CONSTRUCTURE: Benchmarking CONcept STRUCTUre REasoning for Multimodal Large Language Models

Zhiwei Zha[*1,3] **Xiangru Zhu**[1] **Yuanyi Xu**[1] **Chenghua Huang**[1] **Jingping Liu**[2]
Zhixu Li[†4,5] **Xuwu Wang** **Yanghua Xiao**[1] **Bei Yang**[3] **XiaoXiao Xu**[3]
[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2] School of Information Science and Engineering, East China University of Science and Technology [3] Alibaba Group
[4] School of Information, Renmin University of China
[5] Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies, International College (Suzhou Research Institute), Renmin University of China

Abstract

Multimodal Large Language Models (MLLMs) have shown promising results in various tasks, but their ability to perceive the visual world with deep, hierarchical understanding similar to humans remains uncertain. To address this gap, we introduce CONSTRUCTURE, a novel concept-level benchmark to assess MLLMs' hierarchical concept understanding and reasoning abilities. Our goal is to evaluate MLLMs across four key aspects: 1) Understanding atomic concepts at different levels of abstraction; 2) Performing upward abstraction reasoning across concepts; 3) Achieving downward concretization reasoning across concepts; and 4) Conducting multi-hop reasoning between sibling or common ancestor concepts. Our findings indicate that even state-of-the-art multimodal models struggle with concept structure reasoning (e.g., GPT-4o averages a score of 62.1%). We summarize key findings of MLLMs in concept structure reasoning evaluation. Morever, we provide key insights from experiments using CoT prompting and fine-tuning to enhance their abilities.

1 Introduction

> *The basic level is the level in a taxonomy at which things are normally named, in the absence of reasons to the contrary. 'Dog' is a basic level category, 'boxer' a subordinate category, 'quadruped' a superordinate category.*
>
> — John R. Taylor

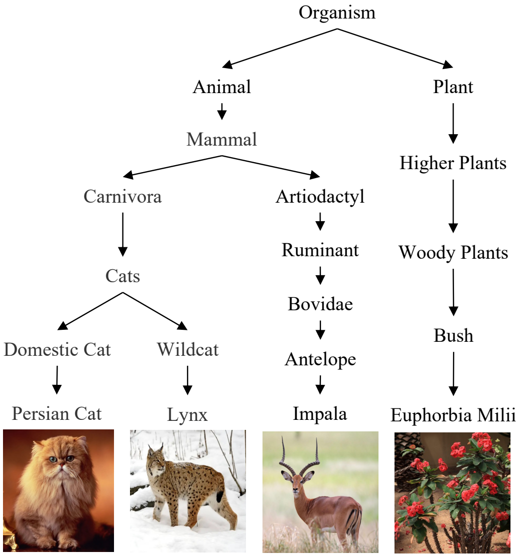


Figure 1: Demo of Concept Structure.

According to the prototype theory (Taylor, 2019) in cognitive science, humans perceive the visual world hierarchically, with basic, subordinate, and superordinate categories. People interpret the world differently based on these conceptual levels in diverse environments. As shown in Figure 1, humans can not only understand basic concepts, but also have a deep concept structure cognition in minds (Medin, 1989; Suresh et al., 2023). This raises the question: do multimodal AI systems, like Multimodal Large Language Models(*e.g.* GPT-4V (OpenAI, 2023)), exhibit similar concept structure cognition ability?

Recently, pretrained multimodal large language models (MLLMs) have transcended the confines of text-only modalities, gaining a deeper conceptual understanding of the world and demonstrating remarkable performance across a spectrum of downstream tasks. As a result, there is a growing importance and challenge in conducting comprehensive evaluations of these models to better understand their capabilities and pinpoint areas for enhancement. Inspired by the cognitive processes underlying human conceptual understanding, we posit that similar deep and structured visual conceptual cognition capabilities are pivotal for the profound

---

[*] zwcha22@m.fudan.edu.cn. Work done during internship at Alibaba Group.

[†] Correspondence to: Zhixu Li<zhixuli@gmail.com>

Findings of the Association for Computational Linguistics: EACL 2024, pages 4954–4968
November 12-16, 2024 ©2024 Association for Computational Linguistics

Benchmark	**Assessment of cognitive related abilities**	**Data Source**	**Answer Type**	**Evaluator**	**Size**
LVLM-eHub (Xu et al., 2023)	Visual Reasoning, Visual Commonsense	Datasets	MC/OE	Metrics/LLMs/User	332k
MME (Fu et al.)	Commonsense Reasoning, Numerical Calculating, Text Translation, Code Reasoning	Datasets	MC	Accuracy	2,194
MMBench (Liu et al., 2023)	Reasoning(*e.g.* Phsycial Relation Reasoning, Social Relation Reasoning)	Datasets/Handcraft/LLMs	MC	Accuracy	2,974
SEED-Bench (Li et al., 2023a)	Visual Reasoning, Spacial Relation	Handcraft/LLMs	MC	Accuracy	19k
MM-Vet (Yu et al., 2023)	Spatial awareness, Knowledge, Math	Datasets/Handcraft	OE	LLMs	218
EgoThink (Cheng et al., 2024)	Scene Reasoning, Planning	Handcraft	OE	LLMs	700
CONSTRUCTURE(Ours)	**Concept Structure Reasoning**	**Datasets/Handcraft**	**MC**	**Accuracy**	**2,064**

Table 1: Comparison Between Comprehensive Benchmarks for MLLMs and Our CONSTRUCTURE.

comprehension of MLLMs. This compelling question demands exploration: Do MLLMs truly grasp and internalize concepts, or do they merely learn the superficial concept alignment through pretraining?

However, addressing this question requires a deeper investigation into the underlying mechanisms and limitations of MLLMs. As shown in Table 1, the deep visual concept structure cognition remains largely unexplored in existing benchmarks. On one hand, these benchmarks seldom consider visual cognitive capabilities. On the other hand, existing benchmarks for visual cognition typically focus solely on assessing visual reasoning ability related to the whole image content, overlooking the evaluation of deep and structural visual concept cognition. For example, SEED-bench (Li et al., 2023a) evaluates spatial relationships or visual reasoning tasks based on the content of images, while MME (Fu et al.) assesses common-sense reasoning, numerical computation, code inference, and text translation. LVLM-eHub (Xu et al., 2023) evaluates visual reasoning and visual common sense. However, all these benchmarks require answering questions based on the entire content of an image. Therefore, there is a pressing need to construct an benchmark specifically designed to assess the deep visual concept structure cognition in MLLMs.

In this work, we propose the novel CONSTRUCTURE benchmark focusing on deep visual concept structure cognition. To fully uncover the deep visual cognitive ability of MLLMs across the overall concept structure beyond the Figure 1, we consider the following four key capabilities, ranging from *atomic visual concenpt understanding*, *concept abstraction reasoning*, *concept concretization reasoning*, to *common ancestor reasoning*. The first capability is to evaluate the atomic concept understanding at different levels. The last three capabilities are to evaluate reasoning ability on concept structure, including upward abstraction reasoning from child concepts to parent concepts, downward concretization reasoning from parent concepts to child concepts and multi-hop reasoning between sibling concepts or common ancestor concepts, respectively.

Based on the our proposed CONSTRUCTURE benchmark, we conduct comprehensive experiments to evaluate concept structure cognition capabilities of fourteen popluar MLLMs (including six API-based MLLMs and eight open-sourced MLLMs). We conclude the main findings as follows:

1) Current MLLMs possess a certain level of conceptual understanding, but their performance in concept structure reasoning is poor. The best model, GPT-4o, only achieved a score of 0.621, indicating significant room for improvement.

2)In concept structure reasoning tasks, MLLMs perform the worst in common ancestor reasoning. The main reasons for errors are inconsistencies in the reasoning process and constraint violations. This demonstrates that adhering to multiple constraints and maintaining consistency in the reasoning process are key challenges to enhancing MLLMs' concept structure cognitive abilities.

3) MLLMs still need to improve their ability to reason about hierarchical relationships in concept structures. For instance, in abstraction and concretization reasoning tasks, their ability to identify more abstract or concrete concepts among multiple concepts is still limited. They often struggle due to a lack of knowledge about related concept hierarchies.

