrating the individual knowledge concepts.

To decompose a composite problem into individual sub-problems according to knowledge concepts, we further select multi-step problems with unique solutions (1.9K) from the 6.5K dataset. These composite problems are gradually decomposed by expert annotators into a one-step problem. After decomposition, we further combine them with the remaining problems (4.6K) to construct a high-quality dataset of 10.9K for assessment. Motivated by human reasoning patterns, WE-MATH also introduces a four-dimensional metric to precisely evaluate the inherent gaps in LMMs' problem-solving abilities, namely *Insufficient Knowledge (IK)*, *Inadequate Generalization (IG)*, *Complete Mastery (CM)*, and *Rote Memorization (RM)*. We also propose a heuristic knowledge concept augmentation (KCA) setting to provide essential knowledge for LMMs' reasoning evaluation, which can further explore the LMMs' ability to understand, organize, and solve problems related to knowledge.

Based on our thorough evaluation of existing LMMs in visual mathematical reasoning, we have also provided comprehensive analysis and insight for future development. We anticipate that WE-MATH can open new pathways for advancements in visual mathematical reasoning. Our main contributions can summarized as follows:

- We propose WE-MATH, the first benchmark that breaks down visual mathematical problems into step-wise sub-problems, exploring process level problem-solving mechanisms of LMMs.
- We establish the first hierarchical knowledge-oriented framework for mathematical reasoning, introducing multiple granularity levels and concepts to systematically investigate LMMs' mathematical problem-solving process.
- We design a fine-grained reasoning diagnosis metric (IK, IG, CM, RM), specifically designed to assess LMMs' step-wise reasoning process, probing into underlying reasoning deficiencies.
- We conduct a comprehensive assessment of 30 different open-source and closed-source multi-modal models with varying parameter sizes. Further quantitative analysis reveals whether these models genuinely understand mathematics or merely rely on rote memorization.

2 Related Work

Assessing mathematical reasoning abilities is crucial for the development of large foundational models (LLMs and LMMs). Early efforts, such as MathQA (Amini et al., 2019), focus on solving mathematical word problems and highlight the importance of operation-based reasoning. Following this, datasets like GSM8K (Cobbe et al.,

Table 1: Comparison between our WE-MATH and existing benchmarks.

Dataset	Step-wise Evaluation	Knowledge Concept	Fine-Grained Metric	Test Set Size	Testmini Subset Size
MMMU-Math (Yue et al., 2023)	-	-	-	540	-
Geometry3K (Lu et al., 2021)	-	-	-	601	-
MathVision (Wang et al., 2024a)	-	✓	-	2,736	304
MathVista (Lu et al., 2023)	-	✓	-	5,141	1,000
MathVerse (Zhang et al., 2024c)	-	✓	✓	15,672	4,728
We-Math (Ours)	✓	✓	✓	10,898	1,740

2021) and MATH (Hendrycks et al., 2021b) set the stage for evaluating text-based mathematical problems at various difficulty levels. Other benchmarks, such as MMLU (Hendrycks et al., 2021a) and MT-Bench (Zheng et al., 2023), also consider mathematical evaluation as a key part of assessing LLMs. Beyond text-only evaluations, datasets like GeoQA (Chen et al., 2021), UniGeo (Chen et al., 2022), and Geometry3K (Lu et al., 2021) have pioneered the evaluation of geometric problems. Recently, several benchmarks (Lu et al., 2023) (Yue et al., 2023) (Wang et al., 2024a) have expanded their scope to cover a broader range of subjects. Additionally, MathVerse (Zhang et al., 2024c) aims to evaluate reasoning paths based on reference answers. However, challenges remain due to the complex nature of mathematical reasoning. In this paper, we introduce WE-MATH designed to evaluate the reasoning abilities of LMMs across a wide range of mathematical categories.

3 WE-MATH Dataset

As shown in Figure 1, WE-MATH is constructed around textbook knowledge units, decomposing composite problem solutions into sub-problems based on the knowledge concepts. To further highlight the differences between WE-MATH and other benchmarks, we provide a detailed comparison in Table 1. WE-MATH consistently demonstrates superior performance across most dimensions. Additionally, we are committed to open-sourcing all the data used in this study.

3.1 Problem Definition

For the visual mathematical reasoning task, given text question Q_i, image I_i and corresponding answer A_i. We define the LMMs evaluation dataset $D_{\text{eval}} = \{(Q_i, I_i, A_i)|K_i, C_i\}_{i=1}^{N}$. where K_i and C_i are two prior constraints for question Q_i. In detail, $K_i = \{k_i\}_{i=1}^{M}$ denote M knowledge concepts within the question. C_i represents the prerequisite conditions needed to solve the problem Q_i (see Figure 2 for example).

3.2 Characteristic

Knowledge-based Decoupling. WE-MATH is designed to explore how LMMs solve problems. Drawing upon that humans tackle problems incrementally by leveraging fundamental knowledge concepts, we break down complex mathematical problems into more manageable sub-problems. We will also employ diverse measurement dimensions for meticulous evaluations.

Hierarchical Knowledge Structure. WE-MATH strictly adheres to the knowledge presented in mathematics textbooks, featuring a rigorous hierarchical and multi-category architecture. It ensures the independence of knowledge concepts within the same level, while establishing logical relationships among concepts at different hierarchical levels.

Fine-grained Fundamental Skills. WE-MATH emphasizes fundamental math skills, believing that complex mathematical reasoning is built upon foundation of basic mathematical reasoning processes. Based on extensive research, mathematical problems are categorized into 5 distinct types, namely *Plane Figures, Solid Figures, Transformation and Motion of Figures, Positions and Directions, Measurements*. These five categories can be decomposed into 12 typical problems, which are further decomposed as 67 knowledge concepts (terminal nodes in the structure). We collect problems according to this tree structure and constrain that each terminal node contains a strict range of 10-40 samples. This rule ensures data balance across domains.

3.3 Data Collection and Annotation

All problems (6.5K) in WE-MATH are sourced from publicly authoritative mathematics websites and subsequently organized based on our defined knowledge structure. We employ three expert anno-

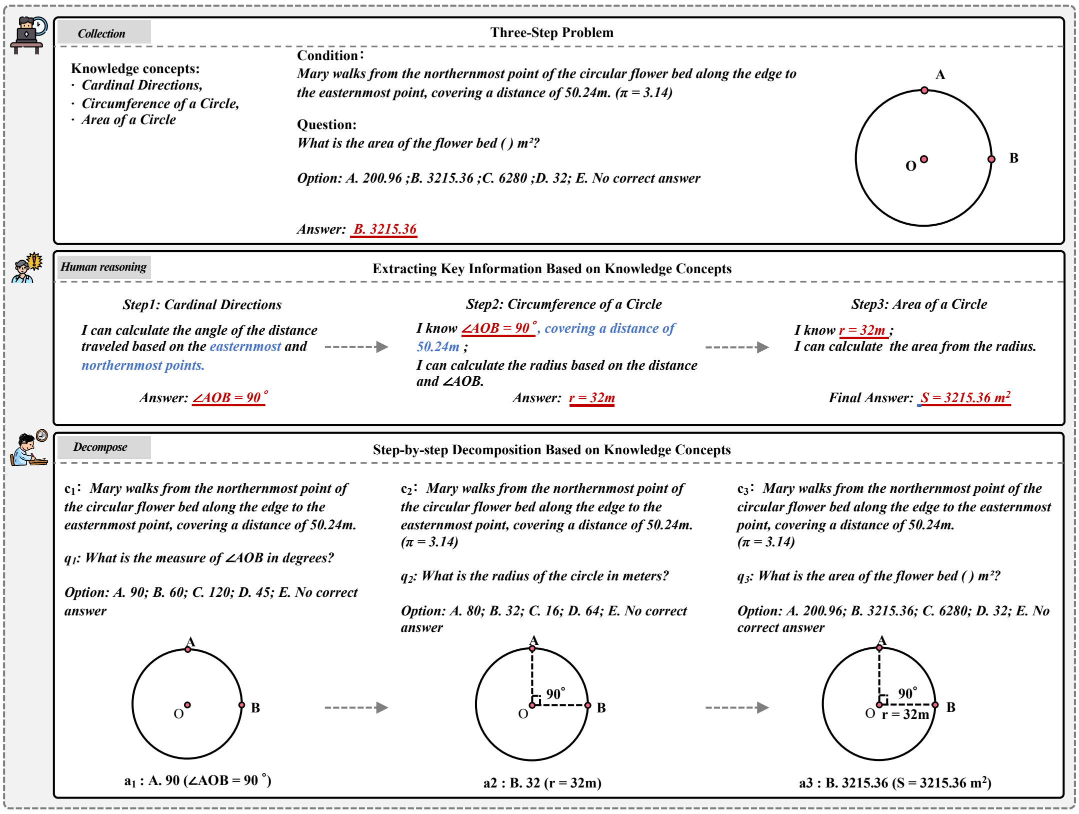


Figure 2: The pipeline of knowledge-based data decomposition (an example of a three-step problem in WE-MATH).

tators to manually label each question with knowledge concepts. Cross-validation is performed to ensure at least two experts have identical annotations for the same question. Samples with notably inconsistent labels will be considered of low quality and subsequently excluded.

To prepare for the subsequent decomposition of problems, we further annotate problem-solving steps based on the knowledge concept labels. We categorize each problem into three distinct classes: "One-Step", "Two-Step", and "Three-Step". This categorization enables us to gain a deeper understanding of how LMMs solve problems. After the annotation, all problems are double-checked by an expert team in terms of four aspects: (1) The consistency between the questions and diagrams; (2) The correctness of the answers to the questions; (3) The alignments between problems and the knowledge concepts; (4) Each problem contains a unique solution path.

We also strictly comply with copyright and licensing rules, ensuring that we refrain from using data from sites that forbid copying and redistribution. Further details about data collection can be found in supplementary materials.

4 WE-MATH Evaluation

4.1 Knowledge-based Decomposition

Inspired by Euclid's Elements (Fitzpatrick, 2008), we argue that the evaluation of mathematical reasoning ability in LMMs essentially involves assessing their mastery of fundamental knowledge concepts. It is quite a natural and objective way to exploit basic knowledge concepts for reasoning evaluation of LMMs. Given an i-th test sample $\{(Q_i, I_i, A_i)|K_i, C_i\} \in D_{\text{WE-MATH}}$ with M concepts $K_i = \{k_i^j\}_{j=1}^M$, we ask human experts to decompose each problem step by step into M sub-problems based on knowledge concepts, which can be formulated as:

$$\{(q_i^j, i_i^j, a_i^j)|k_i^j, c_i^j\}_{j=1}^M = \underset{(Q_i,I_i,A_i)\in D}{\text{Decomp.}} \{(Q_i, I_i, A_i)|K_i, C_i\}, \quad (1)$$

where k_i, c_i denote the individual knowledge and prior condition for the sub-problem. "Decomp." represents the Human decomposition process based on M knowledge concepts. To ensure logical coherence of decomposition, the condition c_i^m is initialized as C_i. Then it is recursively computed by concatenating the answer a_i^{m-1} and condition c_i^{m-1} of the $m-1$-th concept:

$$c_i^m = c_i^{m-1} + a_i^{m-1} \quad \text{for } m = 2, 3, \ldots, M, \quad (2)$$

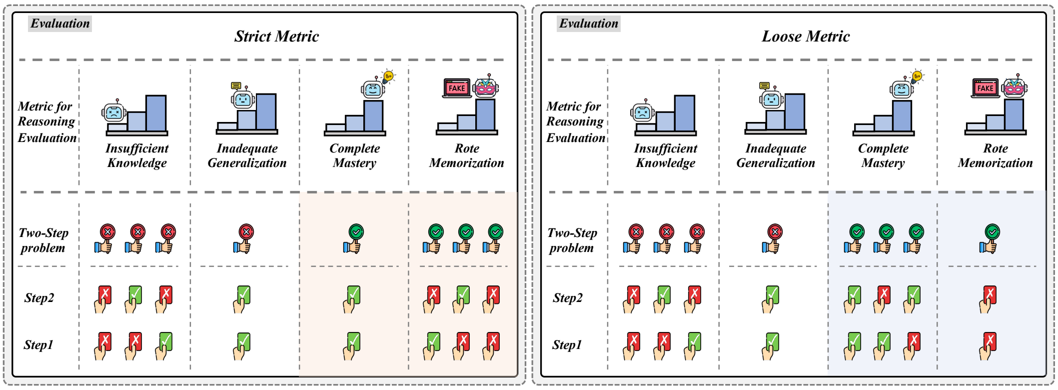


Figure 3: Example of four dimensional metrics for evaluating a two-step problem, using strict and loose metrics.

where "+" denotes the concatenation operation. In addition, the equation $\left\{ \begin{array}{c} q_i^M = Q_i \\ a_i^M = A_i \end{array} \right\}$ must be satisfied, which is also a constraint for logical coherence. Finally, we can obtain the original multi-step problem and M one-step sub-problems for reasoning evaluation. The overall pipeline of knowledge-based data decomposition is shown in Figure 2.

4.2 Metric for Reasoning Evaluation

Based on the decomposed multi-step problems, we further reveal the inherent issues of LMMs in problem-solving process. We feed both the M one-step sub-problems and the original problem into LMMs, and classifying the responses into the following four categories:

Insufficient Knowledge (IK): Part of one-step answers contain errors, and the multi-step answer is wrong. It is reasonable since an insufficient grasp of single knowledge concept may lead to errors in multi-step problem.

Inadequate Generalization (IG): One-Step answers are all correct, but the multi-step answer is incorrect. This is also considered reasonable. Although LMMs are capable of understanding individual knowledge concepts, they may struggle to generalize that knowledge to solve composite problems.

Complete Mastery (CM): One-Step answers are all correct, and the multi-step problem is also answered correctly. This result demonstrates that the model's results are both reliable and accurate.

Rote Memorization (RM): One-Step answers contain errors, but the multi-step problem is answered correctly, which contradicts human logical thinking. If a model can solve composite multi-step problems but fails to answer the one-step problems needed in the process, it raises doubts about the model's reliability.

Considering *IK* and *IG*, it is evident that results falling under the *IG* category are generally more preferred compared to those classified as *IK*. The reason is that *IK* reflects the model's struggle with both single and multiple knowledge concepts, while *IG* shows the model's proficiency in one-step problems. By enhancing the model's generalization ability in the reasoning process, we can potentially shift results from *IG* to *CM*. Therefore, we establish a reasoning capability hierarchy as *IK* $<$ *IG* $<$ *CM*. And we also regard *RM* as an unreasonable scenario (models can solve multi-step problems without mastering one-step problems, which completely contradicts human reasoning intuition).

Moreover, in light of the model's instability, the current criteria for determining whether a result belongs *RM* is strict. We thus propose a more flexible loose metric. As illustrated in Figure 3, only cases where all sub-problems are incorrect will be classified as *RM* under this loose metric. The analysis of three-step cases is presented in the appendix.

Finally, we propose the following metric to judge the reliability of the model's reasoning process: $S_{\text{IK}} = N_{\text{IK}}/N$, $S_{\text{IG}} = N_{\text{IG}}/N$, $S_{\text{CM}} = N_{\text{CM}}/N$, $S_{\text{RM}} = N_{\text{RM}}/\{N_{\text{RM}} + N_{\text{CM}}\}$, where N denotes the total number of samples and N_{IK}, N_{IG}, N_{CM}, N_{RM} represents the number of samples for a specific situation. Therefore, we obtain our final reasoning confidence scores:

$$\text{Score}_{\text{average}} = \alpha S_{\text{IK}} + \beta S_{\text{IG}} + S_{\text{CM}} \tag{3}$$

where α, β denotes the weight for each case. To ensure the reasoning capability hierarchy is "*IK* < *IG* < *CM*", we control the params $\alpha < \beta < 1$, and set the default value of α to 0.0 and β to 0.5.

4.3 Knowledge Concept Augmentation

We have identified that the *IK* issue is the foundation challenge in mathematical reasoning. To heuristically tackle this issue during evaluation, we introduce the knowledge concept augmentation (KCA) setting, which enlists human experts to create knowledge concept cards for LMM's reasoning process. Initially, expert annotators offer precise summaries derived from the definitions in Euclid's Elements (Fitzpatrick, 2008), Wikipedia, and textbooks. Subsequently, these experts further condense the content examined by a series of questions related to a specific knowledge concept, extracting crucial knowledge hints for incorporation into the knowledge cards. After several rounds of review, we confirm the accuracy and utility of each card. Consequently, with a given problem and its respective knowledge concept, LMMs utilize the relevant knowledge cards to deduce the answer.

5 Experiment

Evaluation Protocols. To accelerate the evaluation speed, WE-MATH comprises a *testmini* set with 1740 samples, including 1215 one-step samples, 360 two-step samples, and 165 three-step samples. In subsequent experiments, we utilize the WE-MATH *testmini* subset for evaluation. For automated evaluation, all samples are standardized into a multiple-choice format. We use regex to match the LMMs' predictions and then calculate their accuracy against the ground-truth answers for main results. To avoid LMMs deduce answers from options, we introduce an extra *uncertain* option to mitigate this issue. The results of the entire test set can be found in the supplementary material.

Evaluation Models. We examine the performance of MLLMs across two categories: (a) Closed-source LMMs: GPT-4o (OpenAI, 2024), GPT-4V (OpenAI, 2023), Gemini 1.5 Pro (Team et al., 2023), Qwen-VL-Max (Bai et al., 2023); (b) Open-source LMMs: LLaVA-OneVision-72B, LLaVA-OneVision-7B (Li et al., 2024), InternVL2.5-78B, InternVL2.5-26B, InternVL2.5-8B, InternVL2.5-4B, InternVL-Chat-V1.5 (Chen et al., 2024), Qwen2-VL-72B, Qwen2-VL-7B (Wang et al., 2024b), LLaVA-1.6-13B, LLaVA-1.6-7B (Liu et al., 2023), DeepSeek-VL-1.3B (Lu et al., 2024), Phi3-Vision-4.2B (Abdin et al., 2024), MiniCPM-Llama3-V2.5 (Hu et al., 2023), InternLM-XComposer2.5-7B (Zhang et al., 2024a), GLM-4V-9B (GLM et al., 2024), LongVA (Zhang et al., 2024b), G-LLaVA-13B (Gao et al., 2023a).

5.1 Main Results

Table 2 and Figure 4 show the overall performance of different LMMs on One/Two/Three-Step problems and different problem domains. We have the following observations:

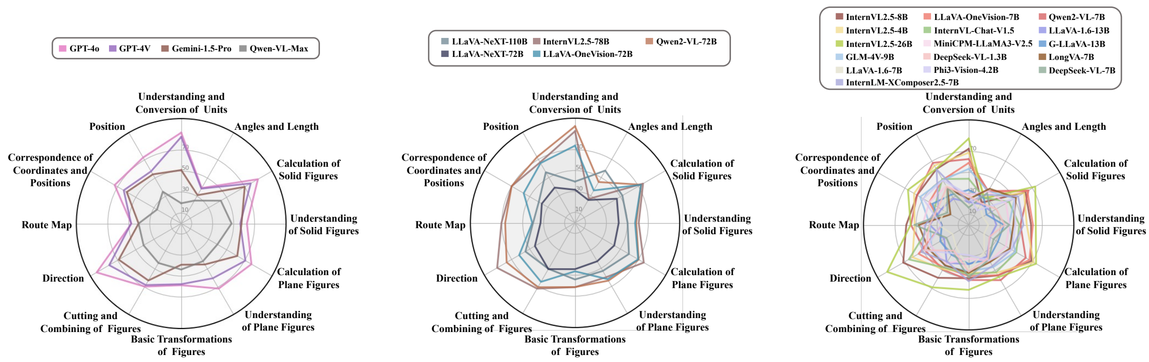
The numbers of knowledge concepts are negatively correlated with performance. Regarding problems of varying complexities (one-step vs. two-step vs. three-step), GPT-4o, InternVL2.5-78B, and Qwen2-VL-72B lead in most settings. However, most LMMs perform significantly worse on multi-step problems compared to one-step problems. For instance, GPT-4o's accuracy drops from 72.8% to 43.6%. This trend is evident in stronger LMMs like InternVL2.5-78B and LLaVA-OneVision-72B. These observations indicate that the number of knowledge concepts in a question correlates positively with its difficulty and negatively with LMMs' performance, reinforcing the rationale for decomposing questions.

Larger parameter scales in LLMs generally achieve better generalization ability. To explore what role LLM plays in LMMs, we conduct pairwise comparisons on the LMMs with the same LLM backbone (e.g. Qwen2-VL-72B vs Qwen2-VL-7B; InternVL2.5-26B vs InternVL2.5-8B). Focusing on the strict metric, we observe that larger parameter scales in LLMs generally perform better, which reveals that the parameter scales in the text decoder is a key factor in achieving the generalization ability in visual mathematical reasoning.

LMMs excel in calculation but struggle with fine-grained visual measurement. Focusing on different math categories, GPT-4o continues to achieve impressive results across various subfields. Moreover, recent LMMs such as InternVL2.5, Qwen2-VL and the LLaVA-OneVision series have also demonstrated competitive performance. However, other LMMs generally struggle with tasks like "Angle Measurement" and "Unit Conversion". Upon analyzing these cases, we reveals that the main challenge for LMMs lies in their inability to perform precise visual angle and unit measurements. Furthermore, most LMMs demonstrate better proficiency in calculation (e.g., *Calculations of Solid Figures*) compared to conceptual understanding (e.g., *Understanding of Solid Figures*), which indicates that most LMMs excel at directly applying formulas based on conditions, but are still

Table 2: **Accuracy scores of LMMs on the *testmini* subset of WE-MATH.** The first 3 columns report the overall performance on one-step, two-step, three-step problems, while the other columns display the result on one-step problems in different problem categories. The highest accuracy for closed-source and open-source LMMs is marked in blue and green respectively. (S1: one-step, S2: two-step, S3: three-step, Mem: Measurement, PF: Plane Figures, SF: Solid Figures, TMF: Transformation and Motion of Figures, PD: Position and Direction. AL: Angles and Length, UCU: Understanding and Conversion of Units, CPF: Calculation of Plane Figures, UPF: Understanding of Plane Figures, CSF: Calculation of Solid Figures, USF: Understanding of Solid Figures, BTF: Basic Transformations of Figures, CCF: Cutting and Combining of Figures, Dir: Direction, Pos: Position, RoM: Route Map, CCP: Correspondence of Coordinates and Positions).

