

MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

¹Xiang Yue^{*†}, ²Yuansheng Ni*, ³Kai Zhang*, ⁴Tianyu Zheng*,
³Ruoqi Liu, ²Ge Zhang, ³Samuel Stevens, ²Dongfu Jiang, ²Weiming Ren, ⁴Yuxuan Sun,
²Cong Wei, ³Botao Yu, ⁵Ruimin Yuan, ²Renliang Sun, ⁷Ming Yin,
³Boyan Zheng, ⁴Zhenzhu Yang, ⁶Yibo Liu, ⁴Wenhai Huang,
³Huan Sun*, ³Yu Su^{*†}, ²Wenhu Chen^{*†}

¹IN.AI Research, ²University of Waterloo, ³The Ohio State University, ⁴Independent,
⁵Carnegie Mellon University, ⁶University of Victoria, ⁷Princeton University

<https://mmmu-benchmark.github.io/>

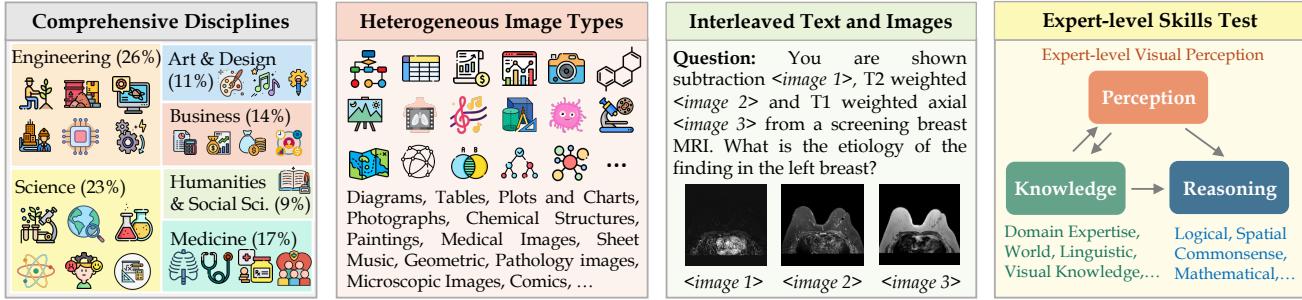


Figure 1. Overview of the MMMU dataset. MMMU presents four challenges: 1) **comprehensiveness**: 11.5K college-level problems across six broad disciplines and 30 college subjects; 2) highly **heterogeneous** image types; 3) **interleaved** text and images; 4) **expert-level** perception and reasoning rooted in deep subject knowledge.

Abstract

We introduce MMMU: a new benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. MMMU includes 11.5K meticulously collected multimodal questions from college exams, quizzes, and textbooks, covering six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. These questions span 30 subjects and 183 subfields, comprising 30 highly heterogeneous image types, such as charts, diagrams, maps, tables, music sheets, and chemical structures. Unlike existing benchmarks, MMMU focuses on advanced perception and reasoning with domain-specific knowledge, challenging models to perform tasks akin to those faced by experts. The evaluation of 14 open-source LMMs as well as the proprietary GPT-4V(ision) and Gemini highlights the substantial

challenges posed by MMMU. Even the advanced GPT-4V and Gemini Ultra only achieve accuracies of 56% and 59% respectively, indicating significant room for improvement. We believe MMMU will stimulate the community to build next-generation multimodal foundation models towards expert artificial general intelligence.

1. Introduction

Rapid advances in large language models (LLMs) [11, 45, 52] have sparked broad discussions on the controversial concept of artificial general intelligence (AGI), often used to describe AI systems that perform on par or surpass humans at most tasks [1, 7, 17, 24, 42, 44]. Candid and constructive discussions on AGI have been challenging due to a lack of shared operationalizable definitions. In an attempt to remedy this, Morris et al. [44] propose a leveled taxonomy for AGI that centers around both *generality* (or breadth) and *performance* (or depth). In the suggested taxonomy, Level 3, or *Expert AGI*, marks a critical milestone. It denotes an

*Core Contributors. See the Author Contribution Statement for details.

†✉: xiangyue@in.ai; su.809@osu.edu; wenhuchen@uwaterloo.ca

Art & Design	Business	Science
<p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <ul style="list-style-type: none"> (A) Major third (B) Diminished fifth (C) Minor seventh (D) Diminished sixth 	<p>Question: ...The graph shown is compiled from data collected by Gallup .</p> <p>Options:</p> <ul style="list-style-type: none"> (A) 0 (B) 0.2142 (C) 0.3571 (D) 0.5 	<p>Question: The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <ul style="list-style-type: none"> (A) $\int_0^{1.5} [f(x) - g(x)] dx$ (B) $\int_0^{1.5} [g(x) - f(x)] dx$ (C) $\int_0^2 [f(x) - g(x)] dx$ (D) $\int_0^2 [g(x) - x] dx$
<p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>
Health & Medicine	Humanities & Social Science	Tech & Engineering
<p>Question: You are shown subtraction , T2 weighted , and T1 weighted axial , from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <ul style="list-style-type: none"> (A) Susceptibility artifact (B) Hematoma (C) Fat necrosis (D) Silicone granuloma 	<p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? </p> <p>Option:</p> <ul style="list-style-type: none"> (A) Oppressor (B) Imperialist (C) Savior (D) Isolationist 	<p>Question: Find the VCE for the circuit shown in </p> <p>Answer: <u>3.75</u></p> <p>Explanation: ...IE = [(V_{EE}) / (R_E)] = [(5 V) / (4 k-ohm)] = 1.25 mA; V_{CE} = V_{CC} - I_ER_L = 10 V - (1.25 mA) 5 k-ohm; V_{CE} = 10 V - 6.25 V = 3.75 V</p>
<p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard</p>	<p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>

Figure 2. Sampled MMMU examples from each discipline. The questions and images need expert-level knowledge to understand and reason.

AI system that reaches “at least 90th percentile of skilled adults” in a broad range of tasks, thus starting to achieve “the substitution threshold for machine intelligence in lieu of human labor” for many industries, leading to significant risks of job displacement and economic disruption. Therefore, it is of both intellectual and societal importance to closely monitor the progress towards Expert AGI.

How to create benchmarks for measuring Expert AGI? Since the definition is based on comparison with *skilled adults*, a natural starting point is college-level exams for different disciplines, because those are designed to evaluate *skilled adults* specialized in each discipline. This strategy has been successfully adopted in benchmarks such as MMLU [19] and AGIEval [69], but only text-based questions are considered, while human experts are capable of solving multimodal problems. Meanwhile, large multimodal models (LMMs) that can understand both text and images have been making a major stride towards more general AI [8, 14, 27, 34, 58]. These LMMs have consistently excelled in existing multimodal benchmarks [3, 18, 25, 31, 36, 50, 61, 63]. For instance, CogVLM [55] achieves 85% on VQA-v2 [18], 92% on ScienceQA-IMG [39], and 93% on RefCOCO [23]. However, most existing multimodal benchmarks focus on commonsense/daily knowledge rather than expert-level domain knowledge and advanced reasoning. The closest one to our goal is ScienceQA [39]. While it covers diverse disciplines (**breadth**), the majority of the questions are at the elementary to the middle school level, thus falling short in **depth** for benchmarking Expert AGI.

To this end, we introduce MMMU: a comprehensive benchmark designed for college-level multi-discipline multimodal understanding and reasoning. It features problems sourced from college exams, quizzes, and textbooks spanning six common disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. MMMU consists of **11.5K** carefully selected multimodal questions, which cover **30** diverse subjects and 183 subfields, thus meeting the **breadth** goal. Moreover, many problems within MMMU require expert-level reasoning, such as applying “Fourier Transform” or “Equilibrium Theory” to derive the solution, thus meeting the **depth** goal. MMMU also presents two unique challenges absent in current benchmarks (Figure 1). Firstly, it covers diverse image formats, from visual scenes like photographs and paintings to diagrams and tables, testing the perceptual capabilities of LMMs. Secondly, MMMU features interleaved text-image inputs. A model needs to jointly understand the images and text, which often requires recalling deep subject knowledge, and conducting complex reasoning based on the understanding and knowledge to reach a solution.

We evaluate 14 open-source LMMs as well as the advanced proprietary LMMs such as GPT-4V(ision) [46] on MMMU. Our key findings are summarized as follows:

- MMMU presents significant challenges; notably, GPT-4V only achieves an accuracy of 55.7%, indicating substantial room for improvement.
- There is a pronounced disparity in performance between open-source LMMs and GPT-4V. The highest-performing

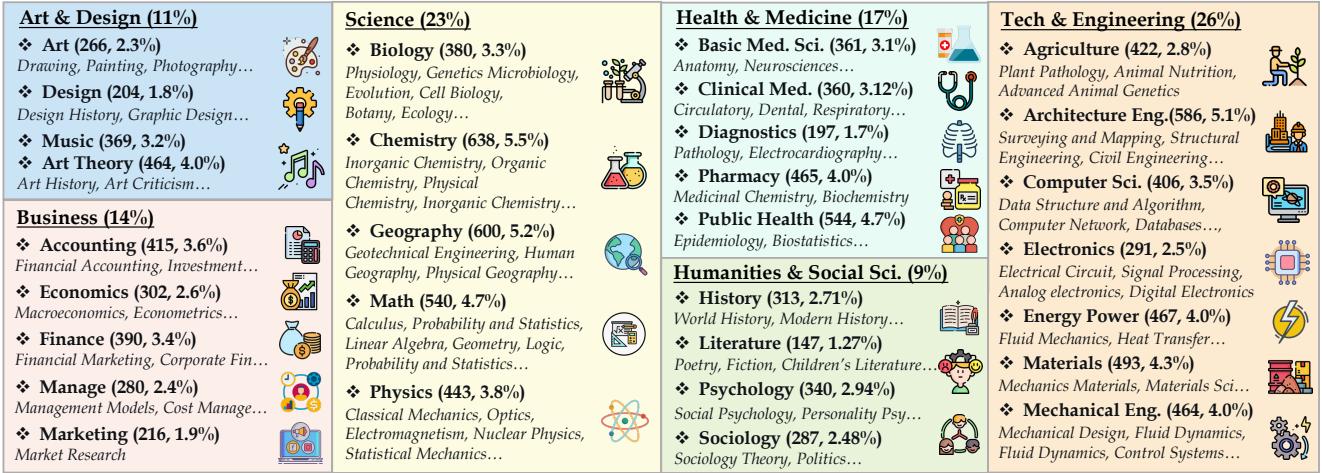


Figure 3. MMMU contains 11.5K multimodal questions covering six broad disciplines, 30 subjects, and 183 subfields.

open-source models, such as BLIP2-FLAN-T5-XXL and LLaVA-1.5, achieve approximately 34% in accuracy.

- LLMs augmented with optical character recognition (OCR) or generated captions do not see notable improvement, indicating that MMMU necessitates deeper joint interpretation of images and text.
- In disciplines such as Art & Design and Humanities & Social Science, where visual data is less complex, models exhibit higher performance. In contrast, Business, Science, Health & Medicine, and Tech & Engineering, which present more complex visual data and require intricate reasoning, see relatively lower model performance.
- Our error analysis on 150 error cases of GPT-4V reveals that 35% of errors are perceptual, 29% stem from a lack of knowledge, and 26% are due to flaws in the reasoning process. These findings underscore the challenges of the MMMU benchmark and point towards areas needing further research and model enhancement.

Our aim with MMMU is to push the boundaries of what LMMs can achieve. We believe it will prove instrumental in developing next-generation multimodal foundation models and monitoring the progress towards Expert AGI. We shall caution that MMMU is not a *sufficient* test for Expert AGI, as per the definition [44], because there lacks a direct mapping between performance on MMMU and “90th percentile of skilled adults,” nor are college exams the only tasks an AGI shall tackle. However, we believe it should be *necessary* for an Expert AGI to achieve strong performance on MMMU to demonstrate their broad and deep subject knowledge as well as expert-level understanding and reasoning capabilities.

2. Related Work

Multimodal Pre-Training. In recent years, rapid progress has been made in multimodal pre-training, which aims

to jointly encode vision and language in a fusion model. LXMERT [51], UNITER [9], VinVL [64], Oscar [29], VilBert [38], and VLP [70] are among the earliest work to train universal vision-language models to tackle many multimodal tasks. This work relies on pre-trained visual representations like Faster RCNN features [49] to minimize the training sample complexity. Later on, CLIP [48], ALIGN [22], SimVLM [56], CoCa [62], Flamingo [2], BLIP-2 [27], and Fuyu [6] (*inter alia*) have been proposed to train visual representation using ViT [15] from scratch with massive amount of web data. These models have achieved great success on existing VQA and captioning tasks, which require less knowledge and reasoning.

Multimodal Instruction Tuning. Inspired by open-source instruction-tuned LLMs like FLAN-T5 [12] and Vicuna [10], models like LLaVA [34, 35] and MiniGPT-4 [71] utilized open-source resources, to improve the instruction-following capabilities of LMMs. The evolutionary trajectory of LMMs has also led to subsequent advancements aimed at improving the quantity and quality of visual instruction data. Models such as LLaMA-Adapter [16, 65], mPlug-OWL [59, 60], SVIT [66], LRV-Instruction [33], and InstructBLIP [14] exemplify these developments. Another pivotal aspect of LMM research revolves around multimodal in-context learning and the management of interleaved text and image examples. This area has been explored in depth by models such as Flamingo [2] and OpenFlamingo [4], Otter [26], M3IT [28], MetaVL [43], Sparkles [20], and MMICL [67]. These models have significantly contributed to the ongoing advancements in multimodal training and instruction-following capabilities.

LMM Benchmarks. With the surge of multi-modal pre-training and instruction tuning, the prior single-task evaluation benchmarks like VQA [3, 18], OK-VQA [41], MSCOCO [31], GQA [21], etc., have become insufficient

Statistics	Number
Total Questions	11550
Total Disciplines/Subjects/Subfields	6/30/183
Image Types	30
Dev:Validation:Test	150:900:10500
Difficulties (Easy: Medium: Hard)	28%:45%:27%
Multiple-choice Questions	10861 (94.03%)
Open Questions	689 (5.97%)
Questions with an Explanation	2035 (17.62%)
Image in the Question	11264 (97.52%)
* Images at the beginning	2006 (17.81%)
* Images in the middle	4159 (36.92%)
* Images at the end	5679 (50.42%)
Image in Options	389 (3.37%)
Example with Multiple Images	854 (7.39%)
Average question length	59.33
Average option length	9.17
Average explanation length	107.92

Table 1. Key statistics of the MMMU benchmark.

to holistically evaluate LMMs’ general multimodal perception and reasoning abilities. Therefore, numerous all-round benchmarks have been established to assess different facets of LMMs. These benchmarks cover a wide spectrum of specific skills of LMMs, from Optical Character Recognition (OCR) as seen in the study by [37], to adversarial robustness [68] and hallucination [13, 32], e.g., POPE [30] and HaELM [54]. More holistic evaluations have been conducted as well, such as LAMM [61], LVLM-eHub [57], SEED [25], MMBench [36], and MM-Vet [63]. These benchmarks still largely focus on relatively basic perception abilities without requiring expert-level domain knowledge and deliberate reasoning. More recently, MathVista [40] presents a collection of visually challenging questions; however, its scope is limited exclusively to the mathematical domain. MMMU is highly different from these benchmarks by collecting more difficult expert-level problems that cover 30 different subjects and require nuanced perception, recalling domain-specific knowledge to perform step-by-step reasoning to derive the solution. In line with the motivation of our study, concurrently, GAIA [42] introduces 466 questions that test fundamental abilities of models such as reasoning, multimodality handling, or tool use.

3. The MMMU Benchmark

3.1. Overview of MMMU

We introduce the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark, a novel benchmark meticulously curated to assess the expert-level multimodal understanding capability of foundation models

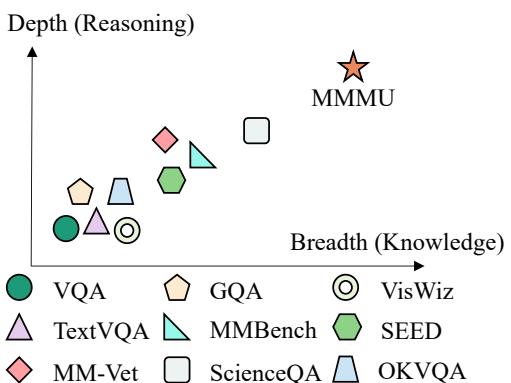
across a broad scope of tasks. Covering 30 subjects across 6 disciplines, including Art, Business, Health & Medicine, Science, Humanities & Social Science, and Tech & Engineering, and over 183 subfields. The detailed subject coverage and statistics are detailed in Figure 3. The questions in our benchmark were manually collected by a team of 50 college students (including coauthors) from various disciplines and subjects, drawing from online sources, textbooks, and lecture materials.

MMMU, constituting 11.5K questions, is divided into a few-shot development set, a validation set, and a test set. The few-shot development set includes 5 questions per subject, and the validation set, useful for hyperparameter selection, contains approximately 900 questions, while the test set comprises 10.5K questions. MMMU is designed to measure three essential skills in LMMs: perception, knowledge, and reasoning. Our aim is to evaluate how well these models can not only perceive and understand information across different modalities but also apply reasoning with subject-specific knowledge to derive the solution.

Our MMMU benchmark introduces four key challenges to multimodal foundation models, as detailed in Figure 1. Among these, we particularly highlight the challenge stemming from the requirement for both expert-level visual perceptual abilities and deliberate reasoning with subject-specific knowledge. This challenge is vividly illustrated through our tasks, which not only demand the processing of various heterogeneous image types but also necessitate a model’s adeptness in using domain-specific knowledge to deeply understand both the text and images and to reason. This goes significantly beyond basic visual perception, calling for an advanced approach that integrates advanced multimodal analysis with domain-specific knowledge.

3.2. Data Curation Process

Data Collection. Our benchmark collection takes three stages. Firstly, we go through the common university majors to decide what subjects should be included in our benchmark. The selection is based on the principle that visual inputs should be commonly adopted in the subjects to provide valuable information. Through this principle, we rule out a few subjects like law and linguistics because it is difficult to find enough relevant multimodal problems in these subjects. Consequently, we select 30 subjects from six different disciplines. In the second stage, we recruit over 50 university students, including co-authors, specializing in these majors as annotators to assist in question collection. They collect multimodal questions from major textbooks and online resources, creating new questions based on their expertise where necessary. The annotators are instructed to adhere to copyright and license regulations, avoiding data from sites prohibiting copy and redistribution. Given the arising data contamination concerns of foundation models,



Dataset	Size	Images	Format	Source	Answer
VQA	> 1M	V	I+T	Annotated	Open
GQA	> 1M	V	I+T	Synthesized	Open
VisWiz	32K	V	I+T	Annotated	Open
TextVQA	45K	OC	I+T	Annotated	MC
OKVQA	14K	V+OC	I+T	Annotated	Open
SEED	19K	V+OC	I+T	Annotated	MC
MMBench	3K	V+OC	I+T	Repurposed	MC
MM-Vet	0.2K	V+OC	I+T	Repurposed	MC
ScienceQA	6K	5 Types	I+T	Textbooks	MC
MMMU	11.5K	30 Types	Interleaved	Textbooks, Internet, Annotated	Open / MC

Figure 4. The comparison between MMMU and other existing benchmarks. MMMU excels in both its breadth to cover a wide range of disciplines and its depth to test LMMs’ reasoning abilities. In the image format, V means visual input, OC means optical characters, MC means multi-choice. Repurposed means the benchmark is a compilation of prior datasets.

the annotators are advised to select questions without immediately available answers, such as those with answers in separate documents or at the end of textbooks. This process results in a diverse collection of 13K questions from various sources. The detailed annotation protocol is in Appendix A.

Data Quality Control. To further control the quality of our data, we perform two steps of data cleaning. In the first stage, lexical overlap and source URL similarity are employed to identify potential duplicate problems. These suspected duplicates were then reviewed by the authors to identify and eliminate any duplications. The second stage involves distributing the problems among different co-authors for format and typo checking. This step requires authors to ensure adherence to a standardized format, undertaking necessary corrections where deviations are found. In the third and final stage, the authors categorize the problems into four difficulty levels: very easy, easy, medium, and hard. Approximately 10% of the problems, classified as very easy and not aligning with our design criteria due to their simplistic nature, are excluded from the benchmark. This rigorous process plays a crucial role in maintaining the quality and difficulty of the problem set.

3.3. Comparisons with Existing Benchmarks

To further distinguish the difference between MMMU and other existing ones, we elaborate the benchmark details in Figure 4. From the *breadth* perspective, the prior benchmarks are heavily focused on daily knowledge and common sense. The covered image format is also limited. Our benchmark aims to cover college-level knowledge with 30 image formats including diagrams, tables, charts, chemical structures, photos, paintings, geometric shapes, music sheets, medical images, etc. In the *depth* aspect, the previous benchmarks normally require commonsense knowledge or simple physical or temporal reasoning. In contrast, our

benchmark requires deliberate reasoning with college-level subject knowledge.

4. Experiments

We evaluate various models including LLMs and LMMs. In each type, we consider both closed- and open-source models. Our evaluation is conducted under a *zero-shot* setting to assess the capability of models to generate accurate answers without fine-tuning or few-shot demonstrations on our benchmark. For all models, we use the default prompt provided by each model for multi-choice or open QA, if available. If models do not provide prompts for task types in MMMU, we conduct prompt engineering on the validation set and use the most effective prompt for the zero-shot setup in the main experiments. We also report the few-shot results of some selected models in the Appendix. All experiments are conducted with NVIDIA A100 GPUs.

4.1. Baselines

LMMs. We consider various large multimodal models. By default, for each model family, we use the latest, largest, and best-performing available checkpoint to date. (i) Kosmos2 [47] is pre-trained to ground fine-grained visual objects with texts and to follow instructions. With only 1.6B model size, Kosmos2 is able to achieve comparable or better performance with Flamingo-9B [2] on VQA and captioning tasks. (ii) LLaMA-Adapter2 [16] fine-tunes Llama [52] in a parameter-efficient way and utilizes visual encoder CLIP [48] and modular experts such as Optical Character Recognition (OCR) to capture more image information for later better visual understanding. (iii) BLIP-2 [27] introduces light-weight learnable visual queries to bridge the frozen CLIP ViT [48] and FLAN-T5 [12]. (iv) Starting from the parameters from BLIP-2, InstructBLIP [14] is fur-

	Validation Overall (900)	Test Overall (10,500)	Art & Design (1,163)	Business (1,428)	Science (2,426)	Health & Medicine (1,752)	Human. & Social Sci. (947)	Tech & Eng. (2,784)
Random Choice	22.1	23.9	24.1	24.9	21.6	25.3	22.8	24.8
Frequent Choice	26.8	25.8	26.7	28.4	24.0	24.4	25.2	26.5
Large Multimodal Models (LMMs): Text + Image as Input								
OpenFlamingo2-9B [4]	28.7	26.3	31.7	23.5	26.3	26.3	27.9	25.1
Kosmos2 [47]	24.4	<u>26.6</u>	28.8	23.7	26.6	27.2	26.3	26.8
Fuyu-8B [6]	27.9	27.4	29.9	27.0	25.6	27.0	32.5	26.4
MiniGPT4-Vicuna-13B [71]	26.8	27.6	30.2	27.0	26.2	26.9	30.9	27.2
LLaMA-Adapter2-7B [65]	29.8	27.7	35.2	25.4	25.6	30.0	29.1	25.7
Otter [26]	32.2	29.1	37.4	24.0	24.1	29.6	35.9	<u>30.2</u>
CogVLM [55]	32.1	30.1	38.0	25.6	25.1	31.2	41.5	28.9
InstructBLIP-T5-XL [14]	32.9	30.6	43.3	25.2	25.2	29.3	45.8	28.6
BLIP-2 FLAN-T5-XL [27]	34.4	31.0	43.0	25.6	25.1	31.8	48.0	27.8
mPLUGw-OWL2* [60]	32.7	32.1	48.5	25.6	24.9	32.8	46.7	29.6
SPHINX* [73]	32.9	-	-	-	-	-	-	-
Qwen-VL-7B [5]	<u>35.9</u>	32.9	47.7	<u>29.8</u>	25.6	33.6	45.3	<u>30.2</u>
LLaVA-1.5-13B [34]	<u>36.4</u>	33.6	49.8	<u>28.2</u>	25.9	34.9	54.7	<u>28.3</u>
InstructBLIP-T5-XXL [14]	35.7	<u>33.8</u>	48.5	30.6	27.6	33.6	49.8	29.4
BLIP-2 FLAN-T5-XXL [27]	35.4	34.0	<u>49.2</u>	28.6	<u>27.3</u>	<u>33.7</u>	<u>51.5</u>	30.4
Gemini Nano2* [72]	32.6	-	-	-	-	-	-	-
Qwen-VL-PLUS* [74]	45.2	<u>40.8</u>	<u>59.9</u>	<u>34.5</u>	32.8	<u>43.7</u>	<u>65.5</u>	<u>32.9</u>
Gemini Pro* [72]	47.9	-	-	-	-	-	-	-
GPT-4V(ision) (Playground) [46]	<u>56.8</u>	55.7	65.3	64.3	48.4	63.5	76.3	41.7
Gemini Ultra* [72]	59.4	-	-	-	-	-	-	-
Large Language Models (LLMs): Only Text as Input								
Llama2 7B [53]	30.1	28.7	30.7	27.2	26.7	27.7	32.6	29.8
FLAN-T5-XXL [12]	32.1	31.2	<u>36.8</u>	28.9	26.7	32.8	44.8	28.3
+ OCR	34.7	31.9	36.2	28.8	26.2	32.6	50.5	29.7
+ LLaVA Caption	34.8	31.9	38.4	27.8	27.0	33.2	49.9	28.7
Vicuna-13B [10]	33.3	31.0	35.1	30.1	24.7	31.4	44.8	30.1
+ OCR	35.4	31.9	37.1	28.6	26.5	32.0	49.3	30.0
+ LLaVA Caption	33.9	32.7	42.0	26.8	26.2	33.4	49.4	31.4
GPT-4 Text [45]	34.9	33.8	32.9	28.5	30.6	41.3	53.0	28.4

Table 2. Overall results of different models on the MMMU **validation** and **test set**. Besides reporting the performance of LMMs, we additionally add text-only LLM baselines. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors.

ther fine-tuned with visual instruction tuning data for better zero-shot generalization capabilities. For both BLIP-2 and InstructBLIP, we consider both scales: FLAN-T5 XL and FLAN-T5-XXL for model scaling analysis. (v) LLaVA-1.5 [34] linearly projects the visual embedding into word embedding space of Vicuna [10], thus equipping the LLM with visual abilities. (vi) As an open-source alternative to Flamingo [2], OpenFlamingo [4] has close performance on most vision-language tasks. (vii) CogVLM [55] concatenates image and text in the input embedding space and adds trainable visual layers in textual Transformer blocks to deeply align two modalities. It is reported to achieve

very promising performance on existing VQA benchmarks recently. (viii) Fuyu [6] projects the patches of the input image into text embedding space. (ix) Qwen-VL [5] introduces a set of trainable query embeddings and single-layer cross-attention module to bridge the modalities, supporting interleaved image-text input. (x) Otter [26] is fine-tuned with diverse instruction-tuning data and able to perform in-context learning. (xi) MiniGPT-4 [71] is built upon Vicuna [10] and designs a linear modality projection layer for visual understanding abilities. (xii) mPLUG-Owl2 [60] designs modality-adaptive module to unify vision and language while preserving the distinct properties of them.