4) MLLMs' performance deteriorates as the concept hierarchy deepens and granularity increases. Therefore, improving fine-grained concept recognition and the ability to reason about related fine-grained concept structures is crucial for enhancement.

Our evaluation results reveal the limitations of MLLMs in concept structure cognition, providing a comprehensive and clear analysis that directs further improvements for MLLMs. Furthermore, we improved MLLMs' concept structure reasoning abilities through few-shot CoT prompting and fine-tuning methods, and elucidated three key insights in the discussion.

2 Related Work

2.1 Multimodal Large Language Models

Building upon the monumental achievements of large language models (LLMs) (Du et al., 2022; OpenAI, 2023; Zhu et al., 2023), recent advancements in multimodal LLMs (MLLMs) have firmly established LLMs as their foundational backbone. Initially, MLLMs commence pre-training using large-scale image-text paired datasets (Yang et al., 2023; Li et al., 2023b) or by integrating random visual and textual data (Ye et al., 2023; Bai et al., 2023). This foundational phase is further enriched by leveraging extensive image-text instruction datasets (Dai et al., 2024). Recent studies (Liu et al., 2024) have increasingly employed finetuning strategies, significantly enhancing MLLMs' capacity to deliver superior performance in various downstream tasks and benchmarks.

2.2 Evaluations of MLLMs

Current benchmarks aim to comprehensively assess MLLMs' capabilities in multimodal cognition tasks like reasoning. For example, LVLM-eHub (Xu et al., 2023) evaluates visual reasoning and common-sense cognition across 47 benchmarks. MME (Fu et al.) covers inference, numerical computation, translation, and code reasoning. SEED-bench (Li et al., 2023a) focuses on visual and spatial reasoning, while MM-VET (Yu et al., 2023) evaluates spatial relationships, knowledge, and math abilities. EgoThink (Cheng et al., 2024) assesses scene reasoning and planning. Our CONSTRUCTURE benchmark specifically targets hierarchical concept structure reasoning in MLLMs' understanding of visual concepts.

2.3 Concept-Related Datasets and Benchmarks

Concept-related datasets like ConceptNet (Speer et al., 2017) and WordNet (Miller, 1995) provide structured taxonomies for words and concepts but are limited to a single modality. In contrast, multimodal datasets such as ImageNet (Deng et al., 2009), M^{2}ConceptBase (Zha et al., 2023), and BabelPic (Calabrese et al., 2020) align visual images with concepts, bridging textual and visual modalities. Datasets like Objects365 (Shao et al., 2019) focus on object detection, while COCO (Lin et al., 2014), CC3M (Sharma et al., 2018), and CC12M (Changpinyo et al., 2021) provide conceptual captions paired with images, playing a key role in training MLLMs. However, benchmarks for evaluating the conceptual understanding of MLLMs are limited. For example, POPE (Li et al., 2023c), based on COCO, evaluates object hallucination in MLLMs but does not assess models' understanding of hierarchical concept structures. In contrast, our CONSTRUCTURE benchmark integrates hierarchical visual concepts and images to evaluate MLLMs' cognition and reasoning abilities related to concept structure, advancing their evaluation beyond surface-level tasks.

3 CONSTRUCTURE Benchmark

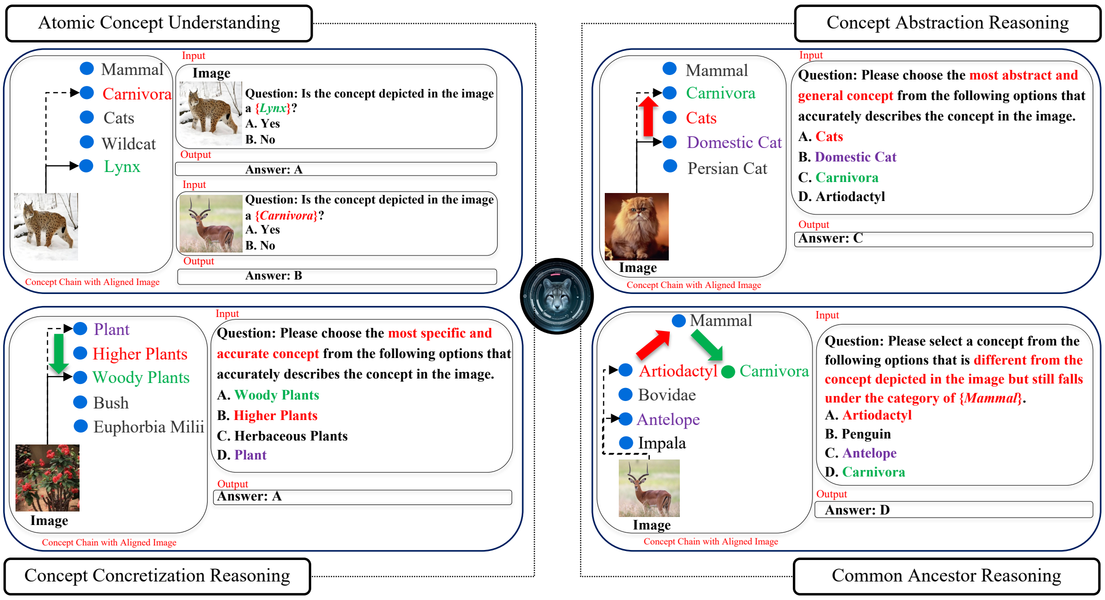
3.1 Problem Formulation

Our objective is to assess concept structure cognition abilities of MLLMs. This task can be formulated as multiple-choice question answering. Given an image I and a question Q with candidate answers $A = \{A_i\}_{i=1}^{m}$, the goal is to select the answer that best corresponds to the question.

To uncover the concept structure cognition ability of MLLMs, we design the following four key tasks, as shown in Figure 2, ranging from *atomic visual concenpt understanding*, *concept abstraction reasoning*, *concept concretization reasoning*, to *common ancestor reasoning*. We explain why these are needed inspired by human concept cognition and introduce how to evaluate with examples.

Atomic Visual Concept Understanding. Human cognition of concepts has different levels of abstraction, and we can not only understand concrete concepts like "Persian Cat", but also abstract concepts like "Mammal". Can MLLMs understand visual concepts at various levels of abstraction? As shown in upper left part of Figure 2, to evaluate atomic visual concept understanding ability, we query MLLMs with a simple discriminant question (*i.e.* "*Is the concept depicted in the image a {concept_name}?*") with true or false options related to specific concepts at various abstraction levels. To answer this question, MLLMs need to have a multi-level understanding of visual concepts.

Concept Abstraction Reasoning. Human beings can categorize concrete concepts into higher-level abstract concepts; for example, they categorize "penguins" and "sharks" as "birds" and "fish", respectively. Furthermore, they can select the most abstract concept from multiple candidates; for instance, identifying "carnivora" as the most abstract concept for "Persian Cat" among other potential choices (*e.g.*, "cats", "domestic cat", "carnivora").


Atomic Concept Understanding
Mammal
Carnivora
Cats
Wildcat
Lynx
Concept Chain with Aligned Image
Input
Image
Output
Concept Abstraction Reasoning
Mammal
Carnivora
Cats
Domestic Cat
Persian Cat
Image
Concept Chain with Aligned Image
Input
Output
Plant
Higher Plants
Woody Plants
Bush
Euphorbia Milii
Image
Concept Chain with Aligned Image
Input
Output
Concept Concretization Reasoning
Mammal
Artiodactyl
Carnivora
Bovidae
Antelope
Impala
Image
Concept Chain with Aligned Image
Input
Output
Common Ancestor Reasoning

Question: Is the concept depicted in the image a {*Lynx*}?
A. Yes
B. No

Answer: A

Question: Is the concept depicted in the image a {*Carnivora*}?
A. Yes
B. No

Answer: B

Question: Please choose the most abstract and general concept from the following options that accurately describes the concept in the image.
A. Cats
B. Domestic Cat
C. Carnivora
D. Artiodactyl

Answer: C

Question: Please choose the most specific and accurate concept from the following options that accurately describes the concept in the image.
A. Woody Plants
B. Higher Plants
C. Herbaceous Plants
D. Plant

Answer: A

Question: Please select a concept from the following options that is different from the concept depicted in the image but still falls under the category of {*Mammal*}.
A. Artiodactyl
B. Penguin
C. Antelope
D. Carnivora

Answer: D

Figure 2: CONSTRUCTURE Benchmark. Each sub-task consists of three components: Input, Output, and Concept Chain with Aligned Image. Concepts in the chains and answer options are consistently color-coded, with green indicating the correct answer and key question requirements highlighted in red. In the Concept Chain with Aligned Image, solid arrows denote correct concepts, dashed arrows indicate incorrect concepts, red arrows represent concept abstraction, and green arrows signify concept concretization.