Model	S1	S2	S3	Mem		PF		SF		TMF		PD			
				UCU	AL	CPF	UPF	CSF	USF	BTF	CCF	Dir	Pos	RoM	CCP
Closed-source															
GPT-4o	72.8	58.1	43.6	86.6	39.1	77.4	71.6	84.5	62.3	58.7	69.4	93.1	72.7	47.5	73.3
GPT-4V	65.5	49.2	38.2	82.5	38.4	70.7	60.2	76.6	56.3	57.8	67.7	79.3	57.5	47.8	63.3
Gemini 1.5 Pro	56.1	51.4	33.9	51.0	31.2	61.8	45.0	70.0	57.5	39.2	62.7	68.8	54.1	40.7	60.0
Qwen-VL-Max	40.8	30.3	20.6	19.4	25.3	39.8	41.4	43.6	48.0	43.8	43.4	41.4	35.1	40.7	26.7
Open-source															
InternVL2.5-78B	68.8	59.7	41.8	87.6	26.5	75.1	60.9	75.9	59.9	61.5	72.6	86.0	66.8	70.3	70.0
Qwen2-VL-72B	68.1	53.1	51.0	92.4	45.1	70.2	63.8	73.0	58.5	61.3	71.0	75.5	72.7	66.8	70.0
InternVL2.5-26B	67.5	55.0	40.6	82.2	29.1	73.1	63.8	74.0	57.3	61.9	68.7	89.5	61.7	55.2	66.7
LLaVA-OneVision-72B	64.0	45.8	35.8	73.8	35.8	69.7	62.2	72.8	57.4	46.3	65.1	61.7	66.0	41.0	56.7
InternVL2.5-8B	60.7	45.6	32.7	72.4	24.6	62.4	59.3	68.8	55.6	50.8	58.2	71.9	61.4	59.3	56.7
Qwen2-VL-7B	59.1	43.6	26.7	62.7	37.2	62.6	60.8	65.7	49.2	52.5	49.2	48.1	68.2	55.0	56.7
InternVL2.5-4B	58.3	42.8	30.3	68.8	30.5	60.9	55.6	71.3	52.7	45.5	48.7	61.7	65.3	51.4	60.0
LLaVA-OneVision-7B	57.5	43.1	39.4	59.0	36.5	66.7	55.4	64.4	61.1	48.6	46.9	55.0	49.5	25.6	43.3
InternVL-Chat-V1.5	49.4	30.6	28.5	44.0	29.8	52.2	52.1	44.2	48.2	47.1	46.8	65.7	50.5	36.5	36.7
InternLM-XComposer2.5-7B	49.0	32.2	23.0	21.7	33.2	54.3	52.1	47.0	45.2	53.7	40.5	51.7	61.1	41.2	33.3
GLM-4V-9B	47.3	37.2	38.2	53.4	37.0	51.3	46.5	50.6	38.2	44.1	45.2	41.0	49.3	36.8	53.3
LongVA-7B	43.5	30.6	28.5	24.5	39.8	45.1	40.8	51.9	42.5	45.6	44.6	44.5	40.7	47.5	20.0
Phi3-Vision-4.2B	42.1	34.2	27.9	28.7	16.0	47.2	38.8	50.0	44.4	28.8	31.2	48.6	49.2	26.4	50.0
MiniCPM-Llama3-V2.5	39.8	31.1	29.7	28.6	37.0	40.8	39.8	41.0	38.6	32.0	42.7	41.0	42.7	44.0	43.3
DeepSeek-VL-1.3B	31.4	27.8	23.0	27.8	23.9	22.8	36.9	30.4	34.2	44.5	28.3	48.1	41.8	37.1	33.3
G-LLaVA-13B	32.4	30.6	32.7	33.3	29.1	32.0	37.9	19.6	33.5	37.1	32.8	31.2	33.2	25.6	40.0
LLaVA-1.6-13B	29.4	25.3	32.7	21.7	23.2	23.4	34.7	25.3	26.4	37.5	41.7	26.9	28.9	37.1	30.0
LLaVA-1.6-7B	23.0	20.8	15.8	18.5	20.5	16.9	29.6	15.6	18.6	42.7	24.1	17.6	43.3	28.9	26.7



(a) Closed-Source Models **(b) Open-Source Models (Larger)** **(c) Open-Source Models (Smaller)**

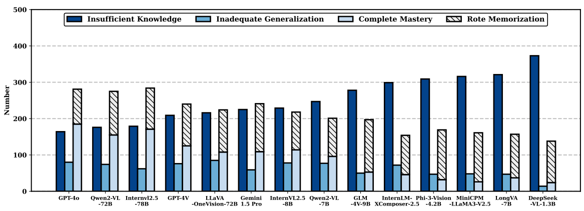
Figure 4: The performance of different LMMs on each category.

limited in understanding and comprehensively applying knowledge.

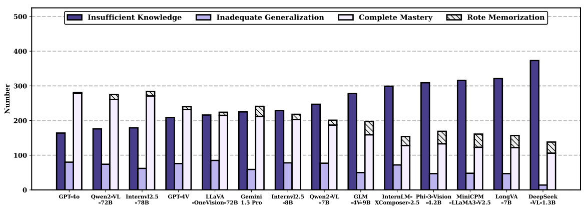
LMMs exhibit strong potential for parameter compression. In terms of different LMMs, InternVL2.5-78B and Qwen2-VL-72B demonstrates performance close to GPT-4o. Surprisingly, despite having smaller parameter scales, the recent InternVL2.5-8B and Qwen2VL-7B demonstrate competitive performance compared to GPT-4V. We attribute this impressive performance largely to their allocation of a greater proportion of parameters to the visual encoder, thereby demonstrating notable capabilities. This underscores the importance of optimizing visual representations and sug-

Table 3: The performance of different LMMs on four-dimensional metrics for reasoning evaluation. The best performance for closed-source and open-source LMMs is marked in blue and green (Avg: Score$_{\text{average}}$).

Model	Strict					Loose				
	Avg (↑)	**IK (↓)**	**IG (↓)**	**CM (↑)**	**RM (↓)**	**Avg (↑)**	**IK (↓)**	**IG (↓)**	**CM (↑)**	**RM (↓)**
Closed-source										
GPT-4o	42.9	31.2	15.2	35.2	34.2	60.6	31.2	15.2	52.3	1.1
GPT-4V	31.1	39.8	14.5	23.8	47.9	51.4	39.8	14.5	44.2	3.3
Gemini-1.5-Pro	26.4	42.9	11.2	20.8	54.8	46.0	42.9	11.2	40.4	12.0
Qwen-VL-Max	10.5	65.1	7.6	6.7	75.5	25.5	65.1	7.6	21.7	20.3
Open-source										
InternVL2.5-78B	38.5	34.1	11.8	32.6	39.8	57.5	34.1	11.8	51.6	4.6
Qwen2-VL-72B	36.6	33.5	14.1	29.5	43.6	56.8	33.5	14.1	49.7	5.1
InternVL2.5-26B	34.6	33.3	16.2	26.5	47.6	56.1	33.3	16.2	48.0	4.9
InternVL2.5-8B	29.1	43.6	14.9	21.7	47.7	46.1	43.6	14.9	38.7	6.9
LLaVA-OneVision-72B	28.7	41.1	16.2	20.6	51.8	49.1	41.1	16.2	41.0	4.0
Qwen2-VL-7B	25.6	47.1	14.7	18.3	52.2	43.0	47.1	14.7	35.6	7.0
InternVL2.5-4B	25.0	48.2	13.0	18.5	52.5	42.1	48.2	13.0	35.6	8.3
LLaVA-OneVision-7B	23.1	45.0	13.1	16.6	60.5	44.9	45.0	13.1	38.3	8.6
InternLM-XComposer2.5-7B	15.6	57.0	13.7	8.8	70.1	31.2	57.0	13.7	24.4	16.9
InternVL-Chat-V1.5	15.0	56.2	13.9	8.0	73.3	32.7	56.2	13.9	25.7	14.0
GLM-4V-9B	14.9	53.0	9.5	10.1	73.1	35.1	53.0	9.5	30.3	19.3
LongVA-7B	11.5	61.1	9.0	7.1	76.4	27.7	61.1	9.0	23.2	22.3
Phi3-Vision-4.2B	10.6	58.9	9.0	6.1	81.1	29.8	58.9	9.0	25.3	21.3
MiniCPM-Llama3-V2.5	9.6	60.2	9.1	5.0	83.9	28.1	60.2	9.1	23.4	23.6
G-LLaVA-13B	6.5	64.2	4.6	4.2	86.6	22.3	64.2	4.6	20.0	36.0
DeepSeek-VL-1.3B	5.9	71.1	2.7	4.6	82.6	21.5	71.1	2.7	20.2	23.2
LLaVA-1.6-13B	5.2	69.1	3.2	3.6	86.9	22.0	69.1	3.2	20.4	26.2
LLaVA-1.6-7B	3.3	78.3	2.5	2.1	89.1	13.8	78.3	2.5	12.6	34.7



(a)



(b)

Figure 5: The performance of different LMMs on four-dimensional metrics under (a) strict or (b) loose metric.

gests that LMMs still have significant potential for parameter compression.

5.2 Knowledge-based Reasoning Analysis

Table 3 and Figure 5 illustrate the results of knowledge-based reasoning evaluation, including four distinct conditions (*IK, IG, CM, RM*). We have the following observations:

IK is the greatest vulnerability of LMMs. All LMMs consistently demonstrate an *Insufficient Knowledge (IK)* issue during the reasoning process, especially in models with smaller parameter scales (LLaVA-1.6-7B, DeepSeek-VL-1.3B). Addressing *IK* is crucial for progressing towards *Inadequate Generalization (IG)* and *Complete Mastery (CM)*. This knowledge gap in solving one-step problems hinders further progress in reasoning about more composite mathematical problems. It also supports the rationale behind our proposed KCA setting.

GPT-4o, InternVL2.5 and Qwen2-VL have gradually shifted from *IK* to *IG*, marking their progression toward the knowledge generalization stage. Focusing on *IK* and *IG*, GPT-4o exhibits a substantial lead in addressing the *IK* issue, but performs poorly in *IG*. Moreover, InternVL2.5 and Qwen2-VL display a similar trend. To gain an insight into the logical relationships between *IK*, *IG*, and *CM* (*IK* $\rightarrow$ *IG* $\rightarrow$ *CM*), we are pleasantly surprised to find that GPT-4o is markedly superior to the open-sourced LLaVA-1.6-13B in *IK* (37.9%), suggesting them successfully converted a considerable amount of *IK* into *IG* issue. This revelation indicates that these models challenges in

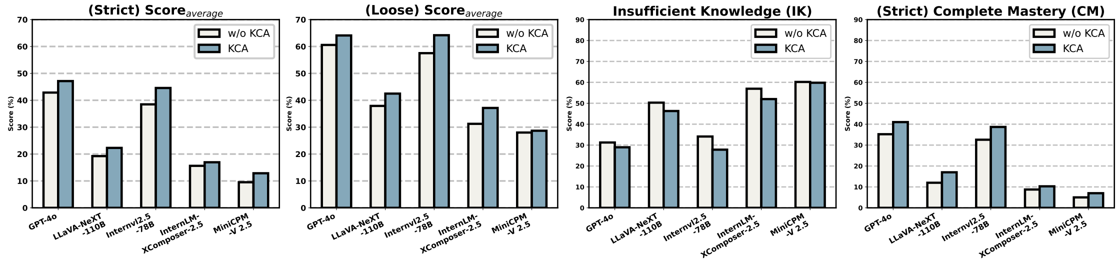


Figure 6: Quantitative Analysis under KCA setting. The left two figures show the impact of KCA on the performance of LMMs under strict and loose settings. The right two figures compare the results between *IK* and *CM*.

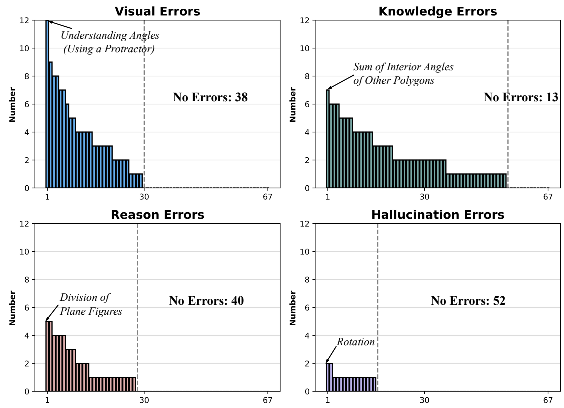


Figure 7: Error analysis of GPT-4o. The definitions of 4 types of errors are listed in supplementary materials.

reasoning have shifted from addressing *Insufficient Knowledge* in one-step problems to the knowledge generalization stage. However, other LMMs remain stuck at the *IK* phase. We argue that it is pointless to compare *IG* without a solid grasp of *IK*, highlighting the significance of our hierarchical metrics (*IK* < *IG* < *CM*).

The unreasonable RM issue remains widespread across most LMMs. GPT-4o achieves a significant lead on the *RM* issue, particularly on the loose metric ($S_{RM} < 2\%$). Futhermore, recent advanced series LMMs such as InternVL2.5, Qwen2-VL, and LLaVA-OneVision have also demonstrated outstanding performance ($S_{RM} < 10\%$). Unfortunately, other LMMs still exhibit nearly 25% S_{RM} on the loose metric. When focusing on the changes in S_{RM} between strict and loose metrics, several models (LLaVA-1.6-7B, GLM-4V-9B, DeepSeek-VL-1.3B, MiniCPM-Llama3-V2.5) show significant variations. This is a beneficial phenomenon, indicating that these models possess a certain ability to solve one-step problems, but their performance fluctuates due to external factors such as prompting templates and hyper-parameters.

5.3 Quantitative Analysis

Assessment under KCA setting. Figure 6 displays the quantitative analysis of the LMMs with our introduced knowledge concept augmentation (KCA) setting. We find that LMMs with different parameter scales show consistent performance improvements on both strict and loose metrics after involving KCA. Additionally, KCA significantly mitigates the *IK* issue and further enhances performance under *CM*, with larger models benefiting more than smaller ones. This aligns with human intuition, as knowledge descriptions primarily address gaps in reasoning knowledge. Moreover, the greater improvements observed in larger models further validate the potential of knowledge augmentation as a direction for future exploration.

Error Anaysis. Figure 7 shows the occurrence of the four types of errors across the 67 knowledge concepts. Knowledge errors are the most frequent, appearing in over 45 knowledge concepts. Notably, although visual errors are the second most common, they are more concentrated in specific concepts (e.g., *"Understanding Angles"* >10), and over 38 concepts have no visual errors. This finding underscores the urgent need to enhance the fine-grained measurement capabilities of LMMs for mathematical reasoning, rather than blindly improving their overall capabilities.

6 Conclusion

In this paper, we propose WE-MATH, a comprehensive benchmark for in-depth analysis of LMMs in visual mathematical reasoning. We pioneeringly decompose composite problems into sub-problems according to the required knowledge concepts and introduce a comprehensive multi-dimensional metric for fine-grained reasoning evaluation. With WE-MATH, we have also thoroughly evaluated existing LMMs in visual mathematical reasoning and provided comprehensive analysis.

Limitations

While WE-MATH is the first work to focus on exploring the problem-solving principles beyond end-to-end performance, it is important to recognize several limitations as follows.

Knowledge-based decomposition of comprehensive questions is applied in WE-MATH. However, due to the limited coverage of advanced function knowledge in elementary school mathematics, this paper does not delve deeply into function-related problems. Furthermore, to ensure that the knowledge concepts remain as independent as possible, WE-MATH currently focuses on elementary school problems. It is also meaningful to further extend the work to address middle school and high school mathematics in the future.

The problems in WE-MATH are all in English. We did not include problems in other languages, so the dataset cannot evaluate models' reasoning abilities in languages other than English. There is significant value in further augmenting WE-MATH with problems spanning a broader range of languages and complexities, including those at the middle school and high school level and within scientific fields.

Ethical Considerations

Ethics Statement We ensure that WE-MATH complies with legal and ethical guidelines throughout its construction process, with no violations. We provide fair compensation to all annotators involved. WE-MATH focuses on elementary mathematics problems, and during its construction, data collection was sourced from publicly available test questions, textbooks, and professional websites. Since mathematics problems inherently have standard answers, they are not subject to cultural differences. Additionally, we guarantee that WE-MATH is solely for academic research purposes, and we uphold the strict prohibition of any commercial use. Additionally, we declare that we will bear full responsibility in the event of any rights violations and confirm the data license.

Copyright and License We strictly comply with the copyright requirements of all datasets used and ensure their usage aligns with the respective licensing agreements. After thorough communication with the expert teaching team involved in WE-MATH, we choose the non-commercial license ("CC BY-NC 4.0") for our open-source usage . This license satisfies our needs during the collection of open-source datasets and the subsequent decomposition and modification of the questions with the expert teaching team.

References

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

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

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong

Zhang, Xizhou Zhu, Lewei Lu, et al. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Richard Fitzpatrick. 2008. Euclid's elements of geometry.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023a. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

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

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

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

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. 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.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

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

OpenAI. 2024. Hello gpt-4o.

R OpenAI. 2023. Gpt-4v (ision) system card. *Citekey: gptvision*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

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

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024b. Long context transfer from language to vision. *Preprint*, arXiv:2406.16852.

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

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A Broaden Impact

Bridging Human-Like Inspiration and Reliability. As previously mentioned, works such as neural networks (LeCun et al., 1998) and attention mechanisms (Vaswani et al., 2017) draw their design inspiration from human thinking patterns. This is fundamentally because the purpose of designing AI is to assist humans. Currently, LMMs have already been helping people in various scenarios, which was unimaginable in the past. Therefore, we firmly believe that a new era is coming, where people will focus not only on the performance of models in specific fields but also on the reliability of a model. In some fundamental scenarios, a reliable model is more important, which is one of the primary motivations behind the creation of WE-MATH. Furthermore, after completing our experiments, we find that in a loose setting, GPT-4o's *RM* metric is only 1.1%, showing us the possibility of a reliable and accurate model emerging in the future.

Fine-grained Evaluation and Versatile Applications. From the model's perspective, WE-MATH can provide LMMs with an assessment of mathematical abilities. Additionally, WE-MATH's *IK*, *IG*, and *CM* metrics offer a fine-grained evaluation of the model's capabilities. Furthermore, the *RM* metric reflects a model's reliability to address our concern of not desiring a model that can solve complex problems but makes errors on sub-problems within the solution process. Ultimately, we introduce the $\text{Score}_{\text{average}}$ metric to quantify the model's overall performance. Moreover, since WE-MATH is constructed from the decomposition of a multi-step problem's necessary solution process, it provides new perspectives for interactive tasks (multi-turn dialogues), self-supervised learning, information extraction, and other tasks. It also offers crucial references and support for the deployment of models in education and other fields.

B More Details on WE-MATH

B.1 Hierarchical Knowledge Structure

Figure 8, 9 shows the detailed hierarchical structure of WE-MATH, which includes 5 levels, 99 nodes, and 67 leaf nodes.

In the initial stages of constructing the benchmark, we aimed to address two key objectives. We believe that the purpose of designing a benchmark is to evaluate the performance of models and provide guidance on areas that need improvement. However, existing benchmarks offer only broad guides in these aspects. Additionally, the core contribution mentioned earlier is that WE-MATH is the first benchmark specifically designed to study the mathematical problem-solving mechanisms of models. Inspired by the learning paradigm of humans, which is based on knowledge concepts, WE-MATH constructs its dataset with knowledge concepts as the basic unit, resulting in evaluations with rigorous scientific accuracy and better guidance.

B.2 Knowledge-based Data Decomposition

Figures 10, 11 illustrate the process of Knowledge-based Data Decomposition.

Collection. In each example, the Collection section presents specific information about each multi-step problem in the dataset.

Human reasoning. The Human reasoning section shows the process required before decomposing each multi-step problem, where educational experts extract the key information needed for each sub-problem based on the reasoning path for the knowledge concepts included in the multi-step problem.

Decompose. The Decompose section uses the

Table 4: Prompt templates for response generations.

Type	Prompt Template
Multiple Choice	Now, we require you to solve a multiple-choice math question. Please briefly describe your thought process and provide the final answer(option). **Question**: <Question> **Option**: <Option> Regarding the format, please answer following the template below, and be sure to include two <> symbols: **<Thought process>**: <<your thought process>> **<Answer>**: <<your option>>
Knowledge Concept Augmentation	Now, we require you to solve a multiple-choice math question. We will provide you with the relevant knowledge concepts of this question for your reference. Please briefly describe your thought process and provide the final answer(option). **Knowledge concept**: <Knowledge concept> **Question**: <Question> **Option**: <Option> Regarding the format, please answer following the template below, and be sure to include two <> symbols: **<Thought process>**: <<your thought process>> **<Answer>**: <<your option>>

key information extracted in the human reasoning section to formulate sub-problems, refine the options, and ultimately achieve the decomposition of a multi-step problem.

It is necessary to further explain that to ensure each sub-problem has a rigorous logical relationship and is independent, the text condition for the first sub-problem is derived from the text condition of the multi-step problem, and the image condition for the first sub-problem is the same as the image condition of the multi-step problem.

Furthermore, in constructing the second sub-problem, two situations may arise. The first situation is where the answer of the first sub-problem is injected as a key condition into the image condition of the second sub-problem, presenting the information visually. The second situation is where the answer of the first sub-problem is injected as a key condition into the text condition of the second sub-problem, while the image condition remains unchanged. In WE-MATH, the vast majority of cases are of the first type. However, for some information that is extremely difficult to present in images, we opt for the second type, presenting the information in text form. To ensure fairness in the decomposition of the problems, only one of these situations will occur in the decomposition of the same multi-step problem. This approach ensures that the question of the final sub-problem will match the original multi-step problem, completing the decomposition.

B.3 Knowledge Concepts Augmentation

Table 4 report the prompt templates in our experiments. We concatenate the textual descriptions into the prompt. Additionally, each knowledge concept description is accompanied by its corresponding visual content, which helps the experimenter understand and facilitates further enhancement when models can incorporate sufficient visual information as part of the prompt in the future. We illustrates the specific content of descriptions for 67 knowledge concepts. For example, as shown in Figure 61, for the knowledge concept "Perimeter of Squares," it is necessary to know that "c=4a", relying solely on textual descriptions is insufficient for understanding this concept, so we include visual information to aid comprehension.

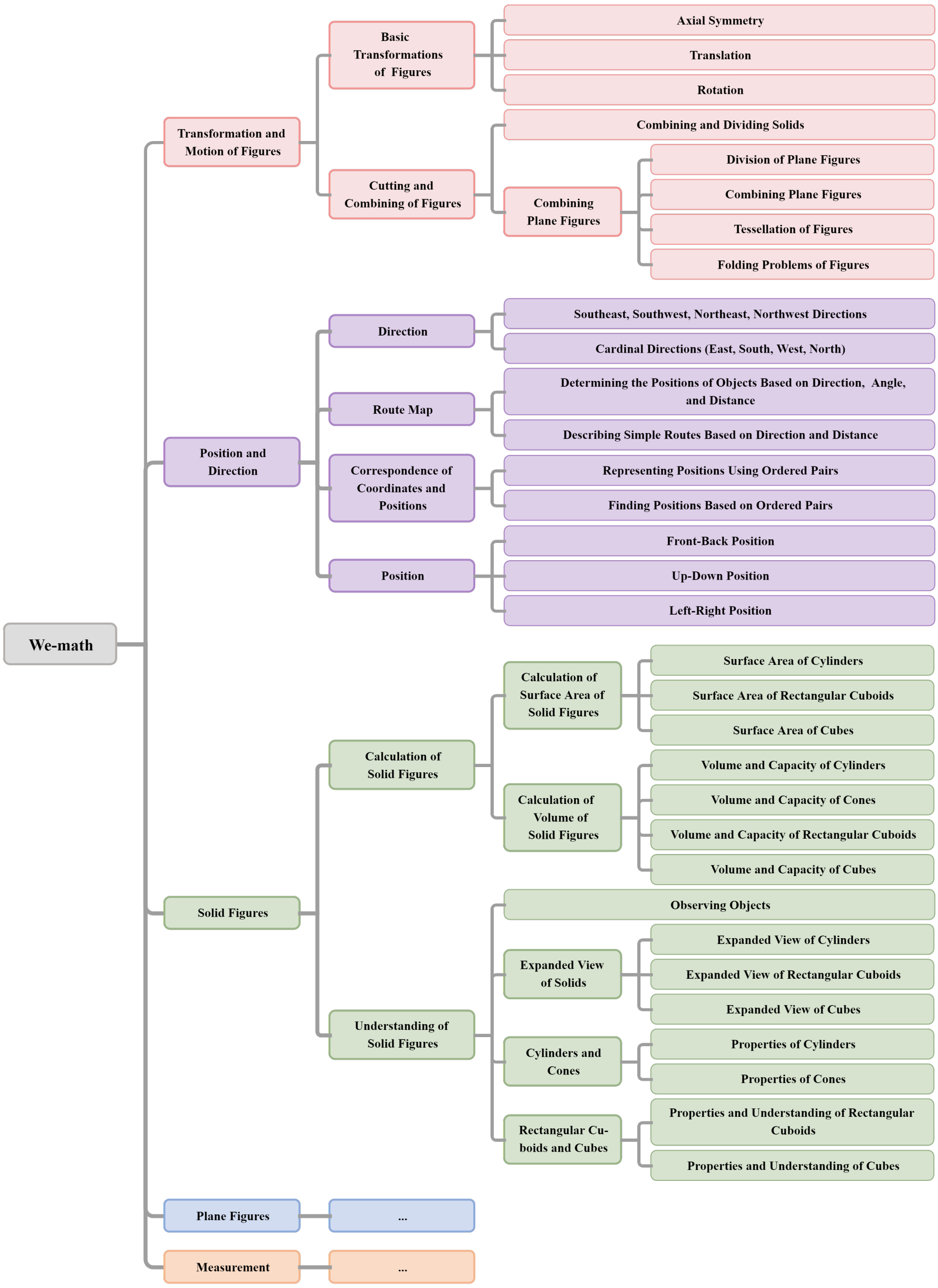


Figure 8: The Hierarchical Knowledge Structure of WE-MATH (1).