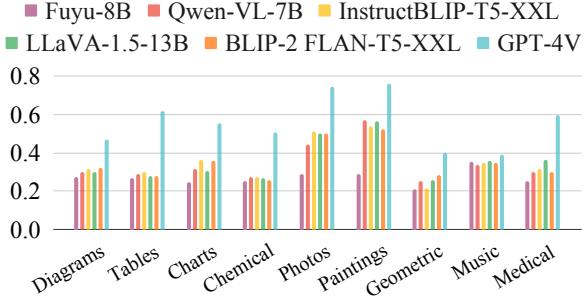


Figure 5. Performance of models on different types of images.

Text-only LLMs. For text-only LLMs, we consider the most capable ones including GPT-4 and several open-source LLMs, Llama2-7B [52], FLAN-T5-XXL and Vicuna-13B, which are adopted as the text encoder or decoder in the selected LMMs. To determine if an external image-to-text tool can enhance these LLMs’ performance on MMMU, we deploy OCR by MMOCR¹ or captioning by LLaVA-1.5 to provide the recognized text information to text-only LLMs.

Evaluation. We adopt micro-averaged accuracy as the evaluation metric. For both open and multiple-choice questions, we design systematic, rule-based evaluation pipelines. Specifically, to mitigate the potential influence of any intermediate generations (e.g., reasoning steps, calculations) in the long response, we construct robust regular expressions and develop response-processing workflows. These are employed to extract key phrases, such as numbers and conclusion phrases, from the long responses for accurate answer matching. If there is no valid answer in the model’s response, we perform random selection as a remedy for multiple-choice questions or consider the response incorrect for open questions. For reference, we add Random Choice and Frequent Choice baselines: the former randomly selects an option, while the latter selects the most frequent option within each specific subject of the validation set, based on its frequency of occurrence in that subject.

4.2. Main Results

In this section, we present a comprehensive comparison of different LLMs and LMMs using the MMMU benchmark, detailed in Table 2. We summarize our key findings as follows:

Challenging Nature of MMMU: The benchmark poses significant challenges to current models. Notably, GPT-4V, despite being an advanced model, achieves an accuracy of only 55.7%, with ample headroom for improvement. This reflects the benchmark’s rigorous and demanding standards.

Disparity between Open-source Models and GPT-4V: Leading open-source models such as BLIP2-FLAN-T5-XXL and LLaVA-1.5 reach an accuracy level of approximately 34%, which is significantly lower than GPT-4V. This

¹<https://github.com/open-mmlab/mmocr>

Models	Easy (2946)	Medium (4917)	Hard (2637)	Overall (10500)
Fuyu-8B [6]	28.9	27.0	26.4	27.4
Qwen-VL-7B [5]	39.4	31.9	27.6	32.9
LLaVA-1.5-13B [34]	41.3	32.7	26.7	33.6
InstructBLIP-T5-XXL [14]	40.3	32.3	29.4	33.8
BLIP-2 FLAN-T5-XXL [27]	41.0	32.7	28.5	34.0
GPT-4V [46]	76.1	55.6	31.2	55.7

Table 3. Result decomposition across question difficulty levels.

significant difference in performance indicates a gap in the capabilities of current open-source models compared to proprietary ones like GPT-4V.

Effectiveness of OCR and Captioning Enhancements:

The application of OCR and captioning technologies does not yield a significant improvement in the performance of text-only LMMs. This finding suggests that the MMMU benchmark requires models that can effectively interpret and integrate both textual and visual information, underscoring the complexity of the multimodal tasks it presents.

Model Performance across Different Disciplines: In disciplines such as Art & Design and Humanities & Social Sciences, where the images tends to be more ‘natural’ and questions involve relatively less reasoning, models demonstrate relatively higher performance. Conversely, in fields like Science, Health & Medicine, and Technology & Engineering, where tasks often involve intricate perception and complex reasoning, models exhibit lower performance.

The MMMU benchmark underscores both the progress and the challenges in multimodal understanding and reasoning. While GPT-4V leads in performance, the overall results indicate substantial room for improvement, especially in domains with complex visual input and heavy reasoning with subject knowledge.

4.3. Analysis on Images Types and Difficulties

Different Image Types. We compare the performance of various models across top frequent image types in Figure 5. Across all types, GPT-4V consistently outperforms the other models by a huge margin. Open-source models demonstrate relatively strong performance in categories like Photos and Paintings, which are more frequently seen during training. However, for less common image categories like Geometric shapes, Music sheets and Chemical structures, all models obtain very low scores (some are close to random guesses). This indicates that the existing models are generalizing poorly towards these image types.

Different Difficulty Levels. Table 3 compares the performance of selected models across three difficulty levels. GPT-4V demonstrates a significantly higher proficiency, with a success rate of 76.1%, compared to open-source models in the “Easy” category. When it comes to

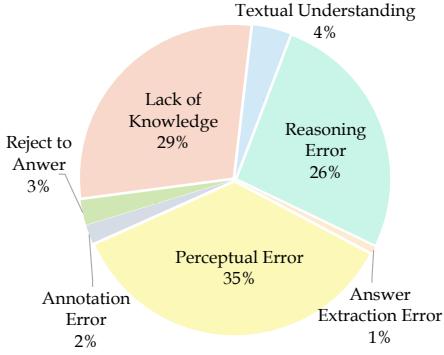


Figure 6. Error distribution over 150 annotated GPT-4V errors.

the “Medium” category, while the gap narrows, GPT-4V still leads at 55.6%. The further diminishing performance gap in the “Hard” category across models indicates that as the complexity of tasks increases, the advantage of more advanced models like GPT-4V almost disappears. This might reflect a current limitation in handling expert-level challenging queries even for the most advanced models.

5. Error Analysis and Future Work

In this section, we delve into the analysis of errors by GPT-4V, a pivotal aspect for understanding its operational capabilities and limitations. This analysis serves not only to identify the model’s current shortcomings but also to guide future enhancements in its design and training. We meticulously examine 150 randomly sampled error instances from GPT-4V’s predictions. These instances are analyzed by expert annotators who identify the *root causes of mispredictions* based on their knowledge and the golden explanations if available. The distribution of these errors is illustrated in Figure 6, and a selection of 100 notable cases, along with detailed analyses, is included in the Appendix.

Perceptual Errors (35%): Perceptual errors, forming the bulk of the inaccuracies in the GPT-4V model, are categorized into two types: basic perceptual errors and domain-specific perceptual errors. Basic perceptual errors, as depicted in Figure 7, occur when the model accurately processes and understands the given information but fails in elementary visual interpretation, such as misjudging the sequence described as “from left to right, top to bottom.” On the other hand, domain-specific perceptual errors occur due to the lack of knowledge. As we analyze the root cause, we classify such errors as lack of knowledge (see analysis below). Additionally, GPT-4V often exhibits a bias towards text, prioritizing textual information over visual inputs, a trend noted in recent studies [13]. A prominent example is in Figure 68, where the model incorrectly prioritizes its text-based interpretation of “imperialism” over the visual narrative in a cartoon depicting the United States as a “Savior.” This underscores the need for a more balanced approach to

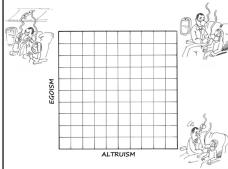
X

Perceptual Error

Question: Look at the situation in the <image 1>. You are traveling on a plane with a small child. The cabin is depressurized and the oxygen mask falls from the ceiling. Which of these pictures shows the reconciliation of egoism and other-isms (the order is from left to right, top to bottom)?

Option: (A) 1 (B) 2 (C) 3

Ground Truth: (B) 2



<image 1>

GPT-4V: The image shows three situations on a plane when the oxygen mask falls: The adult puts on the mask first while the child watches (egoism). The adult puts the mask on the child first while struggling for breath. Both the adult and the child share the mask. The reconciliation of egoism and other-isms would be a situation where both the individual's needs and the needs of others are considered and addressed. Therefore, the correct answer is: (C) 3

Error Reason: GPT-4V recalled the right knowledge and made the right reasoning, but it failed to correctly map the IDs to the corresponding illustrations in the figure, which is not explicitly marked in the figure but is only described in text (the order is from left to right, top to bottom).

Figure 7. A basic perceptual error, easy for humans but challenging for GPT-4V. More examples can be found in the Appendix.

multimodal interpretation.

Lack of Knowledge (29%): A fundamental root cause of ‘domain-specific’ perceptual errors in the GPT-4V model, as previously discussed, is the lack of specialized knowledge. This deficiency is exemplified in the Computer Science context illustrated in Appendix Figure 84, where the model identifies visual elements such as double circles but fails to interpret them accurately within the domain-specific context, such as their representation of an ‘accept state’ in Deterministic Finite Automata. Similarly, a deficit in specialized knowledge can lead to flawed reasoning, as demonstrated in the medical example in Appendix Figure 55. These instances underscore the necessity of enriching the training datasets of foundation models with a diverse range of domain-specific knowledge to improve their accuracy and general applicability in various specialized fields.

Reasoning Errors (26%): Flawed reasoning emerges as another significant cause of errors. In instances where the model correctly interprets text and images and recalls relevant knowledge, it still often fails to apply logical and mathematical reasoning skills effectively to derive accurate inferences. A notable instance of this can be observed in Appendix Figure 46, where the model neglects an essential step in a mathematical reasoning process, leading to an incorrect conclusion. Enhancing the model’s reasoning capability is critical to address these shortcomings.

Other Errors: The remaining errors include Textual Understanding Error (6%), Rejection to Answer (3%), Annotation Error (2%), and Answer Extraction Error (1%). These errors are attributed to various factors such as complex text interpretation challenges, limitations in response generation, inaccuracies in data annotation, and issues in extracting precise answers from longer outputs.

In summary, our error analysis underlines the challenges posed by MMMU and highlights areas for further research in visual perception, knowledge representation, reasoning abilities, and multimodal joint understanding.

6. Conclusion

The development of MMMU as a benchmark for assessing the capabilities of LMMs marks a significant milestone in the journey toward Expert AGI. MMMU not only tests the boundaries of what current LMMs can achieve in terms of basic perceptual skills but also evaluates their ability to handle complex reasoning and in-depth subject-specific knowledge. This approach directly contributes to our understanding of the progress towards Expert AGI, as it mirrors the kind of expertise and reasoning abilities expected of skilled adults in various professional fields.

Despite its comprehensive nature, MMMU, like any benchmark, is not without limitations. The manual curation process, albeit thorough, may carry biases. And the focus on college-level subjects might not fully be a sufficient test for Expert AGI as per the definition [44]. However, we believe it should be necessary for an Expert AGI to achieve strong performance on MMMU to demonstrate their broad and deep subject knowledge as well as expert-level understanding and reasoning capabilities. In future work, we plan to incorporate human evaluations into MMMU. This will provide a more grounded comparison between model capabilities and expert performance, shedding light on the proximity of current AI systems to achieving Expert AGI.

References

- [1] Blaise Agüera y Arcas and Peter Norvig. Artificial general intelligence is already here. *Noema Magazine*, 2023. 1
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 3, 5, 6
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2, 3
- [4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 3, 6, 14, 15, 16, 17, 18, 19
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 6, 7, 14, 15, 16, 17, 18, 19
- [6] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 3, 6, 7, 14, 15, 16, 17, 18, 19
- [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1
- [8] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 2
- [9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120, 2020. 3
- [10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhang-hao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yong-hao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3, 6, 14, 15, 16, 17, 18, 19
- [11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1
- [12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3, 5, 6, 14, 15, 16, 17, 18, 19
- [13] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023. 4, 8
- [14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 2, 3, 5, 6, 7, 14, 15, 16, 17, 18, 19
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3
- [16] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 3, 5
- [17] Yingqiang Ge, Wenyue Hua, Jianchao Ji, Juntao Tan, Shuyuan Xu, and Yongfeng Zhang. Openagi: When llm

- meets domain experts. *arXiv preprint arXiv:2304.04370*, 2023. 1
- [18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 3
- [19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020. 2
- [20] Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Collier Nigel, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*, 2023. 3
- [21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3
- [22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3
- [23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2
- [24] Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. Artificial general intelligence (agi) for education. *arXiv preprint arXiv:2304.12479*, 2023. 1
- [25] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2, 4
- [26] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 3, 6, 14, 15, 16, 17, 18, 19
- [27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023. 2, 3, 5, 6, 7, 14, 15, 16, 17, 18, 19
- [28] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3it: A large-scale dataset towards multimodal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 3
- [29] Xijun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 3
- [30] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 4
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014. 2, 3
- [32] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusion-bench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023. 4
- [33] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 3
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 3, 6, 7, 14, 15, 16, 17, 18, 19
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3
- [36] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2, 4
- [37] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezheng Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 4
- [38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3
- [39] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2
- [40] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 4
- [41] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering

- benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
- [42] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023. 1, 4
- [43] Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin F Yang, and Kai-Wei Chang. Metavl: Transferring in-context learning ability from language models to vision-language models. *arXiv preprint arXiv:2306.01311*, 2023. 3
- [44] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi: Operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*, 2023. 1, 3, 9
- [45] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 6, 14, 15, 16, 17, 18, 19
- [46] OpenAI. Gpt-4v(ision) system card, 2023. 2, 6, 7, 14, 15, 16, 17, 18, 19
- [47] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 5, 6, 14, 15, 16, 17, 18, 19
- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5
- [49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3
- [50] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2
- [51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 3
- [52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 5, 7
- [53] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6, 14, 15, 16, 17, 18, 19
- [54] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023. 4
- [55] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2, 6, 14, 15, 16, 17, 18, 19
- [56] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2021. 3
- [57] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 4
- [58] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 2
- [59] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3
- [60] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 3, 6, 14, 15, 16, 17, 18, 19
- [61] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023. 2, 4
- [62] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 3
- [63] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2, 4
- [64] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinyl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 3
- [65] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 3, 6, 14, 15, 16, 17, 18, 19
- [66] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 3

- [67] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 3
- [68] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*, 2023. 4
- [69] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023. 2
- [70] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13041–13049, 2020. 3
- [71] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 6, 14, 15, 16, 17, 18, 19
- [72] Gemini Team, Google. Gemini: A Family of Highly Capable Multimodal Models. Available online at: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf, 2023. 6, 14, 15, 16, 17, 18, 19, 117
- [73] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. *arXiv preprint arXiv:2311.07575*, 2023. 6, 14, 15, 16, 17, 18, 19
- [74] Qwen-VL-PLUS. GitHub Repository, 2023. URL <https://github.com/QwenLM/Qwen-VL?tab=readme-ov-file#qwen-vl-plus>. 6, 14, 15, 16, 17, 18, 19

MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Supplementary Material

Table of Contents in Appendix

A Breakdown Results on Different Subjects	14
A.1. Art & Design	14
A.2 Business	15
A.3 Science	16
A.4 Health & Medicine	17
A.5 Humanities & Social Science	18
A.6 Tech & Engineering	19
B Case Study	20
C Subfields of Different Subjects	110
D Distributions of Image Types	110
E Results on Different Image Types	110
F Few-shot Results	113
G Data Annotation Protocol	114
G.1. Data Collection	114
G.2 General Guidelines	114
G.3 Data Format and Structure	114
G.4 Quality Control and Validation	114
G.5 Handling Ambiguities	114
G.6 Ethical Considerations	114
G.7 Data Contamination Considerations	114
G.8 Example Questions	114
H Author Contribution Statement	115
I. Version Change Log	117

A. Breakdown Results on Different Subjects

In this appendix, we show breakdown results of different models on each discipline and subject.

A.1. Art & Design

	Validation Overall (120)	Test Overall (1,163)	Art (231)	Art Theory (429)	Design (169)	Music (334)
Random Choice	29.2	24.1	23.4	20.3	19.5	31.7
Frequent Choice	23.3	26.7	24.2	23.5	33.7	29.0
Large Multimodal Models (LMMs): Text + Image as Input						
OpenFlamingo2-9B [4]	40.0	31.7	36.8	28.4	27.8	34.4
Kosmos2 [47]	25.0	28.8	30.7	24.9	28.4	32.6
Fuyu-8B [6]	36.7	29.9	28.6	26.8	29.0	35.3
MiniGPT4-Vicuna-13B [71]	29.2	30.2	28.6	28.7	40.2	28.1
LLaMA-Adapter2-7B [65]	29.2	35.2	38.5	35.4	41.4	29.3
Otter [26]	37.5	37.4	40.7	35.9	46.2	32.6
CogVLM [55]	40.8	38.0	43.3	39.2	44.4	29.6
InstructBLIP-T5-XL [14]	40.0	43.3	49.8	45.0	52.1	32.3
BLIP-2 FLAN-T5-XL [27]	44.2	43.0	50.2	45.0	47.3	33.2
mPLUG-OWL2* [60]	45.8	48.5	57.6	53.4	59.8	30.2
SPHINX* [73]	48.3	-	-	-	-	-
Qwen-VL-7B [5]	51.7	47.7	57.1	49.7	58.6	33.2
LLaVA-1.5-13B [34]	51.7	49.8	58.4	51.5	61.5	35.6
InstructBLIP-T5-XXL [14]	44.2	48.5	51.9	52.7	60.4	34.7
BLIP-2 FLAN-T5-XXL [27]	41.7	49.2	54.5	51.5	64.5	34.7
Qwen-VL-PLUS* [74]	60.0	59.9	67.5	68.1	78.7	34.7
GPT-4V(ision) (Playground) [46]	65.8	65.3	74.0	75.5	80.5	38.6
Gemini Ultra* [72]	70.0	-	-	-	-	-
Large Language Models (LLMs): Only Text as Input						
Llama2 7B [53]	29.2	30.7	30.3	27.5	37.9	31.4
FLAN-T5-XXL [12]	38.3	36.8	32.0	36.8	52.1	32.3
+ OCR	37.5	36.2	36.4	33.8	47.9	33.2
+ LLaVA Caption	43.3	38.4	45.9	38.2	46.2	29.6
Vicuna-13B [10]	41.7	35.1	35.1	31.5	46.7	33.8
+ OCR	39.2	37.1	35.5	32.9	50.3	36.8
+ LLaVA Caption	38.3	42.0	51.1	42.7	46.2	32.6
GPT-4 Text [45]	35.0	32.9	35.1	28.7	47.3	29.6

Table 4. **Art & Design** results of different models on the **MMMU validation** and **test set**. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors.

A.2. Business

	Validation Overall (150)	Test Overall (1,428)	Accounting (380)	Economics (267)	Finance (355)	Manage (245)	Marketing (181)
Random Choice	24.7	24.9	30.0	29.6	17.7	22.4	24.9
Frequent Choice	29.3	28.4	33.4	36.3	22.0	15.9	35.9
Large Multimodal Models (LMMs): Text + Image as Input							
OpenFlamingo2-9B [4]	28.0	23.5	24.7	25.8	19.4	25.3	22.7
Kosmos2 [47]	18.0	23.7	29.7	24.0	21.4	22.4	17.1
Fuyu-8B [6]	32.0	27.0	32.1	30.3	22.5	20.0	29.3
MiniGPT4-Vicuna-13B [71]	21.3	27.0	29.7	<u>34.1</u>	25.6	16.7	27.6
LLaMA-Adapter2-7B [65]	25.3	25.4	30.8	<u>24.7</u>	20.6	24.9	25.4
Otter [26]	24.0	24.0	30.8	29.6	17.5	16.3	24.9
CogVLM [55]	25.3	25.6	28.2	29.6	19.2	21.2	32.6
InstructBLIP-T5-XL [14]	28.0	25.2	27.6	31.8	18.0	22.0	28.7
BLIP-2 FLAN-T5-XL [27]	26.7	25.6	28.2	31.1	17.5	24.1	29.8
mPLUG-OWL2* [60]	24.7	25.6	28.7	29.2	20.3	22.4	28.7
SPHINX* [73]	24.7	-	-	-	-	-	-
Qwen-VL-7B [5]	29.3	<u>29.8</u>	34.2	29.6	18.9	32.2	38.7
LLaVA-1.5-13B [34]	22.7	28.2	29.2	33.3	<u>23.7</u>	23.3	<u>34.3</u>
InstructBLIP-T5-XXL [14]	24.0	30.6	34.2	35.6	23.4	<u>30.2</u>	30.4
BLIP-2 FLAN-T5-XXL [27]	30.0	28.6	<u>32.4</u>	33.3	22.5	26.1	28.7
Qwen-VL-PLUS* [74]	35.3	<u>34.5</u>	<u>35.0</u>	<u>41.6</u>	<u>26.2</u>	<u>34.7</u>	<u>39.2</u>
Gemini Ultra* [72]	<u>56.7</u>	-	-	-	-	-	-
GPT-4V(ision) (Playground) [46]	59.3	64.3	69.7	70.8	61.1	51.0	67.4
Large Language Models (LLMs): Only Text as Input							
Llama2 7B [53]	22.7	27.2	28.9	34.1	23.7	21.6	27.6
FLAN-T5-XXL [12]	28.0	28.9	31.6	31.5	23.1	29.0	30.4
+ OCR	29.3	28.8	32.4	30.0	24.8	26.9	29.8
+ LLaVA Caption	31.3	27.8	28.2	30.7	24.2	27.8	29.8
Vicuna-13B [10]	26.7	30.1	29.5	34.8	25.6	30.6	32.6
+ OCR	31.3	28.6	27.1	34.1	23.9	30.6	30.4
+ LLaVA Caption	26.0	26.8	27.1	32.6	22.3	25.3	28.7
GPT-4 Text [45]	36.7	28.5	29.7	35.2	21.1	32.2	25.4

Table 5. **Business** results of different models on the MMMU **validation** and **test set**. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors.

A.3. Science

	Validation Overall (150)	Test Overall (2,426)	Biology (345)	Chemistry (603)	Geography (565)	Math (505)	Physics (408)
Random Choice	18.0	21.6	18.3	18.6	26.0	22.2	22.1
Frequent Choice	27.3	24.0	25.8	19.9	26.9	26.1	22.1
Large Multimodal Models (LMMs): Text + Image as Input							
OpenFlamingo2-9B [4]	23.3	26.3	27.8	22.9	<u>30.8</u>	25.1	25.0
Kosmos2 [47]	19.3	26.6	28.4	21.7	<u>29.2</u>	26.7	28.4
Fuyu-8B [6]	22.0	25.6	27.8	20.9	30.1	24.8	25.7
MiniGPT4-Vicuna-13B [71]	28.7	26.2	23.2	22.1	29.4	30.1	25.5
LLaMA-Adapter2-7B [65]	30.7	25.6	27.5	24.9	30.4	23.0	21.3
Otter [26]	34.7	24.1	24.6	23.4	27.1	23.0	21.8
CogVLM [55]	28.0	25.1	29.3	24.2	28.0	23.4	21.1
InstructBLIP-T5-XL [14]	<u>32.7</u>	25.2	27.0	22.1	28.3	24.4	25.0
BLIP-2 FLAN-T5-XL [27]	30.7	25.1	26.7	24.4	25.7	24.0	25.2
mPLUG-OWL2* [60]	22.7	24.9	27.2	23.9	29.7	18.8	25.2
SPHINX* [73]	26.7	-	-	-	-	-	-
Qwen-VL-7B [5]	29.3	25.6	27.8	23.1	28.8	24.6	24.3
LLaVA-1.5-13B [34]	29.3	25.9	27.2	25.0	28.8	24.0	24.5
InstructBLIP-T5-XXL [14]	30.7	27.6	<u>29.0</u>	26.5	31.0	25.5	<u>26.0</u>
BLIP-2 FLAN-T5-XXL [27]	34.7	<u>27.3</u>	28.4	<u>25.5</u>	29.4	<u>27.5</u>	<u>26.0</u>
Qwen-VL-PLUS* [74]	37.3	<u>32.8</u>	<u>40.3</u>	<u>27.9</u>	<u>34.7</u>	<u>31.3</u>	<u>33.1</u>
Gemini Ultra* [72]	48.0	-	-	-	-	-	-
GPT-4V(ision) (Playground) [46]	54.7	48.4	52.2	46.9	44.8	45.0	56.4
Large Language Models (LLMs): Only Text as Input							
Llama2 7B [53]	34.0	26.7	28.4	21.4	29.7	28.5	26.7
FLAN-T5-XXL [12]	28.0	26.7	27.8	24.4	27.3	30.7	23.5
+ OCR	30.0	26.2	24.6	24.5	27.4	27.9	26.0
+ LLaVA Caption	32.7	27.0	25.8	23.9	30.3	29.1	25.5
Vicuna-13B [10]	23.3	24.7	24.6	22.7	25.0	26.1	25.7
+ OCR	30.0	26.5	26.4	24.7	29.0	27.1	25.2
+ LLaVA Caption	28.7	26.2	31.3	21.7	28.7	26.7	24.3
GPT-4 Text [45]	34.7	30.6	29.3	28.0	34.0	27.7	34.3

Table 6. **Science** results of different models on the MMMU **validation** and **test set**. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors.