How well do MLLMs perform this kind of visual concept abstraction reasoning process? As shown in upper right part of Figure 2, to evaluate concept abstraction reasoning ablity, we query MLLMs with multiple-choice questions to select the most abstract and general visual concept from candidate options. The image is aligned correctly with several concepts in candidate options, MLLMs need to recognize them and figure out one has the most abstract level. To answer this question, MLLMs need firstly recognize correct options and then reason out the most abstract concept aligned with the image.

Concept Concretization Reasoning. Human beings can concretize abstract concepts to lower-level concrete concepts; for example, recognizing from the category of plants ("higher plants") to fine-grained types ("woody plants"). Additionally, they can identify the most specific concept from multiple options; for instance, classifying "woody plants" as the most specific concept for "Euphorbia Milii" among other potential choices (e.g., "woody plants", "higher plants", "plants"). How well do MLLMs perform this kind of visual concept concretization reasoning process? As shown in lower left part of Figure 2, to evaluate concept concretization reasoning ability, we query MLLMs with multiple-choice questions to select the most specific and accurate concept from candidate options. The image is aligned correctly with several concepts in candidate options, MLLMs need to recognize them and figure out the most concrete one. To answer this question, MLLMs need firstly recognize correct options and then reason out the most concrete concept aligned with the image.

Common Ancestor Reasoning. Human perception of visual concepts follows a hierarchical structure. For instance, we recognize that "dogs" and "cats" have a higher-level common ancestor concept "mammal", and within the category of "dogs", there are subcategories like "pet dog" and "hunting dog". Can MLLMs fully grasp the various hierarchical levels of a visual concept? Can they reason that two visual concepts have a common ancestor concept or are they in a sibling relationship? As shown in lower right part of Figure 2, to evaluate common ancestor reasoning ability, we query MLLMs with multiple-choice questions to select a concept which is different from as well as share the same parent or common ancestor concept with the concept depicted in the image. To answer this question, MLLMs need have the ability to reason across sibling concepts or common ancestor concepts, which indicates that MLLMs need to have the structure cognition ability of the related concepts.

3.2 Data collection

Constructing Concept Chains. We use the Open WordNet (Wang and Bond, 2013) as our data source of concept structure, which encompasses 19.7K isA relations. We clean the raw data and construct an isA concept tree (with a root node). During construction, we drop potential isA relations to avoid the circular dependency. After that, we recursively search for the isA concept chains with a length of 5 as our candidates for the next stage of sample generation.

Collecting Visual Concept Images with Manual Checking. First, since not all concepts in the collected candidate chains are necessarily visual concepts, we leverage M^{2}ConceptBase (Zha et al., 2023), a multimodal knowledge base with extensive concept-image alignments, to select chains where all concepts are visual. Second, we assign images from M^{2}ConceptBase to each of the lowest-level concepts in these chains. Third, we meticulously verify the accuracy of the concept chains, ensuring that the lowest concept in each chain correctly corresponds to its image in the knowledge base. We eliminate any incorrect candidate chains or erroneous isA relations (originating from raw data in Open Wordnet) within the candidate chains. Additionally, we search the Internet for accurate images to replace any lower quality or incorrect ones. In conclusion, we curate a total of 646 chains with lengths ranging from 3 to 5, each with correctly aligned images for the lowest-level concepts.

Sample Construction. For each task, a sample S is composed of an image I, a question Q, a candidate answer set A and an answer A_y: $S = \{I, Q, A, A_y\}$. For each question in every task, we construct (image, question, answer candidates, answer) pairs based on the concept chain. The images are assigned in the previous stage, and the questions are meticulously designed templates, as shown in Figure 2. Next, we generate the answer and different candidate answers using concept chains and concept structure, including neighboring concepts in the concept structure as challenging distractors. For atomic visual concept understanding, we can generate N questions with true or false options for every chain of length N. We randomly assign 50% of concepts paired with correct images ($A_y =$ *"A. Yes"*), and 50% with incorrect images from unrelated concepts ($A_y =$ *"B. No"*). In concept abstraction reasoning, for each chain of length N, we generate $N-1$ questions with options set to current-level concept, lower-level concept, positive upper-level concept (A_y), and negative upper-level concept (in random order). Similarly, in concept concretization reasoning, options include current-level concept, upper-level concept, positive lower-level concept (A_y), and negative lower-level concept. For common ancestor reasoning in chains of length N, options cover current-level concept, upper-level concept, lower-level concept and brother or brother-son concept (A_y), arranged randomly. We conduct rigorous programmatic and manual checks to replace unreasonable options that might lead to multiple correct answers with appropriate options.

Subset	**Train**	**Valid**	**Test**	**Total**
Atomic Visual Concept Understanding	2,168	313	637	3,118
Concept Abstraction Reasoning	1,738	247	490	2,475
Concept Concretization Reasoning	1,717	246	489	3,210
Common Ancestor Reasoning	1,611	225	448	2,284
Total	**7,234**	**1,031**	**2,064**	**10,329**

Table 2: Statistics of CONSTRUCTURE.

3.3 Data Statistics

As shown in Table 2 , our CONSTRUCTURE benchmark has totally 10, 329 samples. We split the benchmark into training, validation, and test sets in a ratio of 7:1:2, totaling 7234, 1031, and 2064 samples, respectively. The test set is exclusively used for evaluating MLLMs, while the training and validation sets are employed for fine-tuning MLLMs to enhance their conceptual structure reasoning capabilities. The breakdown of question types is detailed in Table 2. We further elaborate on the details of the CONSTRUCTURE benchmark in Table 4. Our benchmark comprises 1315 visual concepts, 646 concept chains, and 644 images, with an average concept chain length per sample of 4.93. The breakdown of subsets is detailed in Table 4. As shown in Figure 5, our CONSTRUCTURE benchmark covers a wide range of topics, including "animal", "plant", "device", "clothing", "chemicals", and more.

4 Experiments

4.1 Experimental setups

Multimodal Large Language Models. We evaluate six API-based MLLMs (OpenAI, 2023; Yang et al., 2023; Team et al., 2023; Bai et al., 2023) and eight open-source MLLMs (Li et al., 2023b; Dai et al., 2024; Zhu et al., 2023; Ye et al., 2023;

Model	Concept Understanding	Abstraction Reasoning	Concretization Reasoning	Common Ancestor	Avg. Score
		API-based Models			
gpt-4-vision-preview	0.692	**0.748**	0.601	0.241	0.586
gpt-4o-0513	**0.896**	0.657	**0.663**	0.145	**0.621**
gemini-pro-vision	0.733	0.584	0.486	0.040	0.489
gemini-1.5-pro	0.746	0.692	0.407	0.100	0.513
claude3_sonnet	0.666	0.571	0.501	**0.397**	0.546
Qwen-VL-Max	0.763	0.543	0.631	0.143	0.545
		Open-source Models			
BLIP2	**0.794**	0.484	0.442	**0.326**	**0.535**
InstructBLIP	0.000	0.061	0.143	0.069	0.063
MiniGPT-4	0.455	0.257	0.186	0.172	0.283
mPLUG_Owl	0.495	0.216	0.241	0.237	0.313
VisualGLM	0.281	0.027	0.045	0.016	0.107
Chinese_LLaVA	0.532	**0.531**	0.303	0.201	0.406
LLaVA-1.5	0.670	0.347	0.337	0.246	0.422
Qwen-VL-Chat	0.794	0.316	**0.543**	0.245	0.502

Table 3: Evaluation Results on CONSTRUCTURE Benchmark.