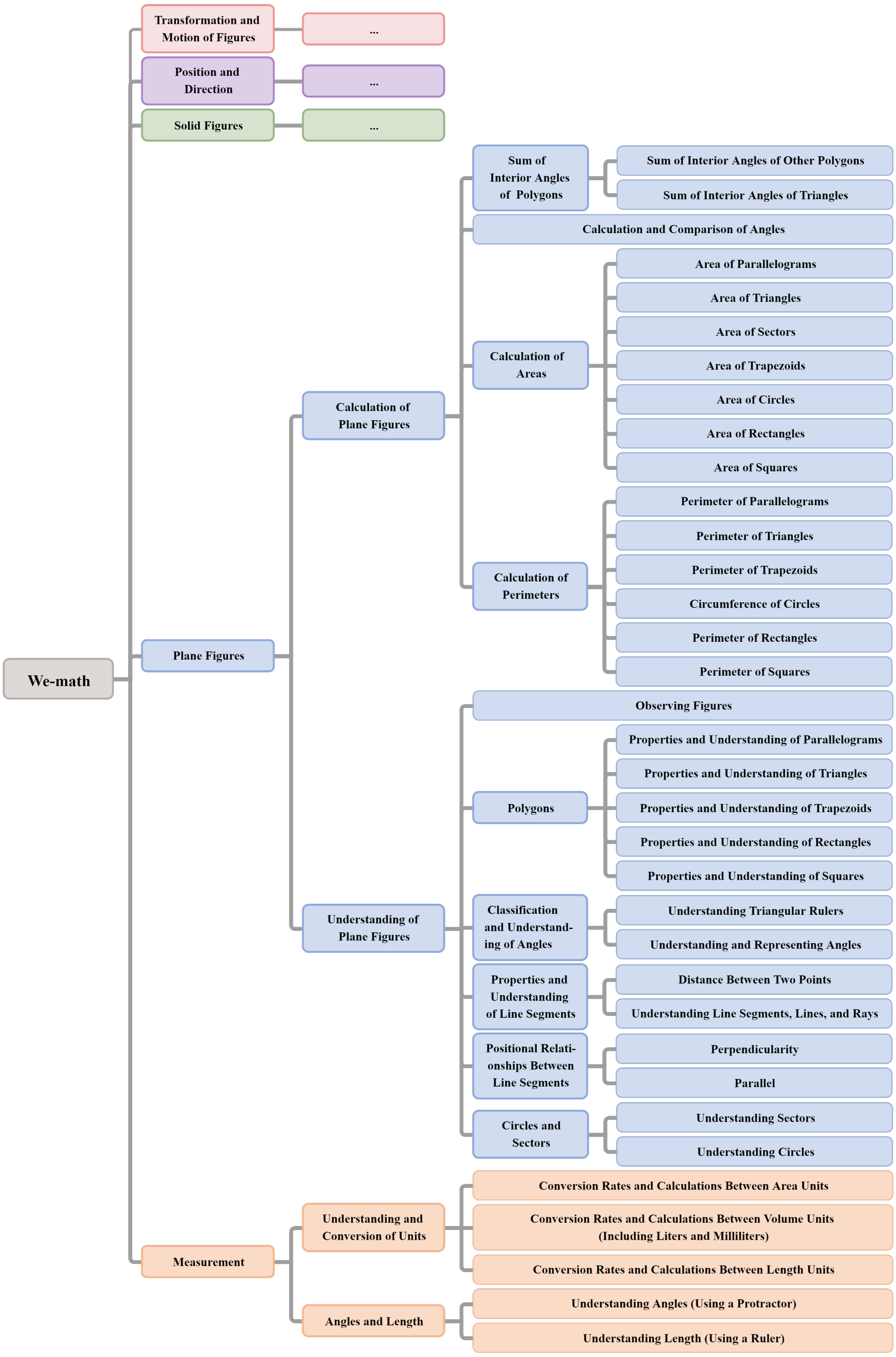


Figure 9: The Hierarchical Knowledge Structure of WE-MATH (2).

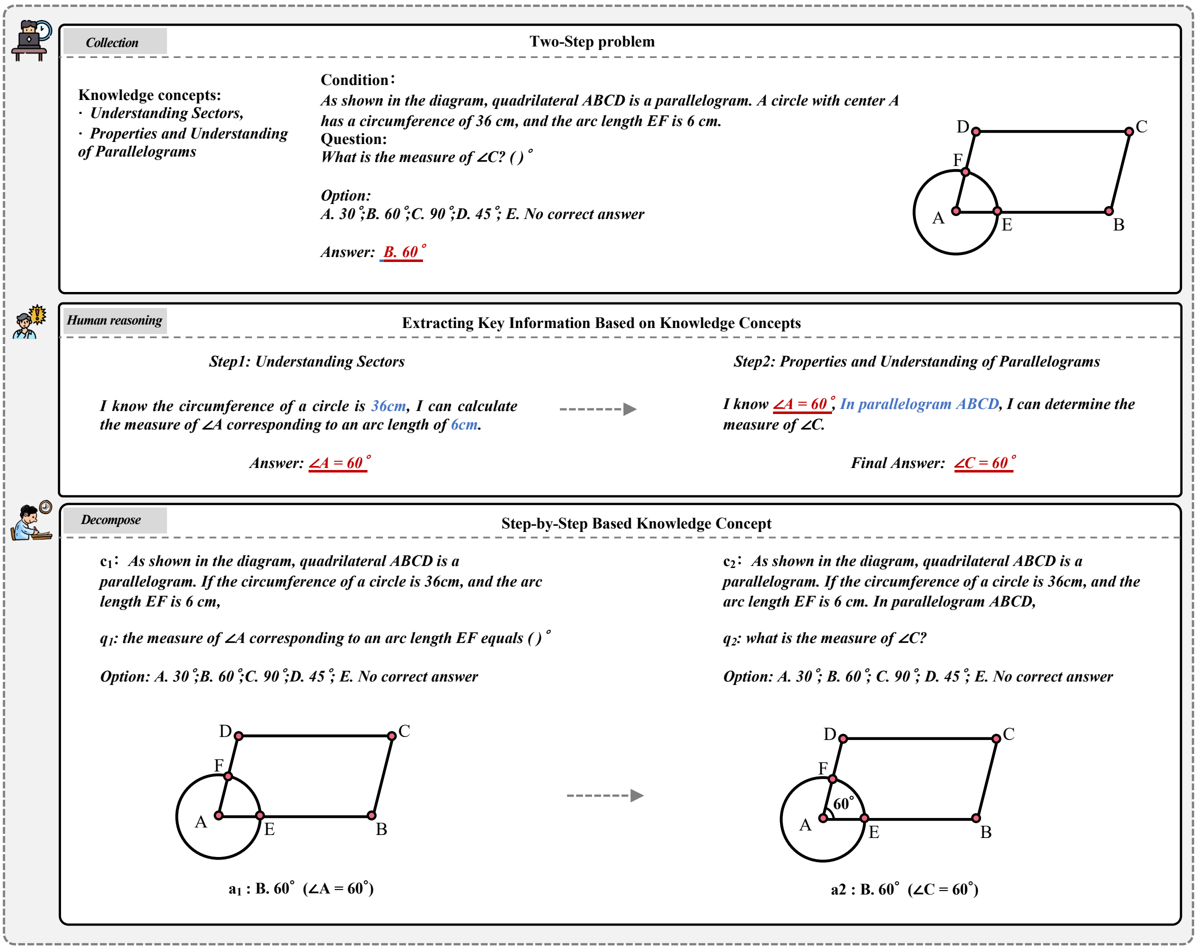


Figure 10: An example of a two-step problem in WE-MATH.

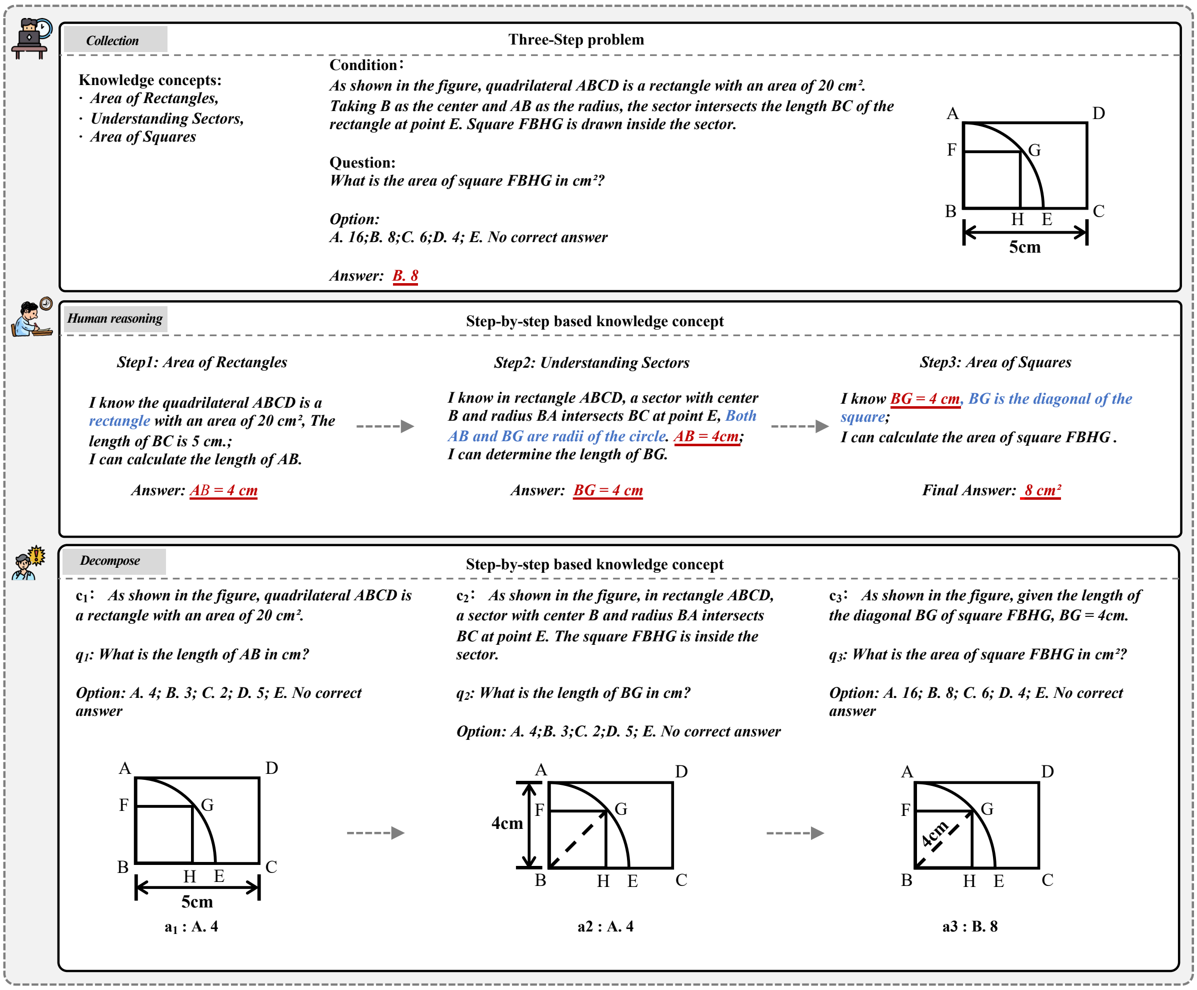


Figure 11: An example of a three-step problem in WE-MATH.

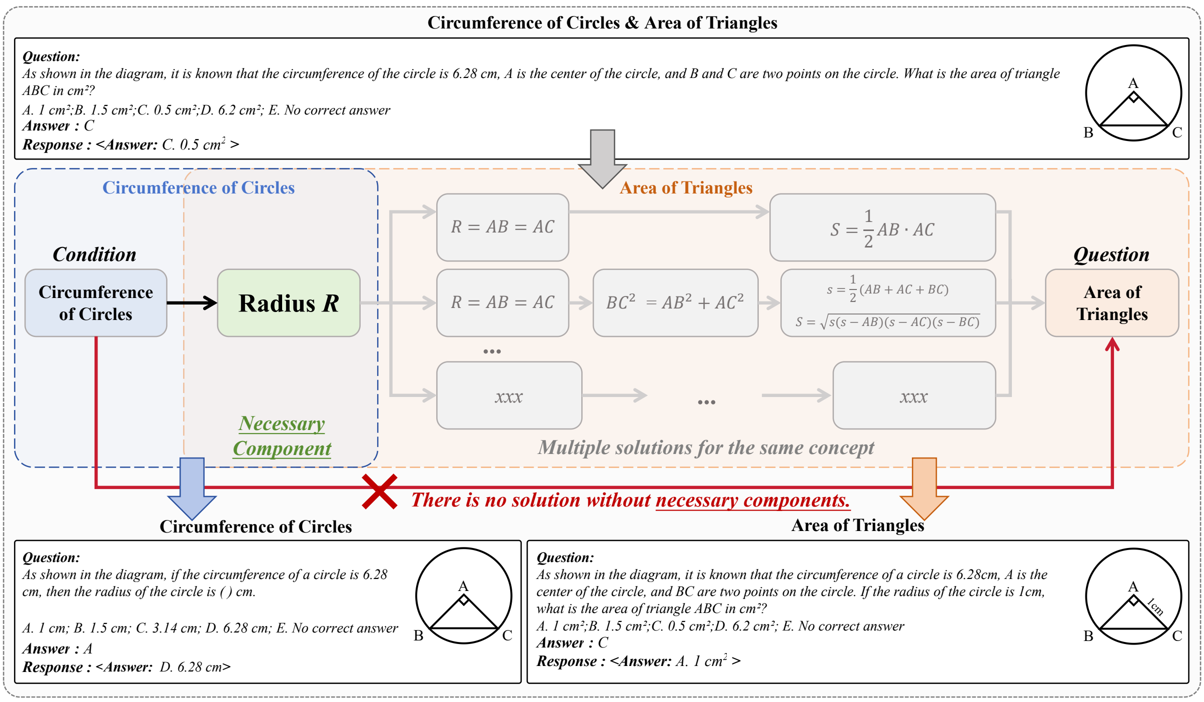


Figure 12: The flowchart for filtering and decomposing decomposable questions.

B.4 Details of Data Collection

With the hierarchical knowledge structure, we select problems with images from publicly authoritative mathematics websites from various countries, including professional exams and practice tests. To ensure comprehensive coverage of fundamental and critical areas in primary math, we select the five most foundational and prevalent domains within the field of primary geometry, including:

- **Plane figures**: Questions involving identification and properties of two-dimensional shapes.
- **Solid figures**: Questions related to the recognition and characteristics of three-dimensional objects.
- **Transformation and motion of figures**: Problems focusing on geometric transformations such as translation, rotation, and reflection.
- **Position and direction**: Questions that involve understanding spatial relationships and directions.
- **Measurement**: Problems requiring the measurement of length, area, volume, and angles.

The selection criteria are as follows: (1) The problems include multiple knowledge concepts and can be decomposed into steps for solution. (2) The problems and images are consistent. (3) The correct answer is unique.

B.5 Details of Data Filtering

During the data filtering stage, to ensure the rigorousness of WE-MATH 's process evaluation and to prevent scenarios where a model solves multi-step problems through alternative methods yet fails to answer one-step problems, we retain only problems where all solution paths pass through a unique intermediate result. This ensures the accuracy of the four-dimensional metric evaluation. To further enhance evaluation efficiency, we sample from these problems to construct the testmini subset.

Specifically, for a multi-step problem, our expert team analyzes its solving process and identifies all possible intermediate conditions. A problem is preliminarily retained if all solution paths require a specific intermediate condition; otherwise, it is discarded.

For instance, as illustrated in Figure 12, in a problem requiring the calculation of a triangle's area, all solution methods necessitate determining

Table 5: Key statistics of WE-MATH.

Statistic	Number
Total samples	**10,898**
Total test set samples	**6,298**
Total testmini set samples	**1,740**
Test set	
- Three-step problems	598
- Two-step problems	1,302
- One-step problems	4,398
Hierachial knowledge structure	
- First-layer nodes	5
- Second-layer nodes	12
- Terminal nodes	67
Question options	
- Total options	44,418
- Average options	4.076
- Proportion of answer A	10,898 (24.5%)
- Proportion of answer B	10,898 (24.5%)
- Proportion of answer C	10,871 (24.5%)
- Proportion of answer D	8,304 (18.7%)
- Proportion of answer E	3,424 (7.7%)
- Proportion of answer F&G	27 (0.06%)
Question length	
- Maximum length (word)	143
- Maximum length (character)	852
- Average length (word)	26.8
- Average length (character)	139.1

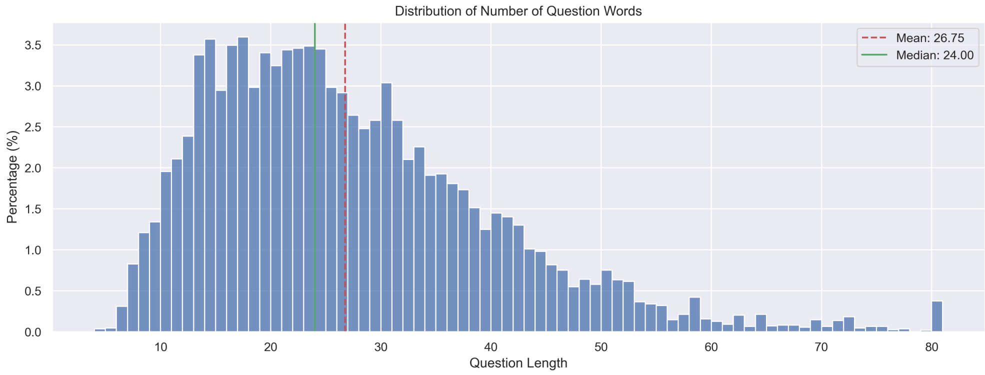


Figure 13: The distribution of the number of words per question in WE-MATH. Questions with a length greater than 80 are categorized as 81 for visualization simplicity.

the radius R as a critical intermediate condition. Thus, solving R becomes a required step. Based on this, we decompose the problem into one-step sub-problems, specifically addressing R. This step is classified under the <*Circumference of Circles*> concept (blue), while the subsequent calculation of the triangle's area corresponds to the <*Area of Triangles*> concept (orange). By ensuring that all required conditions for solving the area include R, we guarantee the necessity of this intermediate step and the relevance of the decomposed sub-question.

This rigorous decomposition and filtering process not only ensures the accuracy of WE-MATH 's evaluation on multi-step problems but also maintains the scientific integrity of the selected evaluation questions. It establishes a robust foundation for reliable assessments using the four-dimensional metric.

B.6 Details of Data Statistics

Full Data Statistics WE-MATH contains a total of 10,898 samples, including both newly collected problems and their decomposed sub-problems. For the 6,500 newly collected samples, these consist of 2,564 one-step problems, 1,900 problems verified to have unique solution paths based on distinct intermediate conditions, and the remaining samples evaluated by experts as not meeting the rigorous decomposition criteria (not applicable to four-dimensional metrics).

As shown in Table 5, focusing on problems that support both knowledge mastery testing and the four-dimensional metric evaluation, the test set includes 1,302 two-step problems and 598 three-step problems, which can be further decomposed into 4,398 one-step problems, resulting in a total of 6,298 test samples.

On this basis, we construct the testmini subset, ensuring that the distribution of two-step and three-step problems aligns with the original collection. The subsequent section E present comprehensive evaluation results for both the full testset subset and the testmini subset.

Question distribution The WE-MATH consists entirely of English questions, as illustrated in Table 5, the average number of words in the English questions within WE-MATH is 26.8, with the maximum number of words in a question reaching 143. Figure 13 further elaborates on the distribution of word counts, highlighting the diverse patterns of the questions.

Advantages of Multiple-Choice Questions In WE-MATH, all problems are presented as multiple-choice questions. Even if some problems did not originally conform to the multiple-choice format during the initial selection, our researchers manually converted them into the format. Using multiple-choice questions offers several advantages:

Standardization: Ensures uniformity across all questions, facilitating consistent assessment and comparison across different hierarchical subjects.

Objective Grading: The use of single correct answers eliminates subjectivity in grading, enhancing the reliability of the evaluation.

Efficiency: Allows for rapid and scalable assessment, suitable for large datasets and automated systems.

Focused Assessment: Carefully designed distractors help in accurately identifying specific knowledge gaps and common misconceptions.

Furthermore, as shown in the experimental results in Table 11, the strict score under the random setting is only 1.1%, further demonstrating that the multiple-choice format does not cause instability into the experiments.

C More Details on the Metrics

Distinguishing Metric. Considering the model's instability, Figure 3, 14 and Figure 15 illustrate the two metrics we propose for distinguishing between *RM* and *CM* metrics. Figure 3 represents the two-step problem, while Figures 14 and Figures 15 represent the three-step problem. Specifically, under the strict metric, if there is any error in the corresponding sub-problems of a multi-step problem that is answered correctly, it is classified as *RM (Rote Memorization)*. Only if all corresponding sub-problems are answered correctly (TTTT, TTT) is it classified as *CM (Complete Master)*. Under the loose metric, it is classified as *RM* only if the model answers all sub-problems incorrectly (FFFT, FFT), otherwise, it is classified as *CM*. Therefore, the $\text{Score}_{\text{average}}$ under the loose metric is slightly higher. We hope to see models like GPT-4o (OpenAI, 2024), InternVL2.5-78B (Chen et al., 2024), Qwen2-VL-72B (Wang et al., 2024c), LLaVA-OneVision-72B (Li et al., 2024) and GPT-4V (OpenAI, 2023), which have already performed nearly perfectly under the loose metric and are far ahead of other models, bring us even greater surprises under the strict metric in the next update.

Metrics' Intrinsic Logic. As shown in Figure 3, 14, 15, it is evident in the Metric for Reasoning Evaluation Section that *IK*, *IG*, and *CM* have a logical relationship. In the early stages of constructing WE-MATH, we recorded all the model's responses and analyzed the answers to each multi-step problem and its corresponding sub-problems. We believe that for both humans and models, a reasonable learning process should involve first mastering each knowledge concept individually and then learning to comprehensively apply them to achieve complete mastery. The situation where the multi-step problem is answered correctly but the sub-problems are answered incorrectly (*RM*) is an unreasonable phenomenon. Therefore, we developed a four-dimensional fine-grained metric to further evaluate the model's performance. Based on this, the reasoning scoring process is formulated as

$$\text{Score}_{\text{average}} = \alpha S_{\text{IK}} + \beta S_{\text{IG}} + S_{\text{CM}} \tag{4}$$

To ensure the reasoning capability hierarchy is "*IK* < *IG* < *CM*", we control the parameters such that $\alpha < \beta < 1$. Considering the lack of knowledge, which fails in one-step problems, as the most critical fundamental flaw of the model, we set the default value of α to 0 to underscore the importance of foundational understanding. Since IG (TTF or TTTF) indicates that the model has already grasped the basic knowledge concepts required to solve the problem, it is given a weight of 0.5. Clearly, CM indicates that the model has fully overcome the above two points, so it is assigned a weight of 1. According to the fine-grained scoring, we aim for the model to master knowledge and truly possess generalization ability, rather than merely memorizing questions.

D More Details on Experiment Setup

D.1 Details of the Evaluated Models

To evaluate the mathematical reasoning abilities of various large language models (LMMs), we selected a total of 30 models. These include 4 proprietary models (GPT-4o, GPT-4V, Gemini 1.5 Pro, Qwen-VL-Max) and 26 Open-source models. Table 7 presents their release dates and specific sources. The selected open-source models encompass both smaller models (1B, 2B) and larger models (110B, 78B, 72B) to ensure the comprehensiveness of the experiment.

Additionally, the release dates of the chosen models span nearly a year, including the recently released (InternVL2.5 series) in December, as well as some earlier models. Through our experiments, we observed that the LMM's community is rapidly evolving. Initially, GPT-4o had a significant leading advantage, but now there are models that are gradually approaching the performance of GPT-4o. The detailed analysis and results are presented in the experimental section E.

D.2 Details of the Model Hyperparameters

For all closed-sourced models with API access, we adopt the generation scheme shown in Table 6 and simply run the inference with CPUs, which typically completes within a day. For all open-source models, we utilize a cluster with 8 NVIDIA A800-SXM4-80GB GPUs to run the inference, and we follow the hyper-parameter settings specified in the model source's inference samples. If no specific instructions are provided, we use the default settings. Table 8 details the specific generation parameters.