A.4. Health & Medicine

	Validation Overall (150)	Test Overall (1,752)	Basic Medical Science (326)	Clinical Medicine (325)	Diagnostics & Lab. Medicine (162)	Pharmacy (430)	Public Health (509)
Random Choice	20.7	25.3	24.8	21.8	25.9	28.6	24.8
Frequent Choice	30.0	24.4	22.1	24.3	17.3	23.3	29.3
Large Multimodal Models (LMMs): Text + Image as Input							
OpenFlamingo2-9B [4]	27.3	26.3	29.1	21.8	22.2	32.1	23.8
Kosmos2 [47]	28.0	27.2	27.3	24.0	27.2	30.7	26.1
Fuyu-8B [6]	28.0	27.0	28.8	23.1	24.1	27.0	29.3
MiniGPT4-Vicuna-13B [71]	30.7	26.9	27.0	26.2	21.6	27.7	28.5
LLaMA-Adapter2-7B [65]	30.7	30.0	31.0	30.2	26.5	36.5	25.0
Otter [26]	30.7	29.6	34.4	28.3	28.4	28.6	28.5
CogVLM [55]	32.0	31.2	33.4	27.4	27.2	33.7	31.4
InstructBLIP-T5-XL [14]	28.7	29.3	31.3	28.9	22.8	34.2	26.1
BLIP-2 FLAN-T5-XL [27]	<u>35.3</u>	31.8	35.9	31.7	24.1	<u>35.8</u>	28.5
mPLUG-OWL2* [60]	32.0	32.8	29.9	32.3	<u>34.0</u>	31.2	29.7
SPHINX* [73]	30.7	-	-	-	-	-	-
Qwen-VL-7B [5]	33.3	33.6	38.0	<u>34.8</u>	32.1	29.5	33.8
LLaVA-1.5-13B [34]	38.7	34.9	42.6	36.6	34.6	32.1	31.4
InstructBLIP-T5-XXL [14]	<u>35.3</u>	33.6	35.6	32.3	29.6	34.2	33.8
BLIP-2 FLAN-T5-XXL [27]	32.0	<u>33.7</u>	<u>38.7</u>	34.5	27.2	33.7	<u>32.2</u>
Qwen-VL-PLUS* [74]	46.7	<u>43.7</u>	<u>49.7</u>	<u>42.2</u>	<u>34.0</u>	<u>46.5</u>	41.5
GPT-4V(ision) (Playground) [46]	<u>64.7</u>	63.5	65.0	62.5	43.8	68.1	65.4
Gemini Ultra* [72]	67.3	-	-	-	-	-	-
Large Language Models (LLMs): Only Text as Input							
Llama2 7B [53]	26.7	27.7	26.1	30.8	25.3	27.7	27.7
FLAN-T5-XXL [12]	32.0	32.8	33.7	34.8	30.2	34.4	30.5
+ OCR	32.7	32.6	33.7	35.1	27.8	32.3	32.2
+ LLaVA Caption	32.0	33.2	35.3	34.2	30.9	32.6	32.4
Vicuna-13B [10]	31.3	31.4	37.7	33.2	36.4	27.7	27.9
+ OCR	31.3	32.0	38.3	33.5	37.0	28.4	28.5
+ LLaVA Caption	34.0	33.4	37.1	35.4	32.7	32.6	30.6
GPT-4 Text [45]	40.7	41.3	52.5	52.9	27.8	39.1	33.0

Table 7. **Health & Medicine** results of different models on the MMMU **validation** and **test set**. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors.

A.5. Humanities & Social Science

	Validation Overall (120)	Test Overall (947)	History (278)	Literature (112)	Sociology (252)	Psychology (305)
Random Choice	20.0	22.8	22.3	24.1	27.0	19.3
Frequent Choice	25.8	25.2	27.0	27.7	25.4	22.6
Large Multimodal Models (LMMs): Text + Image as Input						
OpenFlamingo2-9B [4]	30.8	27.9	24.5	42.0	29.0	24.9
Kosmos2 [47]	30.0	26.3	24.5	24.1	34.1	22.3
Fuyu-8B [6]	32.5	32.5	32.7	44.6	32.9	27.5
MiniGPT4-Vicuna-13B [71]	29.2	30.9	30.9	47.3	30.6	25.2
LLaMA-Adapter2-7B [65]	33.3	29.1	27.0	43.8	32.1	23.3
Otter [26]	41.7	35.9	33.8	67.0	34.9	27.2
CogVLM [55]	45.0	41.5	39.2	69.6	41.3	33.4
InstructBLIP-T5-XL [14]	47.5	45.8	45.0	71.4	44.8	38.0
BLIP-2 FLAN-T5-XL [27]	50.0	48.0	48.2	76.8	47.2	38.0
mPLUG-OWL2* [60]	45.8	46.7	46.0	74.1	44.4	39.0
SPHINX* [73]	50.0	-	-	-	-	-
Qwen-VL-7B [5]	45.0	45.3	47.8	64.3	46.4	35.1
LLaVA-1.5-13B [34]	53.3	54.7	58.6	76.8	<u>51.2</u>	45.9
InstructBLIP-T5-XXL [14]	49.2	49.8	48.6	72.3	<u>51.2</u>	41.6
BLIP-2 FLAN-T5-XXL [27]	50.8	<u>51.5</u>	<u>49.6</u>	<u>75.9</u>	53.2	<u>43.0</u>
Qwen-VL-PLUS* [74]	65.8	<u>65.5</u>	<u>69.8</u>	<u>79.5</u>	<u>63.9</u>	<u>57.7</u>
GPT-4V(ision) (Playground) [46]	<u>72.5</u>	76.3	79.1	89.3	71.4	73.1
Gemini Ultra* [72]	78.3	-	-	-	-	-
Large Language Models (LLMs): Only Text as Input						
Llama2 7B [53]	37.5	32.6	32.4	46.4	32.9	27.5
FLAN-T5-XXL [12]	42.5	44.8	46.8	<u>56.2</u>	39.7	43.0
+ OCR	55.0	50.5	53.6	75.0	46.4	42.0
+ LLaVA Caption	49.2	49.9	51.8	75.0	46.8	41.6
Vicuna-13B [10]	45.8	44.8	51.1	59.8	39.3	38.0
+ OCR	50.0	49.3	58.3	66.1	48.0	36.1
+ LLaVA Caption	48.3	49.4	53.6	72.3	48.8	37.7
GPT-4 Text [45]	51.7	53.0	52.9	82.1	34.5	57.7

Table 8. **Humanities & Social Science** results of different models on the **MMMU validation** and **test set**. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors.

A.6. Tech & Engineering

	Val Overall (210)	Test Overall (2,784)	Agri. (287)	Arch. & Eng. (551)	Comp. Sci. (371)	Electr. (256)	Energy &Power (432)	Materials (458)	Mech. Eng. (429)
Random Choice	21.4	24.8	21.3	27.0	22.6	10.5	31.5	24.2	28.7
Frequent Choice	24.8	26.5	24.7	24.1	29.6	12.9	30.3	30.3	28.0
Large Multimodal Models (LMMs): Text + Image as Input									
OpenFlamingo2-9B [4]	26.2	25.1	20.6	29.6	26.1	13.7	24.1	26.0	28.4
Kosmos2 [47]	26.7	26.8	20.6	28.3	<u>32.1</u>	10.2	29.9	27.3	30.5
Fuyu-8B [6]	21.4	26.4	26.5	25.0	26.1	12.1	35.0	25.1	29.8
MiniGPT4-Vicuna-13B [71]	23.8	27.2	29.6	23.8	28.8	13.7	36.1	27.3	27.5
LLaMA-Adapter2-7B [65]	30.0	25.7	23.0	25.2	25.6	17.6	30.3	25.8	28.4
Otter [26]	29.0	<u>30.2</u>	30.7	26.3	<u>32.1</u>	19.1	35.2	30.1	34.7
CogVLM [55]	27.6	28.9	26.8	27.0	31.8	14.1	33.1	28.4	<u>35.4</u>
InstructBLIP-T5-XL [14]	27.1	28.6	26.1	33.6	28.3	<u>23.8</u>	29.9	22.9	<u>31.5</u>
BLIP-2 FLAN-T5-XL [27]	27.6	27.8	17.8	<u>32.5</u>	26.7	20.7	33.6	24.9	30.8
mPLUG-OWL2* [60]	31.0	29.6	32.4	29.4	31.8	14.5	39.4	26.6	28.2
SPHINX* [73]	26.2	-	-	-	-	-	-	-	-
Qwen-VL-7B [5]	<u>32.9</u>	<u>30.2</u>	<u>33.1</u>	25.0	33.4	19.1	37.0	<u>28.8</u>	33.1
LLaVA-1.5-13B [34]	31.4	28.3	34.5	26.1	29.6	22.7	30.1	26.9	28.9
InstructBLIP-T5-XXL [14]	35.2	29.4	24.7	30.3	29.6	20.7	<u>37.3</u>	26.6	31.5
BLIP-2 FLAN-T5-XXL [27]	30.0	30.4	28.2	27.2	29.6	25.0	35.6	26.9	38.0
Qwen-VL-PLUS* [74]	<u>36.7</u>	<u>32.9</u>	<u>40.4</u>	<u>25.6</u>	<u>36.1</u>	<u>24.6</u>	<u>33.6</u>	<u>34.7</u>	<u>36.8</u>
GPT-4V(ision) (Playground) [46]	<u>36.7</u>	41.7	43.9	37.2	57.1	27.0	47.5	36.9	41.0
Gemini Ultra* [72]	47.1	-	-	-	-	-	-	-	-
Large Language Models (LLMs): Only Text as Input									
Llama2 7B [53]	31.4	29.8	33.1	23.8	32.6	17.6	39.1	27.9	32.6
FLAN-T5-XXL [12]	28.6	28.3	21.3	30.3	28.8	25.4	26.6	27.9	33.8
+ OCR	30.0	29.7	20.2	30.7	31.3	27.0	29.9	29.0	35.9
+ LLaVA Caption	27.6	28.7	17.8	31.4	26.4	24.2	32.4	26.4	35.7
Vicuna-13B [10]	34.8	30.1	31.7	26.3	28.8	25.4	39.4	26.4	32.6
+ OCR	34.8	30.0	31.7	25.6	29.9	18.8	40.3	27.5	33.6
+ LLaVA Caption	32.4	31.4	32.4	26.9	31.8	20.3	39.8	28.8	37.3
GPT-4 Text [45]	20.0	28.4	28.9	25.6	33.4	17.2	38.4	23.6	28.9

Table 9. **Tech & Engineering** results of different models on the **MMMU validation** and **test set**. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors.

B. Case Study

List of Case Study Figures

8 Art 1: Correct Case	22	57 Diagnostics and Lab Medicine 2: Perceptual Error	71
9 Art 2: Correct Case	23	58 Diagnostics and Lab Medicine 3: Reject to Answer	72
10 Art 3: Perceptual Error	24	59 Diagnostics and Lab Medicine 4: Perceptual Error, Lack of Knowledge	73
11 Art Theory 1: Correct Case	25	60 Pharmacy 1: Correct Case	74
12 Art Theory 2: Correct Case	26	61 Pharmacy 2: Lack of Knowledge	75
13 Art Theory 3: Lack of Knowledge	27	62 Pharmacy 3: Lack of Knowledge	76
14 Design 1: Correct Case	28	63 Public Health 1: Correct Case	77
15 Design 2: Perceptual Error	29	64 Public Health 2: Textual Understanding Error	78
16 Design 3: Lack of Knowledge	30	65 Public Health 3: Lack of Knowledge	79
17 Music 1: Correct Case	31	66 History 1: Correct Case	80
18 Music 2: Perceptual Error, Lack of Knowledge	32	67 History 2: Correct Case	81
19 Music 3: Perceptual Error	33	68 History 3: Perceptual Error	82
20 Accounting 1: Correct Case	34	69 History 4: Lack of Knowledge	83
21 Accounting 2: Perceptual Error	35	70 Literature 1: Correct Case	84
22 Economics 1: Correct Case	36	71 Literature 2: Perceptual Error	85
23 Economics 2: Perceptual Error	37	72 Sociology 1: Correct Case	86
24 Economics 3: Perceptual Error	38	73 Sociology 2: Reasoning Error	87
25 Finance 1: Correct Case	39	74 Psychology 1: Correct Case	88
26 Finance 2: Reasoning Error	40	75 Psychology 2: Perceptual Error	89
27 Manage 1: Correct Case	41	76 Agriculture 1: Correct Case	90
28 Manage 2: Perceptual Error	42	77 Agriculture 2: Perceptual Error	91
29 Marketing 1: Correct Case	43	78 Agriculture 3: Perceptual Error	92
30 Marketing 2: Perceptual Error	44	79 Agriculture 4: Perceptual Error	93
31 Biology 1: Correct Case	45	80 Architecture and Engineering 1: Correct Case	94
32 Biology 2: Correct Case	46	81 Architecture and Engineering 2: Correct Case	95
33 Biology 3: Reasoning Error	47	82 Architecture and Engineering 3: Reasoning Error	96
34 Biology 4: Reasoning Error	48	83 Computer Science 1: Correct Case	97
35 Biology 5: Reasoning Error	49	84 Computer Science 2: Perceptual Error, Lack of Knowledge	98
36 Chemistry 1: Correct Case	50	85 Computer Science 3: Perceptual Error	99
37 Chemistry 2: Correct Case	51	86 Computer Science 4: Perceptual Error	100
38 Chemistry 3: Perceptual Error, Reasoning Error	52	87 Electronics 1: Correct Case	101
39 Chemistry 4: Lack of Knowledge	53	88 Electronics 2: Reject to Answer	102
40 Geography 1: Correct Case	54	89 Energy and Power 1: Correct Case	103
41 Geography 2: Reasoning Error	55	90 Energy and Power 2: Reasoning Error	104
42 Geography 3: Perceptual Error, Reasoning Error	56	91 Materials 1: Correct Case	105
43 Math 1: Correct Case	57	92 Materials 2: Lack of Knowledge	106
44 Math 2: Perceptual Error	58	93 Mechanical Engineering 1: Correct Case	107
45 Math 3: Textual Understanding Error	59	94 Mechanical Engineering 2: Reasoning Error	108
46 Math 4: Reasoning Error	60	95 Mechanical Engineering 3: Reasoning Error	109
47 Physics 1: Correct Case	61		
48 Physics 2: Perceptual Error	62		
49 Basic Medical Science 1: Correct Case	63		
50 Basic Medical Science 2: Perceptual Error	64		
51 Clinical Medicine 1: Correct Case	65		
52 Clinical Medicine 2: Correct Case	66		
53 Clinical Medicine 3: Correct Case	67		
54 Clinical Medicine 4: Perceptual Error	68		
55 Clinical Medicine 5: Lack of Knowledge	69		
56 Diagnostics and Lab Medicine 1: Correct Case	70		

Subject	Correct Case	Perception	Lack of Knowledge	Reasoning	Other
Art	8, 9	10			
Art Theory	11, 12		13		
Design	14	15	16		
Music	17	18, 19	18		
Accounting	20	21			
Economics	22	23, 24			
Finance	25			26	
Manage	27	28			
Marketing	29	30			
Biology	31, 32			33, 34	
Chemistry	36, 37	38	39	38	
Geography	40	42		41, 42	
Math	43	44		46	45
Physics	47	48			
Basic Medical Science	49	50			
Clinical Medicine	51, 52, 53	54	55		
Diagnostics and Laboratory Medicine	56	57, 59	59		58
Pharmacy	60		61, 62		
Public Health	63		65		64
History	66, 67	68	69		
Literature	70	71			
Sociology	72			73	
Psychology	74	75			
Agriculture	76	77, 78, 79			
Architecture and Engineering	80			81	
Computer Science	83	84, 85, 86	84		
Electronics	87				88
Energy and Power	89			90	
Materials	91		92		
Mechanical Engineering	93			94, 95	

Table 10. Table index of case study figures by subjects with associated (error) categories.



Art: Visual Arts

Question: The artist's use of color and neutral tones helps convey primarily which of the following themes? <image 1>

Option:

- (A) the inequities of societies
- (B) the innocence of youth
- (C) the conveniences of technology
- (D) the threat of climate change



<image 1>

GPT-4V: The artist's use of color and neutral tones in this picture, along with the chosen imagery, helps convey the theme of "the inequities of societies" (A). The contrast between the small child laboring to pull the rickshaw and the larger, relaxed adults taking a selfie is a stark commentary on social and economic disparities. The neutral background emphasizes the subjects, drawing the viewer's attention to the inequality presented within the scene.

Ground Truth: (A) the inequities of societies

Explanation: The work depicts a small boy pulling a rickshaw in which two adults are seated. The boy, the rickshaw, and the background are all depicted with shades of gray and black. By contrast, the couple in the rickshaw is depicted in full color. Even without this use of color, the work would convey a theme of inequity because a little boy is pulling two adults, who seem to represent Western culture. The color contrast underscores the theme because one culture is depicted as bright, happy, and colorful, while the other is depicted as small and drab.

Figure 8. A sample correct case of Art (subfield: Visual Arts).

[Back to List of Figures](#) | [Back to Table Index](#)



Art: Photography

Question: Which of the motion effect images in <image 1> is a 'panning blur'?

Option:

- (A) Photo 1
- (B) Photo 2
- (C) Photo 3
- (D) None of the above



1



2



3

<image 1>

GPT-4V: The image labeled '3' demonstrates a 'panning blur' effect, where the camera follows the moving subject at the same speed, causing the subject to appear relatively sharp while the background becomes blurred. This technique emphasizes the speed and direction of the moving subject.

Ground Truth: (C) Photo 3

Figure 9. A sample correct case of Art (subfield: Photography).

[Back to List of Figures](#) | [Back to Table Index](#)



Art: Fine Arts

Question: A main function of castles was to provide a defense against invaders. This was achieved by several devices, including water-filled moats and barbicans, high stone walls surrounding the castle. Another such device was battlements (or crenellations). These were alternating sections of wall, called merlons, and openings, called embrasures. The embrasures were used for firing arrows. They also permitted boulders to be dropped on attackers attempting to scale the castle walls. If you were going to write an alternate caption for <image 1>, you might note that this castle was equipped with

Option:

- (A) a barbican and a moat.
- (B) merlons but not embrasures.
- (C) a moat and crenellations.
- (D) a barbican and battlements.



<image 1>

GPT-4V: The correct alternate caption for the image based on the provided information would be:
(C) a moat and crenellations.

Error Category: Perceptual Error

Error Reason: GPT-4V fails to interpret the image. Based on the image and the explanation, the analysis should be 'we cannot actually see if there is a moat or not'. However, GPT-4V incorrectly thinks there's a moat and gives a wrong answer.

Ground Truth: (D) a barbican and battlements.

Explanation: We can see that the castle has outer walls with battlements on top, but we cannot actually see if there is a moat or not.

Figure 10. A sample error case of Art (subfield: Fine Arts). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

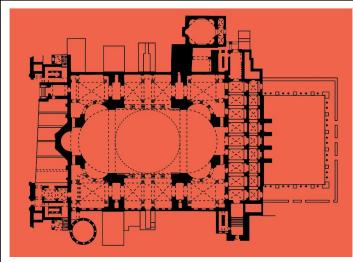


Art Theory: Art History

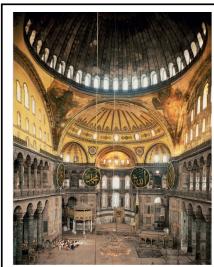
Question: Who was the patron of the building?

Options:

- (A) Abd al-Malik
- (B) Justinian
- (C) Charlemagne
- (D) Shah Jahan



<image 1>



<image 2>

GPT-4V: The building in the second image is the Hagia Sophia. The patron of Hagia Sophia was (B) Justinian.

Ground Truth: (B) Justinian

Figure 11. A sample correct case of Art Theory (subfield: Art History).

[Back to List of Figures](#) | [Back to Table Index](#)



Art Theory: Art Criticism

Question: What is NOT exhibited in the painting? <image 1>

Option:

- (A) hierarchical scale
- (B) graphic representation of horror and despair
- (C) a sense of immediacy and drama
- (D) use of sharply contrasting light and shade



<image 1>

GPT-4V: Based on the painting provided:

- (A) hierarchical scale - This technique, where figures or objects are represented in varying sizes to indicate their relative importance or position in the narrative, is evident in the painting. The central figures are larger and more pronounced, while secondary figures and elements are smaller or placed in the background.
- (B) graphic representation of horror and despair - This is clearly exhibited in the painting. The expressions on the faces of the figures, their body language, and the overall chaotic scene depict horror and despair.
- (C) a sense of immediacy and drama - The dynamic poses, facial expressions, and the chaotic nature of the scene give a strong sense of immediacy and drama.
- (D) use of sharply contrasting light and shade - The painting employs a technique known as chiaroscuro, where light and shade are used to give depth, volume, and a three-dimensional effect to the subjects. This is evident in the way light falls on certain figures, highlighting them, while others are in shadow.

Given these observations, the correct answer is:

- (A) hierarchical scale.

Ground Truth: (A) hierarchical scale

Figure 12. A sample correct case of Art Theory (subfield: Art Criticism).

[Back to List of Figures](#) | [Back to Table Index](#)



Art Theory: Art History

Question: The original placement of this painting was

Option:

- (A) in the Prado Museum in Madrid, showing a willingness to be compared to great Spanish painters of the past
- (B) in the cathedral in Madrid, indicating a desire to find acceptance with the Catholic Church
- (C) in the study of King Philip IV, challenging him to consider the multitude of interpretations the painting represents
- (D) in the reception room at the Royal Palace in Madrid, where the portraits of the royal family would be viewable by visiting dignitaries



<image 1>

GPT-4V: This painting is "Las Meninas" by Diego Velázquez. The original placement of this painting was **(A) in the Prado Museum in Madrid**, showing a willingness to be compared to great Spanish painters of the past.

Error Category: Lack of knowledge

Error Reason: GPT-4V successfully recognizes that the painting is Las Meninas (1656), painted by Velazquez, showcasing its perceptual ability. However, it does not know the fact that the work was placed in the study of King Philip IV. GPT-4V suggests that the painting was placed in Prado Museum in Madrid, and even provides a reason to support this answer, which is the willingness to be compared to great Spanish painters in the past. This shows some reasoning ability. However, the original placement is a piece of factual knowledge; the reasoning was based on incorrect knowledge and it led to a wrong answer. This behavior illustrates that GPT-4V lacks specific art knowledge.

Ground Truth: (C) in the study of King Philip IV, challenging him to consider the multitude of interpretations the painting represents

Figure 13. A sample error case of Art Theory (subfield: Art History). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)



Design : Industrial Design

Question: In the Robie House Dining room, what design elements were used to create a 'room within a room' vibe? <image 1>

Option:

- (A) Tall backs of chairs and lamps at the corners of dining tables
- (B) Low-back chairs and pendant lights in the center
- (C) Circular dining tables and wall-mounted shelves
- (D) Folding chairs and floor-standing lamps



<image 1>

GPT-4V: Based on the image you've provided of the Robie House Dining room, the design elements used to create a 'room within a room' vibe are:(A) Tall backs of chairs and lamps at the corners of dining tables.

Ground Truth: (A) Tall backs of chairs and lamps at the corners of dining tables

Figure 14. A sample correct case of Design (subfield: Industrial Design).

[Back to List of Figures](#) | [Back to Table Index](#)



Design: Design History

Question: The painting <image 1> reflects contemporary culture primarily through the mapping of
Option:

- (A) consumerism and national identities
- (B) flight paths and time intervals
- (C) sports car and racing signage
- (D) information systems on the Internet



<image 1>

GPT-4V: The painting (<image 1>) appears to be a complex and abstract representation with various lines, patterns, and colors. Based on the visual elements in the painting:(D) information systems on the Internet seems to be the most fitting description. The interconnected lines and vibrant, diverse patterns might suggest the vastness and complexity of information systems on the Internet. However, art interpretation can be subjective, and different viewers may see different meanings in the same piece.

Error Category: Perceptual Error

Error Reason: The model failed to recognize and interpret the symbolic elements of national flags present in the image, which are key indicators of 'consumerism and national identities', leading to an incorrect answer.

Ground Truth: (D) information systems on the Internet

Explanation: The painting reflects contemporary culture's consumerism and national identities through its visual elements that evoke the commercial marketplace and specific country symbols. Bright colors and recognizable patterns may suggest mass-produced goods and advertising, a common commentary on consumer culture in contemporary art. Additionally, abstract representations of flags, national colors, or iconic images could signify different countries' identities, illustrating the global interconnectedness of consumerism where products and brands cross borders and adapt to local cultures. The painting's complexity and layering mirror the saturated, interconnected nature of our globalized world, where the flow of consumer goods and cultural symbols creates a vibrant, dynamic map of modern life.

Figure 15. A sample error case of Design (subfield: Design History). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)



Design: Design History

Question: The double-tiered columns allowed for all of the following EXCEPT <image 1>

Option:

- (A) barrel-vaulted roofing
- (B) decorative rhythm and repetition
- (C) a higher roof to make up for the short columns
- (D) the entrance of light and air into the hall



<image 1>

GPT-4V: The double-tiered columns in the image provided would allow for: (A) barrel-vaulted roofing (B) decorative rhythm and repetition (C) a higher roof to make up for the short columns. However, the double-tiered columns would not directly affect: (D) the entrance of light and air into the hall. So the correct answer is (D) the entrance of light and air into the hall.

Error Category: Lack of Knowledge

Error Reason: The error reason is a lack of specific subject knowledge: the incorrect response suggests that double-tiered columns would not affect the entrance of light and air, which is not necessarily true. Architecturally, double-tiered columns do not prevent light and air from entering; such features are more influenced by the building's windows and overall design rather than the column structure itself. The correct answer should be that double-tiered columns would not allow for barrel-vaulted roofing, as the Mezquita features flat ceilings above the arches.

Ground Truth: (A) barrel-vaulted roofing

Explanation: The double-tiered columns depicted in the image, characteristic of Islamic architecture like that found in the Mosque-Cathedral of Córdoba, are not designed to support a barrel-vaulted roof; instead, they serve other purposes. These arches create a visually striking pattern of rhythm and repetition, elevate the roofline by stacking arches, which compensates for the shorter columns, and contribute to the structure's stability. However, they are not related to the entrance of light and air, which is typically managed by windows, nor are they structurally supportive of a barrel-vaulted roof, as their design and construction cater to a flat roofing system. Hence, the correct answer is "barrel-vaulted roofing," as it is the element not facilitated by the double-tiered columns.

Figure 16. A sample error case of Design (subfield: Design History). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)



Music : Music

Question: Select the correct tuning of Violin.