Du et al., 2022; Liu et al., 2024; Bai et al., 2023). All open-source MLLMs are ~7B models. We conduct zero-shot setups in overall evaluation for all MLLMs across our CONSTRUCTURE benchmark. Since our benchmark includes both Chinese and English versions, we evaluate MLLMs with Chinese capabilities in a Chinese context and those without in an English context.

Evaluation Methods. In the CONSTRUCTURE benchmark, test samples are organized in a multiple-choice format. The questions are divided into different sets: atomic concept understanding questions, which include "true" and "false" as the two options, and three other reasoning question sets, each offering four challenging options. Only one of these options is the correct answer. We input the image and the question to query MLLMs to get a response. Following (Huang et al., 2024), we then employ regular expression matching to parse the model outputs and calculate the accuracy by comparing it with the ground truth.

4.2 Main Results

Overall Results. We first present the overall results of the evaluated models on our CONSTRUCTURE benchmark as shown in Table 3. Current MLLMs, whether open-source or proprietary, demonstrate some level of concept understanding, but their ability to comprehend conceptual structures remains limited. The top-performing model, GPT-4o-0513, achieved an average score of only 0.621, with even lower average scores of 0.498 across the three conceptual structure reasoning tasks. There is significant variability in performance across different reasoning tasks, and all MLLMs struggle particularly with Common Ancestor Reasoning, indicating its ongoing challenges. Overall, GPT-4o-0513 and GPT-4-vision-preview stand out among API-based models. Although API-based models generally outperform open-source models, open-source models like BLIP2 and Qwen-VL-Chat achieve competitive results, surpassing models such as Gemini. This indicates that open-source approaches are capable of producing high-performing models in certain areas.

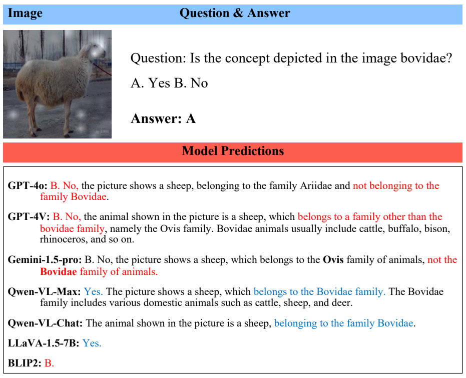


Image **Question & Answer**

Question: Is the concept depicted in the image bovidae?

A. Yes B. No

Answer: A

Model Predictions

GPT-4o: B. No, the picture shows a sheep, belonging to the family Ariidae and not belonging to the family Bovidae.

GPT-4V: B. No, the animal shown in the picture is a sheep, which belongs to a family other than the bovidae family, namely the Ovis family. Bovidae animals usually include cattle, buffalo, bison, rhinoceros, and so on.

Gemini-1.5-pro: B. No, the picture shows a sheep, which belongs to the **Ovis** family of animals, not the **Bovidae** family of animals.

Qwen-VL-Max: Yes. The picture shows a sheep, which belongs to the Bovidae family. The Bovidae family includes various domestic animals such as cattle, sheep, and deer.

Qwen-VL-Chat: The animal shown in the picture is a sheep, belonging to the family Bovidae.

LLaVA-1.5-7B: Yes.

BLIP2: B.

Figure 3: Case study of Atomic Concept Understanding Task.

Results on Atomic Concept Understanding. MLLMs generally perform well on the atomic concept understanding task. Among the API-based models, GPT-4o achieves nearly 90% accuracy, indicating a robust understanding of visual concepts across different levels. In the open-source models, both BLIP2 and Qwen-VL-Chat achieve a score of 0.794, surpassing most API-based models and only slightly behind GPT-4o. We observe that models performing well in the atomic concept understanding task also tend to excel in the other three reasoning tasks, demonstrating a positive correlation between atomic concept understanding and concept structure reasoning performance. InstructBLIP, however, performs poorly across most tasks, frequently outputting incorrect answers or gibberish. VisualGLM also struggles, often disregarding task instructions and merely generating descriptions of the images. Figure 3 below illustrates the performance of different MLLMs on an atomic concept understanding task. In the example, GPT-4o, GPT-4V, and Gemini-1.5-pro correctly identified the image as a sheep but incorrectly answered that a sheep is not a bovine animal (when, in fact, it is). However, Qwen-VL-Max and Qwen-VL-Chat correctly identified that a sheep belongs to the bovine family. This indicates that not all MLLMs possess comprehensive knowledge of concept structures.

Results on Concept Abstraction Reasoning. In the concept abstraction reasoning task, GPT-4V achieved the highest score of 0.748, followed by Gemini-1.5-pro with 0.657. Among open-source models, Chinese_LLaVA performed the best, scoring 0.531. Figure 6 presents the responses of different MLLMs to a specific question in this task. In this example, GPT-4V, Qwen-VL-Chat, LLaVA-1.5, and BLIP2 all provided correct answers. However, both GPT-4o and Gemini-1.5-pro answered incorrectly, misunderstanding the concept abstraction and hierarchical relationship (*i.e.* waterbirds include both swimming birds and wading birds).

Results on Concept Concretization Reasoning. In the concept concretization reasoning task, GPT-4o once again achieved the highest score of 0.663, followed by Qwen-VL-Max with a score of 0.631. Among the open-source models, Qwen-VL-Chat performed the best with a score of 0.543, with BLIP2 coming in second at 0.442. Figure 7 illustrates the performance of different MLLMs on a specific question in this task. In this example, Gemini-1.5-Pro and BLIP2 answered correctly. Gemini-1.5-Pro provided the correct reasoning process, while the other models answered incorrectly: GPT-4o gave an incorrect response without engaging in reasoning, GPT-4v misidentified the concept in the image (it's a Mahi Mahi or dolphinfish, not a saury pike), and Qwen-VL-Max selected an overly broad option due to conservative answering.

Results on Common Ancestor Reasoning. In the common ancestor reasoning task, all MLLMs struggled, indicating it as the most challenging task for MLLMs in concept structure reasoning. The best performer was claude3_sonnet, achieving a score of 0.397. GPT-4V and GPT-4o scored 0.241 and 0.145, respectively. Among the open-source models, BLIP2 scored 0.326, surpassing both GPT-4V and GPT-4o but falling short of claude3_sonnet. Figure 8 illustrates the performance of different MLLMs on a specific question in this task. In this example, the image represents the structural formula of a chemical compound, acetamide. The question requires the model to adhere to two constraints: 1) it must differ from the concept shown in the image, and 2) it must also be a "compound". Only Qwen-VL-Chat provided the correct answer in the example. GPT-4o and Gemini-1.5-pro correctly identified that option B's "solution" is a mixture rather than a "compound" (a pure substance), but they provided an incorrect answer in the final step of reasoning, revealing inconsistency in their reasoning process. GPT-4V gave a hallucinative answer, and its reasoning process was also incorrect. Qwen-VL-Max's answer did not meet the first requirement. Therefore, we observe that MLLMs perform poorly in tasks like Common Ancestor Reasoning due to various reasons, such as errors in understanding images or concepts, inconsistency in generation or reasoning processes, and others.

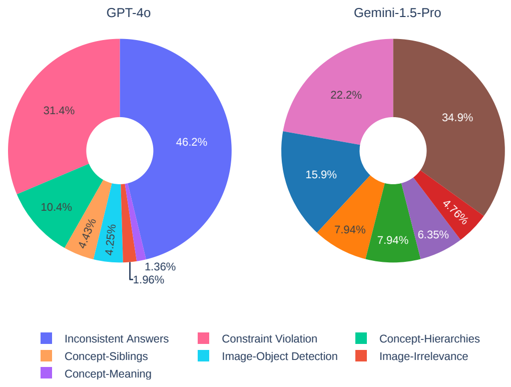


Figure 4: Distribution of Error Types for GPT-4o and Gemini-1.5-Pro.