Table 6: Generating parameters for Closed-Source LMMs.

Model	Generation Setup
GPT-4o	"model" : "gpt-4o", "temperature" : 0, "max_tokens" : 1024
GPT-4V	"model" : "gpt-4-turbo", "temperature" : 0, "max_tokens" : 1024
Gemini 1.5 Pro	"model" : "gemini-1.5-pro-latest", "temperature" : 0, "max_tokens" : 1024
Qwen-VL-Max	"model" : "qwen-vl-max", "temperature" : 0, "max_tokens" : 1024

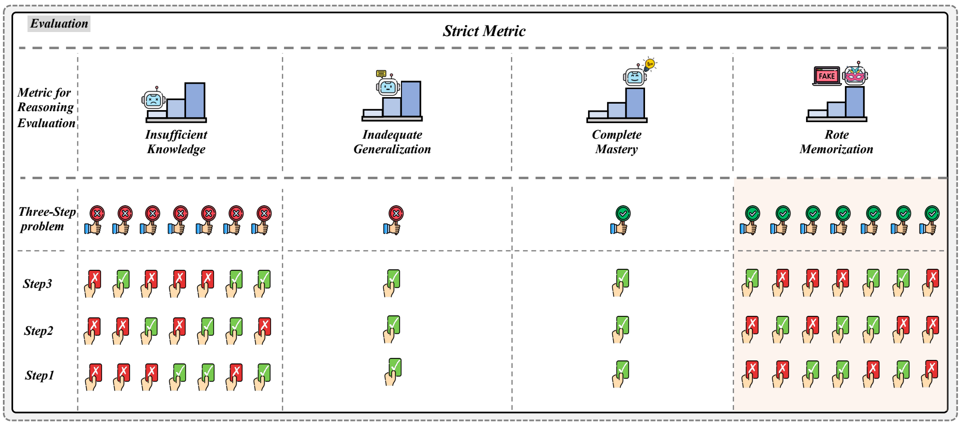


Figure 14: Diagram illustrating strict metric in three-step problem.

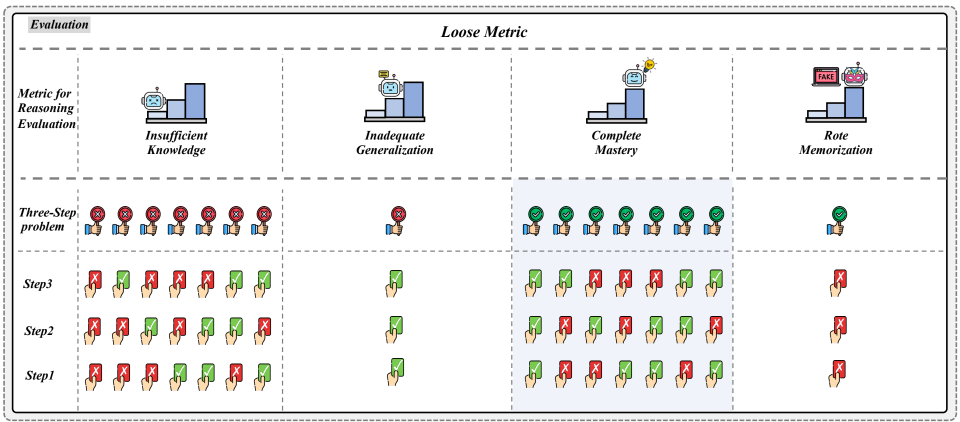


Figure 15: Diagram illustrating loose metric in three-step problem.

Table 7: The release time and model source of LMMs used in WE-MATH.

Model	Release Time	Source
GPT-4o (OpenAI, 2024)	2024-05	https://gpt4o.ai/
GPT-4V (OpenAI, 2023)	2024-04	https://openai.com/index/gpt-4v-system-card/
Gemini 1.5 Pro (Team et al., 2023)	2024-05	https://deepmind.google/technologies/gemini/pro/
Qwen-VL-Max (Bai et al., 2023)	2024-01	https://huggingface.co/spaces/Qwen/Qwen-VL-Max/
Qwen2-VL-72B (Wang et al., 2024c)	2024-09	https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct
Qwen2-VL-7B (Wang et al., 2024c)	2024-09	https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
LLaVA-OneVision-72B (Li et al., 2024)	2024-08	https://huggingface.co/lmms-lab/llava-onevision-qwen2-72b-ov-chat
LLaVA-OneVision-7B (Li et al., 2024)	2024-08	https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov
InternVL2.5-78B (Chen et al., 2024)	2024-12	https://huggingface.co/OpenGVLab/InternVL2_5-78B
InternVL2-Llama3-76B (Chen et al., 2024)	2024-07	https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B
InternVL2.5-26B (Chen et al., 2024)	2024-12	https://huggingface.co/OpenGVLab/InternVL2_5-26B
InternVL2.5-8B (Chen et al., 2024)	2024-12	https://huggingface.co/OpenGVLab/InternVL2_5-8B
InternVL2-8B (Chen et al., 2024)	2024-07	https://huggingface.co/OpenGVLab/InternVL2-8B
InternVL2.5-4B (Chen et al., 2024)	2024-12	https://huggingface.co/OpenGVLab/InternVL2_5-4B
InternVL2-4B (Chen et al., 2024)	2024-07	https://huggingface.co/OpenGVLab/InternVL2-4B
InternVL2-2B (Chen et al., 2024)	2024-07	https://huggingface.co/OpenGVLab/InternVL2-2B
LLaVA-NeXT-110B (Liu et al., 2024a)	2024-05	https://huggingface.co/lmms-lab/llava-next-110b/
LLaVA-NeXT-72B (Liu et al., 2024a)	2024-05	https://huggingface.co/lmms-lab/llava-next-72b/
LLaVA-1.6-13B (Liu et al., 2023)	2024-03	https://huggingface.co/llava-hf/llava-v1.6-vicuna-13b-hf/
LLaVA-1.6-7B (Liu et al., 2023)	2024-03	https://huggingface.co/llava-hf/llava-v1.6-vicuna-7b-hf/
DeepSeek-VL-1.3B (Lu et al., 2024)	2024-03	https://huggingface.co/deepseek-ai/deepseek-vl-1.3b-chat/
DeepSeek-VL-7B (Lu et al., 2024)	2024-03	https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat/
Phi3-Vision-4.2B (Abdin et al., 2024)	2024-05	https://huggingface.co/microsoft/Phi-3-vision-128k-instruct/
MiniCPM-Llama3-V2.5 (Hu et al., 2023)	2024-05	https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5/
InternLM-XComposer2-VL-7B (Dong et al., 2024)	2024-04	https://huggingface.co/internlm/internlm-xcomposer2-vl-7b/
InternVL-Chat-V1.5 (Chen et al., 2023)	2024-04	https://huggingface.co/OpenGVLab/InternVL-Chat-V1-5/
GLM-4V-9B (GLM et al., 2024)	2024-06	https://huggingface.co/THUDM/glm-4v-9b
LongVA (Zhang et al., 2024b)	2024-06	https://huggingface.co/lmms-lab/LongVA-7B
G-LLaVA-13B (Gao et al., 2023a)	2024-03	https://huggingface.co/renjiepi/G-LLaVA-13B/
InternLM-XComposer2.5-7B (Zhang et al., 2024a)	2024-07	https://huggingface.co/internlm/internlm-xcomposer2d5-7b

Table 8: Generating parameters for Open-Source LMMs.

Model	Generation Setup
InternVL2-Llama3-76B	do_sample = False, temperature = 0, max_new_tokens = 1024
InternVL2.5-78B	do_sample = False, max_new_tokens = 1024
Qwen2-VL-72B	do_sample = False, max_new_tokens = 1024
LLaVA-OneVision-72B	do_sample = True, max_length = 1024, top_k = 1
InternVL2.5-26B	do_sample = False, max_new_tokens = 1024
InternVL2-8B	do_sample = False, temperature = 0, max_new_tokens = 1024
InternVL2.5-8B	do_sample = False, max_new_tokens = 1024
Qwen2-VL-7B	do_sample = False, temperature = 0, max_new_tokens = 1024
LLaVA-OneVision-7B	do_sample = True, max_length = 1024, top_k = 1
InternVL2-4B	do_sample = False, temperature = 0, max_new_tokens = 1024
InternVL2.5-4B	do_sample = False, max_new_tokens = 1024
InternVL2-2B	do_sample = False, temperature = 0, max_new_tokens = 1024
LLaVA-NeXT-110B	do_sample = False, temperature = 0, max_new_tokens = 1024
LLaVA-NeXT-72B	do_sample = False, temperature = 0, max_new_tokens = 1024
InternVL-Chat-V1.5	num_beams = 1, do_sample = False, max_new_tokens = 1024
LLaVA-1.6-13B	do_sample = False, temperature = 0, max_new_tokens = 1024
LLaVA-1.6-7B	do_sample = False, temperature = 0, max_new_tokens = 1024
DeepSeek-VL-1.3B	do_sample = False, max_new_tokens = 1024
DeepSeek-VL-7B	do_sample = False, max_new_tokens = 1024
Phi3-Vision-4.2B	do_sample = False, temperature = 0, max_new_tokens = 1024
MiniCPM-Llama3-V2.5	sampling = True, temperature = 0.7
InternLM-XComposer2-VL-7B	do_sample = False
InternLM-XComposer2.5-7B	do_sample = False
GLM-4V-9B	do_sample = True, max_length = 1024, top_k = 1
LongVA-7B	do_sample = False, temperature = 0, max_new_tokens = 1024, num_beams = 1
G-LLaVA-13B	do_sample = True, temperature = 0.2, max_new_tokens = 1024

Table 9: Model architecture of 25 LMMs evaluated on We-Math.

Models	LLM	Vision Encoder
GPT-4o	-	-
GPT-4V	-	-
Gemini 1.5 Pro	-	-
Qwen-VL-Max	-	-
InternVL2-Llama3-76B	Hermes-2-Theta-Llama-3-70B	InternViT-6B-448px-V1-5
InternVL2.5-78B	Qwen2.5-72B-Instruct	InternViT-6B-448px-V2_5
InternVL2.5-26B	internlm2_5-20b-chat	InternViT-6B-448px-V2_5
InternVL2.5-8B	internlm2_5-7b-chat	InternViT-300M-448px-V2_5
InternVL2-8B	InternLM2_5-7b-chat	InternViT-300M-448p
InternVL2.5-4B	Qwen2.5-3B-Instruct	InternViT-300M-448px-V2_5
InternVL2-4B	Phi-3-mini-128k-instruct	InternViT-300M-448px
InternVL2-2B	InternLM2-chat-1_8b	InternViT-300M-448px
Qwen2-VL-72B	Qwen2-72B	CLIP ViT-bigG-P14
Qwen2-VL-7B	Qwen2-7B	CLIP ViT-bigG-P14
LLaVA-OneVision-7B	Qwen2-7B	SigLip-so400m-P14-384
LLaVA-OneVision-72B	Qwen2-72B	SigLip-so400m-P14-384
LLaVA-NeXT-110B	Qwen1.5-110B-Chat	CLIP-ViT-L-P14-336
LLaVA-NeXT-72B	Qwen1.5-72B-Chat	CLIP-ViT-L-P14-336
LLaVA-1.6-13B	Vicuna-13B-v1.5	CLIP-ViT-L-P14-336
LLaVA-1.6-7B	Vicuna-7B-v1-5	CLIP-ViT-L-P14-336
DeepSeek-VL-1.3B	DeepSeek-LLM-1.3B-base	SigLIp-L-P16-384
DeepSeek-VL-7B	DeepSeek-LLM-7B-base	SigLIp-L-P16-384 & SAM-B
Phi3-Vision-4.2B	Phi-3-mini-128K-instruct	CLIP-ViT-L-P14-336
MiniCPM-Llama3-V2.5	Llama3-8B-Instruct	SigLIp-L-P14-384
InternLM-XComposer2-VL-7B	InternLM2-7B-ChatSFT	CLIP-ViT-L-P14-336
InternLM-XComposer2.5-7B	InternLM2-7B-ChatSFT	CLIP-ViT-L-P14-336
InternVL-Chat-V1.5	InternLM2-Chat-20B	InternViT-6B-448px-V1-5
GLM-4V-9B	GLM-9B	EVA_02_CLIP-E-P14
LongVA-7B	Qwen2-7B-Instruct	CLIP-ViT-L-P14-336
G-LLaVA-13B	Vicuna-13B-v1.5	CLIP-ViT-L-P14-336

E More Details on Experiment Results

E.1 Details of Model Performance

The Leaderboard on WE-MATH.

In Figure 16, we present the visualized results of $\text{Score}_{\text{average}}$ under both Strict(a) and Loose(b) metrics. GPT-4o remains in the leading position under both metrics. Among the open-source models, InternVL2.5-78B performs the best under the strict and loose metric.

Notably, recently released model series such as InternVL2.5, Qwen2-VL, and LLaVA-OneVision show strong performance, narrowing the gap with GPT-4o. This contrasts with models released before July, where InternVL2.5-78B is the best-performing open-source model but still has a noticeable gap compared to GPT-4o.

From the leaderboard, it is evident that within the same series of models (InternVL2.5, Qwen2-VL, LLaVA-OneVision, LLaVA-NeXT, LLaVA), there is a clear trend that larger parameter models tend to perform better. However, smaller models within the same series (e.g., Qwen2-VL-7B, InternVL2.5-8B, LLaVA-OneVision-7B) also exhibit impressive performance, even surpassing the closed-source model Qwen-VL-Max, maintaining a leading position among models of similar size. This indicates that optimizing training methods might partially substitute for the performance gains typically achieved by merely increasing the parameter count.

Detailed Performance of Four-Dimensional Metrics. Figure 17, 18, 19 and Figure 20 display the specific performance of LMMs under both loose and strict metric across four metrics. To provide a clearer comparison of model performance across different time periods, we separate the visualization of models released after July from those released before July, with closed-source models included in each figure for reference. Specifically, Figure 17 and 19 showcase open-source models released after July, while Figure 18 and 20 display open-source models released before July. It is evident that in terms of the *IK*, *IG*, *RM*, and *CM* metrics, recent models gradually close the gap with GPT-4o and even surpass closed-source models like GPT-4V, Qwen-VL-Max, and Gemini 1.5 Pro. We speculate that this improvement is related to the enhanced reasoning capabilities of recent models, which benefit from strengthened post-training phases.

Focusing on the *IK* metric, GPT-4o has the fewest instances under both metric, indicating that GPT-4o has the best grasp of the knowledge concepts. Furthermore, for the *IG* metric, we find that InternVL2.5-26B, LLaVA-OneVision-72B, and Qwen2-VL-7B have the highest scores compared to other models. As discussed in the previous Section C, *IG* issues only arise after addressing *IK* issues, which further indicates that GPT-4o and some larger LMMs are progressing to the next stage. Focusing on the *CM* and *RM* metrics, among models with a $\text{Score}_{\text{average}}$ under the strict metric above 30%, GPT-4o continues to show significant leadership. It excels in the *CM* metric, where the number of correctly answered multi-step problems and their corresponding sub-questions is significantly higher than that of other models.

Detailed Performance on Each Category. In Figure 4, we present the performance of open-source and closed-source models under the second-level nodes. In Figure 21 to Figure 50, we detail the specific performance of 25 models across 67 knowledge concepts (based on statistics from one-step problem questions). It is evident that GPT-4o consistently leads in overall performance, but its main issue lies in measurement-related tasks. Notably, some open-source models perform worse on the simpler "Understanding and Conversion of Units" knowledge concepts compared to "Angles and Length" related concepts, while InternVL-Chat-V1.5 and MiniCPM-Llama3-V2.5 exhibit more logically consistent results.

As shown in Figure 4, we present the performance of open-source and closed-source models under the second-level nodes. From Figure 21 to Figure 50, we detail the specific performance of 30 models across 67 knowledge concepts (based on statistics from one-step problem questions). It is evident that GPT-4o consistently leads in a majority of knowledge concepts, but its main issue lies in measurement-related tasks. Notably, some early open-source models perform poorly on the simpler "Understanding and Conversion of Units" knowledge concepts, while recent models have shown significant improvement in this area. Moreover, many models still struggle with "Angles and Length" related concepts, indicating a significant need for further advancements in LMMs within this knowledge concept.

Results on the Test Set. To demonstrate that the testmini set effectively reflects the full test set, we follow the approach used by Mathvista (Lu et al., 2023) and Mathverse (Zhang et al., 2024c) in presenting test set results. We select various sizes of open-source models and two closed-source models (GPT-4o and Qwen-VL-Max). Table 12 and Table 13 report the results under the second-level nodes and the four-dimensional metrics, respectively. The differences between these results and those from the testmini set are minimal, especially for larger models. This indicates that the testmini subset effectively mirrors the test set, serving as a valuable evaluation subset for model development, particularly for those with limited computing resources. In the era of large models, this approach is a common and efficient method.

E.2 Specific Error Analysis

Error Types. To delve into the failure cases of models, we detailed four typical error types in Table 14. Furthermore, to facilitate a better understanding of each error type, we provide examples of each error made by GPT-4o from Figure 51 to Figure 54. Since a single thought process in a problem can involve multiple errors and a single logical error is enough to derail a much larger solution, we consider the first error that occurs in the reasoning steps as the key error and include only this error in our statistics.

Correspondence of Errors in Multi-Step and One-Step Problems. Focusing on Insufficient Knowledge, the errors in multi-step problems often correspond to those in one-step problems. This supports our approach of decomposing problems to accurately associate error types with specific knowledge concepts. Furthermore, we observe a positive correlation between the quantity of knowledge concepts and the errors in the reasoning process. As the complexity of knowledge concepts increases, the difficulty for the model to perform multi-step reasoning also increases, leading to a higher likelihood of visual recognition errors and incorrect application of knowledge concepts.

F Example Demonstration of the Knowledge Concepts

Figure 55 to 64 illustrate the the detailed information of knowledge concepts.

Table 10: **Accuracy scores of LMMs on the *testmini* subset of WE-MATH.** The first 3 columns report the overall performance on one-step, two-step, three-step problems, while the other columns display the result on one-step problems in different problem categories. The highest accuracy for closed-source and open-source LMMs is marked in blue and green respectively. (S1: one-step, S2: two-step, S3: three-step, Mem: Measurement, PF: Plane Figures, SF: Solid Figures, TMF: Transformation and Motion of Figures, PD: Position and Direction. AL: Angles and Length, UCU: Understanding and Conversion of Units, CPF: Calculation of Plane Figures, UPF: Understanding of Plane Figures, CSF: Calculation of Solid Figures, USF: Understanding of Solid Figures, BTF: Basic Transformations of Figures, CCF: Cutting and Combining of Figures, Dir: Direction, Pos: Position, RoM: Route Map, CCP: Correspondence of Coordinates and Positions).

Model	**S1**	**S2**	**S3**	**Mem**		**PF**		**SF**		**TMF**		**PD**			
				UCU	AL	CPF	UPF	CSF	USF	BTF	CCF	Dir	Pos	RoM	CCP
Random	16.5	15.0	21.8	16.9	23.9	17.8	15.1	13.3	17.1	28.1	15.0	24.3	13.6	7.1	16.7
Closed-source															
GPT-4o	72.8	58.1	43.6	86.6	39.1	77.4	71.6	84.5	62.3	58.7	69.4	93.1	72.7	47.5	73.3
GPT-4V	65.5	49.2	38.2	82.5	38.4	70.7	60.2	76.6	56.3	57.8	67.7	79.3	57.5	47.8	63.3
Gemini 1.5 Pro	56.1	51.4	33.9	51.0	31.2	61.8	45.0	70.0	57.5	39.2	62.7	68.8	54.1	40.7	60.0
Qwen-VL-Max	40.8	30.3	20.6	19.4	25.3	39.8	41.4	43.6	48.0	43.8	43.4	41.4	35.1	40.7	26.7
Open-source															
InternVL2.5-78B	68.8	59.7	41.8	87.6	26.5	75.1	60.9	75.9	59.9	61.5	72.6	86.0	66.8	70.3	70.0
InternVL2-Llama3-76B	67.9	53.3	43.6	71.7	39.8	71.3	61.7	73.8	61.5	68.8	63.9	89.5	76.6	62.6	73.3
Qwen2-VL-72B	68.2	53.1	50.9	92.4	45.1	70.2	63.8	72.9	58.5	61.3	71.0	75.5	72.7	66.8	70.0
InternVL2.5-26B	67.5	55.0	40.6	82.2	29.1	73.1	63.8	74.0	57.3	61.9	68.7	89.5	61.7	55.2	66.7
LLaVA-OneVision-72B	64.0	45.8	35.8	73.8	35.8	69.6	62.2	72.8	57.4	46.3	65.1	61.7	65.9	40.9	56.7
InternVL2.5-8B	60.7	45.6	32.7	72.4	24.6	62.4	59.3	68.8	55.6	50.8	58.2	71.9	61.4	59.3	56.7
InternVL2-8B	59.4	43.6	35.2	71.4	20.5	62.0	55.5	67.1	57.3	54.0	60.5	58.6	63.6	44.5	50.0
Qwen2-VL-7B	59.1	43.6	26.7	62.7	37.2	62.6	60.8	65.7	49.2	52.5	49.2	48.1	68.2	55.0	56.7
InternVL2.5-4B	58.3	42.8	30.3	68.8	30.5	60.9	55.6	71.3	52.7	45.5	48.7	61.7	65.3	51.4	60.0
LLaVA-OneVision-7B	57.5	43.1	39.4	59.0	36.5	66.7	55.4	64.4	61.1	48.6	46.9	55.0	49.5	25.6	43.3
InternVL2-4B	50.5	32.5	24.8	44.0	30.5	55.7	47.7	58.0	56.9	38.6	41.4	34.5	53.4	52.2	46.7
InternVL2-2B	38.9	31.1	22.4	28.7	35.8	44.2	38.2	37.5	34.5	36.5	35.4	48.3	51.7	29.4	23.3
LLaVA-NeXT-110B	53.7	36.9	31.5	39.5	57.7	59.5	53.1	52.3	50.2	54.1	50.8	54.8	55.9	40.1	40.0
LLaVA-NeXT-72B	42.9	35.6	30.9	31.6	25.3	43.3	42.4	46.1	41.8	44.2	51.0	44.3	38.9	33.0	36.7
InternVL-Chat-V1.5	49.4	30.6	28.5	44.0	29.8	52.2	52.1	44.2	48.2	47.1	46.8	65.7	50.5	36.5	36.7
LLaVA-1.6-13B	29.4	25.3	32.7	21.7	23.2	23.4	34.7	25.3	26.4	37.5	41.7	26.9	28.9	37.1	30.0
GLM-4V-9B	47.3	37.2	38.2	53.4	37.0	51.3	46.5	50.6	38.2	44.1	45.2	41.0	49.3	36.8	53.3
MiniCPM-Llama3-V2.5	39.8	31.1	29.7	28.6	37.0	40.8	39.8	41.0	38.6	32.0	42.7	41.0	42.7	44.0	43.3
LongVA-7B	43.5	30.6	28.5	24.5	39.8	45.1	40.8	51.9	42.5	45.6	44.6	44.5	40.7	47.5	20.0
InternLM-XComposer2.5-7B	49.0	32.2	23.0	21.7	33.2	54.3	52.1	47.0	45.2	53.7	40.5	51.7	61.1	41.2	33.3
InternLM-XComposer2-VL-7B	47.0	33.1	33.3	31.3	46.5	47.7	42.6	51.4	43.9	41.1	50.6	65.5	53.9	55.2	40.0
LLaVA-1.6-7B	23.0	20.8	15.8	18.5	20.5	16.9	29.6	15.6	18.6	42.7	24.1	17.6	43.3	28.9	26.7
DeepSeek-VL-7B	32.6	26.7	25.5	16.6	35.1	27.3	38.0	24.2	38.6	50.0	30.1	24.5	41.0	51.7	23.3
G-LLaVA-13B	32.4	30.6	32.7	33.3	29.1	32.0	37.9	19.6	33.5	37.1	32.8	31.2	33.2	25.6	40.0
Phi3-Vision-4.2B	42.1	34.2	27.9	28.7	16.0	47.2	38.8	50.0	44.4	28.8	31.2	48.6	49.2	26.4	50.0
DeepSeek-VL-1.3B	31.4	27.8	23.0	27.8	23.9	22.8	36.9	30.4	34.2	44.5	28.3	48.1	41.8	37.1	33.3

Table 11: The performance of different LMMs on four-dimensional metrics for reasoning evaluation. The best performance for closed-source and open-source LMMs is marked in blue and green (Avg: $\text{Score}_{\text{average}}$).