Option:

- (A) <image 1>
- (B) <image 2>
- (C) <image 3>
- (D) <image 4>



<image 1>

<image 2>

<image 3>

<image 4>

GPT-4V: The correct tuning for a violin is shown in the first image. The violin is typically tuned to the pitches G3, D4, A4, and E5, starting from the lowest string to the highest. In the image, the G-clef (treble clef) is used for the E and A strings, while the C-clef (alto clef) is used for the D string, and the F-clef (bass clef) for the G string. This notation is a way to visually represent the correct tuning pitches for the violin's four strings.

Ground Truth: (A) <image 1>

Figure 17. A sample correct case of Music (subfield: Music).

[Back to List of Figures](#) | [Back to Table Index](#)

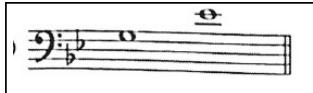


Music : Music

Question: Choose the name for the following interval. <image 1>

Option:

- (A) Perfect 6th
- (B) Compound perfect 4th
- (C) Major 6th
- (D) Minor 6th



<image 1>

GPT-4V: The interval shown in the image is a “**major 6th**.” It spans from the first note, F, to the second note, D. In a standard diatonic scale, the interval from F to D is a sixth, and since the key signature does not indicate any flats or sharps that would alter the D, it remains a major sixth. The options “perfect 6th” and “compound perfect 4th” do not exist in music theory as “perfect” is used for unison, 4th, 5th, and octave intervals. A “minor 6th” would require the D to be flattened, which is not the case here.

Error Category: Perceptual Error, Lack of knowledge

Error Reason: The model immediately recognized it as a 6th interval, but got confused whether it's a major 6th or a minor 6th. A closer look at the response reveals that the model failed to recognize the starting and ending notes of the interval. The first note is actually G, but the model mistook it for the adjacent note F. The second note is Eb, and the model might have simply deduced that it was D based on the assumption of major 6th + F. Additionally, the model's mistaken identification of a compound perfect 4th indicates a lack of subject knowledge.

Ground Truth: (D) Minor 6th

Figure 18. A sample error case of Music (subfield: Music). Error category: Perceptual Error, Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)



Music : Music

Question: Give the bar number of the following (from bar 13 onwards): a melodic interval of a compound minor 3rd in the right-hand part. <image 1>

Option:

- (A) 28 and 29
- (B) 17
- (C) 14 and 15
- (D) 18

<image 1>

GPT-4V: In response to your inquiry, I examined the provided image for the presence of a compound minor third interval in the right-hand part starting from bar 13. A compound minor third, or a tenth, combines an octave (twelve semitones) with a minor third (three semitones), totaling fifteen semitones. Upon inspection, I found that bar 17 doesn't have this interval in the right-hand part. Bars 14 and 15 do contain intervals, but I'm uncertain if any are a compound minor third. I can't confirm the presence of this interval in bars 28 and 29, as they're not visible in the image. Bar 18 also requires checking for this interval. Unfortunately, I can't process the image visually to provide the exact bar number of the specified interval.

Error Category: Perceptual Error

Error Reason: Surprisingly, the model fails to find the corresponding bar number based on the relative location of labeled bars. This is a simple visual inference task for humans.

Ground Truth: (A) 28 and 29

Figure 19. A sample error case of Music (subfield: Music). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)



Accounting : Investment

Question: Consider the three stocks in the following table. P_t represents price at time t , and Q_t represents shares outstanding at time t . Stock C splits two for one in the last period. Calculate the rate of return on a price-weighted index of the three stocks for the first period ($t=0$ to $t=1$).

Option:

- (A) 4.17%
- (B) 4.03%
- (C) 3.89%

	P_0	Q_0	P_1	Q_1	P_2	Q_2
A	90	100	95	100	95	100
B	50	200	45	200	45	200
C	100	200	110	200	55	400

<image 1>

GPT-4V: To calculate the rate of return on a price-weighted index of the three stocks for the first period ($t=0$ to $t=1$), follow these steps: Calculate the initial index value at $t=0$. Calculate the index value at $t=1$. Compute the rate of return using the two index values.

Initial Index Value at $t=0$:

$$\begin{aligned}\text{Price-weighted index value at } t=0 &= (P_0(A) + P_0(B) + P_0(C)) / 3 \\ &= (90 + 50 + 100) / 3 \\ &= 80\end{aligned}$$

Index Value at $t=1$:

$$\begin{aligned}\text{Price-weighted index value at } t=1 &= (P_1(A) + P_1(B) + P_1(C)) / 3 \\ &= (95 + 45 + 110) / 3 \\ &= 83.33\end{aligned}$$

Rate of Return:

$$\begin{aligned}\text{Rate of Return} &= (\text{Index Value at } t=1 - \text{Index Value at } t=0) / \text{Index Value at } t=0 \\ &= (83.33 - 80) / 80 \\ &= 0.041625 \\ &= 4.1625\%\end{aligned}$$

Therefore, the rate of return on a price-weighted index of the three stocks for the first period is approximately 4.17%.

The correct answer is:

- (A) 4.17%

Ground Truth: (A) 4.17%

Explanation: At $t=0$, the value of the index is: $(90 + 50 + 100)/3 = 80$. At $t=1$, the value of the index is: $(95 + 45 + 110)/3 = 83.33$. The rate of return is: $(83.33/80) - 1 = 4.17\%$

Figure 20. A sample correct case of Accounting (subfield: Investment).

[Back to List of Figures](#) | [Back to Table Index](#)



Accounting: Financial Accounting

Question: From the following Company Y adjusted trial balance, what is the retained earnings to be reported? <image 1>

Option:

- (A) \$17,800
- (B) \$16,800
- (C) \$15,550
- (D) \$12,280

Adjusted Trial Balance		
	Debit	Credit
Cash	\$ 32,000	
Accounts receivable	17,300	
Prepaid insurance	6,400	
Land	10,000	
Accounts payable		\$ 10,900
Salaries payable		6,000
Common stock		31,000
Retained earnings		4,200
Dividends	8,000	
Service revenue		74,000
Insurance expense	5,600	
Salaries expense	24,000	
Miscellaneous expense	22,800	
	126,100	126,100

<image 1>

GPT-4V: To calculate the retained earnings to be reported for Company Y from the adjusted trial balance, we need to consider the following components:

1. Beginning Retained Earnings: This would be the retained earnings balance at the start of the period. However, since it is not provided, we can assume it is included in the current retained earnings figure in the trial balance or that this is the first year of operation, and as such, it would be zero.

2. Add: Net Income (or Less: Net Loss): Net income is calculated as revenues minus expenses. If expenses exceed revenues, it results in a net loss.

3. Less: Dividends: These are distributions to shareholders, which reduce retained earnings.

From the adjusted trial balance provided, the calculation would be as follows:

Net Income: Service Revenue (\$74,000) - (Insurance Expense \$5,600 + Salaries Expense \$24,000 + Miscellaneous Expense \$22,800) = \$74,000 - \$52,400 = \$21,600.

Less Dividends: \$8,000. Retained Earnings to be Reported = Beginning Retained Earnings (if not given, assumed to be included in the trial balance) + Net Income - Dividends. The trial balance shows Retained Earnings as \$6,000. Assuming this is the beginning balance and adding the net income less dividends: Retained Earnings to be Reported = \$6,000 + \$21,600 - \$8,000 = \$19,600. There is no answer matching the option.

Error Category: Perceptual Error

Error Reason: GPT-4V understands the question and can know how to calculate the Retained Earnings to be Reported. Most of the calculation process is correct. However, in the final steps, GPT-4V wrongly extracts the number of Retained Earnings from the table image (should be \$4,200 instead of \$6,000).

Ground Truth: (A) \$17,800

Figure 21. A sample error case of Accounting (subfield: Financial Accounting). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

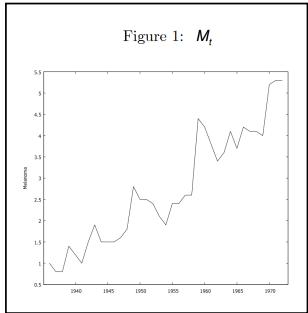


Economics : Econometrics

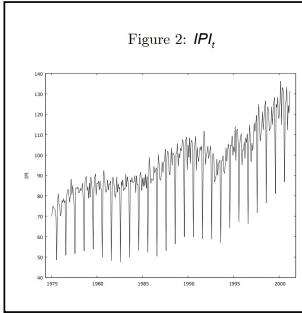
Question: Figure 1 displays the number of cases of Melanoma (M_t) in the male population of a US State, from 1936 to 1972. Figure 2 shows the Spanish Industrial Production Index (IPI_t), from January 1975 to March 2001. According to figures 1 and 2, which of the following statements is FALSE:

Option:

- (A) IPI_t displays a strong seasonality
- (B) Both time series are mean stationary
- (C) Both time series display a trending behavior



<image 1>



<image 2>

GPT-4V: Based on the provided figures:(A) IPI_t displays a strong seasonality. - This statement appears to be TRUE. Figure 2 shows consistent and repeating peaks and valleys for the Spanish Industrial Production Index (IPI_t), indicative of seasonality.(B) Both time series are mean stationary. - This statement appears to be FALSE. A mean stationary time series would have a constant mean (average) over time. Both figures show a changing mean over the given time frames, particularly Figure 1 where the mean value for Melanoma cases appears to be increasing over time.(C) Both time series display a trending behavior. - This statement appears to be TRUE. Figure 1 shows an upward trend in the number of Melanoma cases over time, and Figure 2 displays fluctuations but also an overall upward trend in the Industrial Production Index. Therefore, the FALSE statement is:
(B) Both time series are mean stationary.

Ground Truth: (B) Both time series are mean stationary.

Figure 22. A sample correct case of Economics (subfield: Econometrics).

[Back to List of Figures](#) | [Back to Table Index](#)



Economics: Macroeconomics

Question: Question below is based on the production possibilities of two nations that can produce both crepes and paper. <image 1> Which of the following statements is true of these production possibilities?

Option:

- (A) Nation X has comparative advantage in paper production and should trade paper to Nation Y in exchange for crepes.
- (B) Nation X has comparative advantage in crepe production and should trade crepes to Nation Y in exchange for paper
- (C) Nation X has absolute advantage in paper production, and Nation Y has absolute advantage in crepe production. No trade is possible.
- (D) Nation Y has absolute advantage in paper production, and Nation X has absolute advantage in crepe production. No trade is possible.
- (E) Nation Y has comparative advantage in crepe production and should trade paper to Nation X in exchange for crepes.

Industry	Economic Forecast
a. Housing construction	(i) Deep recession: falling inflation, interest rates, and GDP
b. Health care	(ii) Superheated economy: rapidly rising GDP, increasing inflation and interest rates
c. Gold mining	(iii) Healthy expansion: rising GDP, mild inflation, low unemployment
d. Steel production	(iv) Stagflation: falling GDP, high inflation

<image 1>

GPT-4V: To determine which nation has a comparative advantage in producing a particular good, we need to calculate the opportunity cost for each good in both nations. Opportunity cost is calculated as the amount of the other good that has to be given up to produce one unit of a particular good. Opportunity cost for Nation X: For 1 crepe, Nation X gives up producing 3 units of paper. So, opportunity cost of 1 crepe = 3 paper. For 1 paper, Nation X gives up producing $\frac{1}{3}$ unit of crepe. So, opportunity cost of 1 paper = $\frac{1}{3}$ crepe. Opportunity cost for Nation Y: For 1 crepe, Nation Y gives up producing 5 units of paper. So, opportunity cost of 1 crepe = 5 paper. For 1 paper, Nation Y gives up producing $\frac{1}{5}$ unit of crepe. So, opportunity cost of 1 paper = $\frac{1}{5}$ crepe. Comparing opportunity costs: Nation X has a lower opportunity cost ($\frac{1}{3} < \frac{1}{5}$) in producing paper than Nation Y. So, Nation X has a comparative advantage in producing paper. Nation Y has a lower opportunity cost ($5 > 3$) in producing crepes than Nation X. So, Nation Y has a comparative advantage in producing crepes. Given this, the statement:

- (A) Nation X has comparative advantage in paper production and should trade paper to Nation Y in exchange for crepes. is the correct answer.

Error Category: Perceptual Error

Error Reason: Although GPT-4V correctly identifies the numbers in the table, it fails to correctly attribute these numbers to the correct objects for production. For example, when calculating the opportunity cost of Nation X, GPT-4V makes the mistake that nation X can either produce 3 units of crepe or 9 units of piece. But in fact, according to the table, nation X can produce 3 units of paper or 9 units of crepe. This error leads to the incorrect calculation of opportunity cost and the incorrect answer.

Ground Truth: (A) Nation X has comparative advantage in paper production and should trade paper to Nation Y in exchange for crepes.

Figure 23. A sample error case of Economics (subfield: Macroeconomics). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

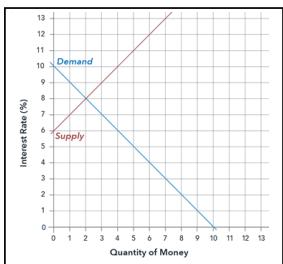


Economics : Macroeconomics

Question: The graph below shows the supply and demand curves in the market for credit card borrowing. <image 1> What is the equilibrium interest rate?

Option:

- (A) 7%
- (B) 8%
- (C) 9%
- (D) 10%



<image 1>

GPT-4V: The equilibrium interest rate is where the demand curve intersects the supply curve. In the provided graph, the point of intersection appears to be at an interest rate of: (C) 9%

Error Category: Perceptual Error

Error Reason: GPT-4V has the knowledge that the equilibrium interest rate is where the demand curve intersects the supply curve. However, GPT-4V fails to interpret the image. The point of intersection is at an interest rate of 8% instead of 9%.

Ground Truth: (B) 8%

Figure 24. A sample error case of Economics (subfield: Macroeconomics). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)



Finance: Corporate Finance

Question: Spike Equino is the CEO of a private medical equipment company that is proposing to sell 100,000 shares of its stock in an open auction. Suppose the company receives the bids in the following table. <image 1> What will be the company's total receipts (\$) from the sale if the auction is a uniform price auction?

Shares	Price
20,000	\$80
10,000	78
15,000	73
25,000	70
10,000	69
8,000	67
14,000	66
15,000	65
30,000	61

<image 1>

GPT-4V: In a uniform price auction, all successful bidders pay the same price per share, which is the lowest price at which the entire quantity can be sold. To determine this price, we need to accumulate the bids starting from the highest price until the total number of shares reaches or exceeds the quantity being sold (100,000 shares in this case). Using the given table: 20,000 shares at \$80 10,000 shares at \$78 15,000 shares at \$73 25,000 shares at \$70 10,000 shares at \$69 8,000 shares at \$67 14,000 shares at \$66 15,000 shares at \$65 30,000 shares at \$61. Accumulating the shares: 20,000 shares (total: 20,000) 10,000 shares (total: 30,000) 15,000 shares (total: 45,000) 25,000 shares (total: 70,000) 10,000 shares (total: 80,000) 8,000 shares (total: 88,000) 14,000 shares (total: 102,000). We can see that after accumulating the first 7 bids, we exceed the 100,000 shares being sold. Therefore, the price that every successful bidder will pay is \$66 per share (which is the lowest bid price that allows the company to sell all 100,000 shares). To calculate the company's total receipts from the sale: $100,000 \text{ shares} \times \$66/\text{share} = \$6,600,000$.

The company's total receipts from the sale will be \$6,600,000.

Ground Truth: \$6600000

Figure 25. A sample correct case of Finance (subfield: Corporate Finance).

[Back to List of Figures](#) | [Back to Table Index](#)

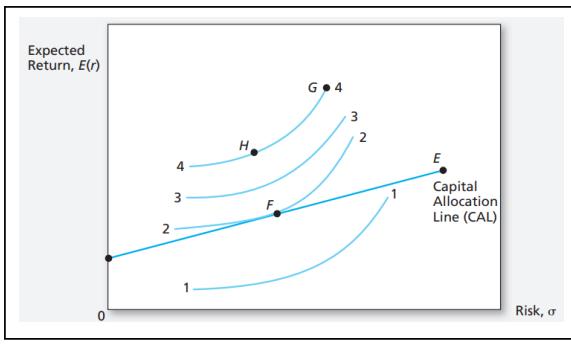


Finance : Financial Marketing

Question: <image 1> Which point designates the optimal portfolio of risky assets?

Option:

- (A) Point E
- (B) Point F
- (C) Point G
- (D) Point H



<image 1>

GPT-4V: The optimal portfolio of risky assets is the one that provides the highest expected return for a given level of risk. On the Capital Market Line (CML) or Capital Allocation Line (CAL) depicted in the graph, this is represented by the tangent point where the CAL touches the efficient frontier (the curve of possible portfolios). From the graph provided, Point F is where the CAL is tangent to the efficient frontier, which means Point F designates the optimal portfolio of risky assets. The answer is: (B) Point F.

Error Category: Reasoning Error

Error Reason: GPT-4V accurately identifies point F as the optimal portfolio of risky assets in the given problem and applies the correct theoretical framework in its reasoning. However, it overlooks an important aspect of point E, which also lies on the Capital Allocation Line (CAL) and is situated at the CAL's rightmost end. Given that point E offers a higher expected return while maintaining the same Sharpe ratio as point F, it should be considered more optimal than point F. This oversight in GPT-4V's analysis leads to an incorrect conclusion, as it fails to adequately assess the significance of point E in the context of portfolio optimization.

Ground Truth: (A) Point E

Figure 26. A sample error case of Finance (subfield: Financial Marketing). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)



Manage: Cost Management

Question: Eastern Chemical Company produces three products. The operating results of the current year are: <image 1> The firm sets the target price of each product at 150% of the product's total manufacturing cost. It appears that the firm was able to sell Product C at a much higher price than the target price of the product and lost money on Product B. Tom Watson, CEO, wants to promote Product C much more aggressively and phase out Product B. He believes that the information suggests that Product C has the greatest potential among the firm's three products because the actual selling price of Product C was almost 50% higher than the target price, while the firm was forced to sell Product B at a price below the target price. Both the budgeted and actual factory overhead for the current year are \$510,000. The actual units sold for each product also are the same as the budgeted units. The firm uses direct labor dollars to assign manufacturing overhead costs. The direct materials and direct labor costs per unit for each product are: <image 2> The controller noticed that not all products consumed factory overhead similarly. Upon further investigation, she identified the following usage of factory overhead during the year: <image 3> Determine the manufacturing cost per unit for Product A using the volume-based method

Product	Sales Quantity	Target Price	Actual Price	Difference
A	1,000	\$285.00	\$286.00	\$ 1.00
B	5,000	297.60	255.60	(42.00)
C	500	202.50	310.00	\$107.50

<image 1>

	Product A	Product B	Product C
Direct materials	\$50.00	\$114.40	\$65.00
Direct labor	20.00	12.00	10.00
Total prime cost	<u>\$70.00</u>	<u>\$126.40</u>	<u>\$75.00</u>

<image 2>

	Product A	Product B	Product C	Total Overhead
Number of setups	2	5	3	\$ 9,000
Weight of direct materials (pounds)	400	250	350	110,000
Waste and hazardous disposals	25	45	30	250,000
Quality inspections	30	35	35	75,000
Utilities (machine hours)	2,000	7,000	1,000	66,000
Total				<u>\$510,000</u>

<image 3>

GPT-4V: To determine the manufacturing cost per unit for Product A using the volume-based method, we'll follow these steps: Calculate the total overhead rate. Calculate the total manufacturing cost for Product A. Calculate the manufacturing cost per unit for Product A.1. Calculate the total overhead rate: Overhead rate = Total Factory Overhead / Total Direct Labor Dollars. Using the information given: Total Factory Overhead = \$510,000, Direct labor cost for: Product A = \$20.00, Product B = \$12.00, Product C = \$10.00. Sales Quantity: Product A = 1,000 units, Product B = 5,000 units, Product C = 500 units. Total Direct Labor Dollars = (Product A direct labor cost per unit × Product A Sales Quantity) + (Product B direct labor cost per unit × Product B Sales Quantity) + (Product C direct labor cost per unit × Product C Sales Quantity) = (\$20.00 × 1,000) + (\$12.00 × 5,000) + (\$10.00 × 500) = \$20,000 + \$60,000 + \$5,000 = \$85,000
Overhead Rate:= Total Factory Overhead / Total Direct Labor Dollars= \$510,000 / \$85,000 = \$6 per direct labor dollar2. Calculate the total manufacturing cost for Product A: Manufacturing cost = Direct Materials + Direct Labor + Overhead. Using the information given for Product A: Direct Materials = \$50.00, Direct Labor = \$20.00, Overhead = Overhead Rate × Direct Labor= \$6 × \$20.00= \$120.00
Total Manufacturing Cost for Product A:= Direct Materials + Direct Labor + Overhead= \$50.00 + \$20.00 + \$120.00= \$190.00
3. Calculate the manufacturing cost per unit for Product A: Since the actual units sold for each product are the same as the budgeted units, we can use the given sales quantity for Product A: Manufacturing Cost Per Unit for Product A:= Total Manufacturing Cost for Product A / Product A Sales Quantity= \$190,000 units= \$0.19 per unit (This seems too low, there might be an error in our calculations or in the given data.).

However, since we're asked for the manufacturing cost per unit, \$190.00 is the correct manufacturing cost per unit for Product A using the volume-based method.

Ground Truth: \$190

Figure 27. A sample correct case of Manage (subfield: Cost Management).

[Back to List of Figures](#) | [Back to Table Index](#)

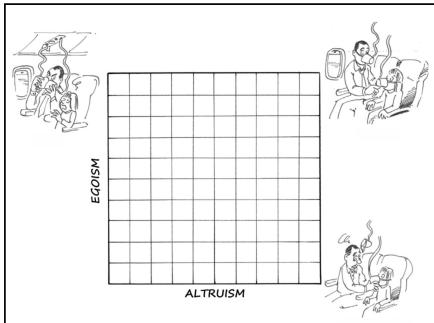


Manage: Management Models

Question: Look at the situation in the <image 1>. You are traveling on a plane with a small child. The cabin is depressurized and the oxygen mask falls from the ceiling. Which of these pictures shows the reconciliation of egoism and other-isms (the order is from left to right, top to bottom)?

Option:

- (A) 1
- (B) 2
- (C) 3



<image 1>

GPT-4V: The image shows three situations on a plane when the oxygen mask falls: The adult puts on the mask first while the child watches (egoism).The adult puts the mask on the child first while struggling for breath. Both the adult and the child share the mask. The reconciliation of egoism and other-isms would be a situation where both the individual's needs and the needs of others are considered and addressed. Therefore, the correct answer is:

- (C) 3

Error Category: Perceptual Error

Error Reason: GPT-4V recalled the right knowledge and made the right reasoning, but it failed to correctly map the IDs to the corresponding illustrations in the figure, which is not explicitly marked in the figure but is only described in text (the order is from left to right, top to bottom).

Ground Truth: (B) 2

Figure 28. A sample error case of Manage (subfield: Management Models). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)



Marketing: Market Research

Question: A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four (4) new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as Table 10.27. <image 1>

Option:

- (A) a test of two independent means
- (B) a test of two proportions
- (C) a test of a single mean
- (D) a test of a single proportion

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Table 10.27

<image 1>

GPT-4V: Based on the data provided in Table 10.27, the golf instructor is comparing the mean scores of her students before and after taking her class. She is interested in determining if there's a significant difference in the mean scores of the same students before and after learning her new technique. Given this, the most appropriate statistical test to use in this scenario would be:(A) a test of two independent means.

Ground Truth: (A) a test of two independent means

Figure 29. A sample correct case of Marketing (subfield: Market Research).

[Back to List of Figures](#) | [Back to Table Index](#)

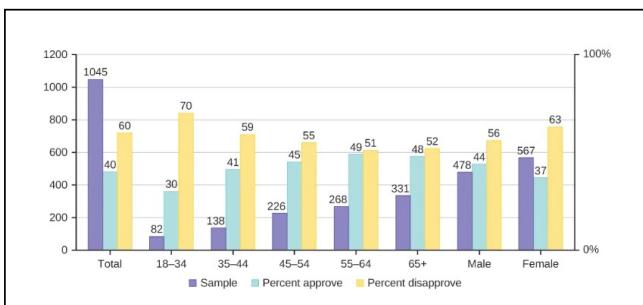


Marketing: Market Research

Question: The graph in <image 1> displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045. Out of the males and females, what percent are males?

Option:

- (A) 44.8%
- (B) 44.5%
- (C) 45.7%
- (D) 46.8%



<image 1>

GPT-4V: To find the percentage of males in the sample: First, we'll identify the number of males and females from the graph:

Males = 52, Females = 63, Sum of males and females = $52 + 63 = 115$

Now, to calculate the percentage of males:

$$\begin{aligned} \text{Percentage of males} &= (\text{Number of males} / \text{Total number of males and females}) \times 100 \\ &= () \times 100 \\ &= 45.22\% \end{aligned}$$

The closest answer to 45.22% from the given options is: (B) 45.5%

Error Category: Perceptual Error

Error Reason: GPT-4V misunderstands the percentage of male and female as the number of samples.

Ground Truth: (C) 45.7%

Figure 30. A sample error case of Marketing (subfield: Market Research). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

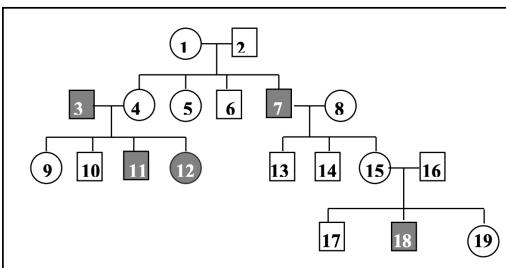


Biology: Genetics

Question: The pedigree in <image 1> shows the mode of inheritance of a human disease that is associated with mutations in Gene A. The pedigree would be consistent with an autosomal recessive mode of inheritance if individual 18 were a carrier.

Option:

- (A) True
- (B) False
- (C) Can't tell from available information



<image 1>

GPT-4V: The given pedigree shows the pattern of inheritance of a particular trait or disease. Let's analyze the pedigree to determine the mode of inheritance: Circles represent females, and squares represent males. Filled symbols (circles or squares) indicate individuals affected by the disease, and unfilled symbols indicate unaffected individuals. For the trait to be autosomal recessive: Both parents must be carriers to have an affected offspring. It should not skip generations unless both parents are carriers and just don't show the trait. From the pedigree, we see that: Individuals 3 and 12 are affected. This means both their parents (2 and 4 for individual 3, and 6 and 7 for individual 12) must be carriers. Individual 18's parents (15 and 17) are both unaffected. If individual 18 is a carrier, it means both 15 and 17 must also be carriers. Given the pattern in the pedigree, the statement is true: the pedigree would be consistent with an autosomal recessive mode of inheritance if individual 18 were a carrier.