4.3 Analysis

Error Type Analysis. To further analyze the error type distribution of MLLMs, we examined two models: GPT-4o and Gemini-1.5-Pro. As shown in Figure 4, we finally categorize errors into seven types, focusing on the capabilities of MLLMs regarding images, concepts, and language. The error types are as follows: 1) Image-Irrelevance: This error occurs when models provide answers that do not take the content of the provided images into account. 2)Image-Object Detection: This refers to models failing to recognize key objects or misidentifying them. For example, a cow might be incorrectly detected as an antelope. 3)Concept-Hierarchies: models have difficulties in understanding hierarchical relationships between concepts and identifying which one is in a higher level. 4) Concept-Meaning: models fail to comprehend the meanings of options because they don't know which concepts include the objects in the images. 5) Concept-Siblings: models have difficulties in recognizing sibling relationships between concepts. For example, the models doesn't recognize that "antelopes" and "yaks", which are hyponyms of "bovine animals", are sibling nodes, distinct from "deer". 6) Constraint Violation: models have difficulties in understanding the constraints of a given task. Most of errors in common ancestor reasoning fall into this category. 7) Inconsistent Answers: models produce conflicting answers or multiple answers inside the analysis of answers, unable to reach a definitive conclusion.

Analyzing the Impact of Concept Abstraction Levels. We analyze the impact of conceptual abstraction levels on model performance in understanding conceptual structures. We find that as the level of concept abstraction increases, the model performance declines. See detailed analyses in Appendix A.3.

4.4 CoT Reasoning and Finetuning

This section explores CoT reasoning and finetuning methods to enhance the reasoning abilities of MLLMs. We conducted experiments on the last three reasoning tasks in our CONSTRUCTURE benchmark. We will first present the overall results, followed by a detailed discussion of each experimental setting.

Overall Results. As shown in Table 5, across the three concept structure reasoning tasks, GPT-4V emerged as the top performer among all baselines, achieving an average score of 0.537, followed by GPT-4o at 0.498. We conducted separate experiments with Zero-shot CoT prompting and Few-shot CoT prompting based on GPT-4o. The results revealed that Few-shot CoT significantly enhanced GPT-4o's concept structure reasoning capabilities, increasing its average score from 0.498 to 0.699—an improvement of over 20 points. This improvement surpassed GPT-4v by more than 16 points. Notably, in the Common Ancestor Reasoning task, the score rose from 0.145 to 0.529, marking an impressive increase of nearly 38 percentage points. Conversely, Zero-shot CoT did not yield significant improvements. Furthermore, through fine-tuning on the constructed training and validation sets, we achieved the most substantial performance enhancement with Qwen-VL-Chat, reaching a score of 0.74. This performance surpassed even the best baseline, GPT-4V, which scored 0.537. We also conduct experiments using four non-multimodal LLMs with a known ground truth concept that specifies the concept present in the image. As observed, these non-multimodal LLMs demonstrate significantly poor performance, often falling below random guessing levels (~0.25). This stark difference highlights that the capability for concept structure reasoning is significantly enhanced by multimodal input, as these 'blind' LLMs lack the contextual integration provided by multimodal data.

Zero-shot CoT. We evaluate and report on two empirically derived zero-shot CoT prompts and two multimodal zero-shot CoT prompts. Our experiments reveal that these prompts do not enhance model performance on conceptual structure reasoning tasks. Specifically, we randomly sample 200 error cases from the GPT-4o conceptual structure reasoning tasks and apply five empirically designed zero-shot CoT prompts.(see detailed prompt in Appendix B). Based on the error correction rate, we select the two best-performing zero-shot CoT prompts for evaluation on the full test set and report the results. As shown in Table 5, although these zero-shot CoT prompts achieve up to a 30% error correction rate in the sampled error cases, they lead to a performance decline when applied to the full test set, with scores dropping from 0.498 to 0.479 and 0.451, respectively. This indicates that while these prompts slightly improve performance on error cases, particularly in the Common Ancestor reasoning task (from 0.145 to 0.165), they negatively impact Concretization Reasoning (from 0.662 to 0.575 and 0.495), leading to an overall decline in performance. This suggests that prompts like "Let's think step by step." do not significantly help with concept structure reasoning problems.

For the multimodal CoT prompt, we evaluate using the m-cot and v-cot methods (Singh et al., 2023; Chen et al., 2024). As shown in the results, the multimodal CoT with m-cot achieved the highest score of 0.507 in zero-shot settings. However, the improvement over the zero-shot method (0.498) is minimal, and the score remains significantly lower than the few-shot CoT performance (0.699). This result demonstrates that even multimodal zero-shot CoT methods struggle to assist with concept structure reasoning tasks. We speculate that this is because these tasks primarily evaluate MLLMs' understanding of relationships between concepts, which requires a more refined reasoning process to solve the inference between visual inputs and conceptual relationships. Since the model has not learned to reason from multimodal inputs to concepts or to infer the transitions between different levels of concepts, zero-shot CoT fails to be effective.

Few-shot CoT. We use few-shot CoT prompts written by human experts for each question type (detailed prompts can be found in the Appendix B). The human-designed few-shot CoT samples thoroughly address issues identified in previous error analyses, such as GPT-4o not following question instructions and having inconsistent reasoning processes. The CoT prompts instruct the model to answer questions according to the given instructions and to analyze each option one by one. In the samples, the prompts first break down the multiple constraints of the question and then analyze whether each option meets these constraints. By guiding GPT-4o through this reasoning paradigm, its ability to solve conceptual structure reasoning problems is significantly enhanced. This demonstrates that following a proper reasoning process and possessing strong reasoning capabilities are crucial for improving the model's concept structure reasoning.

Fine-tuning. Additionally, we conduct fine-tuning experiments on several MLLMs (including mPLUG-owl, LLaVA-1.5, and Qwen-VL-Chat) using our training and validation sets. The fine-tuned models, trained on the specialized concept structure reasoning dataset, achieve significant improvements of 20-30 percentage points, with Qwen-VL-Chat achieving a top performance score of 0.740. The results show that fine-tuning can improve concept structure reasoning capabilities, indicating that MLLMs may lack certain concept-related knowledge. To further understand the impact of atomic concept understanding data, we perform an ablation study by removing the atomic concept understanding subset from the training data and training for the same number of epochs. These results show a decrease in model performance, demonstrating that the inclusion of atomic concept understanding data is beneficial for enhancing the model's concept structure reasoning abilities.

Discussion. We summarize three key insights from experiments on enhancing MLLMs' concept structure reasoning abilities: 1) **Zero-CoT prompting does not significantly improve MLLMs' performance in concept structure reasoning.** This may be because step-by-step reasoning prompts can increase the likelihood of the model focusing on erroneous interference items, leading to reasoning errors. 2) **Few-shot CoT prompting that integrates human expert reasoning process priors can significantly enhance MLLMs' concept structure reasoning abilities.** The few-shot CoT used specifically designed problem-solving processes as prompts, providing more explicit reasoning guidance compared to the general CoT prompts used in v-cot and m-cot. We hypothesize that multimodal concept structure reasoning requires deeper visual concept metacognition, which cannot be easily addressed by generic problem-solving processes like v-cot and m-cot. Instead, it requires profound conceptual knowledge and related reasoning abilities. Proper conceptual reasoning prompts can better guide MLLMs in concept structure reasoning. 3) **Fine-tuning based on concept structure reasoning data can significantly improve MLLMs' performance.** This indicates that MLLMs may lack certain concept structure-related knowledge, highlighting the significant value of our proposed concept structure reasoning dataset.

5 Conclusion

We introduce the CONSTRUCTURE benchmark, which evaluates MLLMs' cognitive and reasoning abilities in tasks like atomic concept understanding, concept abstraction reasoning, concept concretization reasoning, and common ancestor reasoning. Our findings highlight significant challenges in concept structure reasoning for MLLMs. The top-performing model, GPT-4o, achieved an average score of 0.621, indicating room for improvement. We summarize current evaluations of MLLMs in concept structure reasoning, analyze reasons for their underperformance, and provide key insights from experiments using CoT prompting and fine-tuning to enhance their abilities. Our discoveries offer crucial guidance for advancing MLLMs' cognitive capabilities in concept structure reasoning.