Model	Strict					Loose				
	Avg (↑)	**IK (↓)**	**IG (↓)**	**CM (↑)**	**RM (↓)**	**Avg (↑)**	**IK (↓)**	**IG (↓)**	**CM (↑)**	**RM (↓)**
Random	1.1	81.7	1.1	0.6	96.7	6.9	81.7	1.1	6.3	63.3
Closed-source										
GPT-4o	42.9	31.2	15.2	35.2	34.2	60.6	31.2	15.2	52.3	1.1
GPT-4V	31.1	39.8	14.5	23.8	47.9	51.4	39.8	14.5	44.2	3.3
Gemini-1.5-Pro	26.4	42.9	11.2	20.8	54.8	46.0	42.9	11.2	40.4	12.0
Qwen-VL-Max	10.5	65.1	7.6	6.7	75.5	25.5	65.1	7.6	21.7	20.3
Open-source										
InternVL2.5-78B	38.5	34.1	11.8	32.6	39.8	57.5	34.1	11.8	51.6	4.6
InternVL2-Llama3-76B	36.9	33.9	15.8	29.0	42.4	56.3	33.9	15.8	48.4	3.8
Qwen2-VL-72B	36.6	33.5	14.1	29.5	43.6	56.8	33.5	14.1	49.7	5.1
InternVL2.5-26B	34.6	33.3	16.2	26.5	47.6	56.1	33.3	16.2	48.0	4.9
InternVL2.5-8B	29.1	43.6	14.9	21.7	47.7	46.1	43.6	14.9	38.7	6.9
LLaVA-OneVision-72B	28.7	41.1	16.2	20.6	51.8	49.1	41.1	16.2	41.0	4.0
Qwen2-VL-7B	28.7	41.1	16.2	20.6	51.8	49.1	41.1	16.2	41.0	4.0
LLaVA-OneVision-7B	23.1	45.0	13.1	16.6	60.5	44.9	45.0	13.1	38.3	8.6
InternVL2-8B	26.6	45.5	13.5	19.8	51.6	44.9	45.5	13.5	38.1	7.0
InternVL2.5-4B	25.0	48.2	13.0	18.5	52.5	42.1	48.2	13.0	35.6	8.3
InternVL2-4B	17.4	59.8	10.1	12.4	58.9	31.5	59.8	10.1	26.5	12.0
InternVL2-2B	9.1	64.0	7.6	5.3	81.2	25.1	64.0	7.6	21.3	24.8
LLaVA-NeXT-110B	19.2	50.3	14.5	12.0	66.0	37.9	50.3	14.5	30.7	13.0
LLaVA-NeXT-72B	13.4	58.9	7.1	9.9	71.0	31.5	58.9	7.1	28.0	17.9
InternVL-Chat-V1.5	15.0	56.2	13.9	8.0	73.3	32.7	56.2	13.9	25.7	14.0
LLaVA-1.6-13B	5.2	69.1	3.2	3.6	86.9	22.0	69.1	3.2	20.4	26.2
GLM-4V-9B	14.9	53.0	9.5	10.1	73.1	35.1	53.0	9.5	30.3	19.3
MiniCPM-Llama3-V2.5	9.6	60.2	9.1	5.0	83.9	28.1	60.2	9.1	23.4	23.6
LongVA-7B	11.5	61.1	9.0	7.1	76.4	27.7	61.1	9.0	23.2	22.3
InternLM-XComposer2.5-7B	15.6	57.0	13.7	8.8	70.1	31.2	57.0	13.7	24.4	16.9
InternLM-XComposer2-VL-7B	12.7	56.4	10.5	7.4	77.6	31.0	56.4	10.5	25.7	22.4
G-LLaVA-13B	6.5	64.2	4.6	4.2	86.6	22.3	64.2	4.6	20.0	36.0
LLaVA-1.6-7B	3.3	78.3	2.5	2.1	89.1	13.8	78.3	2.5	12.6	34.7
DeepSeek-VL-7B	6.3	69.1	4.6	4.0	84.8	21.0	69.1	4.6	18.7	29.0
Phi3-Vision-4.2B	10.6	58.9	9.0	6.1	81.1	29.8	58.9	9.0	25.3	21.3
DeepSeek-VL-1.3B	5.9	71.1	2.7	4.6	82.6	21.5	71.1	2.7	20.2	23.2

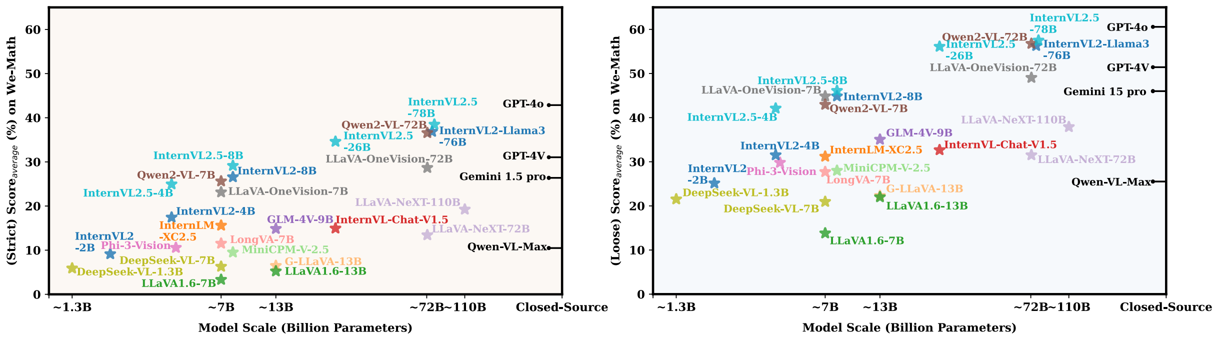


Figure 16: The Leaderboard of different LMMs under the strict and loose metric (average score %). "∼" represents an approximate estimate of the total parameters nums in LMMs.

Table 12: **Accuracy scores of LMMs on the *test set* of WE-MATH.**

Model	S1	S2	S3	Mem		PF		SF		TMF		PD			
				UCU	AL	CPF	UPF	CSF	USF	BTF	CCF	Dir	Pos	RoM	CCP
Closed-source															
GPT-4o	73.0	57.8	44.8	86.1	46.5	77.8	68.5	80.3	66.0	57.3	72.5	96.1	80.8	62.6	72.2
Qwen-VL-Max	40.8	30.9	21.9	19.5	23.9	39.8	42.4	43.4	47.5	43.4	39.4	43.8	40.6	40.9	21.9
Open-source															
Qwen2-VL-72B	67.8	52.8	48.5	95.9	48.1	69.1	63.7	74.3	58.6	59.3	68.7	84.5	67.4	66.9	68.9
InternVL2-8B	59.6	43.9	35.0	75.0	15.8	62.6	56.9	66.4	57.9	58.0	57.8	61.6	59.4	38.5	54.7
LLaVA-OneVision-7B	57.9	42.6	40.1	61.8	38.3	66.1	57.4	64.6	61.8	54.5	46.5	52.8	50.7	25.7	37.8
LLaVA-1.6-13B	29.2	23.9	32.6	25.0	22.1	22.0	35.5	27.9	24.3	34.5	42.6	26.8	23.6	42.5	33.7
Phi3-Vision-4.2B	42.0	35.2	32.9	24.3	15.8	46.8	38.4	51.5	42.8	26.4	31.3	55.7	54.7	27.5	38.7
DeepSeek-VL-1.3B	31.4	30.3	20.1	30.9	22.9	21.6	38.4	28.4	34.7	44.2	27.5	43.9	42.3	42.0	31.2

Table 13: The performance of different LMMs on four-dimensional metrics for reasoning evaluation across the entire *test set*.

Model	Strict					Loose				
	Avg (↑)	**IK (↓)**	**IG (↓)**	**CM (↑)**	**RM (↓)**	**Avg (↑)**	**IK (↓)**	**IG (↓)**	**CM (↑)**	**RM (↓)**
Closed-source										
GPT-4o	43.4	33.5	12.8	37.0	31.1	59.8	33.5	12.8	53.4	0.5
Qwen-VL-Max	10.9	64.8	7.1	7.3	73.9	26.1	64.8	7.1	22.5	19.7
Open-source										
Qwen2-VL-72B	36.1	34.5	14.1	29.1	43.6	56.2	34.5	14.1	49.2	4.5
LLaVA-OneVision-7B	23.1	44.8	13.4	16.4	60.7	45.0	44.8	13.4	38.3	8.4
InternVL2-8B	26.6	45.9	13.1	20.1	51.0	44.4	45.9	13.1	37.9	7.7
Phi3-Vision-4.2B	11.1	56.8	8.7	6.7	80.5	31.7	56.8	8.7	27.3	20.8
LLaVA-1.6-13B	4.8	69.9	3.5	3.1	88.3	21.4	69.9	3.5	19.7	26.1
DeepSeek-VL-1.3B	6.4	69.5	3.5	4.6	82.9	22.0	69.5	3.5	20.3	25.1

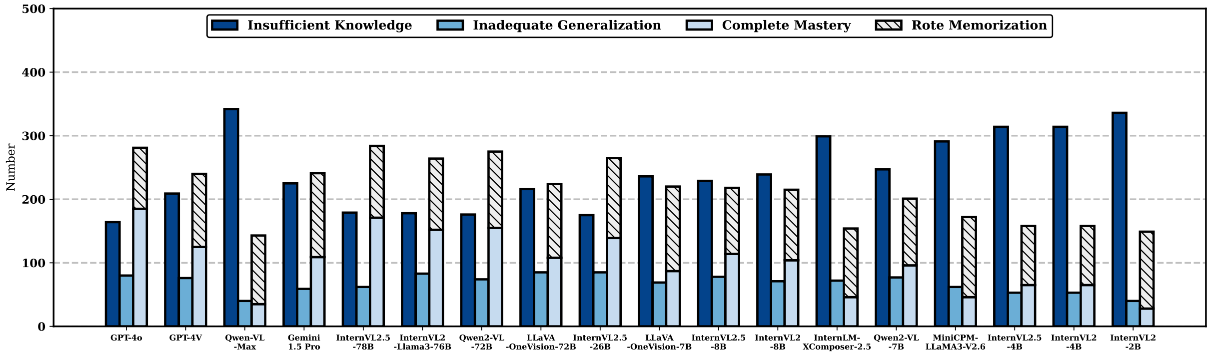


Figure 17: The performance of different LMMs (including closed-source models and closed-source models after July) on four-dimensional metrics under strict metric.

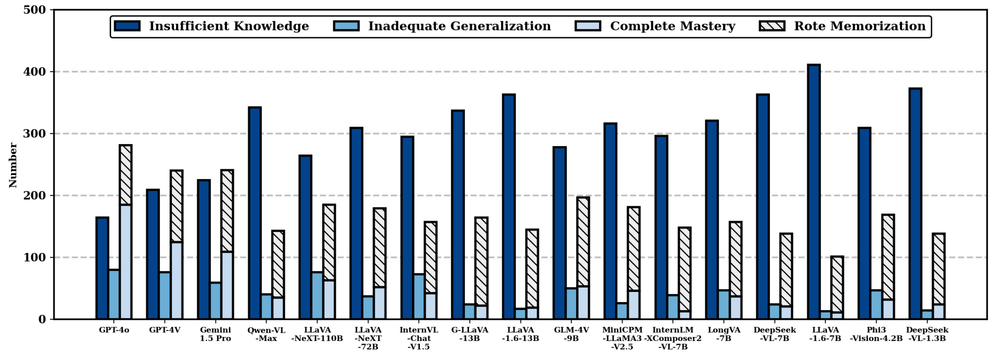


Figure 18: The performance of different LMMs (including closed-source models and closed-source models before July) on four-dimensional metrics under strict metric.

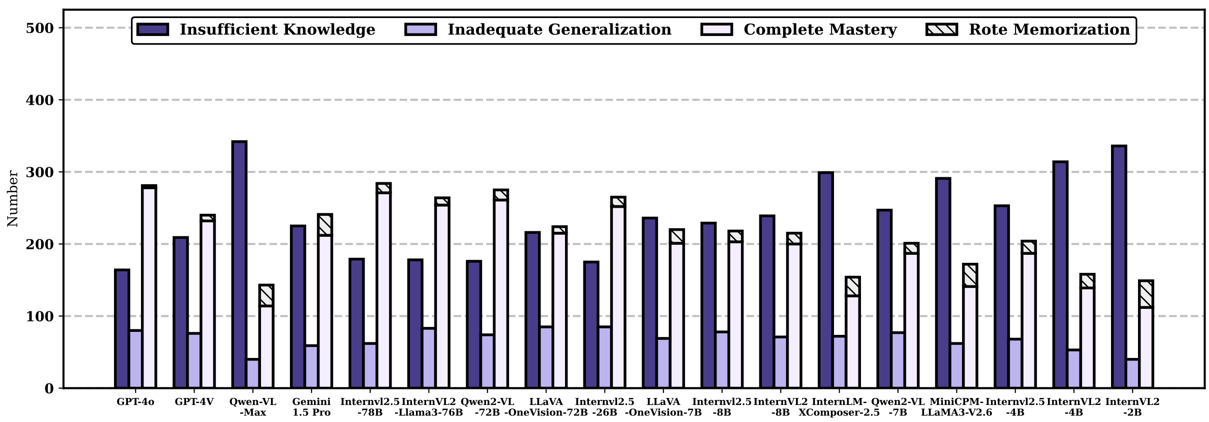


Figure 19: The performance of different LMMs (including closed-source models and closed-source models after July) on four-dimensional metrics under loose metric.

Table 14: Detailed Descriptions of Error Types.

Error Type	Explanation
Knowledge Error	For a specific knowledge concept, the model is unclear or confused about it, or it misuses another knowledge concept to solve the problem.
Reason Error	Errors that occur in the logical reasoning process while using knowledge concepts to solve the problem step by step.
Visual Error	Errors in visual perception, where the model incorrectly identifies shapes or numbers in an image.
Hallucination	The thought process introduces factors that are not consistent with the facts, which are not mentioned in the context of the image or question.

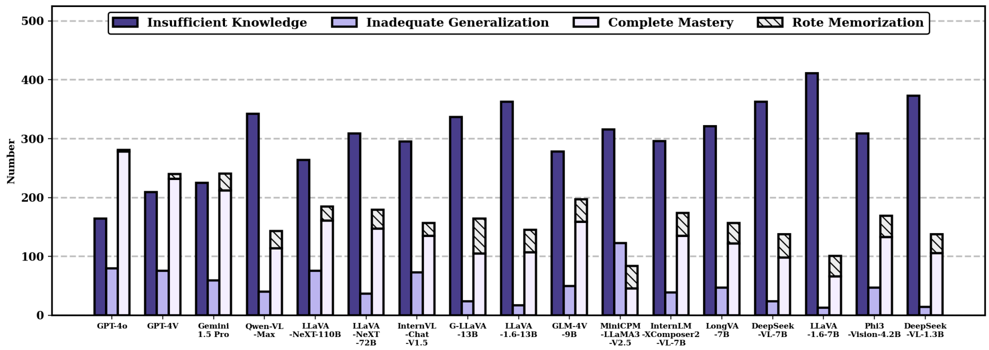


Figure 20: The performance of different LMMs (including closed-source models and closed-source models before July) on four-dimensional metrics under loose metric.

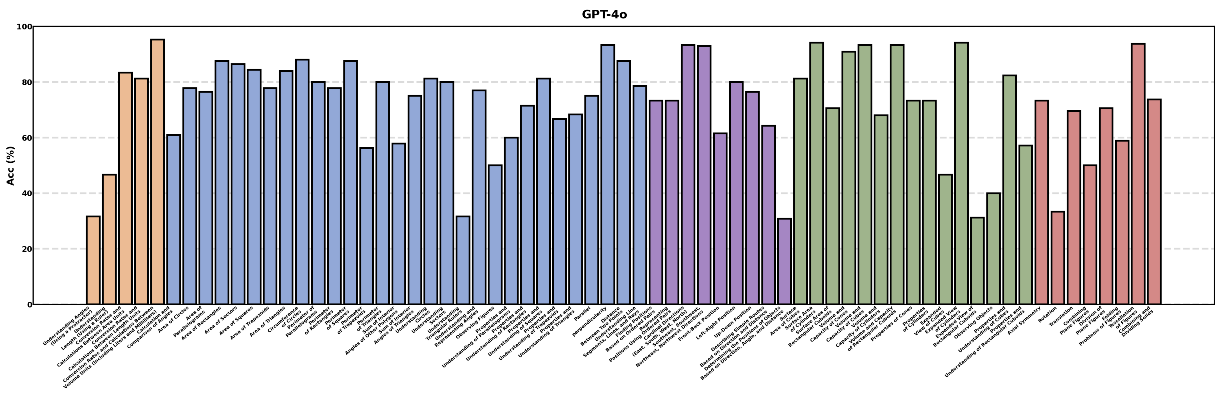


Figure 21: Detailed performance of GPT-4o across 67 knowledge concepts.

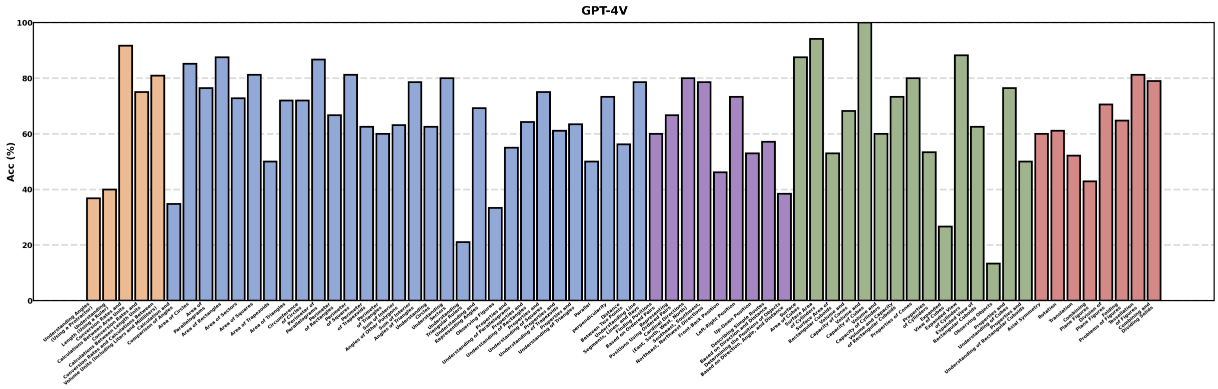


Figure 22: Detailed performance of GPT-4V across 67 knowledge concepts.

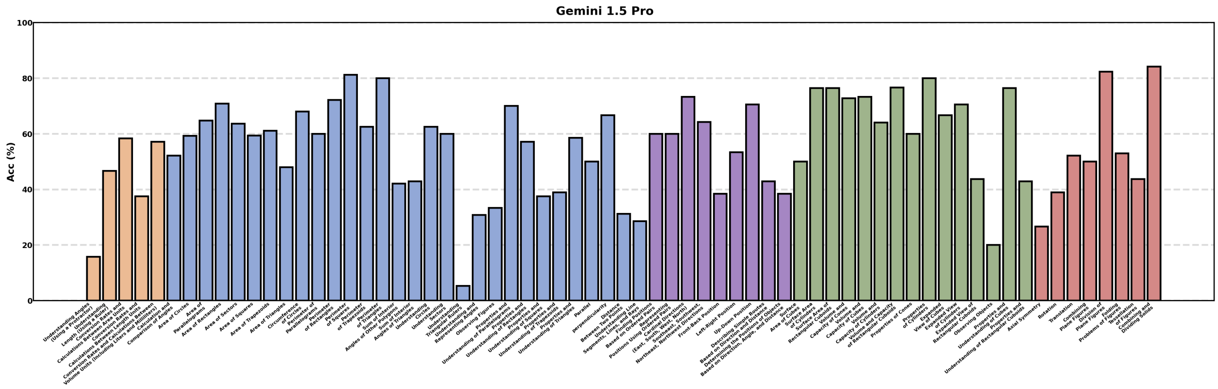


Figure 23: Detailed performance of Gemini 1.5 Pro across 67 knowledge concepts.

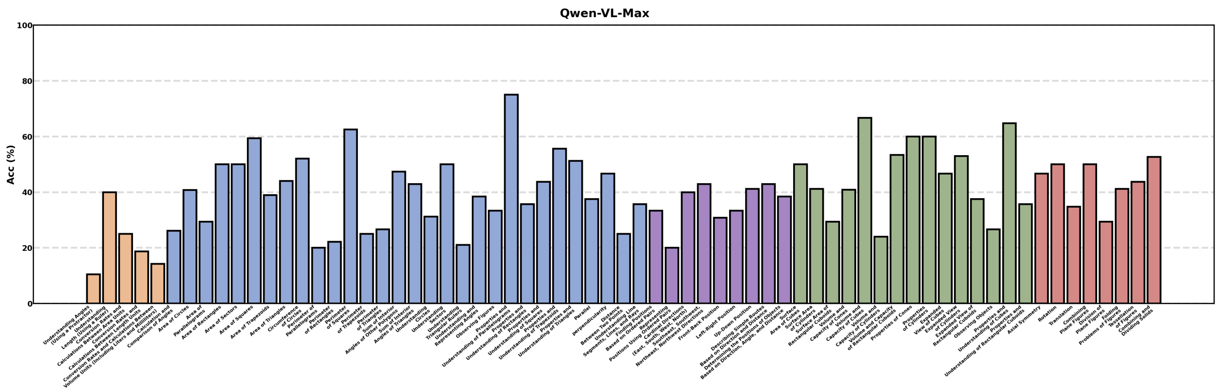


Figure 24: Detailed performance of Qwen-VL-Max across 67 knowledge concepts.

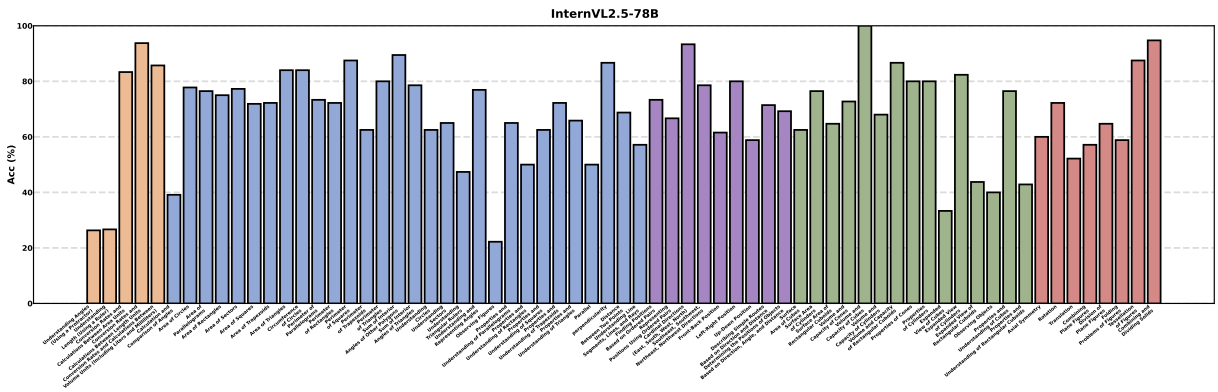


Figure 25: Detailed performance of InternVL2.5-78B across 67 knowledge concepts.

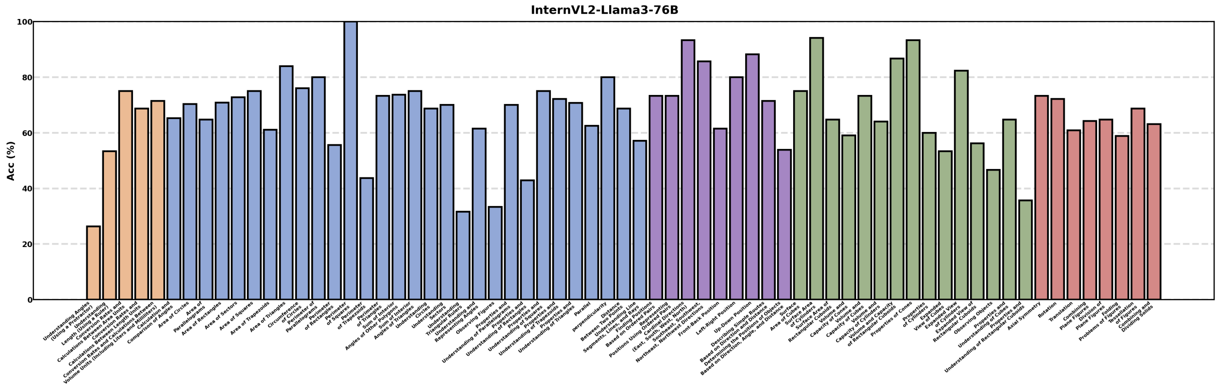


Figure 26: Detailed performance of InternVL2-Llama3-76B across 67 knowledge concepts.