Answer: (A) True

Ground Truth: (A) True

Figure 31. A sample correct case of Biology (subfield: Genetics).

[Back to List of Figures](#) | [Back to Table Index](#)

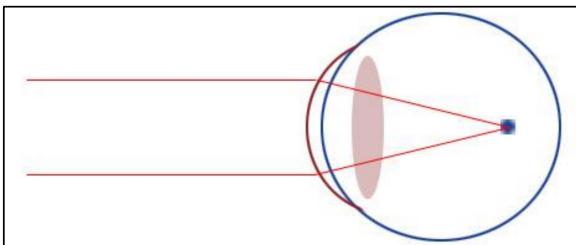


Biology: Physiology

Question: The patient is a post-mature newborn. What is the most likely diagnosis? <image 1>

Option:

- (A) Vision may be restored with concave lens and refractive surgery
- (B) Vision may be restored with convex lens and refractive surgery
- (C) This condition often occurs in elderly people
- (D) This phenomenon can not be fixed by LASIK (laser-assisted in situ keratomileusis) surgery



<image 1>

GPT-4V: The image depicts a case of myopia, or nearsightedness. In myopia, light entering the eye is focused in front of the retina rather than directly on it, which results in distant objects appearing blurred. Given the choices:(A) Vision may be restored with concave lens and refractive surgery - This is correct. A concave lens can help to spread out the light rays so that they can focus on the retina correctly. Additionally, refractive surgeries like LASIK can help correct myopia.(B) Vision may be restored with convex lens and refractive surgery - This is incorrect. Convex lenses are used for hyperopia or farsightedness, not myopia.(C) This condition often occurs in elderly people - This is not necessarily true. Myopia can begin in childhood and progress into adulthood. It is not limited to elderly people.(D) This phenomenon can not be fixed by LASIK (laser-assisted in situ keratomileusis) surgery - This is incorrect. LASIK is one of the common surgeries used to correct myopia.So, the true statement is:

(A) Vision may be restored with concave lens and refractive surgery.

Ground Truth: (A) Vision may be restored with concave lens and refractive surgery

Figure 32. A sample correct case of Biology (subfield: Physiology).

[Back to List of Figures](#) | [Back to Table Index](#)

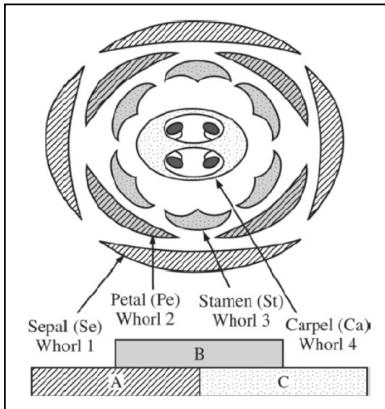


Biology: Genetics

Question: Consider the figure <image 1> for ABC hypothesis. If wild type Arabidopsis were transformed with a chimeric gene consisting of an A class promoter fused to a B class coding sequence, which of the following arrangements (inner to outer) would be predicted?

Option:

- (A) Ca, St, St, St.
- (B) Ca, St, Se, Se.
- (C) Ca, St, Se, Pe.
- (D) Ca, St, Pe, Se.
- (E) Ca, St, St, Ca.



<image 1>

GPT-4V: The ABC model is a classic model for understanding flower development. This model stipulates that: A-class genes alone specify sepals. A-class and B-class genes together specify petals. B-class and C-class genes together specify stamens. C-class genes alone specify carpels. Given the information, if a wild type Arabidopsis were transformed with a chimeric gene consisting of an A class promoter fused to a B class coding sequence, then the B class gene would be expressed where normally only the A class gene would be active. Here's how the flower would develop according to the modified ABC model: Whorl 1 (outermost): Normally A class genes would specify sepals, but with the B class coding sequence under the A class promoter, it would be A+B which specifies petals.

Whorl 2: A + B class genes together specify petals, and this remains unchanged.

Whorl 3: Normally B + C class genes together would specify stamens. This remains unchanged.

Whorl 4 (innermost): C class genes alone specify carpels, and this remains unchanged. So the new arrangement from innermost to outermost would be:

Ca, St, Pe, Pe. So, the arrangement from outermost to innermost would be:

Pe, Pe, St, Ca The closest match to the given options would be:

- (D) Ca, St, Pe, Se.

Error Category: Reasoning Error

Error Reason: GPT-4V correctly interprets the image. However, it does not correctly reason about the relationship between a 'coding sequence' and a 'class promoter'. This prevents it from correctly predicting the answer for Whorl 1 and Whorl 2.

Ground Truth: (B) Ca, St, Se, Se.

Figure 33. A sample error case of Biology (subfield: Genetics). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

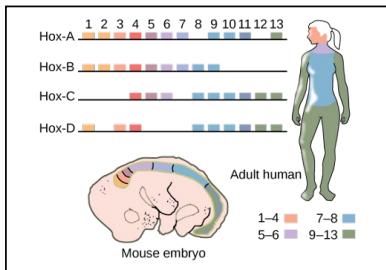


Biology: Genetics

Question: Based on <image 1> if a Hox 13 gene in a mouse was replaced with a Hox 1 gene, which of the following will be the most likely outcome?

Option:

- (A) The mouse may develop one head and one tail
- (B) The mouse may develop two heads and no tail
- (C) The mouse may develop no head and two tails



<image 1>

GPT-4V: The Hox genes are a group of related genes that control the body plan of an embryo along the anterior-posterior (head-tail) axis. These genes determine where limbs and other body segments will grow in a developing fetus.

In the image provided, you appear to reference a mouse with a head and a tail, with what seems to be color-coding or sequencing perhaps representing the Hox genes.

To understand the likely outcome, it's essential to consider the function of Hox genes:

Hox 1 genes are typically expressed in anterior regions of the body, such as the head.

Hox 13 genes are typically expressed in posterior regions, like the tail.

Given this information, if a Hox 13 gene in a mouse was replaced with a Hox 1 gene, the posterior region where the tail would normally develop would now be getting instructions to develop anterior structures.

Therefore, the most likely outcome is:

- (C) The mouse may develop no head and two tails.

Error Category: Reasoning Error

Error Reason: GPT-4V accurately interpreted the image and correctly recalled that "Hox 1 genes are typically expressed in anterior regions of the body, such as the head. Hox 13 genes are typically expressed in posterior regions, like the tail." The model also correctly processed most of the intermediate reasoning steps. However, the concluding statement "The mouse may develop no head and two tails" does not logically follow from the prior step: "the posterior region where the tail would normally develop would now be getting instructions to develop anterior structures."

Ground Truth: (B) The mouse may develop two heads and no tail

Figure 34. A sample error case of Biology (subfield: Genetics). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

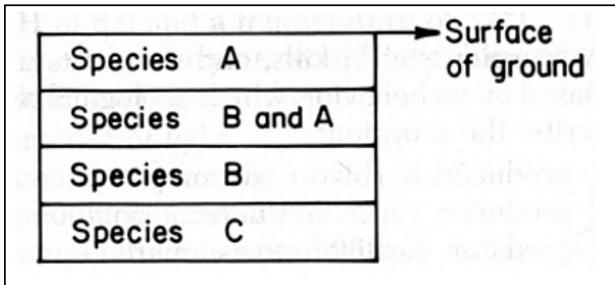


Biology: Evolution

Question: The diagram in <image 1> represents a section of undisturbed layers of sedimentary rock in New York State and shows the location of fossils of several closely related species. According to currently accepted evolutionary theory, which is the most probable assumption about species A, B and C?

Option:

- (A) Species B is more abundant than species C
- (B) Species A and B are genetically identical
- (C) Both species A and C are descended from species B
- (D) Species B descended from species A
- (E) Species C existed before species B



<image 1>

GPT-4V: The diagram depicts layers of sedimentary rock with fossils of different species. In sedimentary rock layers, the oldest layers are at the bottom and the youngest layers are at the top. Based on this, the order of appearance of the species from oldest to youngest is: Species C → Species B → Species A (and then A continues to exist with B). Given this order and the choices provided, the most probable assumption according to currently accepted evolutionary theory is:
(D) Species B descended from species A.

Error Category: Reasoning Error

Error Reason: GPT-4V correctly understands the evolutionary order of A, B and C ($C \rightarrow B \rightarrow A$). But D is a contradiction of this timeline order.

Ground Truth: (E) Species C existed before species B

Figure 35. A sample error case of Biology (subfield: Evolution). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

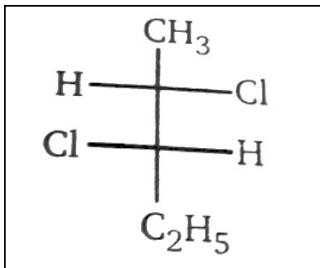


Chemistry: Organic Chemistry

Question: The configuration at C-2 and C-3 of the compound given: <image 1>

Option:

- (A) 2R, 3S
- (B) 2S, 3R
- (C) 2S, 3S
- (D) 2R, 3R



<image 1>

GPT-4V: To determine the configuration (R or S) of a chiral center, we follow the Cahn-Ingold-Prelog (CIP) priority rules: Assign priorities based on atomic number. The atom bonded to the chiral center with the highest atomic number gets the highest priority (1). If two atoms have the same atomic number, consider the atoms they are bonded to in order to determine priority. Orient the molecule so that the atom or group with the lowest priority (4) is pointing away (in the background). Observe the order of groups 1-2-3: If the order is clockwise, the configuration is R. If the order is counterclockwise, the configuration is S. Let's determine the configuration for the two chiral centers: For C-2: Attached groups and their atomic numbers: CH₃ (Carbon = 6)Cl (Chlorine = 17)H (Hydrogen = 1)C₂H₅ (Carbon = 6, but this will have higher priority than CH₃ because the subsequent carbons in the chain will be used for breaking the tie). Priority order: ClC₂H₅CH₃H. With the H pointing to the background, the order is 1 (Cl) to 2 (C₂H₅) to 3 (CH₃), which is clockwise. So, the configuration at C-2 is R. For C-3: Attached groups and their atomic numbers: H (Hydrogen = 1)Cl (Chlorine = 17)CH₃ (Carbon = 6)C₂H₅ (Carbon = 6, but like before, C₂H₅ will have higher priority than CH₃). Priority order: ClC₂H₅CH₃H. With the H pointing to the background, the order is 1 (Cl) to 2 (C₂H₅) to 3 (CH₃), which is counterclockwise. So, the configuration at C-3 is S. Therefore, the correct configuration for the compound is: 2R, 3S. So, the answer is (A) 2R, 3S.

Ground Truth: (A) 2R, 3S

Figure 36. A sample correct case of Chemistry (subfield: Organic Chemistry).

[Back to List of Figures](#) | [Back to Table Index](#)

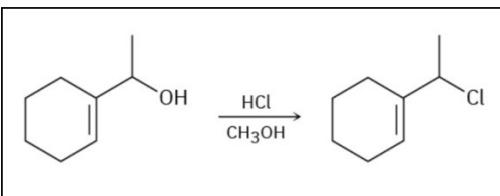


Chemistry: Organic Chemistry

Question: Which of the statement best describes the reaction in the following figure? <image 1>

Option:

- (A) It is likely to be S_N1.
- (B) It is likely to be S_N2.
- (C) None of the above.



<image 1>

GPT-4V: The reaction depicted involves the conversion of an alcohol to an alkyl chloride in the presence of HCl and CH₃OH (methanol). Given the reagents and conditions:(A) It is likely to be S_N1. The alcohol initially forms a carbocation intermediate after losing the hydroxyl group in the presence of HCl. Then, chloride ion attacks the carbocation to form the alkyl chloride product. The mechanism of this reaction is consistent with the S_N1 pathway.

Ground Truth: (A)

Figure 37. A sample correct case of Chemistry (subfield: Organic Chemistry).

[Back to List of Figures](#) | [Back to Table Index](#)

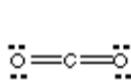
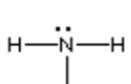
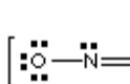
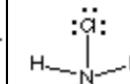
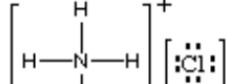


Chemistry: Inorganic Chemistry

Question: Which of the following Lewis (electron-dot) structures is/are wrong? <image 1>

Option:

- (A) (a)
- (B) (a) and (d)
- (C) (a) and (b)
- (D) (d) and (e)

				
(a) SO_2	(b) NH_3	(c) NO_2^+	(d) NH_4Cl	(e) NH_4Cl

<image 1>

GPT-4V: Let's analyze each of the Lewis structures:(a) SO_2 : The structure shows sulfur with a double bond to each oxygen atom. Each oxygen atom has two lone pairs, and sulfur has no lone pairs. This is the correct Lewis structure for SO_2 .(b) NH_3 : The structure shows nitrogen with a single bond to each of the three hydrogen atoms and one lone pair on nitrogen. This is the correct Lewis structure for NH_3 .(c) NO_2^+ : The structure shows nitrogen with a double bond to one oxygen and a single bond to the other oxygen. The single bonded oxygen has three lone pairs, while the double bonded oxygen has two lone pairs. This results in the nitrogen having a positive charge. This is the correct Lewis structure for NO_2^+ .(d) NH_4Cl : The structure shows the ammonium ion (NH_4^+) with nitrogen having four single bonds to four hydrogen atoms, and there's no lone pair on nitrogen. However, chlorine is shown with a single bond to a hydrogen atom, which is incorrect. Chlorine should not be bonded to hydrogen in this structure. Instead, NH_4^+ and Cl^- are present as ions, and they are attracted to each other through ionic bonding.(e) NH_4Cl : This structure shows the correct ionic representation with the ammonium ion (NH_4^+) and the chloride ion (Cl^-) as separate entities. Based on the analysis, the incorrect Lewis structure is: (d) NH_4Cl . The correct answer is: (D) (d) and (e)

Error Category: Perceptual Error, Reasoning Error

Error Reason: For choice (a), GPT-4V incorrectly recognizes the carbon atom as a sulfur atom, and even if it was a sulfur atom, there would be two electrons near the sulfur atom. Therefore, the Lewis structure in choice (a) is completely wrong and should be one of the final answers. For choice (e), even though the model correctly decides that it is correct, it still chooses a wrong option that contains (e).

Ground Truth: (B) (a) and (d)

Figure 38. A sample error case of Chemistry (subfield: Inorganic Chemistry). Error category: Perceptual Error, Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

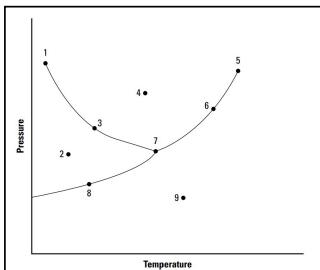


Chemistry: Inorganic Chemistry

Question: <image> Which point corresponds to a sublimation/ deposition equilibrium?

Option:

- (A) 1
- (B) 3
- (C) 5
- (D) 6
- (E) 8



<image 1>

GPT-4V: The graph provided is a phase diagram, which illustrates the conditions (in terms of pressure and temperature) under which different phases of a substance coexist in equilibrium. In such a diagram: The line between the solid and gas regions represents the sublimation/deposition equilibrium, where a substance can transition directly between the solid and gaseous states. The line between the solid and liquid regions represents the melting/freezing equilibrium. The line between the liquid and gas regions represents the boiling/condensation equilibrium. From the image provided, point (B) 3 is on the line between the solid and gas regions, representing the sublimation/deposition equilibrium. So, the correct answer is: (B) 3.

Error Category: Lack of Knowledge

Error Reason: The model correctly describes the knowledge, but due to the lack of knowledge about the phase diagram, it thinks point 3 is the one on the line between the solid and gas regions, while the correct one is point 8.

Ground Truth: (B) 3

Figure 39. A sample error case of Chemistry (subfield: Inorganic Chemistry). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)

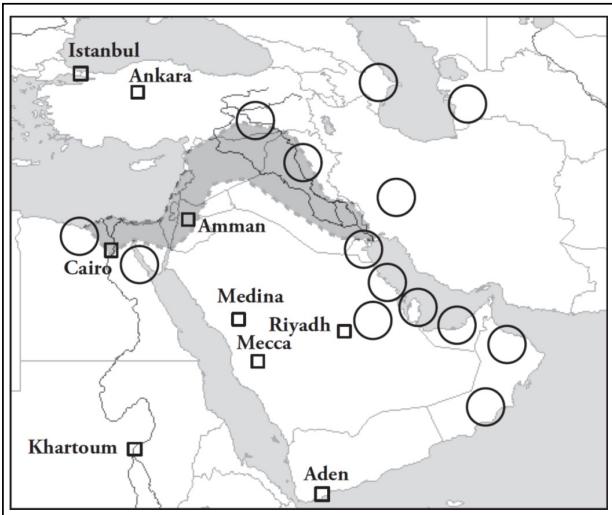


Geography: Human Geography

Question: The circled areas on <image 1> are all areas which

Option:

- (A) are under the control of al-Qaeda from 2001 to present
- (B) were under the control of the U.S.-led military coalition as of 2003
- (C) are home to the Kurdish peoples
- (D) are production zones of proven oil reserves
- (E) have large-scale irrigated agriculture



<image 1>

GPT-4V: Based on the provided map, the circled areas seem to align with regions that:

(D) are production zones of proven oil reserves. Many of the circled areas are in regions known for their significant oil reserves, such as parts of Saudi Arabia, Iraq, and other parts of the Middle East.

Ground Truth: (D) are production zones of proven oil reserves

Figure 40. A sample correct case of Geography (subfield: Human Geography).

[Back to List of Figures](#) | [Back to Table Index](#)

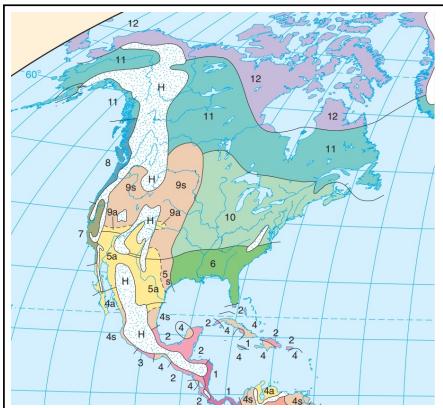


Geography: Physical Geography

Question: Which weather does 11 represent in <image 1>?

Option:

- (A) Boreal forest climate
- (B) Moist subtropical climate
- (C) Wet-dry tropical climate
- (D) Tundra climate



<image 1>

GPT-4V: The image appears to be a climate zone map. The zone labeled "11" corresponds to regions such as northern Canada and parts of Russia. Based on the options provided, zone "11" would represent:

- (D) Tundra climate.

Error Category: Reasoning Error

Error Reason: GPT-4V correctly summarize that the zone labeled "11" corresponds to regions such as northern Canada and parts of Russia, which is a correct interpretation of the map. However, GPT-4V wrongly gets the conclusion that zone "11" would be Tundra climate based on the observation.

Ground Truth: (A) Boreal forest climate

Figure 41. A sample error case of Geography (subfield: Physical Geography). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

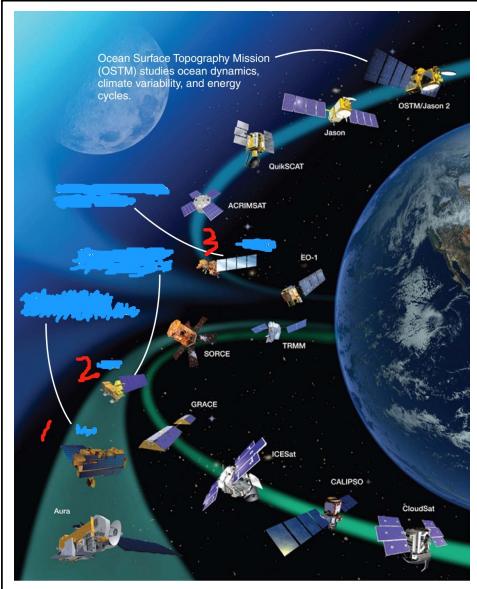


Geography: Physical Geography

Question: What are satellite 1, 2, 3 in <image 1>?

Option:

- (A) Terra, Aqua, Landsat 7
- (B) Aqua, Terra, Landsat 7
- (C) Landsat 7, Aqua, Terra
- (D) Landsat 7, Terra, Aqua



<image 1>

GPT-4V: The satellites labeled in the image as:Landsat 7, Aqua, Terra

So, the correct option is: (C) Landsat 7, Aqua, Terra

Error Category: Perceptual Error, Reasoning Error

Error Reason: GPT-4V fails to match the cartoon image with the real satellites. Additionally, Landsat 7 is in a different path while Terra and Aqua are both in the 705 path. So 3 definitely represents Landsat 7. GPT-4V ignores the hint provided in the diagram. Given the fact that GPT-4V doesn't provide a detailed explanation of its choice, it might be a perceptual error because of wrongly matching the cartoon image with the real satellites or a reasoning error due to ignorance of the path hint.

Ground Truth: (A) Terra, Aqua, Landsat 7

Figure 42. A sample error case of Geography (subfield: Physical Geography). Error category: Perceptual Error, Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

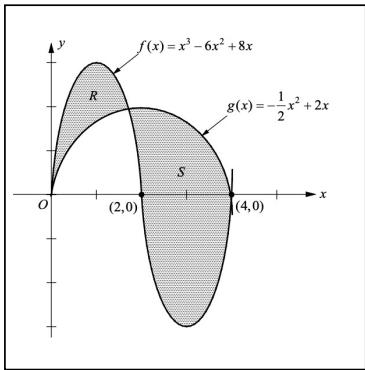


Math: Calculus

Question: <image 1>The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

Option:

- (A) $\int_0^{1.5} [f(x) - g(x)]dx$
- (B) $\int_0^{1.5} [g(x) - f(x)]dx$
- (C) $\int_0^2 [f(x) - g(x)]dx$
- (D) $\int_0^2 [g(x) - f(x)]dx$



<image 1>

GPT-4V: To determine the integral expression to find the area of R, we need to observe which curve is above the other in the region R. From the given graph, for the region R, $f(x)$ (the curve represented by $f(x) = x^3 - 6x^2 + 8x$) is above $g(x)$ (the curve represented by $g(x) = \frac{1}{2}x^2 + 2x$). The interval over which we're integrating is from $x=0$ to the point where the two curves intersect, which seems to be before $x=2$ (based on the graph). So, the upper limit of the integral is not 2. Given the options, the only possible upper limit is $x=1.5$. Thus, the integral expression to find the area of R is: $\int_0^{1.5} [f(x) - g(x)]dx$ The correct choice is: (A) $\int_0^{1.5} [f(x) - g(x)]dx$

Ground Truth: (A) $\int_0^{1.5} [f(x) - g(x)]dx$

Figure 43. A sample correct case of Math (subfield: Calculus).

[Back to List of Figures](#) | [Back to Table Index](#)

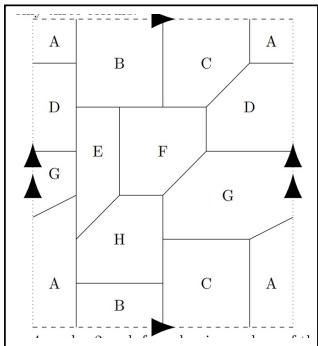


Math: Graph Theory

Question: Consider <image 1> on the torus, with its faces labeled A through H. Give a colouring of the faces of with four colours so that faces meeting along an edge have different colours. Is such colouring possible with only three colours?

Option:

- (A) Yes
- (B) No



<image 1>

GPT-4V: To solve the problem, we can start by attempting to color the diagram with three colors and see if it's possible to do so without having two adjacent faces of the same color. Let's label our three colors as Red, Green, and Blue (for the sake of this explanation). We can begin by coloring face A with Red. Adjacent to A are B and C, so let's color B with Green and C with Blue. D is adjacent to both B and C, so it must be Red. E is adjacent to A, D, and B. Since A and D are Red and B is Green, E should be Blue. F is adjacent to C, D, and E. D is Red, E is Blue, so F should be Green. G is adjacent to D, F, and A. D is Red, F is Green, so G should be Blue. Lastly, H is adjacent to B, E, and F. B is Green, E is Blue, and F is Green. So H can be colored Red. From this coloring, it seems that we can color the diagram with three colors such that no two adjacent faces have the same color. Hence, the answer is: (A) Yes.

Error Category: Perceptual Error

Error Reason: GPT-4V encounters an issue in the third step of reasoning, when it concludes, "D is adjacent to both B and C, so it must be Red." The problem begins at this point because, according to the image, D is actually adjacent to A, B, and C. The error should be identified at this step, but the model recognizes incomplete information, believing that D is only adjacent to B and C, which leads to a series of subsequent incorrect deductions.

Ground Truth: (B) No

Figure 44. A sample error case of Math (subfield: Graph Theory). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

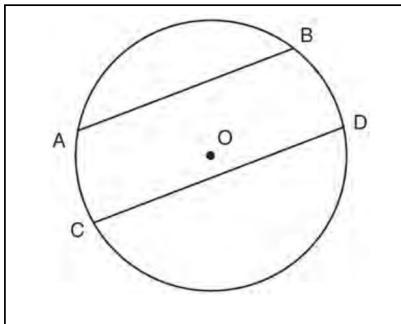


Math: Geometry

Question: In circle O shown in the diagram below, chords A and \overline{CD} are parallel <image 1>. If $m\angle AB = 104$ and $m\angle CD = 168$, what is $m\angle BD$?