Limitation

Since our concept chains are based on a Chinese taxonomy, there may be some language bias during translation into English. Despite extensive manual checks, our data annotations might still contain a few inaccuracies due to errors in the raw data, influenced by the annotators' understanding of the correct answers. Additionally, we primarily evaluate the capabilities of MLLMs within a Chinese context. Given the varying proficiency of different models in Chinese and English, the results may exhibit some variations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62072323), Shanghai Science and Technology Innovation Action Plan (No. 22511104700), and Alibaba Group through the Alibaba Innovation Research Program.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Agostina Calabrese, Michele Bevilacqua, Roberto Navigli, et al. 2020. Fatality killed the cat or: Babelpic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4680–4686. Association for Computational Linguistics.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1254–1262.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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

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

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

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

Douglas L Medin. 1989. Concepts and conceptual structure. *American psychologist*, 44(12):1469.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, and Gust Verbruggen. 2023. Assessing gpt4-v on structured reasoning tasks. *arXiv preprint arXiv:2312.11524*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. 2023. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 722–738.

John R. Taylor. 2019. 2. prototype theory. In Claudia Maienborn, Klaus Heusinger, and Paul Portner, editors, *Semantics - Theories*, pages 29–56. De Gruyter Mouton, Berlin, Boston.

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*.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

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*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhiwei Zha, Jiaan Wang, Zhixu Li, Xiangru Zhu, Wei Song, and Yanghua Xiao. 2023. M2conceptbase: A fine-grained aligned multi-modal conceptual knowledge base. *arXiv preprint arXiv:2312.10417*.

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*.

A Appendix

A.1 Dataset Statistics

Dataset statistics details are shown in Table 4. The top-level concept frequency distribution is shown in Figure 5.

A.2 Case Study of Zero-shot Evaluation

See case study of concept abstraction reasoning task in Figure 6. See case study of concept concretization reasoning task in Figure 7. See case study of common ancestor reasoning task in Figure 7.

A.3 Evaluation Analylis Details.

Since each question in our benchmark is generated based on concepts within a conceptual chain, effectively examining the conceptual structure surrounding a specific level of abstraction, including superordinate, subordinate, and sibling concepts. To quantify this, we designate the most abstract concept level as 0, with subsequent subordinate levels labeled as 1, 2, 3, and so on, up to a maximum of 5. Based on this framework, we categorize the questions into five types, each corresponding to a different level of conceptual abstraction. We then evaluate and compare the performance of both API-based models and open-source models across these abstraction levels. Figures 9 and 10 show the performance of API-based MLLMs and open-source MLLMs on questions at different levels of conceptual abstraction, respectively. We observed a consistent trend across both types of models: as the level of concept abstraction increases, model

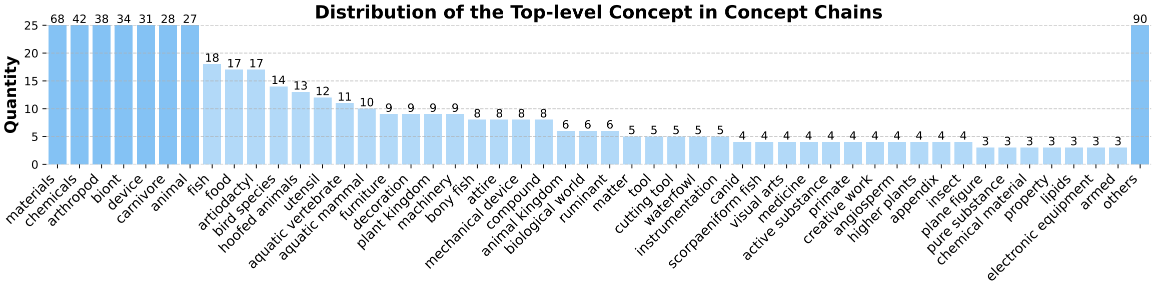


Figure 5: Top-level Concept Frequency Distribution.

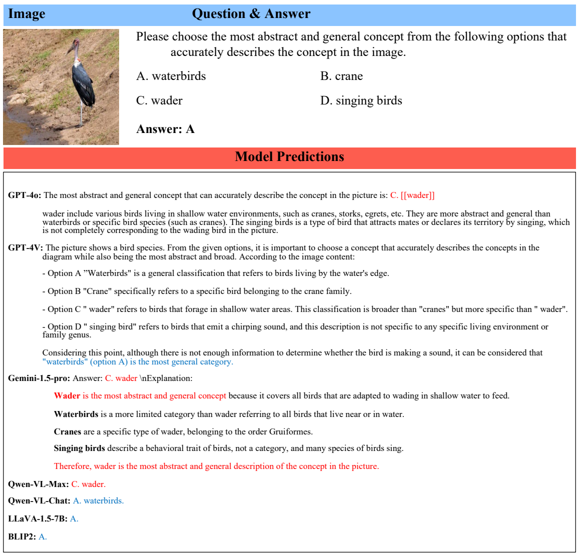


Image
Question & Answer
Model Predictions

Please choose the most abstract and general concept from the following options that accurately describes the concept in the image.

A. waterbirds B. crane

C. wader D. singing birds

Answer: A

GPT-4o: The most abstract and general concept that can accurately describe the concept in the picture is: C. [[wader]]

wader include various birds living in shallow water environments, such as cranes, storks, egrets, etc. They are more abstract and general than waterbirds or specific bird species (such as cranes). The singing birds is a type of bird that attracts mates or declares its territory by singing, which is not completely corresponding to the wading bird in the picture.

GPT-4V: The picture shows a bird species. From the given options, it is important to choose a concept that accurately describes the concepts in the diagram while also being the most abstract and broad. According to the image content:

- Option A ”Waterbirds" is a general classification that refers to birds living by the water's edge.

- Option B "Crane" specifically refers to a specific bird belonging to the crane family.

- Option C " wader" refers to birds that forage in shallow water areas. This classification is broader than "cranes" but more specific than " wader".

- Option D " singing bird" refers to birds that emit a chirping sound, and this description is not specific to any specific living environment or family genus.

Considering this point, although there is not enough information to determine whether the bird is making a sound, it can be considered that "waterbirds" (option A) is the most general category.

Gemini-1.5-pro: Answer: C. wader \nExplanation:

Wader is the most abstract and general concept because it covers all birds that are adapted to wading in shallow water to feed.

Waterbirds is a more limited category than wader referring to all birds that live near or in water.

Cranes are a specific type of wader, belonging to the order Gruiformes.

Singing birds describe a behavioral trait of birds, not a category, and many species of birds sing.

Therefore, wader is the most abstract and general description of the concept in the picture.

Qwen-VL-Max: C. wader.

Qwen-VL-Chat: A. waterbirds.

LLaVA-1.5-7B: A.

BLIP2: A.

Figure 6: Case study of Concept Abstraction Reasoning Task.

Subset	# of Concepts	# of Chains	Avg. Chain Length	# of Images
Train	1,312	638	4.94	636
Valid	1,144	534	4.93	521
Test	1,280	615	4.91	604
Total	**1,315**	**646**	**4.93**	**644**

Table 4: Detail Statistics of CONSTRUCTURE.

performance declines. This result indicates that MLLMs perform better in understanding and reasoning about more abstract concepts compared to more concrete ones. Based on prior analysis, models struggle with recognizing and reasoning about the hierarchical structure of fine-grained, concrete concepts, often lacking the necessary knowledge or understanding of these detailed concepts. This finding highlights a direction for further improving MLLMs by enhancing their capability to perceive and comprehend the structure of fine-grained concepts.

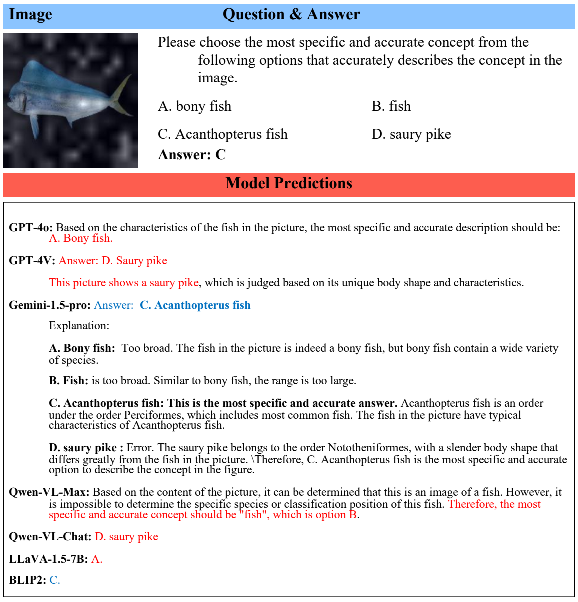


Image
Question & Answer
Model Predictions

Please choose the most specific and accurate concept from the following options that accurately describes the concept in the image.