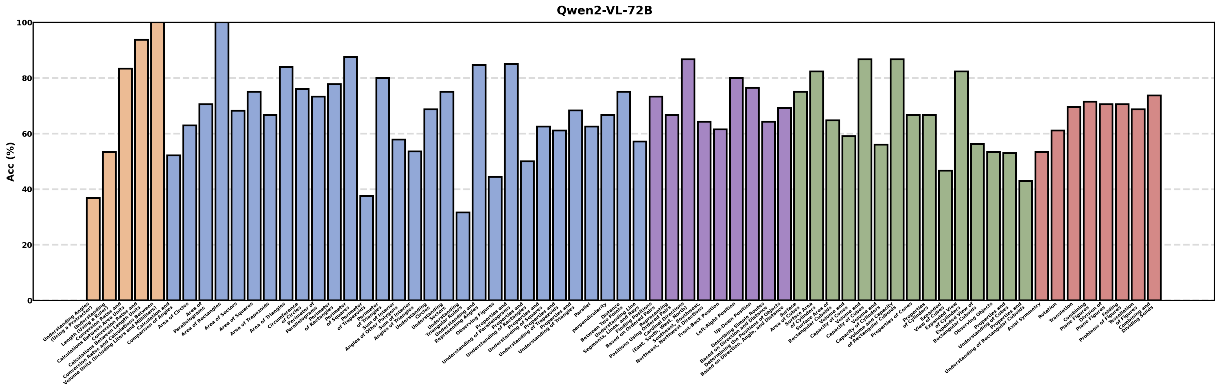


Figure 27: Detailed performance of Qwen2-VL-72B across 67 knowledge concepts.

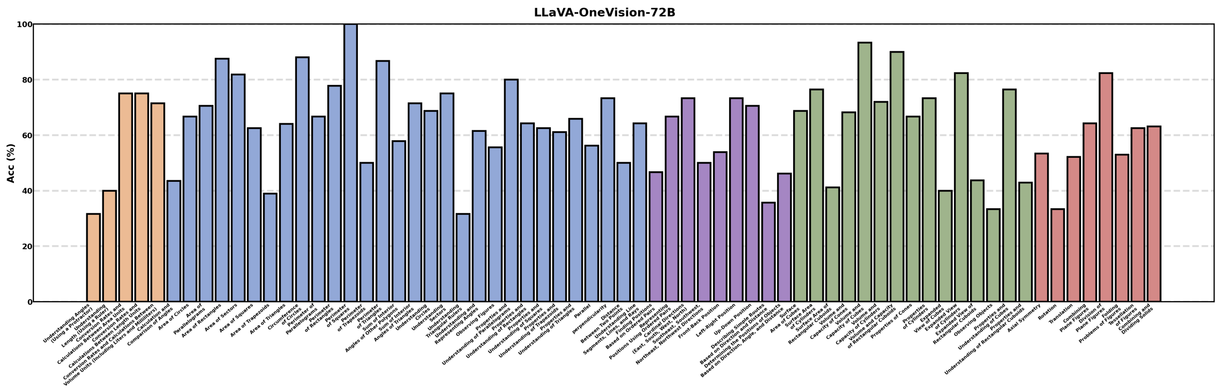


Figure 28: Detailed performance of LLaVA-OneVision-72B across 67 knowledge concepts.

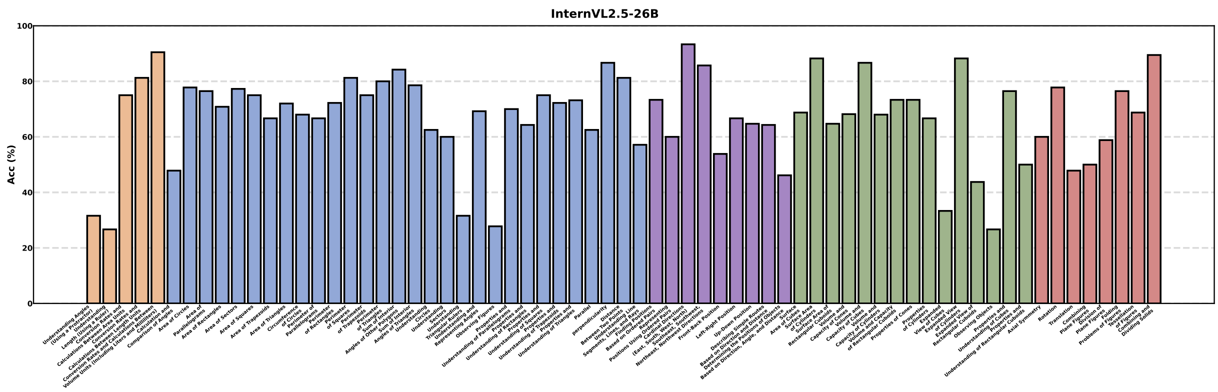


Figure 29: Detailed performance of InternVL2.5-26B across 67 knowledge concepts.

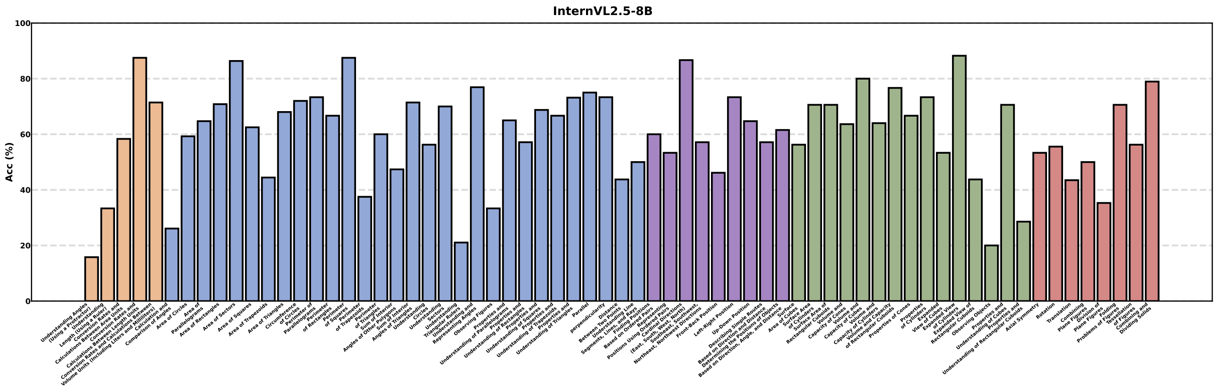


Figure 30: Detailed performance of InternVL2.5-8B across 67 knowledge concepts.

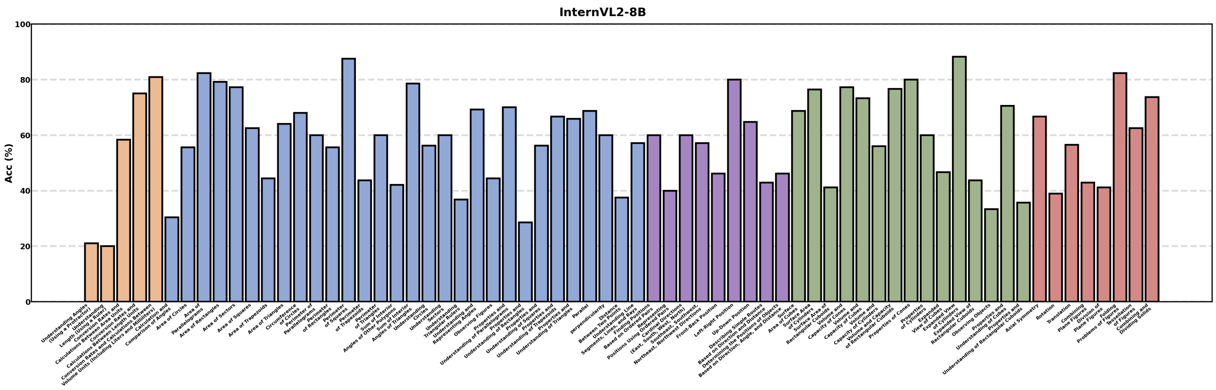


Figure 31: Detailed performance of InternVL2-8B across 67 knowledge concepts.

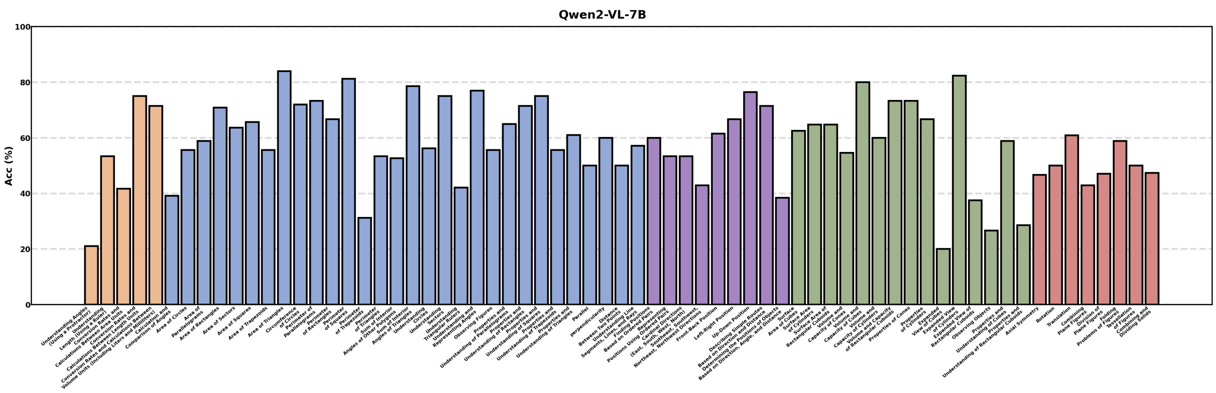


Figure 32: Detailed performance of Qwen2-VL-7B across 67 knowledge concepts.

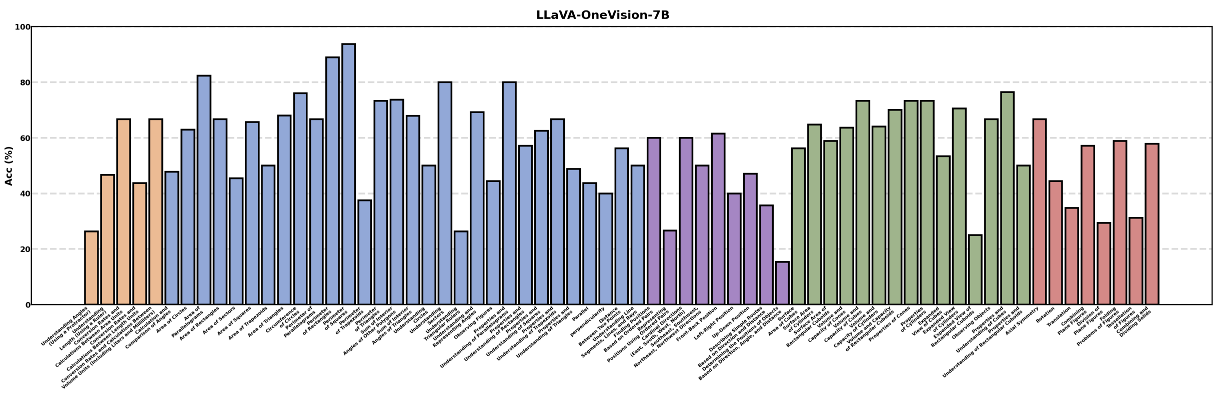


Figure 33: Detailed performance of LLaVA-OneVision-7B across 67 knowledge concepts.

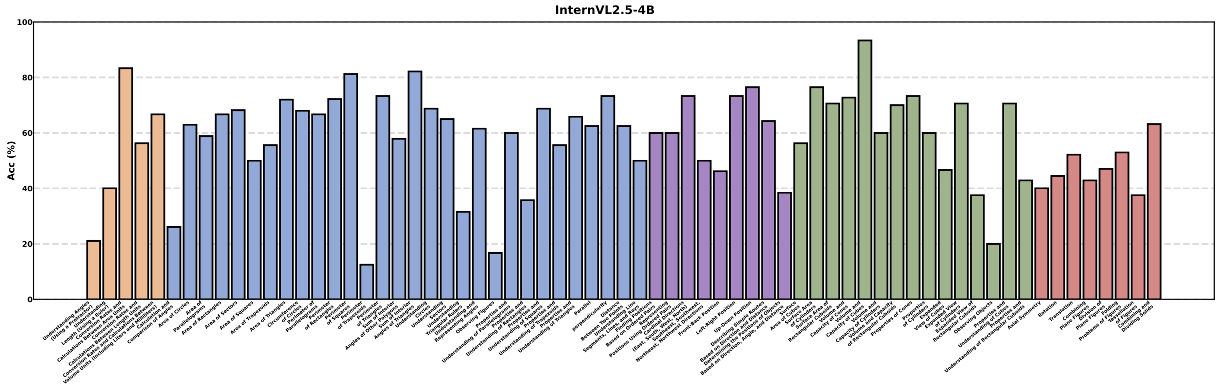


Figure 34: Detailed performance of InternVL2.5-4B across 67 knowledge concepts.

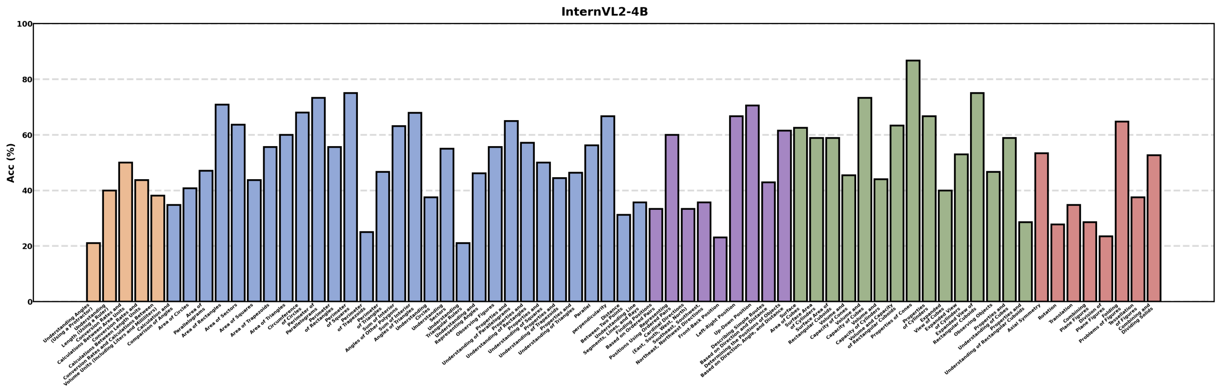


Figure 35: Detailed performance of InternVL2-4B across 67 knowledge concepts.

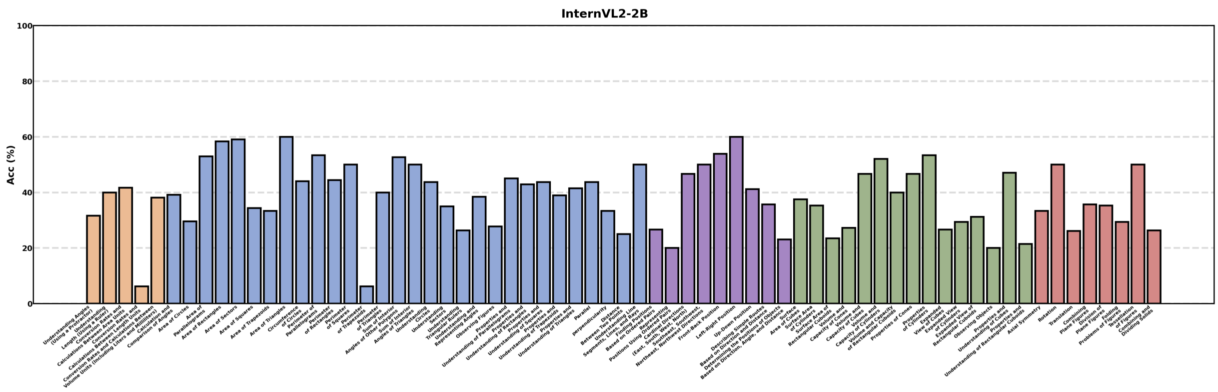


Figure 36: Detailed performance of InternVL2-2B across 67 knowledge concepts.

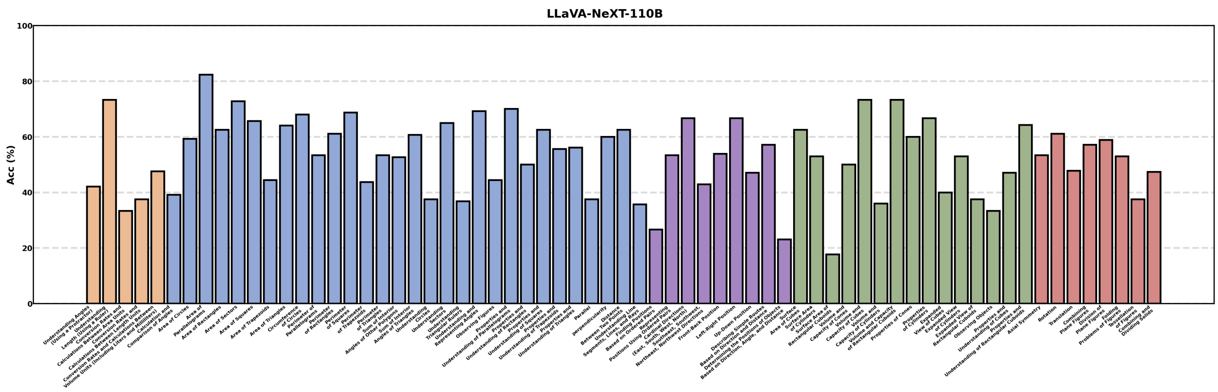


Figure 37: Detailed performance of LLaVA-NeXT-110B across 67 knowledge concepts.

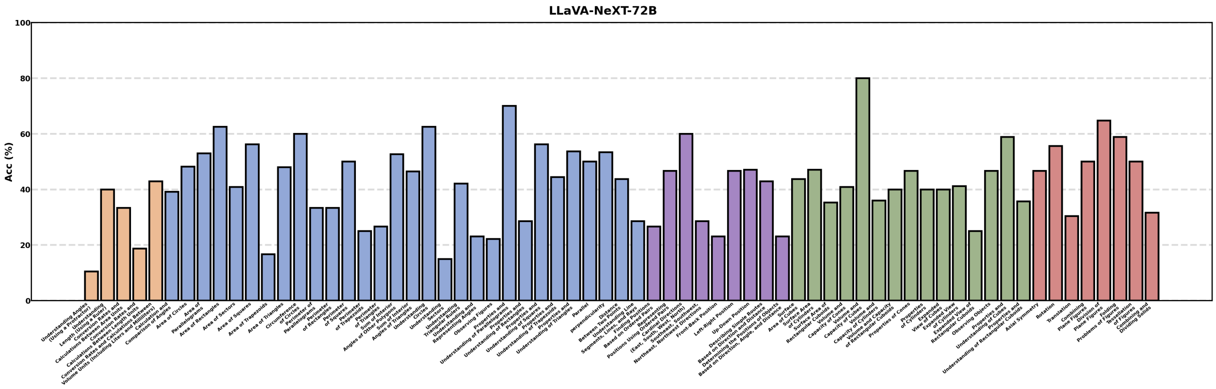


Figure 38: Detailed performance of LLaVA-NeXT-72B across 67 knowledge concepts.

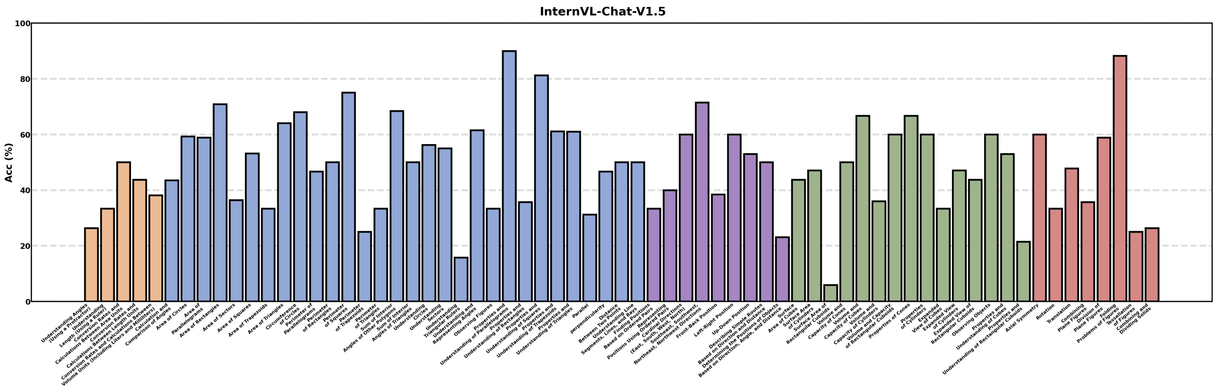


Figure 39: Detailed performance of InternVL-Chat-V 1.5 across 67 knowledge concepts.

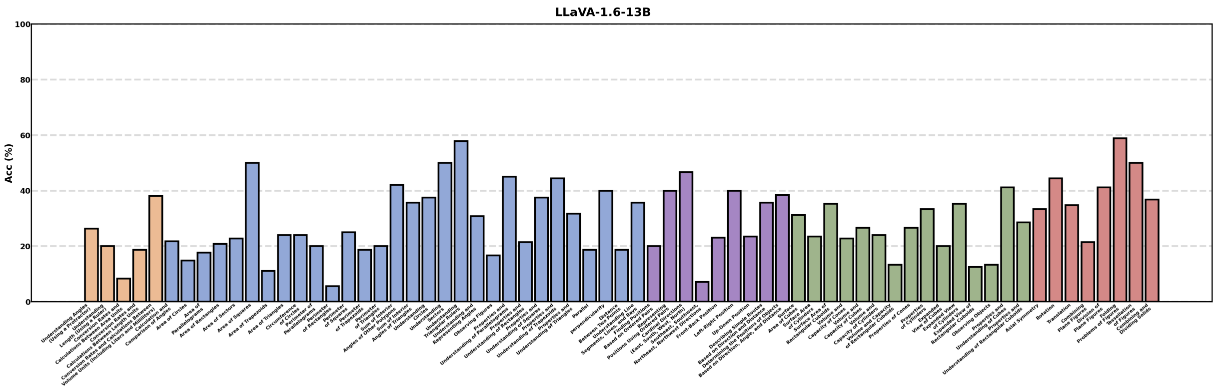


Figure 40: Detailed performance of LLaVA-1.6-13B across 67 knowledge concepts.

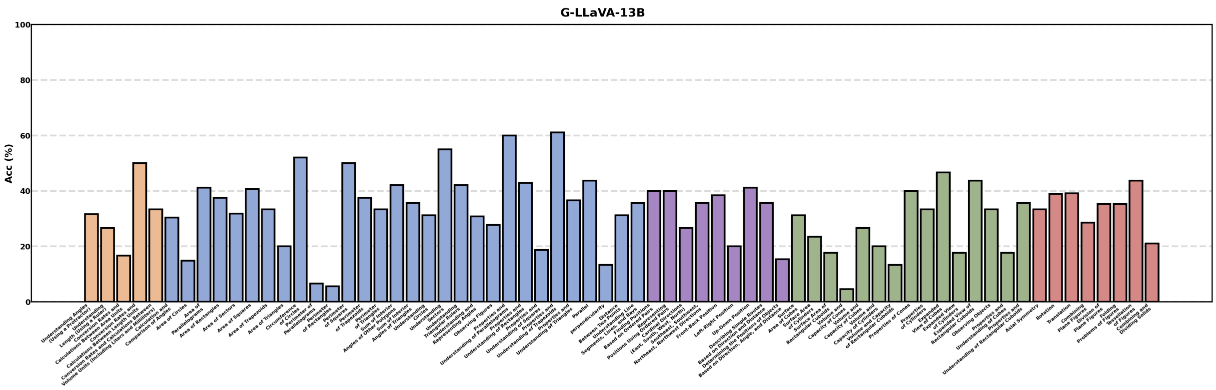


Figure 41: Detailed performance of G-LLaVA-13B across 67 knowledge concepts.