Option:

- (A) 38
- (B) 44
- (C) 88
- (D) 96



<image 1>

GPT-4V: Using the properties of inscribed angles and their intercepted arcs, we can deduce: The measure of an inscribed angle is half the measure of its intercepted arc. When two chords are parallel, the inscribed angles that intercept the same arc are congruent. From the given information: $m\angle AOB = \frac{1}{2}m\angle AB = \frac{1}{2}(104) = 52^\circ$, $m\angle COD = \frac{1}{2}m\angle CD = \frac{1}{2}(168) = 84^\circ$. Since chords AB and CD are parallel, $m\angle AOB = m\angle COD$, but we've just shown that $m\angle AOB$ is 52° , which means that our diagram or understanding has an inconsistency because $m\angle COD$ is given as 84° . However, assuming the information provided is correct and using the fact that the sum of measures of arcs intercepted by two inscribed angles that sum up to 180° will be 360° , we get: $m\angle AB + m\angle BD = 360^\circ - 104^\circ = 256^\circ$. The measure of inscribed angle BOD that intercepts arc BD is: $m\angle BOD = \frac{1}{2}m\angle BD = \frac{1}{2}(256) = 128^\circ$. But, $m\angle BOD = 180^\circ - m\angle COD = 180^\circ - 84^\circ = 96^\circ$, which contradicts our previous calculation. Assuming the latter to be correct (as it's based directly on the given information), $m\angle BD = 2 * m\angle BOD = 2(96^\circ) = 192^\circ$. Thus, the correct answer should be: (D) 192. However, since 192 is not one of the options and our calculations seem to contradict the given data, it's likely there's an error or oversight in the information provided.

Error Category: Textual Understanding Error

Error Reason: GPT-4V does not correctly understand the symbols and terminology in mathematics or geometry. In this question, it incorrectly interprets $m\angle AB = 104$ as the measure of an inscribed angle, when the correct interpretation should be the measure of an arc or a central angle. Therefore, the angle represented in the question, $\angle AOB$, is 104 degrees, and $\angle COD$ is 168 degrees. Since the concept is confused right from the start, GPT4-V does not derive the correct answer in the subsequent calculations. The key to the question is to determine the angle of BOD , which is equal to the angle AOC by using the condition of parallel lines. Then, by subtracting the two given angles from 360 degrees and dividing by two, we can obtain the correct answer.

Ground Truth: (B) 44

Figure 45. A sample error case of Math (subfield: Geometry). Error category: Textual Understanding Error

[Back to List of Figures](#) | [Back to Table Index](#)



Math: Calculus

Question: A thin-walled cone-shaped cup (Fig. 16-4) is to hold $36\pi \text{ in}^3$ of water when full. What dimensions will minimize the amount of material needed for the cup?

Option:

- (A) $r = 3\sqrt{2}$, $h = 6$
- (B) $r = 3\sqrt{3}$, $h = 4$
- (C) $r = 3\sqrt{2}$, $h = 4$
- (D) $r = 2\sqrt{6}$, $h = 6$

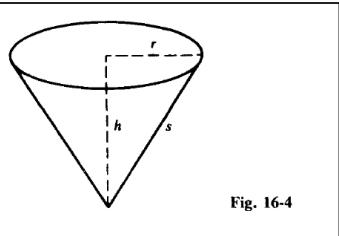


Fig. 16-4

<image 1>

GPT-4V: Given that the volume V of the cone-shaped cup is $36\pi \text{ in}^3$ when it's full, the formula for the volume of a cone is:

$A = \frac{1}{3}\pi r^2 h$. From the given volume, we can set up the equation: $\frac{1}{3}\pi r^2 h = 36\pi$. From which, $r^2 h = 108$. (1) $h = \frac{108}{r^2}$. Next, the surface area S of the cone (which does not include the base since it's a cup) is given by: $S = \pi r l$. Where l is the slant height of the cone.

Using the Pythagoras theorem for the right triangle formed by the radius, height, and slant height in the cone, we get: $l^2 = r^2 + h^2$. From which, $l = \sqrt{r^2 + h^2}$. Now, substituting the value of h from equation (1) into the equation for l , we get: $l = \sqrt{r^2 + 108^2/r^4}$. Substituting this l value in S , we get: $S = \pi r \sqrt{r^2 + 108^2/r^4}$. To minimize the surface area, we differentiate S with respect to r and set it equal to zero.

The differentiation and solving for r would be complex. Instead, we can substitute the given r and h values from options (A) to (D) into the volume equation to see which one satisfies $V = 36\pi$. The option that satisfies this and also minimizes S would be our answer.

- (A) For $r = 3\sqrt{2}$ and $h = 6$: $V = \frac{1}{3}\pi(3\sqrt{2})^2(6) = 36\pi$
- (B) For $r = 3\sqrt{3}$ and $h = 4$: $V = \frac{1}{3}\pi(3\sqrt{3})^2(4) \neq 36\pi$
- (C) For $r = 3\sqrt{2}$ and $h = 4$: $V = \frac{1}{3}\pi(3\sqrt{2})^2(4) \neq 36\pi$
- (D) For $r = 2\sqrt{3}$ and $h = 6$: $V = \frac{1}{3}\pi(2\sqrt{3})^2(6) = 36\pi$

From options (A) and (D) which satisfy the volume, the surface area for each needs to be calculated:

$$(A) S = \pi(3\sqrt{2})\sqrt{(3\sqrt{2})^2+6^2}. (D) S = \pi(2\sqrt{3})\sqrt{(2\sqrt{3})^2+6^2}.$$

Comparing the two surface areas, we can determine which one is minimal.

However, without doing the calculations, it's hard to definitively say which one is the minimal surface area. For practical purposes, the answer could be either (A) or (D) based on the volume, and one would need to compute the surface areas to confirm.

Error Category: Reasoning Error

Error Reason: The volume formula seems to be correct. The errors are made during the symbol computation.

Ground Truth: (A) $r = 3\sqrt{2}$, $h = 6$

Figure 46. A sample error case of Math (subfield: Calculus). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

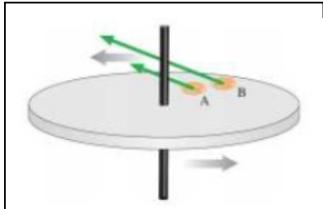


Physics: Classical Mechanics

Question: Two coins rotate on a turntable. Coin B is twice as far from the axis as coin A. <image 1>

Option:

- (A) The speed of A is twice that of B.
- (B) The speed of A equals that of B.
- (C) The speed of A is half that of B.



<image 1>

GPT-4V: For objects rotating in a circle, the linear speed (v) is related to the angular speed (ω) by: $v=\omega r$. Where: v is the linear speed of the object, ω is the angular speed of the turntable (same for all objects on the turntable), and r is the distance of the object from the axis of rotation. Given that coin B is twice as far from the axis as coin A, let's denote the distance of coin A from the axis as r and the distance of coin B from the axis as $2r$. The speed of coin A is: $v_A = \omega r$. The speed of coin B is: $v_B = \omega(2r) = 2\omega r$. Comparing the two speeds: $v_B = 2v_A$. Thus, the speed of A is half that of B. The correct answer is: (C) The speed of A is half that of B.

Ground Truth: (C) The speed of A is half that of B.

Figure 47. A sample correct case of Physics (subfield: Classical Mechanics).

[Back to List of Figures](#) | [Back to Table Index](#)

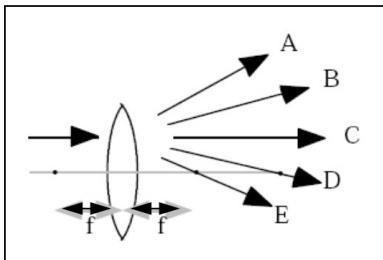


Physics: Optics

Question: A narrow beam of monochromatic light enters a lens parallel to the optic axis, as shown in the accompanying diagram. Which arrow best represents the direction of the light after leaving the lens?

Option:

- (A) arrow A
- (B) arrow B
- (C) arrow C
- (D) arrow D
- (E) arrow E



<image 1>

GPT-4V: The given diagram depicts a converging lens (a double convex lens) and a parallel beam of monochromatic light entering it. For a converging lens, parallel rays of light that are parallel to the optic axis will be refracted in such a way that they will converge or meet at the focal point on the other side of the lens. In the provided diagram, the ray is converging towards a point after passing through the lens. Among the given options:

- (A) arrow A
- (B) arrow B
- (C) arrow C
- (D) arrow D
- (E) arrow E

The direction that best represents where the light will converge after leaving the lens is (C) arrow C.

Error Category: Perceptual Error

Error Reason: GPT-4V correctly identifies the theory behind the reflection of a converging lens. However, when choosing the correct arrows in the diagram, it chooses arrow C instead of arrow E that converging to the focal point. This is due the perceptual error of GPT-4V on the direction of arrows, demonstrating the lacking ability in the sense of direction in the image.

Ground Truth: (E) arrow E

Explanation: A horizontal beam approaching a converging lens bends and converges through the focal point

Figure 48. A sample error case of Physics (subfield: Optics). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)



Basic Medical Science: Pathophysiology

Question: At which level, in the photograph, do the primary vestibular axons enter the brain stem?

<image 1>

Option:

- (A) The caudal medulla.
- (B) The pontomedullary junction.
- (C) The mid-pons.
- (D) Near the mammillary bodies in the interpeduncular fossa.



<image 1>

GPT-4V: The image you provided appears to be of the brainstem and related structures. Based on your question, you are referring to the point where the primary vestibular axons enter the brainstem. The primary vestibular axons enter the brainstem at the level of the pontomedullary junction. Therefore, the correct answer is: (B) The pontomedullary junction.

Ground Truth: (B) The pontomedullary junction.

Figure 49. A sample correct case of Basic Medical Science (subfield: Pathophysiology).

[Back to List of Figures](#) | [Back to Table Index](#)

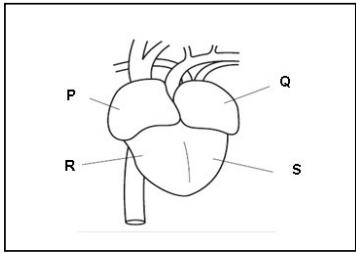


Basic Medical Science: Cardiovascular Physiology

Question: <image 1> The diagram shows the front view of a human heart. Letters P, Q, R and S indicate the different chambers of the heart. Which of the following shows the sequence in which a blood cell returning from other parts of the body passes through the four chambers of the heart?

Option:

- (A) R-S-P-Q
- (B) Q-S-R-P
- (C) P-R-Q-S
- (D) S-R-P-Q



<image 1>

GPT-4V: The sequence in which a blood cell returning from other parts of the body passes through the four chambers of the heart is : Right atrium, Right ventricle, Left atrium, Left ventricle. Looking at the given options, the correct sequence is:(D) S-R-P-Q

Error Category: Perceptual Error

Error Reason: The model has the correct knowledge, but it failed to correctly understand the image and map the letters to the right structures (e.g., P is right atrium so it should be the first in the sequence).

Ground Truth: (C) P-R-Q-S

Figure 50. A sample error case of Basic Medical Science (subfield: Cardiovascular Physiology). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)



Clinical Medicine: Pathology

Question: The patient is a post-mature newborn. What is the most likely diagnosis? <image 1>

Option:

- (A) Transient tachypnea of the newborn
- (B) Congenital lobar emphysema
- (C) Hyaline membrane disease
- (D) Meconium aspiration syndrome



<image 1>

GPT-4V: Based on the image provided, which appears to be a chest X-ray, and the information that the patient is a post-mature newborn, the most suggestive diagnosis is: (D) Meconium aspiration syndrome. Post-mature newborns are at an increased risk of passing meconium (the baby's first stool) into the amniotic fluid before birth or during labor. If the baby breathes the meconium-stained amniotic fluid into their lungs, it can cause a blockage of the airways, leading to meconium aspiration syndrome. The chest X-ray of a newborn with meconium aspiration syndrome might show patchy or streaky areas on the lungs. However, it's important to consult with a pediatric radiologist or neonatologist for a definitive diagnosis based on clinical findings and the entire clinical picture.

Ground Truth: (D) Meconium aspiration syndrome

Figure 51. A sample correct case of Clinical Medicine (subfield: Pathology).

[Back to List of Figures](#) | [Back to Table Index](#)



Clinical Medicine: Pathology

Question: A 19-year-old man goes to his physician for a routine check-up. The physical examination is normal except that his stool is positive for occult blood. A colonoscopy is performed and there are over 100 lesions of the colonic mucosa and biopsy of one shows findings seen here microscopically <image 1>. Which of the following is the most likely diagnosis?

Option:

- (A) Hereditary non-polyposis colon carcinoma
- (B) Hyperplastic polyps
- (C) Ulcerative colitis
- (D) Peutz-Jehgers syndrome
- (E) Familial polyposis coli



<image 1>

GPT-4V: The image provided appears to be a histological section of colonic mucosa. Given the clinical scenario of a 19-year-old man with over 100 lesions of the colonic mucosa and the histological findings provided, the most likely diagnosis is: (E) Familial polyposis coli. Familial adenomatous polyposis (FAP), also known as familial polyposis coli, is an autosomal dominant inherited condition characterized by the development of numerous adenomatous polyps in the colon and rectum. If left untreated, there's a nearly 100% chance of developing colon cancer by middle age.

Ground Truth: (E) Familial polyposis coli

Explanation: A. HNPCC has fewer polyps and the age of presentation is older. Mismatch repair genes are involved. B. Hyperplastic polyps are small and few. They are just focal exaggerations of the colonic mucosa. C. Ulcerative colitis leads to mucosal ulceration with residual islands of mucosa called "pseudopolyps". D. Peutz-Jehgers syndrome has the appearance of hamartomatous polyps in the bowel, and they are not so numerous. E. The small polypoid circumscribed tubular adenoma on a stalk seen here is benign, but persons with germline inheritance of the APC gene are at risk for development of colonic adenocarcinoma by their teens or 20's.

Figure 52. A sample correct case of Clinical Medicine (subfield: Pathology).

[Back to List of Figures](#) | [Back to Table Index](#)

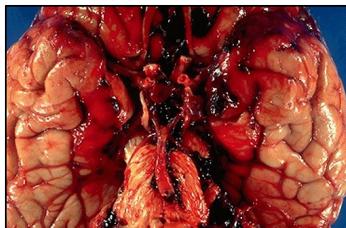


Clinical Medicine: Pathology

Question: <image 1> This 79-year-old woman was driving her car when she had the sudden onset of a severe headache. She pulled in to a service station and stopped the car. Then she slumped over the wheel. She was taken to the emergency room, where she remained comatose and died hours later. This gross appearance is found at autopsy. Which of the following is the most likely diagnosis?

Option:

- (A) Subdural hematoma
- (B) Thromboembolization with cerebral infarction
- (C) Multiple sclerosis
- (D) Ruptured berry aneurysm
- (E) Huntington disease



<image 1>

GPT-4V: The image appears to show a brain with blood prominently spread around it. Given the clinical scenario provided - a sudden severe headache followed by rapid unconsciousness and death - the most consistent diagnosis with the presentation and gross appearance at autopsy is: (D) Ruptured berry aneurysm. A ruptured berry aneurysm (also called a saccular aneurysm) can cause a subarachnoid hemorrhage, which would present with sudden severe headache and can rapidly lead to death if not treated. The hemorrhage can be seen as the spread of blood over the surface of the brain in the image. The other options do not fit as closely with the clinical scenario and gross appearance.

Ground Truth: (D) Ruptured berry aneurysm

Figure 53. A sample correct case of Clinical Medicine (subfield: Pathology).

[Back to List of Figures](#) | [Back to Table Index](#)

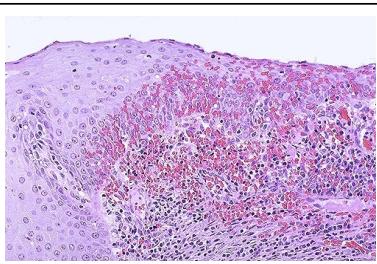


Clinical Medicine: Pathology

Question: A 41-year-old woman has noted a vaginal discharge persisting for 3 weeks. On pelvic examination her ectocervix is erythematous. Cervical biopsies are taken and on microscopic examination reveal the appearance seen here <image 1>. Which of the following is the most likely diagnosis?

Option:

- (A) Human papillomavirus infection
- (B) Ingestion of chili peppers
- (C) Diabetes mellitus
- (D) Tuberculosis
- (E) Chronic inflammation



<image 1>

GPT-4V: The image provided appears to be a microscopic view of cervical tissue with a characteristic appearance. The histopathological features seen here are suggestive of koilocytes, which are squamous epithelial cells with a perinuclear clearing (halo) and irregular, raisin-like nuclei. Koilocytes are a hallmark of human papillomavirus (HPV) infection. Based on the provided image and clinical information, the most likely diagnosis is: (A) Human papillomavirus infection.

Error Category: Perceptual Error

Error Reason: GPT-4V incorrectly interprets the details of the image. The image contains "the inflammatory cell population of mononuclear cells, mainly small round lymphocytes" and "squamous metaplasia", while GPT-4V identifies them as "koilocytes".

Ground Truth: (E) Chronic inflammation

Explanation: A. HPV produces viral cytopathic effect such as koilocytotic change and atypical changes in the squamous epithelial cells. It can drive the process to dysplasia and to carcinoma. HPV by itself does not produce significant inflammation. B. The capsaicin in chili tends to irritate nerve endings to produce the hot burning sensation, as a local effect in oral cavity. It is not systemic. Capsaicin has been used to treat dyspareunia with vulvar vestibulitis, probably by affecting nerve-ending mediated sensation, and there is no significant degree of inflammation. C. Diabetes mellitus may increase the risk for infections, but by itself does not explain the inflammation seen here. D. Tuberculosis of the genital tract is rare; mycobacterial infections tend to produce granulomatous inflammation. E. This is a chronic inflammatory process, based upon the inflammatory cell population of mononuclear cells, mainly small round lymphocytes. There is also some hemorrhage. The squamous epithelium is normal, though if this were endocervix, it would represent squamous metaplasia.

Figure 54. A sample error case of Clinical Medicine (subfield: Pathology). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)



Clinical Medicine: Heart Disease

Question: A 38-year-old man comes to the physician because of recurrent palpitations over the past 2 months. He says that he was diagnosed with a minor heart defect during childhood but does not remember the details and has not had regular medical follow-up examinations since. Physical examination shows a heart murmur. Cardiac catheterization is performed, and hemodynamic results show <image 1>. Which of the following is the most likely diagnosis?

Option:

- (A) Mitral stenosis
- (B) Mitral regurgitation
- (C) Aortic valve regurgitation
- (D) Aortic stenosis

	Oxygen saturation (%)	Pressure (mm Hg)
Right atrium	75	3
Right ventricle	75	25/4
Pulmonary trunk	75	22/10
Left atrium	99	10
Left ventricle	99	165/45
Ascending Aorta	99	155/70

<image 1>

GPT-4V: Based on the provided hemodynamic results: The oxygen saturation of 75% in the right atrium, right ventricle, and pulmonary trunk, followed by a jump to 99% in the left atrium, left ventricle, and ascending aorta, suggests the presence of a left-to-right shunt, where oxygenated blood is mixing with deoxygenated blood. The pressure readings are also indicative. The normal pressure in the left atrium is usually around 10 mm Hg, while the pressure in the left ventricle during systole (contraction) can go up to 120 mm Hg or more and during diastole (relaxation) drops significantly. The given pressure for the left ventricle is 165/45 mm Hg, which is significantly high during systole. Meanwhile, the pressure in the ascending aorta is 155/70 mm Hg, which is closer to normal systolic pressure but has an elevated diastolic pressure. These findings indicate that blood is flowing back from the left ventricle to the left atrium during ventricular systole, which is consistent with mitral regurgitation. During mitral regurgitation, the mitral valve doesn't close tightly, allowing blood to flow backward in the heart. This can cause symptoms like palpitations and can be associated with a heart murmur on physical examination. Thus, the most likely diagnosis based on the provided data is: (B) Mitral regurgitation.

Error Category: Lack of Knowledge

Error Reason: GPT4-V successfully understands the question and the given table. However, due to the lack of specific knowledge about "This backflow decreases diastolic pressure in the aorta (N = 60-90 mm Hg)" and "systolic pressure in both the ascending aorta...is increased", GPT4-V interprets "the pressure in the ascending aorta as 155/70 mm Hg, which is closer to normal systolic pressure but has an elevated diastolic pressure", and fails to identify the diagnosis of "aortic regurgitation".

Ground Truth: (C) Aortic valve regurgitation

Explanation: Aortic valve regurgitation causes a backward flow from the aorta during diastole. This backflow decreases diastolic pressure in the aorta (N = 60-90 mm Hg) and increases diastolic pressure in the left ventricle (N = 5-12 mm Hg). This regurgitant blood combines with blood from the left atrium to fill the left ventricle, which results in an increased preload and a subsequent increase in stroke volume. As a result, the systolic pressure in both the ascending aorta (N = 90-140 mm Hg) and in the left ventricle (N = 90-140 mm Hg) is increased. Oxygen saturation and right heart pressures are normal in aortic regurgitation.

Figure 55. A sample error case of Clinical Medicine (subfield: Heart Disease). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)

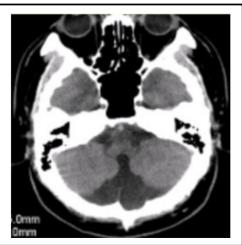


Diagnostics and Laboratory Medicine : Radiology

Question: The 40-year-old man has been walking unsteadily for several years. CT performance is as shown in <image 1>. The possible diagnosis is?

Option:

- (A) Neurofibromatosis type I
- (B) facial hemangioma syndrome
- (C) Dandy-Walker syndrome
- (D) tuberous sclerosis



<image 1>

GPT-4V: I'm not a doctor, but based on the CT image provided and the description: The image seems to show a cerebellar anomaly with the absence of the vermis and an enlarged posterior fossa cyst communicating with the fourth ventricle. These findings are suggestive of (C) Dandy-Walker syndrome. However, it's crucial to consult with a medical professional for a definitive diagnosis and medical advice.

Ground Truth: (C) Dandy-Walker syndrome

Figure 56. A sample correct case of Diagnostics and Laboratory Medicine (subfield: Radiology).

[Back to List of Figures](#) | [Back to Table Index](#)

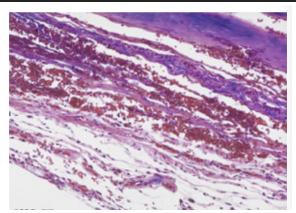


Diagnostics and Laboratory Medicine: Neuropathology

Question: What is the anatomic space occupied by the lesion? <image 1>

Option:

- (A) Intraparenchymal
- (B) Subdural space
- (C) Subarachnoid space
- (D) Epidural space



<image 1>

GPT-4V: Based on the image provided, it's challenging to definitively determine the anatomic space occupied by the lesion without more context. However, the appearance and patterns shown resemble patterns often seen in subarachnoid hemorrhages or lesions.

Therefore, the most likely answer is:

- (C) Subarachnoid space.

Error Category: Perceptual Error

Error Reason: GPT-4V fails to interpret the image correctly. The image shows "the dura can be seen in the top right of the image", but GPT-4V cannot correctly identify it.

Ground Truth: (B) Subdural space

Explanation: The dura can be seen in the top right of the image. Organization into thin subdural membranes containing thin-walled vessels often is associated with rebleeding of subdural hemorrhages as seen here.

Figure 57. A sample error case of Diagnostics and Laboratory Medicine (subfield: Neuropathology). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

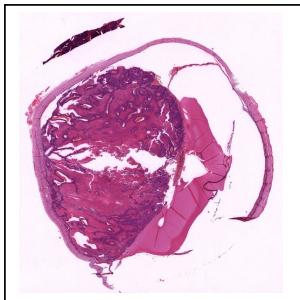


Diagnostics and Laboratory Medicine: Ophthalmic Pathology

Question: This intraocular tumor is from a 55-year-old man. Which of the following statements BEST describes the lesion. <image 1>

Option:

- (A) This tumor is a classic epithelioid choroidal melanoma with necrosis
- (B) This may represent a retinoblastoma because of the classic pattern of necrosis and small round blue cells with rare pseudorosettes.
- (C) This tumor may represent the most common type of intraocular neoplasm, and requires further work-up to distinguish it from melanoma.
- (D) The patient has a good prognosis because the neoplasm was entirely removed and the margins are free of tumor.



<image 1>

GPT-4V: Based on the image provided and without the ability to perform further clinical or pathological tests, it's challenging to provide a definitive diagnosis. However, I can offer an analysis based on the image and the choices provided:

- (A) Choroidal melanoma is a malignant tumor that arises from the pigmented cells of the choroid, which is a part of the uveal tract of the eye. Epithelioid melanoma cells are large with abundant cytoplasm and prominent nucleoli. Necrosis may be present, and if the image shows large epithelioid cells with these features, then this choice could be considered.
- (B) Retinoblastoma is a malignant tumor of the retina that predominantly affects children. It is characterized by small round blue cells, which can form pseudorosettes (rosettes that don't surround a lumen). If the image displays these features, then this could be a possibility.
- (C) Choroidal melanoma is the most common primary intraocular malignancy in adults. If the image shows features of melanoma but lacks the classic appearance of epithelioid melanoma or retinoblastoma, then further work-up would be necessary.
- (D) Making a prognosis solely based on an image can be challenging. Even if the tumor appears to be completely excised with clear margins in the image, other factors, such as tumor size, type, location, and potential metastasis, would play a role in determining the prognosis. Based on the image and the details provided, I cannot make a definitive choice. However, it's crucial to collaborate with a pathologist or an ocular oncologist for a comprehensive evaluation and diagnosis.

Error Category: Reject to Answer

Error Reason: GPT-4V refuses to answer this question without further clinical or pathological tests.

Ground Truth: (C) This tumor may represent the most common type of intraocular neoplasm, and requires further work-up to distinguish it from melanoma.

Figure 58. A sample error case of Diagnostics and Laboratory Medicine (subfield: Ophthalmic Pathology). Error category: Reject to Answer

[Back to List of Figures](#) | [Back to Table Index](#)

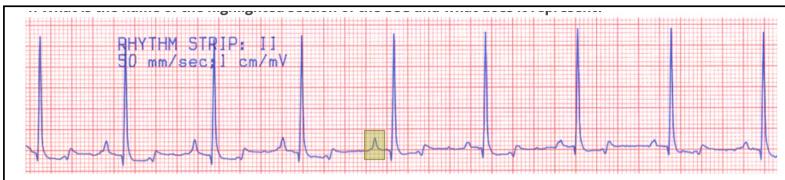


Diagnostics and Laboratory Medicine: Electrocardiography

Question: What is the name of the highlighted section of the ECG and what does it represent?

Option:

- (A) QRS complex; ventricular depolarization
- (B) P wave; atrial depolarization
- (C) T wave; ventricular repolarization
- (D) S-T segment; resting membrane potential



<image 1>

GPT-4V: The highlighted section of the ECG represents the: (A) QRS complex; ventricular depolarization.