A. bony fish B. fish

C. Acanthopterus fish D. saury pike

Answer: C

GPT-4o: Based on the characteristics of the fish in the picture, the most specific and accurate description should be: A. Bony fish.

GPT-4V: Answer: D. Saury pike

This picture shows a saury pike, which is judged based on its unique body shape and characteristics.

Gemini-1.5-pro: Answer: **C. Acanthopterus fish**

Explanation:

A. Bony fish: Too broad. The fish in the picture is indeed a bony fish, but bony fish contain a wide variety of species.

B. Fish: is too broad. Similar to bony fish, the range is too large.

C. Acanthopterus fish: This is the most specific and accurate answer. Acanthopterus fish is an order under the order Perciformes, which includes most common fish. The fish in the picture have typical characteristics of Acanthopterus fish.

D. saury pike : Error. The saury pike belongs to the order Nototheniformes, with a slender body shape that differs greatly from the fish in the picture. \Therefore, C. Acanthopterus fish is the most specific and accurate option to describe the concept in the figure.

Qwen-VL-Max: Based on the content of the picture, it can be determined that this is an image of a fish. However, it is impossible to determine the specific species or classification position of this fish. Therefore, the most specific and accurate concept should be "fish", which is option B.

Qwen-VL-Chat: D. saury pike

LLaVA-1.5-7B: A.

BLIP2: C.

Figure 7: Case study of Concept Concretization Reasoning Task.

A.4 Experimental Results of CoT

See Experimental Results of Chain-of-Thoughts Reasoning in Table 5.

B Prompts

We present specific CoT reasoning prompts. For Zero-shot CoT in Table 5, prompt_1 is the first one in the list, and prompt_2 is the last one in the list.

Zero CoT Prompts.

1. *Let's think step by step.*
2. *Let's analyze each option one by one.*
3. *Let's analyze each option according to the requirements of the question.*
4. *Let's first identify the possible concepts in the image, and then analyze them one by one.*

Model	Abstraction Reasoning	Concretization Reasoning	Common Ancestor	Avg. Score
MLLMs (Zero-shot)				
gpt-4-vision-preview	**0.748**	0.602	**0.241**	**0.537**
gpt-4o-0513	0.657	**0.662**	0.145	0.498
Qwen-VL-Max	0.543	0.631	0.143	0.444
mPLUG-Owl	0.216	0.241	0.237	0.225
Qwen-VL-Chat	0.316	0.543	0.245	0.372
LLaVA-1.5	0.347	0.337	0.246	0.310
BLIP2-flant5-xl	0.484	0.442	0.326	0.419
LLMs + Oracle Concept				
LLaMA2	0.045	0.039	0.036	0.040
LLaMA3-Instruct	0.204	0.037	0.313	**0.181**
Qwen-Chat	0.122	0.065	0.089	0.093
Qwen1.5-Chat	0.018	0.010	0.020	0.016
Zero-shot CoT				
gpt-4o-0513 + *prompt_1*	0.669	0.575	0.165	0.479
gpt-4o-0513 + *prompt_2*	0.669	0.495	0.165	0.451
gpt-4o-0513 + *v_cot*	0.649	0.544	0.150	0.456
gpt-4o-0513 + *m_cot*	0.726	0.574	0.192	0.507
Qwen-VL-Chat + *v_cot*	0.182	0.303	0.163	0.217
Qwen-VL-Chat + *m_cot*	0.365	0.466	0.192	0.345
Few-shot CoT				
gpt-4o-0513(1-shot)	**0.859**	0.695	0.529	**0.699**
gpt-4o-0513(2-shot)	0.826	**0.723**	0.493	0.686
Finetuning				
mPLUG-owl-sft	**0.595**	0.419	**0.353**	**0.446**
LLaVA-1.5-sft	**0.724**	0.618	**0.627**	**0.639**
Qwen-VL-Chat-sft	0.857	**0.720**	0.622	0.737
Qwen-VL-Chat-sft w. atomic	**0.859**	0.716	**0.636**	**0.740**

Table 5: Results of Chain-of-Thoughts Experiments.

5. *Let’s analyze by comparing the image with each option.*
6. *Let’s analyze by comparing the image with each option according to the requirements of the question.*

Few CoT Prompt for Concept Abstraction Reasoning.

“<img>This is an image of a silver fox.</img>
Prompt: Please select from the following options the one that correctly describes the concept in the image and is the most abstract and general concept.
A. Bear
B. Fox
C. Vixen
D. Canid
Answer: Let’s analyze each option according to the prompt requirements. First, the prompt requires 1) selecting an option that correctly describes the concept in the image, and 2) is the most abstract and general concept. Next, we analyze each option:
A. Bear - The concept in the image is a silver fox, which is different from a bear, so this option does not meet the requirements.
B. Fox - The concept in the image is a silver fox, which is a type of fox, so this option meets the first requirement.
C. Vixen - The concept in the image is a fox, but it’s difficult to determine if it’s a vixen or a male fox, so this option does not meet the first requirement.
D. Canid - The concept in the image is a fox, which belongs to the Canidae family, so this option meets the first requirement. Now, between option B. Fox and option D. Canid, we select the most abstract and general concept. A fox belongs to the Canidae family, so Canid is the more abstract concept.
Therefore, the correct answer is D. Canid.”

Few CoT Prompt for Concept Concretization Reasoning.

“<img>This is an image of a humpback whale.</img>
Prompt: Please select from the following options the one that correctly describes the concept in the image and is the most specific and accurate concept.
A. Sirenians
B. Cetaceans
C. Aquatic mammals

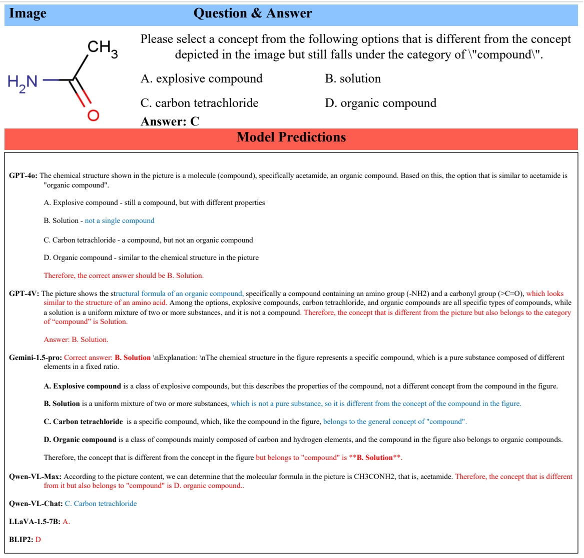


Image **Question & Answer**

Please select a concept from the following options that is different from the concept depicted in the image but still falls under the category of \"compound\".

A. explosive compound B. solution

C. carbon tetrachloride D. organic compound

Answer: C

Model Predictions

GPT-4o: The chemical structure shown in the picture is a molecule (compound), specifically acetamide, an organic compound. Based on this, the option that is similar to acetamide is "organic compound".

A. Explosive compound - still a compound, but with different properties

B. Solution - not a single compound

C. Carbon tetrachloride - a compound, but not an organic compound

D. Organic compound - similar to the chemical structure in the picture

Therefore, the correct answer should be B. Solution.

GPT-4V: The picture shows the structural formula of an organic compound, specifically a compound containing an amino group (-NH2) and a carbonyl group (>C=O), which looks similar to the structure of an amino acid. Among the options, explosive compounds, carbon tetrachloride, and organic compounds are all specific types of compounds, while a solution is a uniform mixture of two or more substances, and it is not a compound. Therefore, the concept that is different from the picture but also belongs to the category of "compound" is Solution.

Answer: B. Solution.