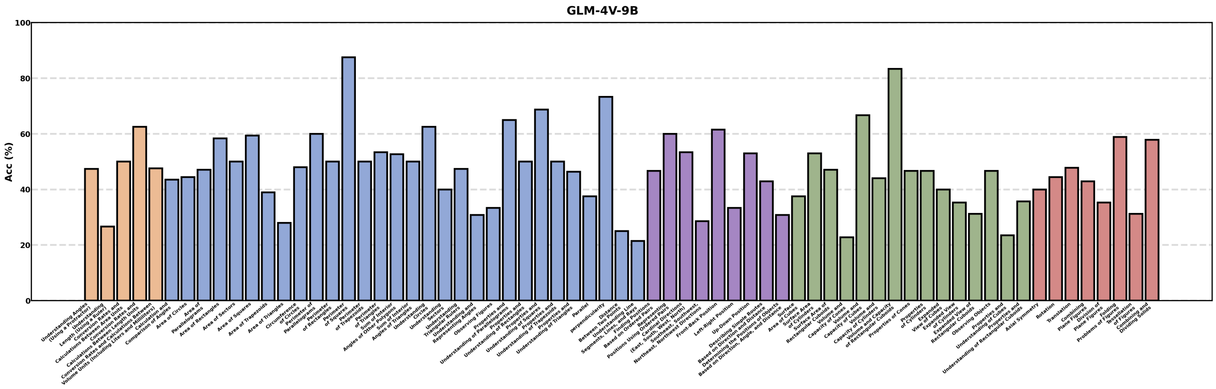


Figure 42: Detailed performance of GLM-4V-9B across 67 knowledge concepts.

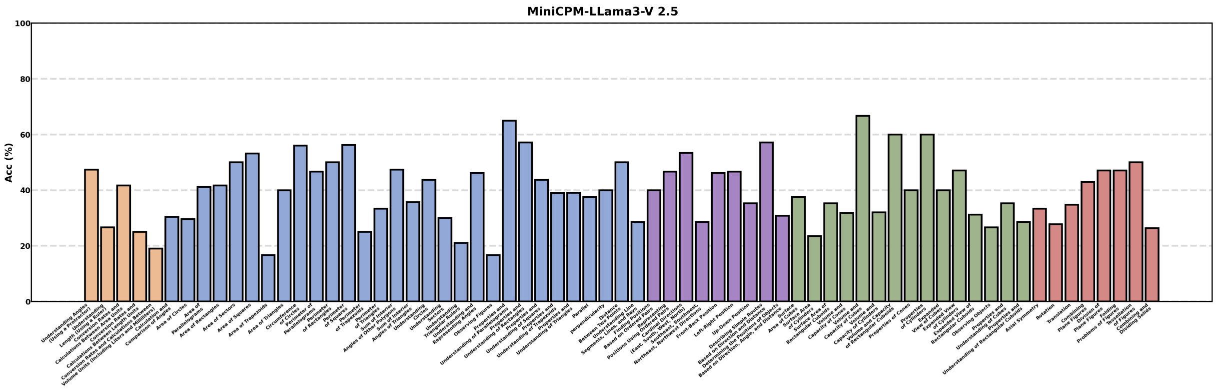


Figure 43: Detailed performance of MiniCPM-Llama3-V2.5 across 67 knowledge concepts.

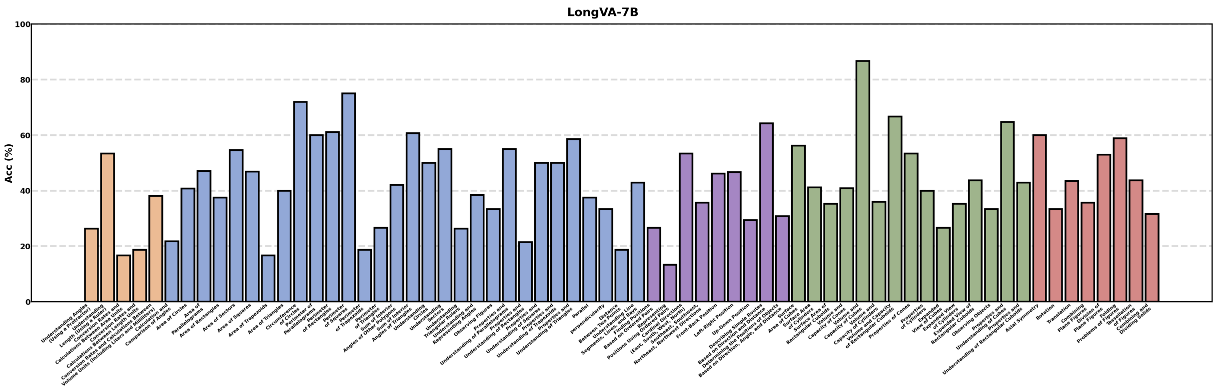


Figure 44: Detailed performance of LongVA-7B across 67 knowledge concepts.

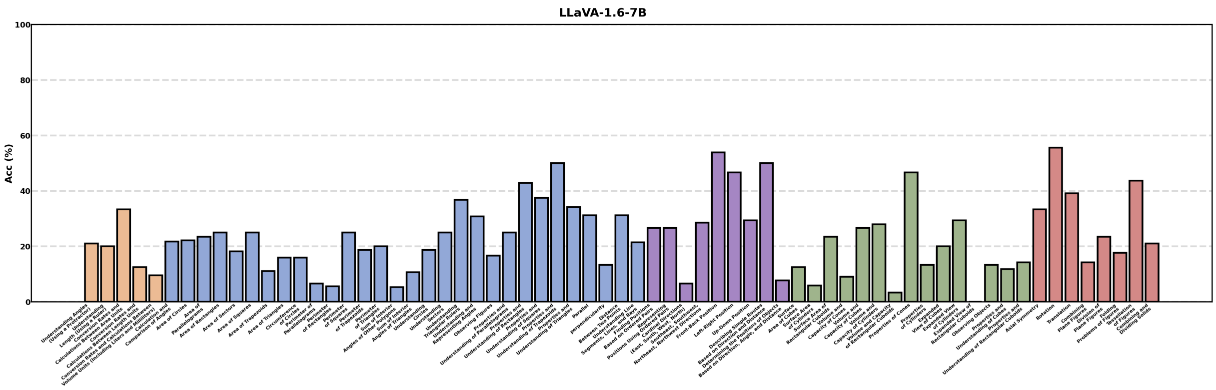


Figure 45: Detailed performance of LLaVA-1.6-7B across 67 knowledge concepts.

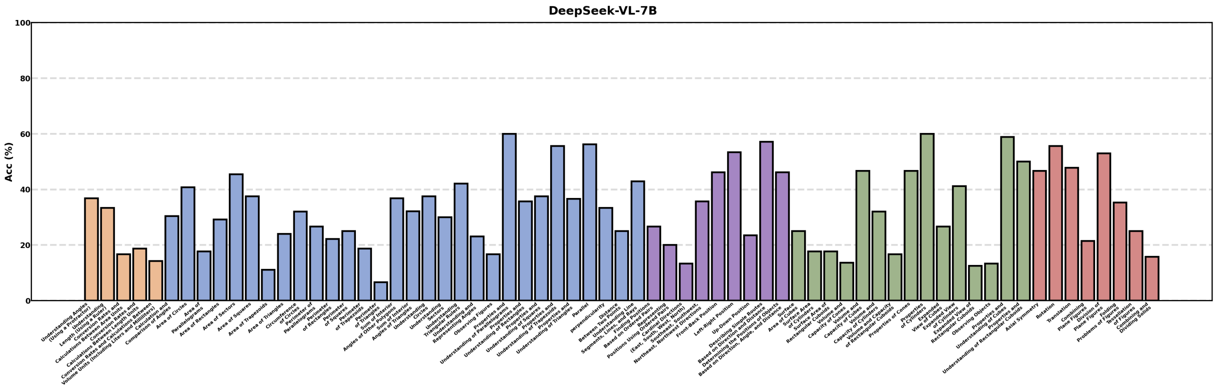


Figure 46: Detailed performance of DeepSeek-VL-7B across 67 knowledge concepts.

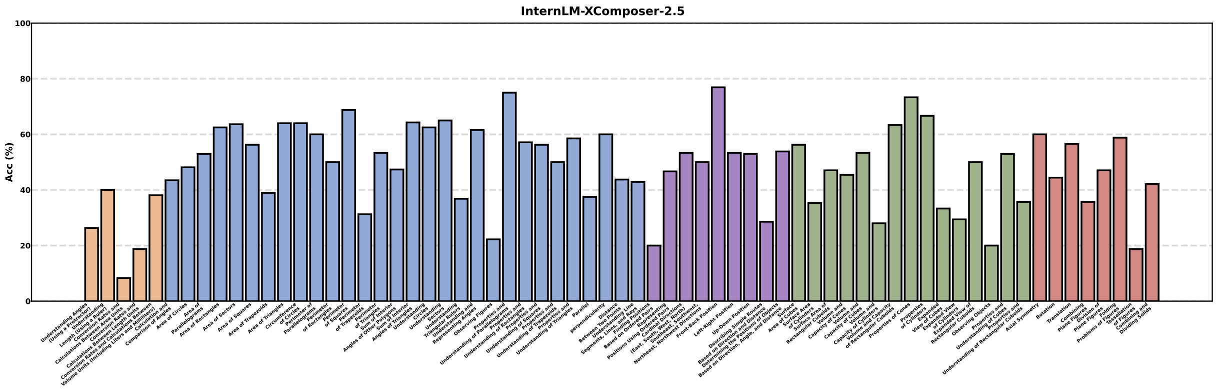


Figure 47: Detailed performance of InternLM-XComposer-2.5-7B across 67 knowledge concepts.

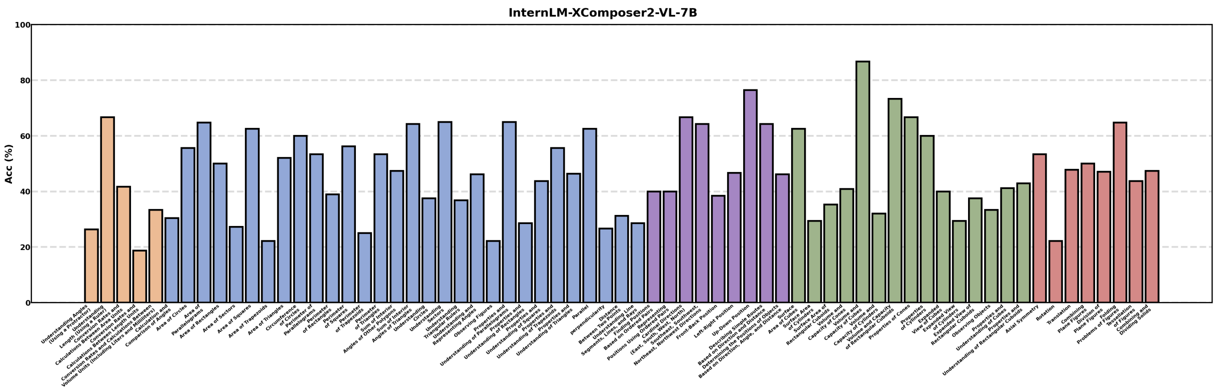


Figure 48: Detailed performance of InternLM-XComposer2-VL-7B across 67 knowledge concepts.

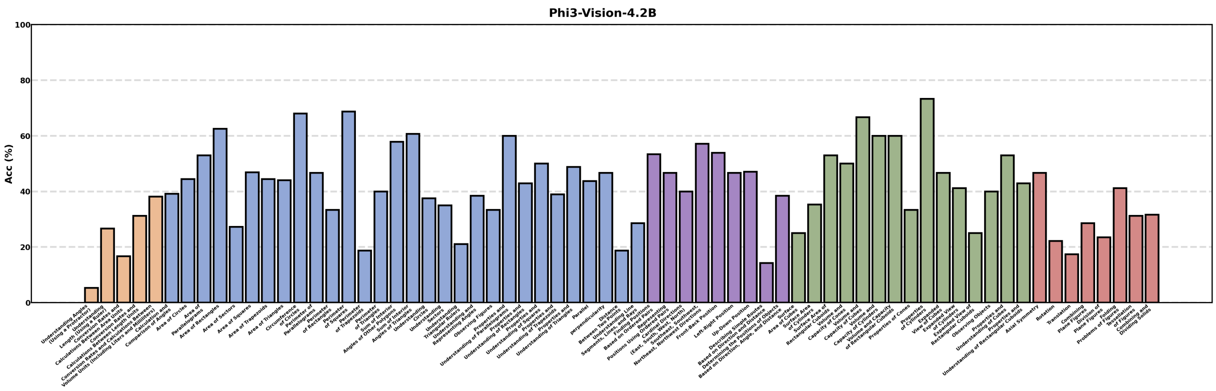


Figure 49: Detailed performance of Phi3-Vision-4.2B across 67 knowledge concepts.

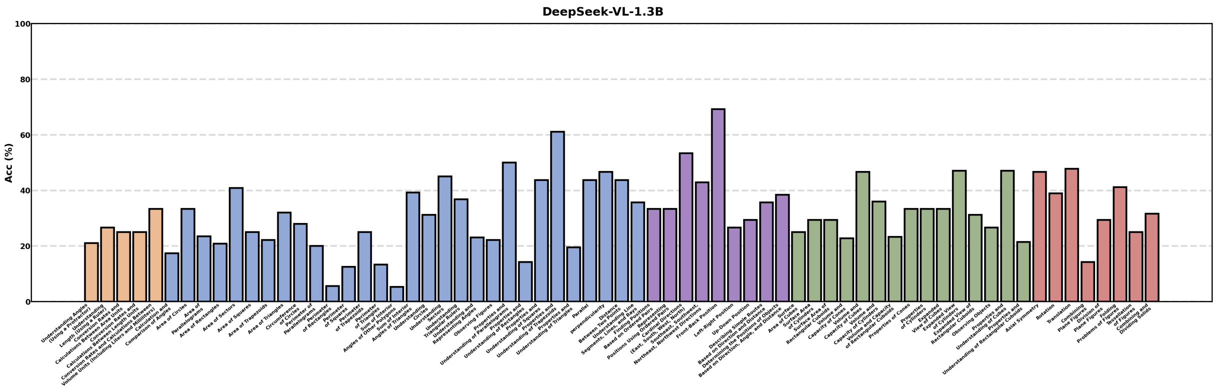
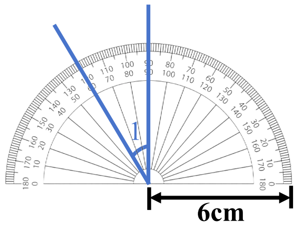


Figure 50: Detailed performance of DeepSeek-VL-1.3B across 67 knowledge concepts.

Visual Error

Question:

As shown in the diagram, using a protractor to measure the angle, what is the size of ∠1? ( )



A. 30°;B. 60°;C. 55°;D. 90°; E. No correct answer

Answer : *A*

Knowledge Concept :

Understanding Angles (Using a Protractor)

Response for GPT-4o:

<Thought process>: *To determine the size of ∠1, I will use the protractor shown in the diagram. The protractor is aligned with the baseline at 0°. The angle ∠1 is measured from the baseline to the inclined line. By observing the protractor, the inclined line intersects the protractor at the 60° mark. Therefore, the size of ∠1 is 60°.*

<Answer>: *B. 60°*

Visual Error

Question:

As shown in the diagram, how do you transform figure A into figure B?

A. Translate 10 units to the right; B. Translate 9 units to the right;
C. Translate 8 units to the right; D. Translate 7 units to the right;
E. No correct answer

Answer : *A*

Knowledge Concept : *Translation*

Response for GPT-4o:

<Thought process>: *To determine how to transform figure A into figure B, I need to count the number of units figure A needs to be translated to the right to align with figure B. By counting the grid squares horizontally from the leftmost point of figure A to the leftmost point of figure B, I can determine the correct translation distance.*

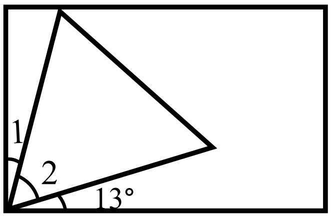
<Answer>: *C. Translate 8 units to the right.*

Figure 51: Specific examples of Visual Error.

Reason Error

Question:

As shown in the diagram, there is an equilateral triangle within a rectangle. If ∠2 = 60°, then the degree measure of ∠1 is ( ).



A. 13°;B. 12°;C. 17°;D. 30°; E. No correct answer

Answer : *C*

Knowledge Concept :

Calculation and Comparison of Angles

Response for GPT-4o:

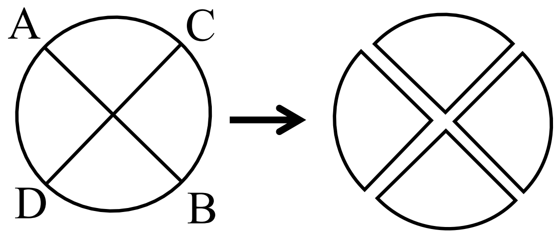
<Thought process>: *Given that the triangle is equilateral, all its internal angles are 60°. Since ∠2 is given as 60°, it confirms the triangle's equilateral nature. The angle adjacent to ∠2 within the rectangle is 13°. The sum of angles around a point is 360°. Therefore, the sum of the angles ∠1, ∠2, and the 13° angle should be 180° (since they form a straight line). Thus, ∠1 + 60° + 13° = 180°. Solving for ∠1, we get ∠1 = 180° - 60° - 13° = 107°. However, this does not match any of the given options.*

<Answer>: *E. No correct answer*

Reason Error

Question:

As shown in the diagram, there is a circular cake with a radius of 5 cm, which is divided into four parts. Compared to the original circumference of the cake, the additional edge length is ( )



A. 4AB+4CD; B. AB+CD; C. 2AB+2CD; D. No correct answer

Answer : *C*

Knowledge Concept : *Translation*

Response for GPT-4o:

<Thought process>: *The original circumference of the cake is given by the formula \(2\pi r\), where \(r\) is the radius. For a radius of 5 cm, the original circumference is \(2\pi \times 5 = 10\pi\) cm. When the cake is divided into four parts, each part will have two additional edges: one along the radius (AB or CD) and one along the arc. The additional edge length for each part is AB + CD. Since there are four parts, the total additional edge length is 4(AB + CD).*

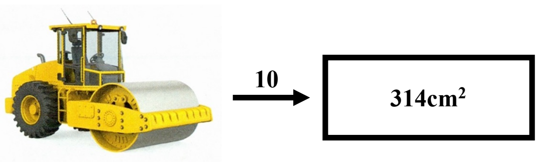
<Answer>: *A. 4AB+4CD*

Figure 52: Specific examples of Reason Error.

Knowledge Error

Question:

The diagram shows a cylindrical roller. The area covered by the roller after rolling 10 circles is equivalent to ( ) dm².(π = 3.14)



A. 3.14;B. 62.8;C. 314;D. 31.4; E. No correct answer

Answer : *A*

Knowledge Concept :

Conversion Rates and Calculations Between Area Units

Response for GPT-4o:

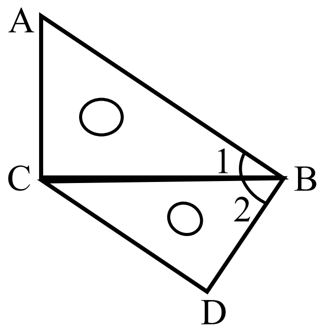
<Thought process>: *To determine the area covered by the cylindrical roller after rolling 10 circles, we need to calculate the area covered in one complete revolution and then multiply it by 10. The area covered in one revolution is the circumference of the roller times its width. Given that the total area covered after 10 circles is 314 cm², we can divide this by 10 to find the area covered in one circle, which is 31.4 cm². Therefore, the correct answer is 31.4 dm².*

<Answer>: *D. 31.4*

Knowledge Error

Question:

As shown in the diagram, two triangular boards are arranged to form a shape. ∠1 + ∠2 = ( )



A. 120°; B. 160°; C. 145°; D. 90°; E. No correct answer

Answer : *D*

Knowledge Concept : *perpendicularity*

Response for GPT-4o:

<Thought process>: *The two triangular boards are arranged such that they share a common side, CB. The angles ∠1 and ∠2 are adjacent to each other and form a straight line along CB. Since a straight line measures 180°, the sum of ∠1 and ∠2 must be 180°.*

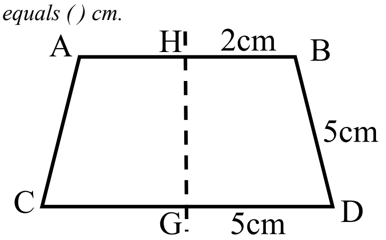
<Answer>: *E. No correct answer*

Figure 53: Specific examples of Knowledge Error.

Hallucination

Question:

As shown in the diagram, HG is the axis of symmetry for trapezoid ABCD. Then, AC equals ( ) cm, GC equals ( ) cm, and AH equals ( ) cm.



A. 5, 4, 2;B. 5, 5, 2;C. 5, 4, 3;D. 5, 5, 3;E. No correct answer

Answer : *B*

Knowledge Concept : *Symmetry*

Response for GPT-4o:

<Thought process>: *Given that HG is the axis of symmetry for trapezoid ABCD, it means that the trapezoid is symmetric about this axis. Therefore, the lengths on either side of the axis are equal.*

1. Since HG is the axis of symmetry, AC and BD are equal in length. Given that BD is 5 cm, AC must also be 5 cm.

2. Since H is the midpoint of AB, and AB is 4 cm (2 cm on each side of H), GC must be half of AC. Therefore, GC is 5 cm / 2 = 2.5 cm.

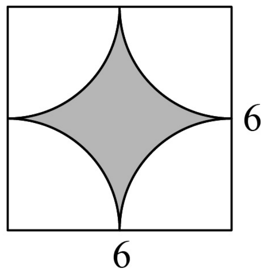
3. Since H is the midpoint of AB, AH is half of AB. Therefore, AH is 2 cm.

<Answer>: *E. No correct answer*

Hallucination

Question:

The white part in the diagram is composed of 4 identical sectors, which can be assembled into a circle. The area of the white part is (   ).



A. π;B. 3π;C. 9π;D. 36π; E. No correct answer

Answer : *D*

Knowledge Concept : *Area of a Circle*

Response for GPT-4o:

<Thought process>: *The diagram shows a square with a side length of 6 units. The white part is composed of 4 identical sectors, which can be assembled into a circle. The radius of each sector is 3 units (half of the side length of the square). The area of one sector is (1/4)πr², where r is the radius. Therefore, the area of one sector is (1/4)π(3)² = (1/4)π(9) = (9/4)π. Since there are 4 sectors, the total area of the white part is 4 * (9/4)π = 9π.*

<Answer>: *E. No correct answer*

Figure 54: Specific examples of Hallucination.