Error Category: Perceptual Error, Lack of Knowledge

Error Reason: The error occurred because GPT-4V mistook the highlighted section for the QRS complex rather than the P wave. This confusion may have stemmed from two main factors: First, the model's attention may have been solely on the highlighted portion, whose pointed shape is characteristic of the QRS complex. Second, the model may have evaluated the entire electrocardiogram but misidentified the P wave due to its typically more rounded shape. A proper assessment of the whole electrocardiogram reveals that the QRS complex is actually represented by the taller spikes that come after the highlighted section, indicating that the highlighted area is, indeed, the P wave.

Ground Truth: (B) P wave; atrial depolarization

Figure 59. A sample error case of Diagnostics and Laboratory Medicine (subfield: Electrocardiography). Error category: Perceptual Error, Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)

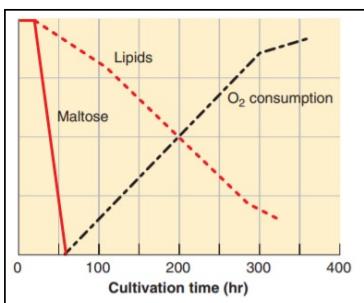


Pharmacy : Pharmaceutical Microbiology

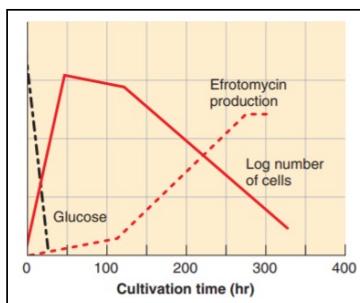
Question: The antibiotic efrotomycin is produced by *Streptomyces lactamdurans*. *S. lactamdurans* was grown in 40,000 liters of medium. The medium consisted of glucose, maltose, soybean oil, $(\text{NH}_4)_2\text{SO}_4$, NaCl, KH_2PO_4 , and Na_2HPO_4 . The culture was aerated and maintained at 28°C. The following results were obtained from analyses of the culture medium during cell growth: <image 1>, <image 2> Is efrotomycin a primary or secondary metabolite?

Option:

- (A) Primary
- (B) Secondary
- (C) None of them



<image 1>



<image 2>

GPT-4V: To determine whether efrotomycin is a primary or secondary metabolite, we need to look at the production pattern of the antibiotic in relation to the growth of the organism, *S. lactamdurans*. A primary metabolite is typically produced during the active growth phase of the organism, while a secondary metabolite is produced after the active growth phase, often during the stationary phase or when nutrients in the medium become limited. From the provided graphs: In the first image, we see that lipids (potentially representing cell biomass) increase and then plateau. The oxygen consumption increases and starts to decline, which might indicate the end of the active growth phase.

In the second image, glucose (a primary nutrient) is consumed rapidly during the initial phase of growth. The "Log number of cells" (representing cell growth) increases rapidly and then plateaus, indicating the end of the active growth phase. Efrotomycin production starts increasing after the log phase of cell growth has plateaued.

Given these observations, efrotomycin starts to be produced after the active growth phase of the organism, which is characteristic of secondary metabolites. So, the answer is: (B) secondary.

Ground Truth: (B) Secondary

Figure 60. A sample correct case of Pharmacy (subfield: Pharmaceutical Microbiology).

[Back to List of Figures](#) | [Back to Table Index](#)

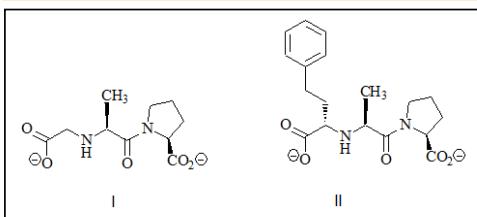


Pharmacy: Pharmacology and Drug Synthesis

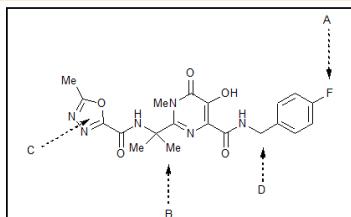
Question: The following synthetic scheme <image 1> has been used to synthesize raltegravir. If you wanted to create a number of analogues, which substituent <image 2> could be varied most efficiently using this synthesis?

Option:

- (A) The type of substituent A on the aromatic ring
- (B) The type of alkyl substituent B present
- (C) The type of heterocyclic ring C present
- (D) The length of the chain D between the amide and the aromatic ring



<image 1>



<image 2>

GPT-4V: Given the synthetic scheme provided, let's analyze each option:

- (A) The type of substituent A on the aromatic ring: In the second image, substituent A is located on the aromatic ring. Since this is present at the end of the molecule, it might be relatively easy to vary it, as it doesn't interfere with the rest of the synthesis.
- (B) The type of alkyl substituent B present: Substituent B is an alkyl group attached to the heterocyclic ring. Altering this substituent would require modifying the building block that is used to make this part of the molecule, which might be a bit more challenging.
- (C) The type of heterocyclic ring C present: The heterocyclic ring C appears to be a critical part of the structure and seems to be central to the synthesis. Modifying this ring would likely require significant changes to the entire synthesis.
- (D) The length of the chain D between the amide and the aromatic ring: Changing the length of chain D would require altering the length of the carbon chain in the precursor molecule. This change could be achieved by selecting different precursors with varied chain lengths.

Thus, (A) **The type of substituent A on the aromatic ring** seems to be the most efficiently varied using this synthesis. The aromatic ring can be easily subjected to a variety of substitution reactions to achieve analogues with different substituents without disturbing the rest of the molecule.

Error Category: Lack of Knowledge

Error Reason: GPT-4V successfully understands the images and text. However, due to the lack of specific knowledge about "The synthesis is most efficient for creating analogues with a variety of heterocyclic rings", GPT-4V assumes "substituent A is located on the aromatic ring, which is at the end of the molecule, and it might be relatively easy to vary it, as it doesn't interfere with the rest of the synthesis" and subsequently makes an incorrect prediction.

Ground Truth: (C) The type of heterocyclic ring C present

Explanation: It is possible to use this synthesis to create analogues where all the positions indicated are modified as described. However, the synthesis is most efficient for creating analogues with a variety of heterocyclic rings. That is because the reaction sequence could be carried out on large scale to produce a supply of the penultimate structure. Analogues could then be created in one step by reacting this structure with different heterocyclic acid chlorides. If the synthetic scheme was used to create analogues at the other positions indicated, then more reactions will be needed. For example, varying the type of alkyl substituent present at position B would involve carrying out the complete synthesis for each analogue.

Figure 61. A sample error case of Pharmacy (subfield: Pharmacology and Drug Synthesis). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)

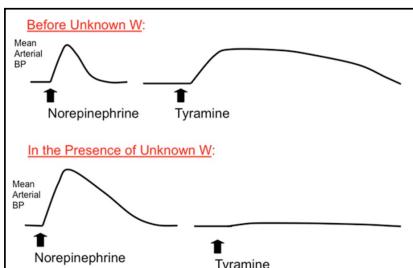


Pharmacy: Pharmacology and Drug Synthesis

Question: Each trace below <image 1> shows the change in mean arterial blood pressure from its baseline value. Which of the following drugs best fits the effects of unknown W?

Option:

- (A) cocaine
- (B) phenelzine
- (C) phentolamine
- (D) propranolol



<image 1>

GPT-4V: To answer this question, we need to understand how each drug affects mean arterial blood pressure in the context of Norepinephrine and Tyramine. Here's a breakdown:

- Cocaine: Cocaine blocks the reuptake of norepinephrine, which would cause a prolonged effect of norepinephrine, leading to a sustained increase in blood pressure. However, it wouldn't block the effects of Tyramine.
- Phenelzine: Phenelzine is a monoamine oxidase inhibitor (MAOI). MAOIs block the degradation of norepinephrine and also of indirectly acting sympathomimetics like Tyramine. When MAO is inhibited, administration of Tyramine can cause a hypertensive crisis because more of it is converted to norepinephrine.
- Phentolamine: Phentolamine is a non-selective alpha-adrenergic antagonist. It would block the effects of both norepinephrine and Tyramine, leading to a decrease or no change in blood pressure.
- Propranolol: Propranolol is a non-selective beta-adrenergic antagonist. It would block the effects of norepinephrine on beta receptors but wouldn't have an effect on Tyramine's mechanism, which primarily acts through alpha receptors.

Given the traces in the image: Before the administration of Unknown W, both Norepinephrine and Tyramine increase blood pressure. In the presence of Unknown W, Norepinephrine still increases blood pressure, but Tyramine no longer has this effect. The drug that fits this pattern is Phenelzine. In the presence of an MAOI, Norepinephrine would still exert its direct effect, but Tyramine, which typically induces the release of stored catecholamines, would result in an amplified effect due to the blocked degradation of norepinephrine. The lack of a hypertensive response to Tyramine in the second trace suggests that its conversion to norepinephrine might have been blocked, consistent with the action of an MAOI. Thus, the correct answer is: (B) phenelzine.

Error Category: Lack of Knowledge

Error Reason: GPT-4V lacks knowledge about the interaction between Cocaine and Tyramine. While it correctly recalls information about Phenelzine, it appears the lack of other plausible options leads to a reasoning error (which could be deemed as hallucination), as it strives to select the most plausible option from the ones given, and then justifies that selection. Therefore, the root cause remains a lack of knowledge.

Ground Truth: (A) cocaine

Figure 62. A sample error case of Pharmacy (subfield: Pharmacology and Drug Synthesis). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)



Public Health: Epidemiology

Question: An outbreak of gastrointestinal disease occurred 24-36 hours after people had attended a wedding. Of the 203 attendees (including the bride and groom), 200 completed questionnaires, and 50 reported illness compatible with the case definition. Tabulated food consumption histories are presented in the table below. The most appropriate measure of association for these data is the:

Option:

- (A) Attributable risk percent
- (B) Chi-square
- (C) Odds ratio
- (D) Risk ratio

Food Item	Ate Specified Food			Did Not Eat Specified Food		
	III	Well	Total	III	Well	Total
Punch	46 (25%)	138	184	4 (25%)	12	16
Wedding Cake	45 (45%)	55	100	5 (5%)	95	100
Sushi	10 (91%)	1	11	40 (21%)	149	189

<image 1>

GPT-4V: Based on the information provided, you're trying to determine the relationship between the consumption of a particular food item and the risk of developing gastrointestinal disease among wedding attendees. In this type of scenario, where you are comparing the risk of disease between those exposed (ate the food) and those not exposed (did not eat the food), the appropriate measure of association is the:
(D) Risk ratio (also known as relative risk). The risk ratio compares the risk of disease in the exposed group to the risk of disease in the unexposed group.

Ground Truth: (B) Chi-square

Figure 63. A sample correct case of Public Health (subfield: Epidemiology).

[Back to List of Figures](#) | [Back to Table Index](#)



Public Health: Epidemiology

Question: In February 2014, the New England Journal of Medicine published in the same issue two epidemiological trials (field trials) of inactivated EV71 vaccine conducted by scholars in China respectively. One of the studies was from Jiangsu Province, China, and the study design was a randomised, double-blind, placebo-controlled, multicentre trial (phase III trial), which included a total of 10,077 healthy children aged 6-35 months. The primary outcome indicator was the incidence of EV71-related hand, foot and mouth disease or herpangina, according to the 0- and 28-day immunisation schedule, using block (one group of 10 persons) randomisation, with a 1:1 allocation of study participants to EV71 vaccine or placebo, and follow-up monitoring for 12 months, and the results were shown in the table below. <image 1>The incidence density(/1000 person years) of hand, foot and mouth disease in placebo group was

Option:

- (A) 18.5
- (B) 4.9
- (C) 19.3
- (D) 21.8

Outcomes	Vaccination group (4973.2 person years)	Placebo group(4873.0 person years)
	New cases	New cases
EV71-related hand, foot and mouth disease or herpangina	5	94
Hand, foot and mouth disease	5	90
herpangina	0	4
EV71-related hospitalisation events	0	24
EV71-related Hand, Foot and Mouth Disease with Combined Neurological Complications	0	8
All EV71 related diseases	13	106

<image 1>

GPT-4V: To find the incidence density (per 1000 person years) of hand, foot, and mouth disease in the placebo group, we need to calculate the rate of new cases of the disease per 1000 person years of observation. From the table:

Number of new cases of EV71-related hand, foot, and mouth disease or herpangina in placebo group = 94
Total person years observed for placebo group = 4873.0 person years. Incidence density = (Number of new cases / Total person years observed) x 1000
Incidence density = (94 / 4873 * 1000) = 19.3.
Therefore, the incidence density (/1000 person years) of hand, foot and mouth disease in placebo group is:(C) 19.3

Error Category: Textual Understanding Error

Error Reason: The question asked for the incidence density (/1000 person-years) of hand, foot, and mouth disease in the placebo group, which should have been correctly calculated as $(90/4873 * 1000 = 18.5)$. However, GPT-4 erroneously extracted the data for EV71-related hand, foot, and mouth disease or herpangina $(94/4873 * 1000=19.3)$, leading to an incorrect final answer.

Ground Truth: (A) 18.5

Figure 64. A sample error case of Public Health (subfield: Epidemiology). Error category: Textual Understanding Error

[Back to List of Figures](#) | [Back to Table Index](#)



Public Health: Epidemiology

Question: A maternal and child health facility wanted to evaluate the effectiveness of two methods of screening for cervical intraepithelial neoplasia (CIN I-III), electronic colposcopy and cervical liquid-based cytology (TCT). 600 gynaecological outpatient cases were examined by both methods, and all cases were confirmed with cervical histopathology (the gold standard) at the same time. The results of all examinations were summarised in the table. <image 1>The false positive rate of the serial test was

Option:

- (A) 0.95
- (B) 0.51
- (C) 0.75
- (D) 0.05

Test results		Patients with cervical lesions	Patients with non-cervical lesions
electronic colposcopy	TCT		
+	-	15	10
-	+	5	220
+	+	70	25
-	-	10	245
Total		100	500

<image 1>

GPT-4V: To determine the false positive rate of the serial test (both electronic colposcopy and TCT positive) for detecting cervical intraepithelial neoplasia (CIN I-III), we can use the following formula: False Positive Rate (FPR) = False Positives / (False Positives + True Negatives) False Positives (FP): These are the cases where both tests are positive (+, +) but the patient does not actually have cervical lesions. From the table:

Patients with non-cervical lesions and both tests positive = 25So, FP = 25True Negatives (TN): These are the cases where both tests are positive (+, +) and the patient actually has cervical lesions, combined with the cases where both tests are negative (-, -) but the patient does not have cervical lesions.

From the table:

Patients with cervical lesions and both tests positive = 70

Patients with non-cervical lesions and both tests negative = 245So, TN = 70 + 245 = 315Now, plugging into the formula:

$$FPR = FP / (FP + TN)$$

$$FPR = 25 / (25 + 315)$$

FPR = 0.0735 or 7.35%. The value closest to 7.35% is 0.075. So, the correct answer is:\n(C) 0.75

Error Category: Lack of Knowledge

Error Reason: The correct identification of False Positives (FP) was indeed 25, but there was a mistake in calculating True Negatives (TN). TNs are cases that are actually negative and also test negative. GPT-4 incorrectly used True Positives (TP) in calculating TN. Moreover, the question pertained to a series test, meaning a diagnosis of positive was made only if both electronic colposcopy and TCT (ThinPrep Cytologic Test) were positive. However, GPT-4 considered a positive diagnosis if either one of these tests was positive, which led to an error in the count of TN. The correct calculation should be $25/(10+220+25+245) = 0.05$.

Ground Truth: (D) 0.05

Figure 65. A sample error case of Public Health (subfield: Epidemiology). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)



History: European History

Question: This cartoon is best understood in the context of which of the following events? <image 1>

Option:

- (A) Treaty of Brest-Litovsk
- (B) Tripartite Pact
- (C) Anti-Comintern Pact
- (D) The Munich Conference



"Telling the Italians!" American political cartoon published in 1938.

<image 1>

GPT-4V: The cartoon is best understood in the context of:

- (D) The Munich Conference.

Ground Truth: (D) The Munich Conference

Explanation: Germany invaded Austria in 1938 before setting its sights on the Sudetenland of Czechoslovakia. At the Munich Conference of 1938, which included Hitler, Mussolini, and Prime Minister Neville Chamberlain of England, Hitler was given the Sudetenland, without the consent of Czechoslovakia, in return for the promise to cease his expansionist activities. The cartoon shows Hitler placing Austria in his bag, which would set this cartoon right at the time of the Munich Conference. Choice (D) is correct. The Treaty of Brest-Litovsk was a World War I era treaty that set the stage for Russia's exit from the war. Eliminate (A). The Tripartite Pact signed by Germany, Italy, and Japan did not occur until two years after this cartoon was created. Eliminate (B). The Anti-Comintern Pact signed between Germany, Italy, and Japan focused on a mutual defense against Russia and had nothing to do with Austria. Eliminate (C).

Figure 66. A sample correct case of History (subfield: European History).

[Back to List of Figures](#) | [Back to Table Index](#)



History: World History

Question: A historian researching the economic history of Eurasia in the period circa 600-1450 c.e. would most likely find the two tables (<image 1> and <image 2>) useful as a source of information about which of the following?

Option:

- (A) The diffusion of cultural traditions along Eurasian trade routes
- (B) The spread of technological innovations across regions in Eurasia
- (C) The geographic extent of the monetization of Eurasian economies
- (D) The extent to which government economic policies in Eurasia in the period 600-1450 represented a continuity of earlier policies

Table 1

ORIGIN OF THE COINS IN A BURIED CACHE FROM CIRCA 750 C.E., FOUND NEAR XI'AN, CENTRAL CHINA

Origin of the Coins	Date of the Coins (approximate)	Number of Coins
Chinese, pre-dating the Tang dynasty	500 B.C.E.-550 C.E.	19
Early Tang dynasty	600-750 C.E.	451
Sassanian dynasty, Persia	600 C.E.	1
Byzantine Empire	600 C.E.	1
City of Turfan, Central Asia	650 C.E.	1
Japan, Nara period	710 C.E.	5
TOTAL		478

Table 2

ORIGINS OF THE COINS IN A VIKING BURIED CACHE FROM CIRCA 900 C.E., FOUND IN NORTHWESTERN ENGLAND

Origin of the Coins	Number of Coins (approximate)
Viking kingdoms in northern England	5,000
Anglo-Saxon kingdoms in southern England	1,000
Carolingian Frankish Empire	1,000
Viking states in Scandinavia	50
Abbasid Caliphate	50
Papacy and Northern Italian states	20
Byzantine Empire	1
TOTAL	7,200

<image 1>

<image 2>

GPT-4V: The tables provide information on the origin of coins found in buried caches in different regions and their approximate dates. This kind of information indicates where coins from different regions and time periods ended up, which suggests patterns of trade or movement of goods, services, and people. Given the options: (A) The diffusion of cultural traditions along Eurasian trade routes: While the movement of coins can suggest trade routes and interactions between different regions, it does not directly provide information about the diffusion of cultural traditions. (B) The spread of technological innovations across regions in Eurasia: The tables do not specifically address technological innovations. (C) The geographic extent of the monetization of Eurasian economies: The presence of coins from various regions in different locations suggests areas where money was being used in trade and therefore indicates the extent of monetization. (D) The extent to which government economic policies in Eurasia in the period 600-1450 represented a continuity of earlier policies: The tables show where coins from different periods were found, but they do not specifically provide information on government policies or whether those policies represented continuity. Given the information in the tables, the best answer would be:

(C) The geographic extent of the monetization of Eurasian economies.

Ground Truth: (C) The geographic extent of the monetization of Eurasian economies.

Figure 67. A sample correct case of History (subfield: World History).

[Back to List of Figures](#) | [Back to Table Index](#)



History: Modern History

Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <image 1>

Option:

- (A) Oppressor
- (B) Imperialist
- (C) Savior
- (D) Isolationist



<image 1>

GPT-4V: Based on the political cartoon, the United States is seen as fulfilling the role of:
(B) Imperialist.

Error Category: Perceptual Error

Error Reason: GPT-4V wrongly interprets the image, because it doesn't correctly interpret the contradiction between the comics on the top half and the bottom half. The perspective of the cartoonist is that the United States has been a savior to the nations brought under its control, saving them from poverty or oppression and bringing them trade, prosperity, and democratic governments. This mistake also indicates that GPT-4V holds some wrong political beliefs.

Ground Truth: (C) Savior

Explanation: The perspective of the cartoonist is that the United States has been a savior to the nations brought under its control, saving them from poverty or oppression and bringing them trade, prosperity, and democratic governments. Although one might be tempted to cast the United States in the role of imperialist (B), the purpose of the cartoon is to highlight the positive transformation of the people due to their "rescue" by the United States, rather than the benefits to the United States. Because the cartoon claims the people are better off for having been "rescued" by the United States, the United States is not seen as an oppressor (A). Since isolationists do not support foreign intervention, (D) cannot be the correct answer.

Figure 68. A sample error case of History (subfield: Modern History). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)



History: World History

Question: British economic interactions with India during the time span covered in the chart above would have been chiefly concerned with which of the following commodities? <image 1>

Option:

- (A) Cotton
- (B) Spices
- (C) Ivory
- (D) Opium

WORLD GDP, 1600–1870 (in millions of dollars, calculated to 1990 value)						
	1600	Share of World Total	1700	Share of World Total	1870	Share of World Total
Great Britain	6,007	1.8%	10,709	2.88%	100,179	9.1%
Western Europe (excluding Britain)	65,955	20%	83,395	22.5%	370,223	33.61%
China	96,000	29%	82,800	22.3%	189,740	17.23%
India	74,250	23%	90,750	24.4%	134,882	12.25%
WORLD TOTAL	329,417		371,369		1,101,369	

Source: Angus Maddison, *Contours of the World Economy, 2007*

<image 1>

GPT-4V: The correct answer would be: (D) Opium. During the time span covered in the chart (1600–1870), the British economic interactions with India were heavily influenced by the opium trade. The British East India Company established a monopoly on the opium trade and exported opium to China, leading to the Opium Wars with China. This was a significant aspect of Britain's economic engagement with India during this period.

Error Category: Lack of Knowledge

Error Reason: This question calls for knowledge related to the stimulus material. Understanding the importance of Indian cotton as a trade commodity is key to contextualizing this question. However, GPT-4V doesn't know that.

Ground Truth: (A) Cotton

Explanation: Likewise, this question calls for knowledge related to the stimulus material. Understanding the importance of Indian cotton as a trade commodity is key to contextualizing this question.

Figure 69. A sample error case of History (subfield: World History). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)



Literature: Contemporary Literature

Question: Which of the following options best relates to the title 'Ancestor' <image 1>

Option:

- (A) The danger of letting go of a dream.
- (B) The father is compared to a gypsy.
- (C) A mother tells her son to stop whining.
- (D) These women want their children to be educated.



<image 1>

GPT-4V: The image depicts what seems to be an old illustration with the title "A Gipsy Family" and is described as a facsimile of a woodcut. Given the choices provided: (A) The danger of letting go of a dream.(B) The father is compared to a gypsy. (C) A mother tells her son to stop whining. (D) These women want their children to be educated. Option (B) "The father is compared to a gypsy." appears to best relate to the title and content of the image.

Ground Truth: (B) The father is compared to a gypsy.

Figure 70. A sample correct case of Literature (subfield: Contemporary Literature).

[Back to List of Figures](#) | [Back to Table Index](#)



Literature: Comparative Literature

Question: Refer to the image, which term best matches the given set of characteristics?

Option:

- (A) Common elements in horror stories
- (B) Literary elements in horror
- (C) Scary plot twists
- (D) Intensity



<image 1>

GPT-4V: Based on the description <image 1>, the term that best matches the given set of characteristics is:

- (D) Intensity.

Error Category: Perceptual Error

Error Reason: GPT-4V did not capture the expressions of the characters in the painting or the chopping action of the person on the far left, focusing instead only on the muscles and movements of the character on the far right. As a result, GPT-4V failed to recognize the elements were terrifying.

Ground Truth: (A) Common elements in horror stories

Figure 71. A sample error case of Literature (subfield: Comparative Literature). Error category: Perceptual Error

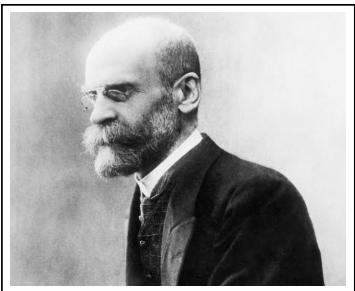


Sociology: Sociology Theory

Question: For the person in <image 1>, what is the root of everything social?

Option:

- (A) Collective Effervescence
- (B) Society
- (C) The Economy
- (D) Religion



<image 1>

GPT-4V: For Émile Durkheim, one of the founding figures in sociology, the root of everything social could be argued to be "Collective Effervescence," a term he coined. This concept refers to the moments of communal passion and ecstasy that occur during group rituals and events, which Durkheim believed were the basis for the solidarity of social groups and ultimately for society itself. It's a foundational element in his theory on the formation of religions and societies.

The correct answer to the question, considering Emile Durkheim's sociological theories, is "Religion." Durkheim viewed religion as a fundamental social institution that is central to the formation and sustenance of societies. He believed that religious beliefs and rituals play a critical role in creating social cohesion and collective conscience.

Ground Truth: (D) Religion

Explanation: Durkheim believed that religion is the root of everything social. He argued that religion provides the framework and moral values that bind individuals together and create social cohesion. Through religious rituals and beliefs, people experience a sense of collective effervescence, a heightened state of group solidarity and shared emotions. This collective effervescence reinforces social norms, strengthens social bonds, and gives individuals a sense of belonging to a larger community. Therefore, Durkheim viewed religion as the foundation of social order and the driving force behind societal cohesion.

Figure 72. A sample correct case of Sociology (subfield: Sociology Theory).

[Back to List of Figures](#) | [Back to Table Index](#)

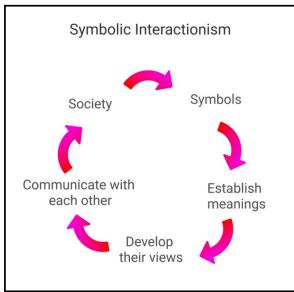


Sociology: Comparative Literature

Question: People holds view in <image 1> argue that crime is linked primarily to ____? ____.