Gemini-1.5-pro: Correct answer: **B. Solution** \nExplanation: \nThe chemical structure in the figure represents a specific compound, which is a pure substance composed of different elements in a fixed ratio.

A. Explosive compound is a class of explosive compounds, but this describes the properties of the compound, not a different concept from the compound in the figure.

B. Solution is a uniform mixture of two or more substances, which is not a pure substance, so it is different from the concept of the compound in the figure.

C. Carbon tetrachloride is a specific compound, which, like the compound in the figure, belongs to the general concept of "compound".

D. Organic compound is a class of compounds mainly composed of carbon and hydrogen elements, and the compound in the figure also belongs to organic compounds.

Therefore, the concept that is different from the concept in the figure but belongs to "compound" is **B. Solution**.

Qwen-VL-Max: According to the picture content, we can determine that the molecular formula in the picture is CH3CONH2, that is, acetamide. Therefore, the concept that is different from it but also belongs to "compound" is D. organic compound..

Qwen-VL-Chat: C. Carbon tetrachloride

LLaVA-1.5-7B: A.

BLIP2: D

Figure 8: Case study of Common Ancestor Reasoning Task.

D. Baleen whales
Answer: Let's analyze each option according to the prompt requirements. First, the prompt requires 1) selecting an option that correctly describes the concept in the image, and 2) is the most specific and accurate concept. Next, we analyze each option:
A. Sirenians - The concept in the image is a humpback whale, not a sirenian, so this option does not meet the first requirement.
B. Cetaceans - The concept in the image is a humpback whale, which is a cetacean, so this option meets the first requirement.
C. Aquatic mammals - The concept in the image is a humpback whale, which is an aquatic mammal, so this option meets the first requirement.
D. Baleen whales - The concept in the image is a humpback whale, which is a type of baleen whale, so this option meets the first requirement. Now, between options B. Cetaceans, C. Aquatic mammals, and D. Baleen whales, the most specific concept is Baleen whales.
Therefore, the correct answer is D. Baleen whales."

Few CoT Prompt for Common Ancestor Reasoning.

"<img>This is an image of a butterflyfish.</img>
Prompt: Please select from the following options the concept that is different from the image but belongs to the same 'Perciformes' group.
A. Sturgeon
B. Surgeonfish
C. Crocodile
D. Perciformes
Answer: Let's analyze each option according to the prompt requirements. First, the prompt requires selecting a concept that 1) is different from the image, and 2) belongs to the 'Perciformes' group. Next, we analyze each option:
A. Sturgeon - The concept in the image is a butterflyfish, which is different from a sturgeon, so this option meets the first requirement, but sturgeon belongs to the order Acipenseriformes, not Perciformes, so this option does not meet the second requirement.
B. Surgeonfish - The concept in the image is a butterflyfish, not a surgeonfish, so this option meets the first requirement. Surgeonfish belong to the order Acanthuriformes, which is not Perciformes, so this option does not meet the second requirement.
C. Crocodile - Crocodiles are reptiles and do not belong to Perciformes, so this option does not meet the second requirement.
D. Perciformes - The concept in the image is a butterflyfish, which belongs to the Perciformes order, so this option meets the second requirement. Now, between options A. Sturgeon and D. Perciformes, we select the concept that is different but belongs to the same group, which is D. Perciformes.
Therefore, the correct answer is B. Surgeonfish."

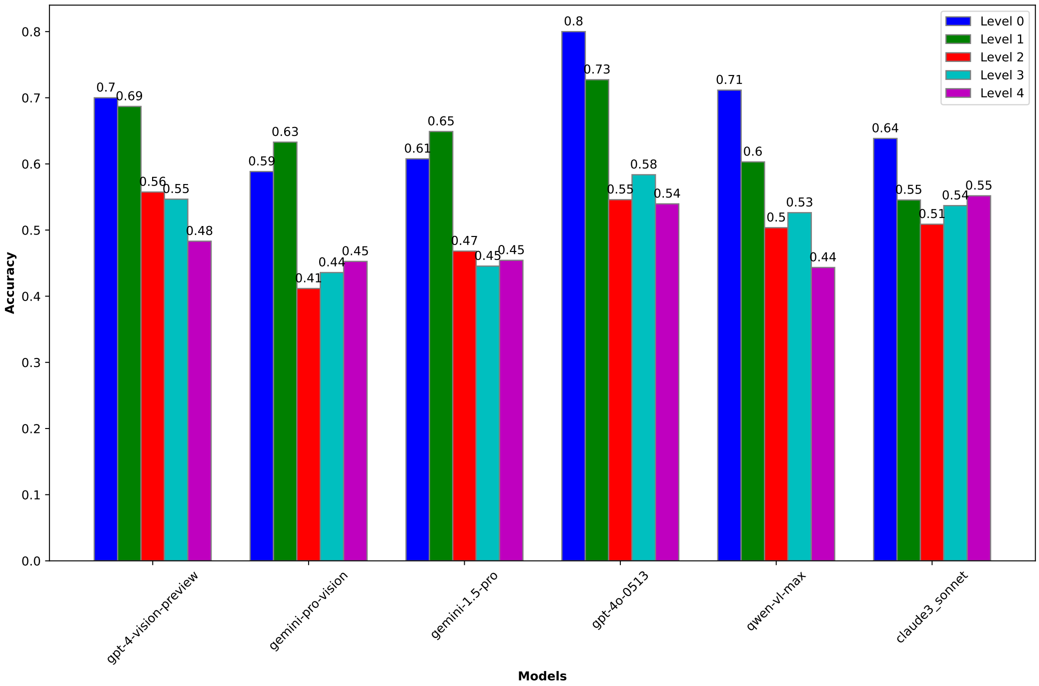


Figure 9: Level-wise Accuracy of API-based MLLMs.

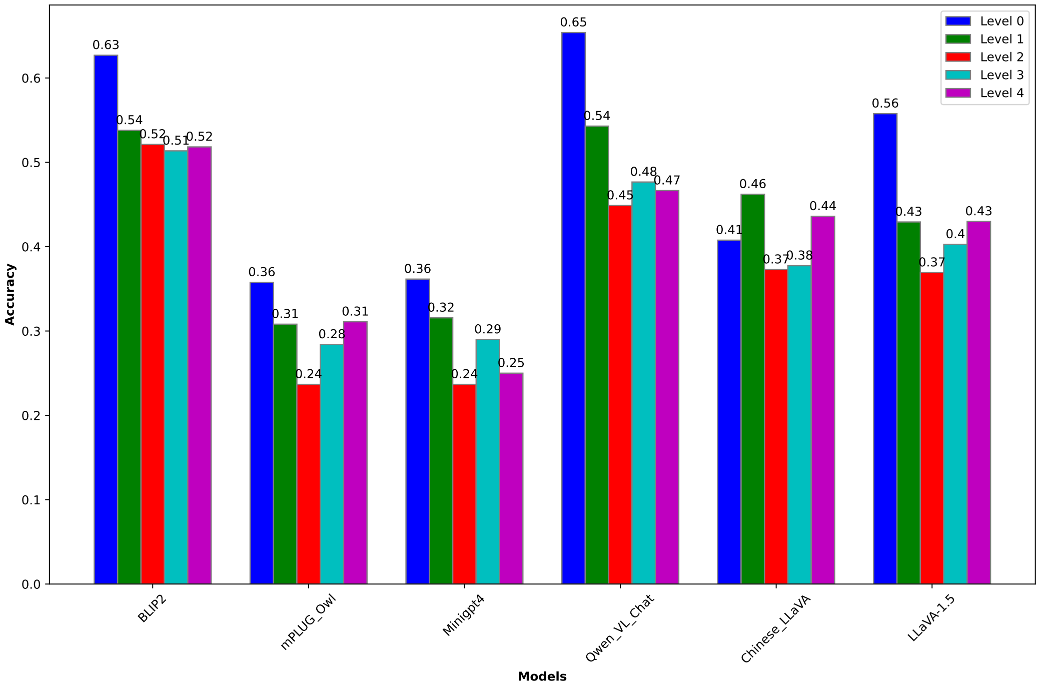


Figure 10: Level-wise Accuracy of Open-source MLLMs.