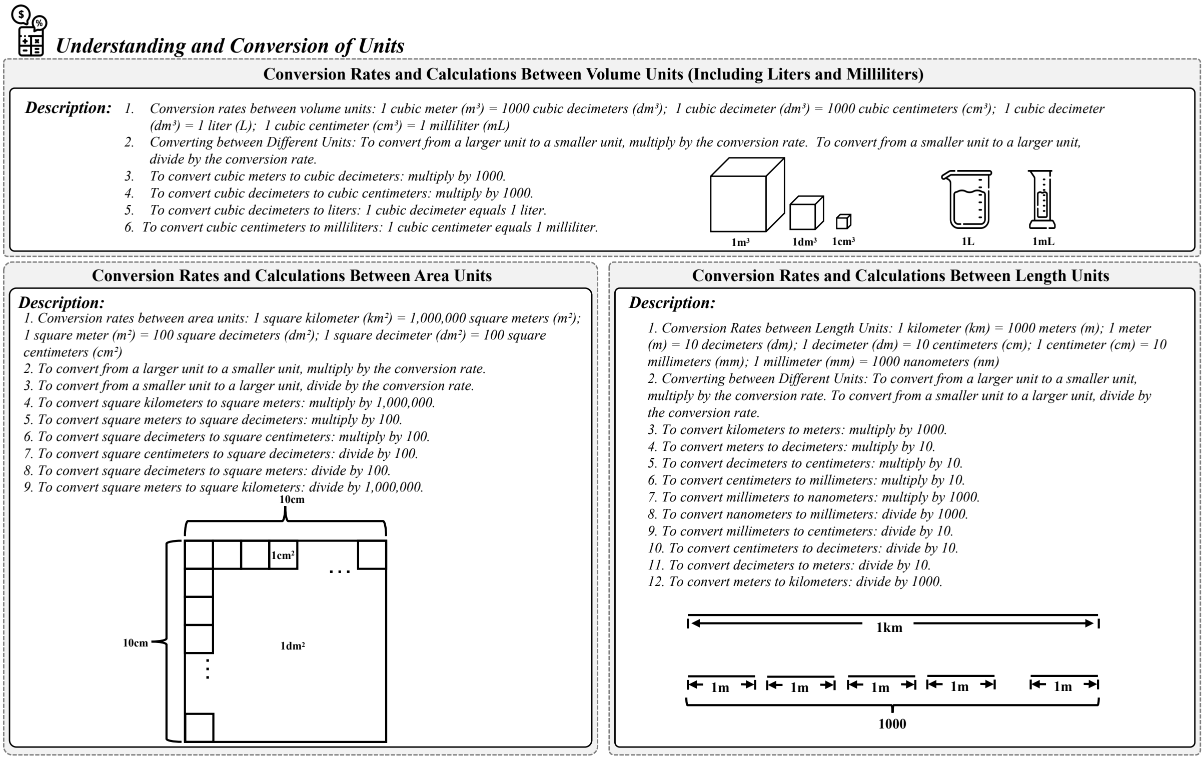


Figure 55: The description of the knowledge concept "Understanding and Conversion of Units".

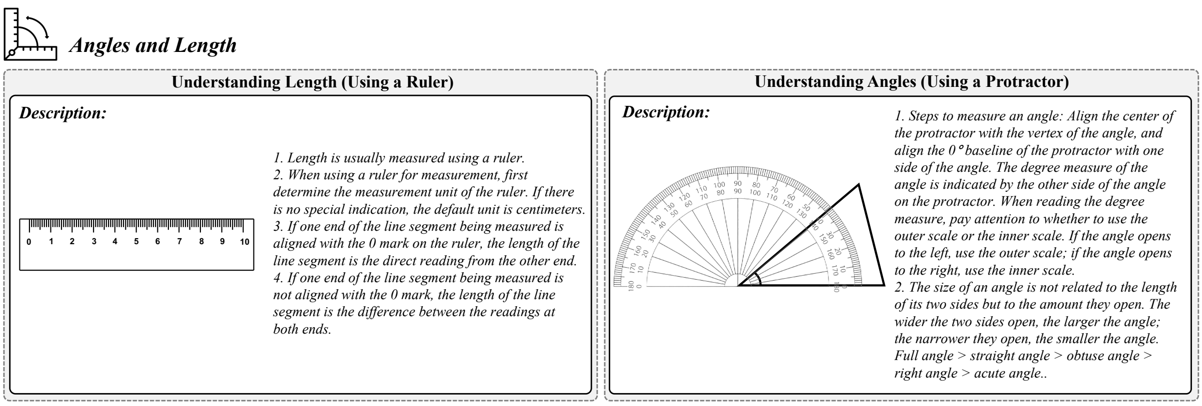


Figure 56: The description of the knowledge concept "Angles and Length".

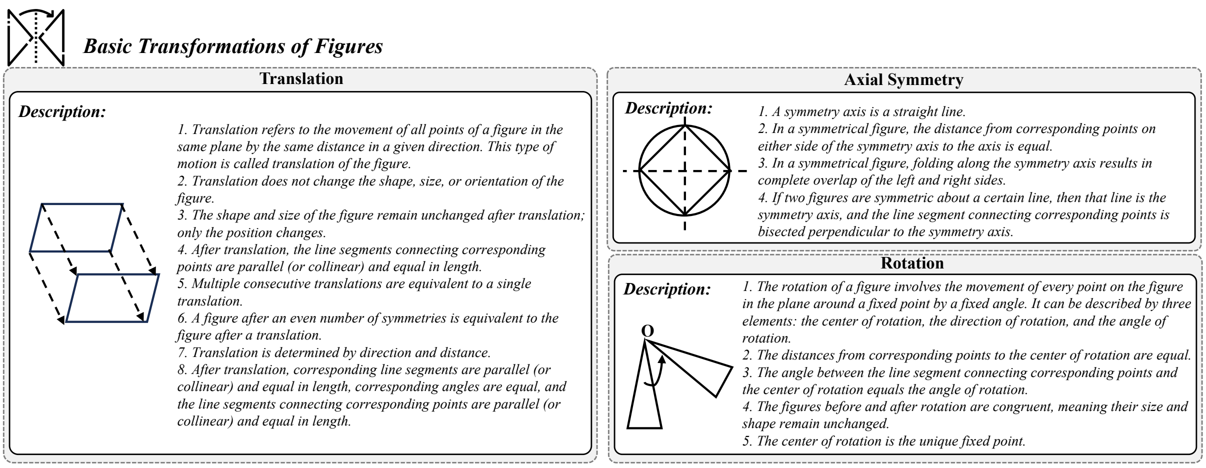


Figure 57: The description of the knowledge concept "Basic Transformations of Figures".

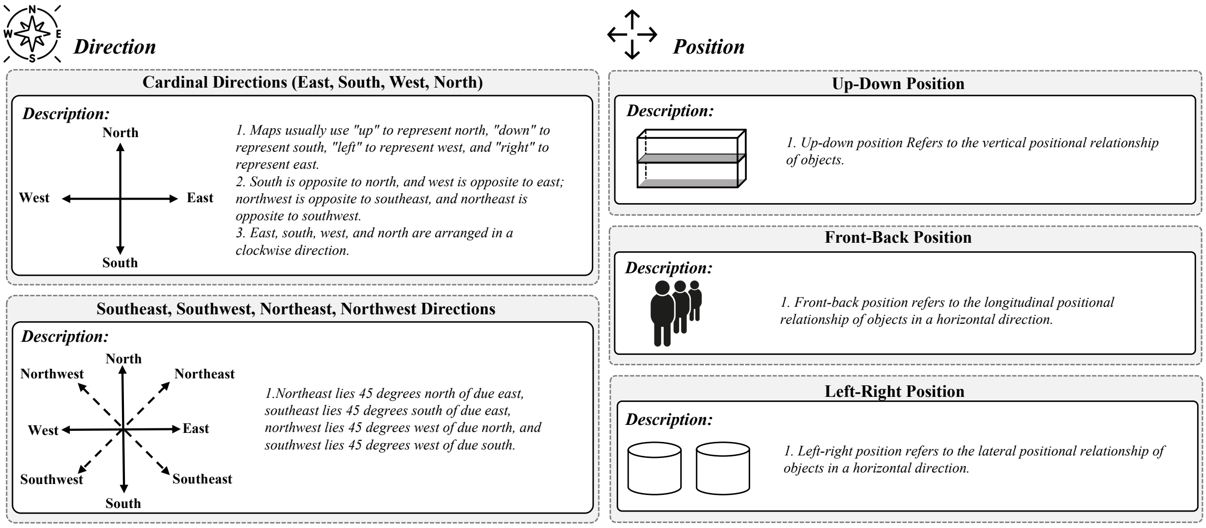


Figure 58: The description of the knowledge concepts "Direction" and "Position".

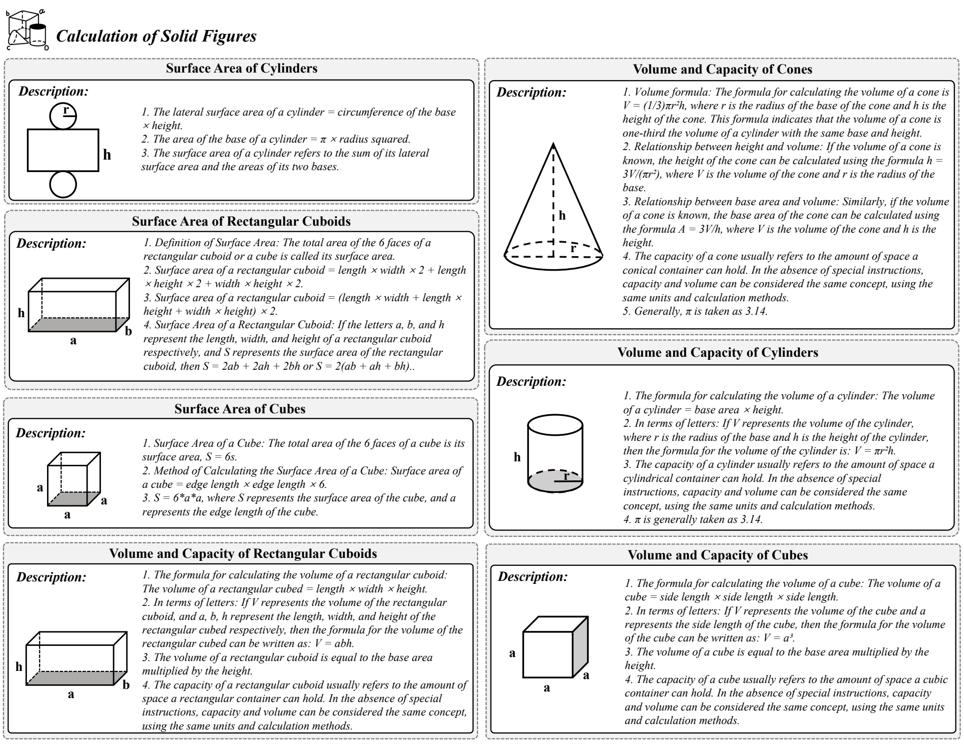


Figure 59: The description of the knowledge concept "Calculation of Solid Figures".

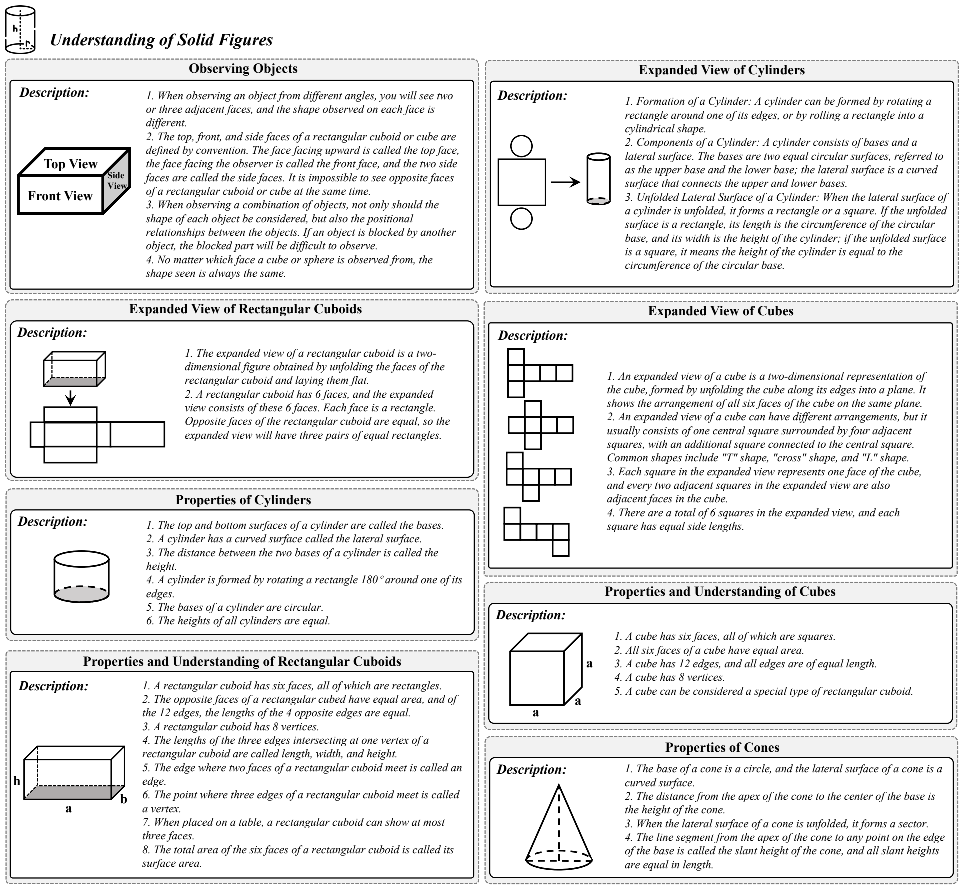


Figure 60: The description of the knowledge concept "Understanding of Solid Figures".

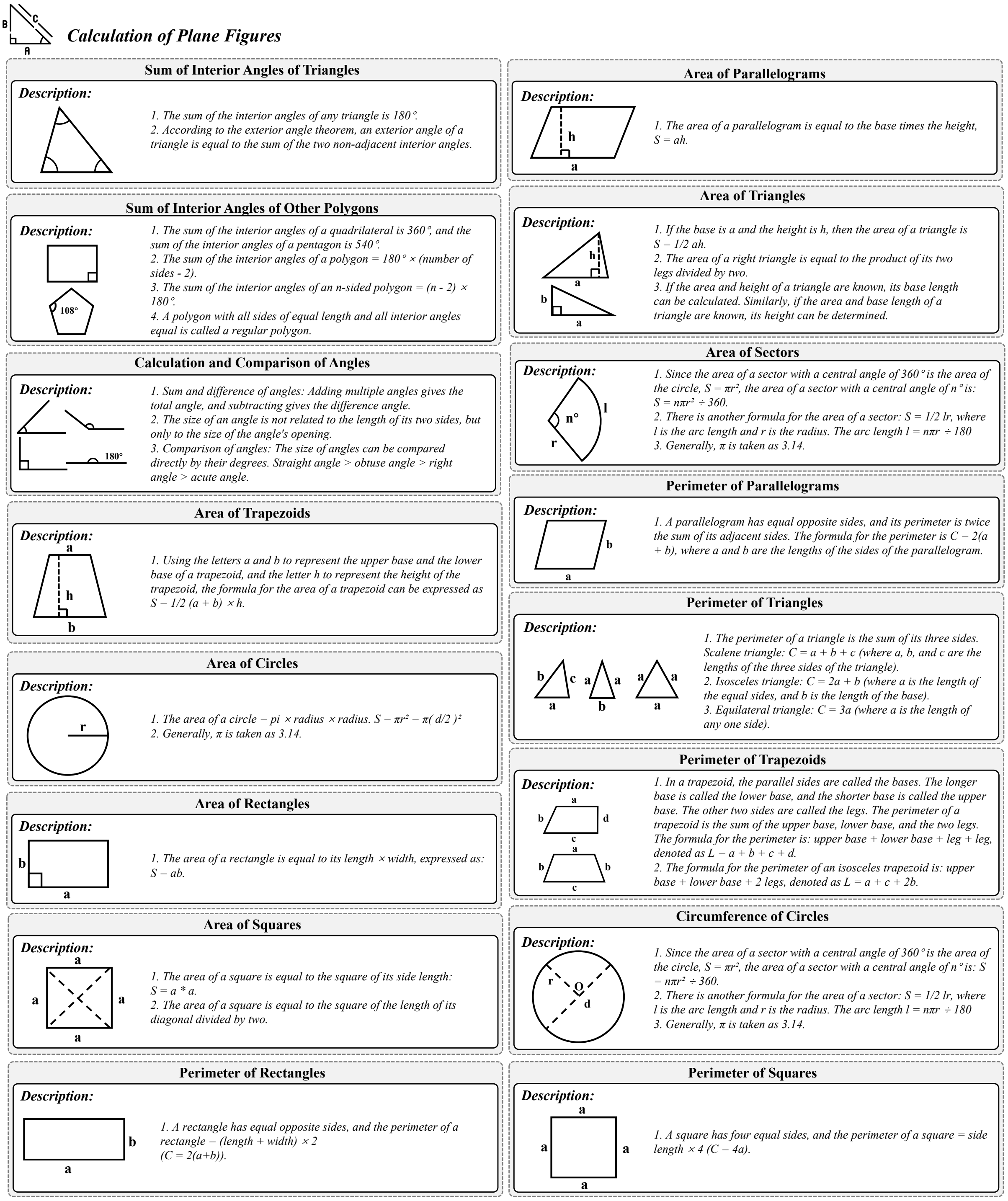


Figure 61: The description of the knowledge concept "Calculation of Plane Figures".

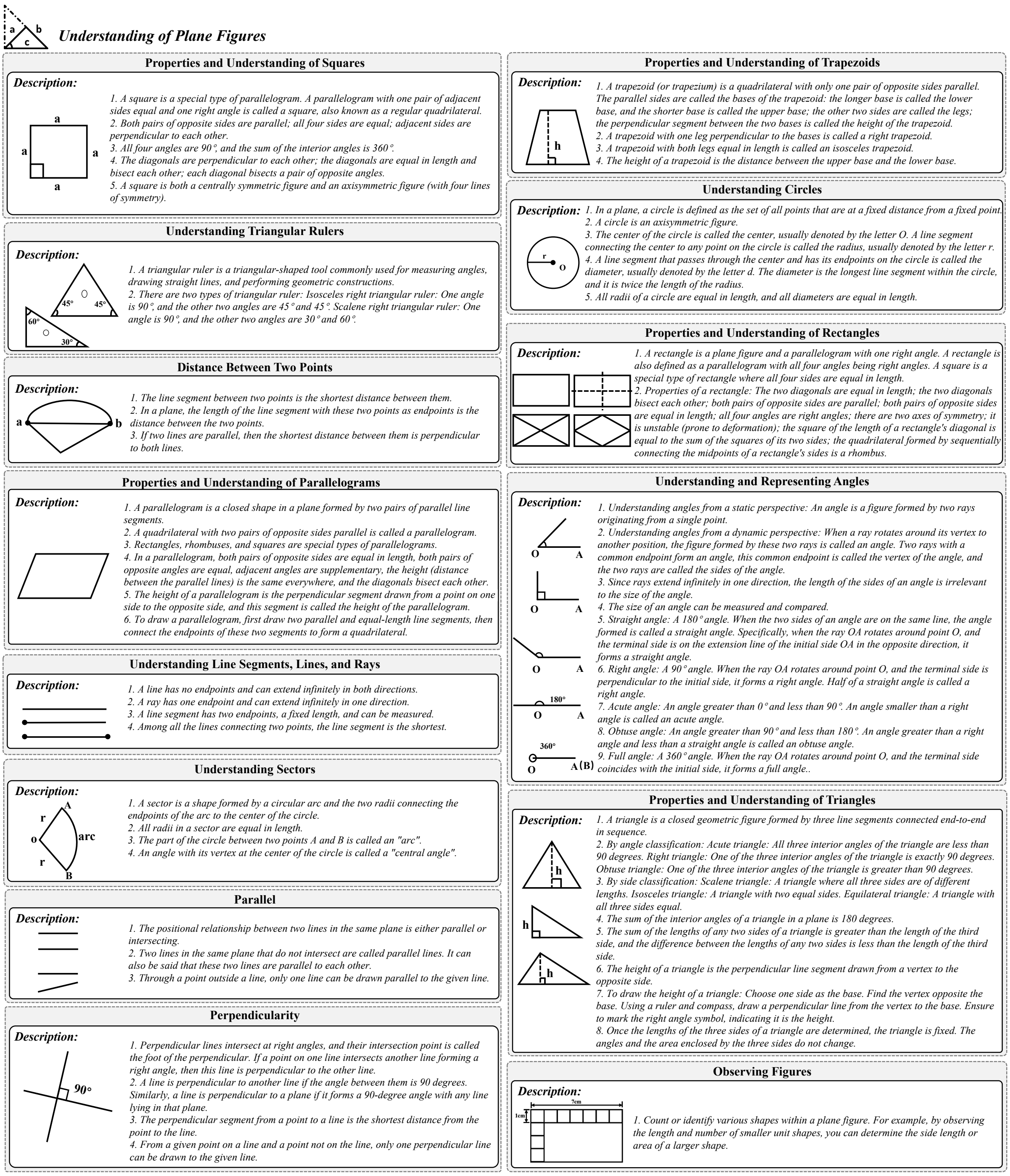
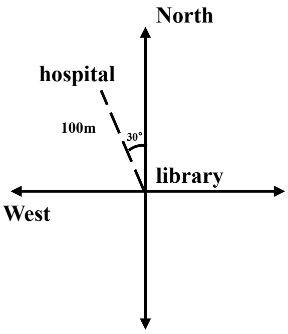


Figure 62: The description of the knowledge concept "Understanding of Plane Figures".

Route Map

Determining the Positions of Objects Based on Direction, Angle, and Distance

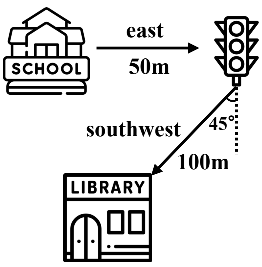
Description:



1. Determining Direction: Common basic directions include east, west, south, and north, as well as intermediate directions such as southeast, southwest, northeast, and northwest. In mathematics, "up" usually represents north, "down" represents south, "left" represents west, and "right" represents east to indicate directions.
2. Determining Angle: Angles are used to describe the relative position between objects or the deviation of an object from a standard direction. For example, "30 degrees west of north" describes a direction that is 30 degrees to the left of the north direction.
3. Determining Distance: Distance refers to the straight-line length from one point to another. When describing a position, it is necessary to specify the distance along a certain direction from a reference point.
4. Describing Position: To describe the position of an object, first determine a reference point, then describe the direction, angle, and distance from the reference point to the object based on its actual location. For example, "The hospital is located 100 meters in the direction 30 degrees west of north from the library."

Describing Simple Routes Based on Direction and Distance

Description:



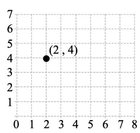
1. Determining Direction: Commonly used basic directions include east, west, south, and north, as well as intermediate directions such as southeast, southwest, northeast, and northwest. In teaching, the convention is to use "up" for north, "down" for south, "left" for west, and "right" for east to identify directions.
2. Determining Angles: Angles are used to describe the relative position between objects or the deviation of an object from a standard direction. For example, ”45 degrees west by south" describes a direction 45 degrees to the left of due south.
3. Determining Distance: Distance refers to the straight-line length from one point to another. When describing a route, it is necessary to indicate how far to travel from the starting point along a certain direction.
4. Describing Routes: To describe a simple route, first determine the starting position, then describe the direction, angle, and distance from the starting point to the destination based on the actual path taken. For example, "Starting from the school, walk 50 meters east to the traffic light, then walk 100 meters in a direction 45 degrees south by west to reach the library."

Figure 63: The description of the knowledge concept "Route Map".

Correspondence of Coordinates and Positions

Finding Positions Based on Ordered Pairs

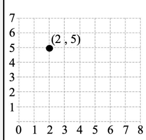
Description:



1. Representation of an Ordered Pair: An ordered pair consists of two numbers, with the first number representing the position on the horizontal axis (column) and the second number representing the position on the vertical axis (row). For example, the ordered pair (2, 4) represents the position in column 2, row 4.
2. Method of Determining Position: To find a position based on an ordered pair, first identify a reference point on the plane. Then move horizontally according to the first number in the ordered pair, and move vertically according to the second number. The final position reached is the point represented by the ordered pair.

Representing Positions Using Ordered Pairs

Description:



1. Definition of an Ordered Pair: An ordered pair is a combination of two numbers, typically used to represent the position of a point in a Cartesian coordinate system. The two numbers in the ordered pair represent values on different directions or axes.
2. Composition of an Ordered Pair: In an ordered pair (x, y), the first number x represents the column number, which is the horizontal position; the second number y represents the row number, which is the vertical position. This notation follows the "horizontal first, then vertical" convention.
3. Notation of an Ordered Pair: When using an ordered pair to represent a point's position, the numbers or letters representing the column and row need to be enclosed in parentheses, for example, (2, 5) represents the position at column 2, row 5.
4. Uniqueness of an Ordered Pair: Each ordered pair uniquely determines a position, and conversely, a position can be uniquely represented by an ordered pair.

Figure 64: The description of the knowledge concept "Correspondence of Coordinates and Positions".