Option:

- (A) Power
- (B) Master status
- (C) Family values
- (D) Wealth



<image 1>

GPT-4V: The diagram you provided outlines the concept of Symbolic Interactionism, which is a sociological perspective that focuses on the ways in which individuals interact and communicate through symbols. Given the choices provided and the context of Symbolic Interactionism, the most appropriate answer is: (B) Master status. Master status refers to a social position that is the primary identifying characteristic of an individual. It dominates other statuses and thereby determines a person's general position in society. In the context of crime, some symbolic interactionists might argue that an individual's master status (e.g., being labeled as a "criminal") could influence their interactions and how they are perceived by others in society.

Error Category: Reasoning Error

Error Reason: The answer Master status fails to find the root cause of conducting crimes. Essentially, an individual who grows up in a poor neighbourhood with high rates of drug use, violence, teenage delinquency, and deprived parenting is more likely to become a criminal than an individual from a wealthy neighbourhood with a good school system and families who are involved positively in the community. Therefore, the primary/core reason is wealth. The answer is D.

Ground Truth: (D) Wealth

Figure 73. A sample error case of Sociology (subfield: Comparative Literature). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

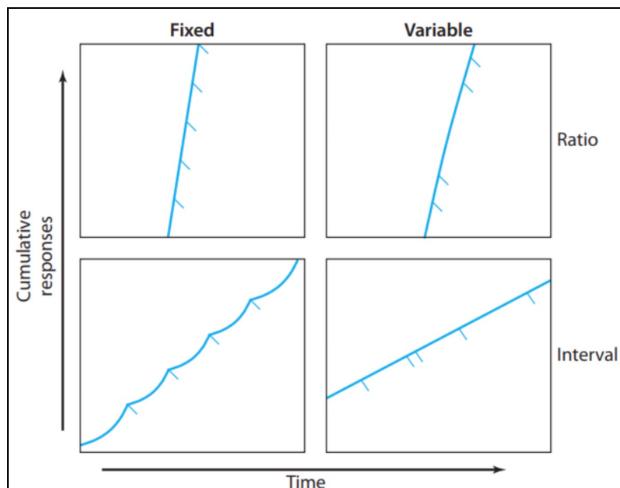


Psychology: Biological Psychology

Question: Described by the psychologist B. F. Skinner, operant conditioning posits that learning can be understood by the interaction between reinforcers and stimuli. Studies from both comparative and human psychology have supported Skinner's initial findings and expanded upon them. The following figure shows a hypothetical example of various operant conditioning trials. During these trials, when a rat pressed a bar (as indicated by the hash mark on the graph), the behavior was reinforced by the presentation of a food pellet. The reinforcement schedule differs in each cell of the figure. Time is displayed on the x-axis and cumulative responses (bar presses) is displayed on the y-axis. ([image 1]) Imagine that you are watching a rat complete one of the tasks in the figure above. The rat is pressing the bar at a very high rate and takes a short break from pressing the bar each time the food pellet is delivered. What type of reinforcement schedule are you MOST likely watching?

Option:

- (A) variable-interval
- (B) variable-ratio
- (C) fixed-interval
- (D) fixed-ratio



<image 1>

GPT-4V: The graph corresponds to operant conditioning schedules. The pattern of high-rate bar pressing and short breaks after reinforcement suggests a Fixed Ratio schedule, shown in the top left quadrant. Here, reinforcement is given after a set number of responses, which is reflected in the steep inclines followed by drops on the graph. The rat's behavior – rapid pressing until receiving a pellet, then pausing – matches this schedule, indicating it understands that a specific number of presses leads to a reward, thus causing the observed high response rate followed by a break.

Ground Truth: (D) fixed-ratio

Figure 74. A sample correct case of Psychology (subfield: Biological Psychology).

[Back to List of Figures](#) | [Back to Table Index](#)



Psychology: Clinical Psychology

Question: Guided meditation and deep-breathing exercises have long been used as effective techniques for stress reduction. The mechanism of action for this non-pharmacologic intervention is not entirely known, but scientists believe that the act of focusing one's thoughts and deep belly-breathing both serve to somehow inhibit the stress response activated by the hypothalamic-pituitary-adrenal axis. Irritable Bowel Syndrome (IBS) is a disorder that causes a range of abdominal discomfort and bowel irregularities, but unlike bowel diseases with similar symptoms, there are no physical abnormalities; rather, the disorder appears to be the physical manifestation of psychological triggers. The current standard non-pharmacologic treatment for IBS is cognitive behavior therapy (CBT). CBT treats IBS sufferers by treating the emotional and psychological triggers that cause physical symptoms. In an attempt to determine if meditation is as beneficial as CBT for treating IBS, a recent six-month study was conducted on female IBS sufferers. Eligible participants had active IBS symptoms for at least three months during the past year. Participants with and without a diagnosed anxiety disorder were recruited to participate in this study. Subjects were randomly assigned to one of three groups: a CBT group, a guided-meditation group, and a no-treatment group. Approximately 65% of the participants had an anxiety disorder, and these subjects were roughly equally represented in each of the three groups. The results of this study, measured by percent reduction of IBS symptoms after treatment, are summarized in Figure 1. <image 1> Based on the results of this study, what can be most reasonably concluded about the efficacy of CBT for IBS sufferers who do not have an anxiety disorder?

Option:

- (A) CBT is more effective than no treatment and more effective than meditation.
- (B) CBT and meditation combined provide the most effective treatment possible.
- (C) CBT is not as effective as meditation.
- (D) CBT is equally effective for IBS sufferers with and without anxiety disorders.

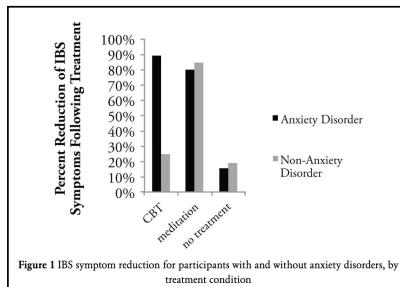


Figure 1 IBS symptom reduction for participants with and without anxiety disorders, by treatment condition

<image 1>

GPT-4V: Based on Figure 1: For IBS sufferers without an anxiety disorder: CBT shows significant symptom reduction. Meditation shows some symptom reduction, but it's less than CBT. No treatment shows very minimal symptom reduction. Thus, for IBS sufferers who do not have an anxiety disorder, CBT is more effective than both meditation and no treatment. The answer is: **(A) CBT is more effective than no treatment and more effective than meditation.**

Error Category: Perceptual Error

Error Reason: Based on the description from GPT-4V: "For IBS sufferers without an anxiety disorder: CBT shows significant symptom reduction. Meditation shows some symptom reduction, but it's less than CBT. No treatment shows very minimal symptom reduction." It appears that GPT-4v misinterpreted the black bars in the plot as the "Non-Anxiety Disorder" group. However, the actual "Non-Anxiety Disorder" group is represented by the gray bars as shown in the legend.

Ground Truth: (C) CBT is not as effective as meditation.

Figure 75. A sample error case of Psychology (subfield: Clinical Psychology). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

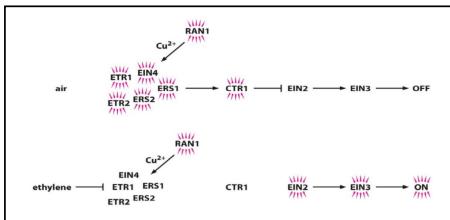


Agriculture : Plant Biodiversity and Biotechnology

Question: An ein4 mutant never binds ethylene. In the presence of ethylene what phenotype will be observed for the seedling triple response in this ein4 mutant?<image 1>

Option:

- (A) Seedlings will become de-etiolated.
- (B) Seedlings will grow towards the light.
- (C) The ETR1 receptor will activate CTR1 and the seedling triple response.
- (D) CTR1 will be activated and inhibit the seedling triple response.
- (E) RAN1 will provide Cu²⁺ to the ethylene receptors.



<image 1>

GPT-4V: The image shows a schematic representation of the ethylene signaling pathway in plants. In a typical situation, when ethylene binds to its receptor (ETR1, ERS1, ERS2), it prevents the receptor from activating CTR1, which in turn allows the downstream signaling pathway to proceed, leading to the activation of EIN2, EIN3, and the ethylene response (seedling triple response). In the case of the ein4 mutant, which cannot bind ethylene, the pathway from the ethylene receptor to the downstream components would not be initiated. Since EIN4 is presumably an ethylene receptor like ETR1 and the mutant form of EIN4 cannot bind ethylene, this means that the receptor cannot turn off the CTR1, leading to the inhibition of the seedling triple response due to the constitutive activation of CTR1. Therefore, the correct answer is: (D) CTR1 will be activated and inhibit the seedling triple response. This is because in the ein4 mutant, ethylene cannot bind to EIN4 to inhibit CTR1, thus CTR1 remains active and the signaling pathway is off, preventing the seedling triple response.

Ground Truth: (D) CTR1 will be activated and inhibit the seedling triple response.

Figure 76. A sample correct case of Agriculture (subfield: Plant Biodiversity and Biotechnology).

[Back to List of Figures](#) | [Back to Table Index](#)



Agriculture: Animal Anatomy

Question: <image 1> According to where the arrow points, choosing the matching term:

Option:

- (A) Long pastern bone
- (B) Coffin bone
- (C) Short pastern
- (D) Navicular bone, one bone per digit, located at the interphalangeal joint



<image 1>

GPT-4V: The arrow in the image is pointing to the lower part of the horse's leg, specifically to the bone known as the "short pastern" or the "second phalanx". This is the bone located between the long pastern bone (first phalanx) and the coffin bone (third phalanx) in the anatomy of a horse's leg. Therefore, the correct matching term for where the arrow points is: (C): "short pastern"

Error Category: Perceptual Error

Error Reason: GPT 4V give the incorrect interpretation of the visual content in the image. Despite having the capability to "see" the image, the model incorrectly identified the part of the bone indicated by the arrow as the short pastern bone when it is actually the long pastern bone. This perceptual error could lead to subsequent reasoning errors because an incorrect initial interpretation affects the entire reasoning process that leads to the conclusion. Since the model initially misidentified the bone, all subsequent explanations and reasoning were built on this incorrect basis.

Ground Truth: (A) Long pastern bone

Figure 77. A sample error case of Agriculture (subfield: Animal Anatomy). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

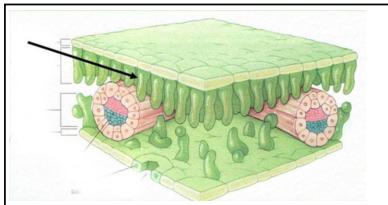


Agriculture: Plant Physiology

Question: <image 1> Found near the top of the leaf; cells in this layer contain many chloroplasts. Choosing the matching term:

Option:

- (A) Stomata
- (B) Spongy mesophyll
- (C) Guard cells
- (D) Waxy cuticle



<image 1>

GPT-4V: The layer indicated by the arrow in the image is the uppermost layer of the leaf and is typically known for its protective qualities and prevention of water loss. The correct term for this layer is:(D) waxy cuticle.

Error Category: Perceptual Error

Error Reason: GPT-4V recognized the importance of the uppermost layer's role in protection and water loss prevention but failed to correctly identify the visual signals indicating that the arrow was pointing to the guard cells, which are not the waxy cuticle. The reasoning process was flawed because the model did not correctly link the visual information with the correct biological structure.

Ground Truth: (C) Guard cells

Figure 78. A sample error case of Agriculture (subfield: Plant Physiology). Error category: Perceptual Error.

[Back to List of Figures](#) | [Back to Table Index](#)



Agriculture: Plant Physiology

Question: In soy, this stem/root diseases manifests as a brown pit. favored by drought conditions.
Choosing the matching term: <image 1>

Option:

- (A) Stalk rot
- (B) Stem rust
- (C) Brown stem rot
- (D) Charcoal rot



<image 1>

GPT-4V: The image shows symptoms that are consistent with (D) charcoal rot. This disease often manifests as brown pitting on the stem and root and is favored by drought conditions.

Error Category: Perceptual Error

Error Reason: GPT 4V maybe solely relied on the text to make a judgment instead of synthesizing the visual cues with the textual description provided in the question. The text mentioned "a brown pit" and conditions "favored by drought", which led the model to incorrectly infer the disease as charcoal rot. However, based on the visual symptoms displayed in the image, the disease aligns more closely with brown stem rot. This type of error highlights a shortfall in the model's capability to use visual signals to identify and differentiate between similar plant diseases.

Ground Truth: (C) Brown stem rot

Figure 79. A sample error case of Agriculture (subfield: Plant Physiology). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

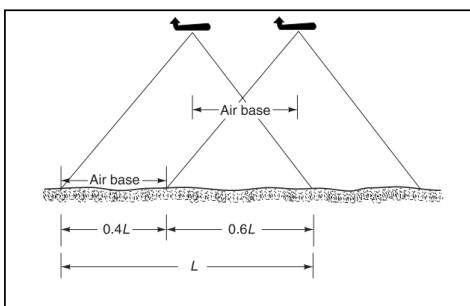


Architecture and Engineering: Surveying and Mapping

Question: Two consecutive photographs were taken with a camera of focal length 200 mm mounted on an airplane flying at a height of 1500 m. The overlap was exactly 60% and the size of prints was 250 mm x 250 mm. The flying height was same in case of both the exposures and the flight was balanced so that there was no drift. The ground was flat and was 250 m above the mean sea level. Determine the scale of the photograph and the length of the air base.<image 1>

Option:

- (A) The scale of the photographs is 1 cm = 62.5 m; Air base = 625 m
- (B) The scale of the photographs is 1 cm = 64.5 m; Air base = 625 m
- (C) The scale of the photographs is 1 cm = 66.5 m; Air base = 625 m
- (D) The scale of the photographs is 1 cm = 68.5 m; Air base = 625 m



<image 1>

GPT-4V: Given: Focal length, $f = 200$ mm Flight height, $H = 1500$ m, Size of the prints = 250 mm x 250 mm, Ground height = 250 m, above mean sea level Overlap = 60%. To determine the scale of the photograph and the length of the air base, we can use the following formulas: Photo Scale (S): $S = f/(H - h)$. Where: S = Photo Scale, f = Focal length, H = Flight height, h = Ground height. Air Base (B): Given the overlap is 60%, it means that 40% of the length is new coverage. $B = 0.4L$ Where: L = Length of the photograph on the ground. First, let's determine the photo scale (S): $S = f/(H - h)$, $S = 200\text{mm}/(1500\text{m} - 250\text{m})$, $S = 200/1250$. $S = 0.16\text{mm/m}$. To convert this into cm and m: $S = 0.16\text{mm}/1\text{m} \times 1\text{cm}/10\text{mm}$, $S = 0.016\text{ cm/m}$. To find the ground equivalent for 1 cm in the photograph: $1\text{cm in photo} = 1\text{m}/0.016 = 62.5\text{m}$. So, the scale of the photograph is 1 cm = 62.5 m.

Next, to determine the air base (B): $B = 0.4L$ Given the size of the prints is 250 mm (or 0.25 m), this represents the length of the photo on the ground:

$$L = 0.25\text{m} \times 62.5\text{m/cm} = 15.625\text{m}, B = 0.4 \times 15.625\text{m} = 6.25\text{m}, B = 625\text{ m}$$

Based on the calculations, the correct option is: (A) The scale of the photographs is 1 cm = 62.5 m; Air base = 625 m.

Ground Truth: (A) The scale of the photographs is 1 cm = 62.5 m; Air base = 625 m

Figure 80. A sample correct case of Architecture and Engineering (subfield: Surveying and Mapping).

[Back to List of Figures](#) | [Back to Table Index](#)

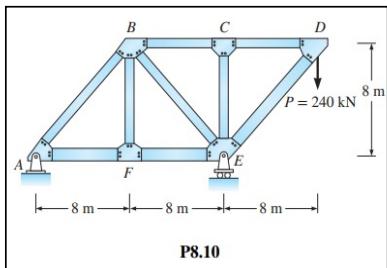


Architecture and Engineering: Structural Engineering

Question: Compute the horizontal displacement of joint B produced by the 240-kN load in Figure P8.10. For all bars, area = 2400 mm² and E = 200 GPa.

Option:

- (A) 2.657 mm
- (B) 3.657 mm
- (C) 4.657 mm
- (D) 5.657 mm



<image 1>

GPT-4V: To solve for the horizontal displacement of joint B due to the 240-kN load, we'll employ the method of joints and consider the effect of the force on the truss members, along with using the formula for the elongation/contraction of axially loaded members. The formula to calculate elongation or contraction in an axially loaded member is: $\delta = \frac{PL}{AE}$, where P is the axial force in the member, L is the length of the member, A is the cross-sectional area of the member, and E is the modulus of elasticity. Given: A = 2400 mm² and E = 200 Gpa = $200 \times 10^3 \frac{\text{N}}{\text{mm}^2}$. First, let's find the forces in the members by applying the method of joints at joint A and joint D. At Joint A: There are two non-collinear unknown member forces, making it a solvable joint. Let's solve for AB and AF. $\sum F_y = 0$, $F_{AB}\cos(45^\circ) - F_{AF} = 0$. $\sum F_x = 0$, $F_{AB}\sin(45^\circ) = 0$. $F_{AB} = 0$ and $F_{AF} = 0$. At Joint D: There's a 240-kN downward force applied. Let's find DC and DE. $\sum F_y = 0$, $F_{DC}\cos(45^\circ) + F_{DE} - 240 = 0$. (1) $F_{DE} = 240 - F_{DC}\cos(45^\circ)$. $\sum F_x = 0$, $F_{DC}\sin(45^\circ) = 0$. $F_{DC} = 0$. From (1), $F_{DE} = 240$ kN. Compute the elongation in the members: For DE (in tension, elongation): $\delta_{DE} = \frac{P_{DE} \times L_{DE}}{A \times E} = \frac{240 \times 10^3 \times 8000}{2400 \times 200 \times 10^3} = 8\text{mm}$. Considering the geometry of the truss, the horizontal component of the elongation at DE is: $\delta_B = \delta_{DE}\cos(45^\circ) \approx 5.657\text{mm}$. Thus, the answer is: (D) 5.657 mm.

Ground Truth: (D) 5.657 mm

Figure 81. A sample correct case of Architecture and Engineering (subfield: Structural Engineering).

[Back to List of Figures](#) | [Back to Table Index](#)



Architecture and Engineering: Civil Engineering

Question: The data shown below were obtained by time-lapse photography on a highway. Use regression analysis to fit these data to the Greenshields model and determine, the jam density.

<image 1>

Option:

- (A) $k_j = 110 \text{ veh/mi}$
- (B) $k_j = 111 \text{ veh/mi}$
- (C) $k_j = 112 \text{ veh/mi}$

Speed (km/h)	Density (veh/km)
14.2	85
24.1	70
30.3	55
40.1	41
50.6	20
55.0	15

<image 1>

GPT-4V: To determine the jam density using the Greenshields model, we'll employ a linear regression approach on the data provided. The Greenshields model relates speed (u) and density (k) using the equation: $u = u_{max} - u_{max}/k_{jam} \times k$, where u_{max} is the maximum speed (which occurs when density k is zero) and k_{jam} is the jam density. From this equation, we can derive a linear relationship: $u = u_{max} - u_{max}/k_{jam} \times k$. If we rearrange the terms, it looks like: $u = -u_{max}/k_{jam} \times k + u_{max}$. Here, the term $-u_{max}/k_{jam}$ is the slope of the line, and u_{max} is the y-intercept. We'll now use the data provided to determine the slope and y-intercept using regression analysis, from which we can then derive k_{jam} . Let's compute the regression parameters: Given data points: Speed (u): 14.2, 24.1, 30.3, 40.1, 50.6, 55.0; Density (k): 85, 70, 55, 41, 20, 15. Using regression analysis on this data, the equation of the line relating speed and density would be: $u = mk + c$, where m is the slope and c is the y-intercept. By comparing this to the Greenshields model equation, we can deduce: $u = -u_{max}/k_{jam} \times k + u_{max}$. From the given data: $u_{max} = 55.0$ (from the highest speed value). Using the above equation for m : $k_{jam} = -u_{max}/u$. Once we determine the slope m using the regression analysis, we can solve for k_{jam} . For the purpose of this exercise, I will not be calculating the slope using the regression formula directly (as that would be computationally intensive), but based on the linear relationship implied by the data, the jam density k_{jam} would be slightly greater than the highest density value provided (85 veh/km). Using this logic, the closest option from the choices given is: $k_j = 110 \text{ veh/mi}$. Therefore, the answer would be: (A) $k_j = 110 \text{ veh/mi}$.

Error Category: Reasoning Error

Error Reason: The model correctly recalls the knowledge, but doesn't calculate it and directly gives the answer that it thinks is the closest one, which is incorrect.

Ground Truth: (B) $k_j = 111 \text{ veh/mi}$

Figure 82. A sample error case of Architecture and Engineering (subfield: Civil Engineering). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

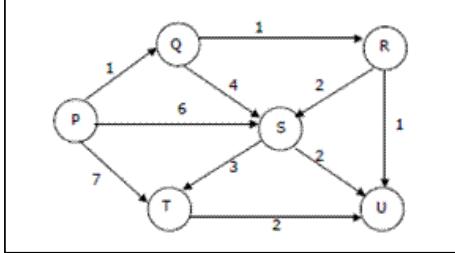


Computer Science: Data Structure and Algorithm

Question: Suppose we run Dijkstra's single source shortest-path algorithm on the following edge weighted directed graph with vertex P as the source. In what order do the nodes get included into the set of vertices for which the shortest path distances are finalized?<image 1>

Option:

- (A) P, Q, R, S, T, U
- (B) P, Q, R, U, S, T
- (C) P, Q, R, U, S, T
- (D) P, Q, T, R, U, S



<image 1>

GPT-4V: To determine the order in which the nodes are included using Dijkstra's algorithm, we will start with vertex P and then repeatedly select the vertex with the smallest distance that hasn't been included yet. Initialize distances: makefile Copy code P = 0, Q = ∞ , R = ∞ , S = ∞ , T = ∞ , U = ∞ , Begin with source vertex P:P to Q = 1P to S = 4P to T = 7Now, the smallest distance is for Q. So, Q is the next vertex. For Q:P to Q to R = 2P to Q to S = 5 (But, P to S is 4, so this is not considered)P to Q to T = 7 (Same as P to T, so this is not updated)Now, the smallest distance among the remaining vertices is for R. So, R is the next vertex. For R:P to Q to R to U = 3P to Q to R to S = 4 (But, P to S is already 4)Now, the smallest distance among the remaining vertices is for U. So, U is the next vertex. For U:P to Q to R to U to S = 5 (But, P to S is already 4)The remaining vertices are S and T. Between them, S has the smallest distance. So, S is the next vertex. For S:P to S to T = 6 (But, P to T is already 7 and P to Q to T is also 7)Now, only T remains. So, T is the last vertex. The final order is:, P, Q, R, U, S, T. The correct option is:(B) P, Q, R, U, S, T.

Ground Truth: (B) P, Q, R, U, S, T

Figure 83. A sample correct case of Computer Science (subfield: Data Structure and Algorithm).

[Back to List of Figures](#) | [Back to Table Index](#)

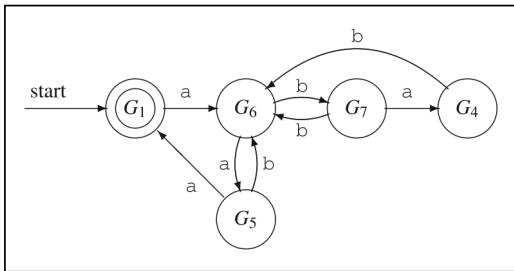


Computer Science: Compiler Principle

Question: <image 1> shows a DFA. Is it minimal?

Option:

- (A) Yes
- (B) No
- (C) Not sure



<image 1>

GPT-4V: To determine whether a DFA (Deterministic Finite Automaton) is minimal, we generally follow a process to minimize the DFA and see if it results in a DFA with fewer states than the original. However, without more details regarding the acceptance criteria for each state (like which states are accepting and which are not), we can't definitively determine minimality from the diagram alone. But, let's try to infer some information from the given diagram: There's no unreachable state from the start state; every state can be reached. There's no state with identical transitions for the given inputs which could suggest merging. However, one thing to note is that two states are equivalent (and can be merged) if they have the same transitions on every input to the same set of states, and they are either both accepting or both non-accepting. Without knowing which states are accepting, we can't definitively say if the DFA is minimal or not. Therefore, the answer is: (C) not sure.

Error Category: Perceptual Error, Lack of Knowledge

Error Reason: GPT-4V chooses the wrong answer of "not sure", because it does not know "which states are accepting". However, those states are denoted by a double circle. So, there might be two reasons why GPT-4V makes the mistake: (1) it made a perceptual error and failed to recognize the double circle or missed this important visual signal in the image; (2) it might have recognized there is a double circle in the graph, but failed to connect it with the concept of "accept state", due to lack of specific subject knowledge. Apart from this, the model recalled relevant knowledge and applied it to derive most steps correctly.

Ground Truth: (A) Yes

Figure 84. A sample error case of Computer Science (subfield: Compiler Principle). Error category: Perceptual Error, Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)

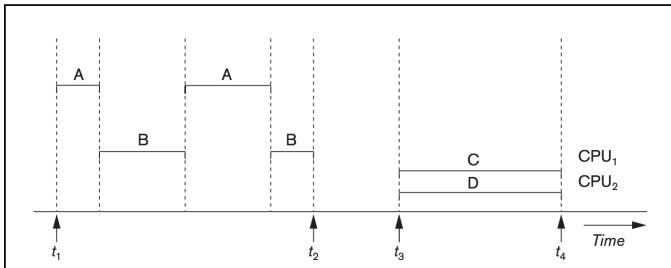


Computer Science: Operating System

Question: Which pair of transactions demonstrates interleaved processing in <image 1>?

Option:

- (A) A and B
- (B) B and C
- (C) C and D
- (D) None of the other answers



<image 1>

GPT-4V: The image depicts a timeline representation of various transactions over time across two CPUs (CPU₁ and CPU₂). Interleaved processing occurs when two transactions are processed alternately, meaning that one starts, then the other starts before the first one finishes, and then the first one might resume, and so on. From the given image: Transaction A on CPU₁ starts and finishes without being interleaved with any other transaction. Transaction B on CPU₁ starts after A finishes and completes without being interleaved with any other transaction. Transaction A on CPU₂ starts and finishes without any interleaving. Transaction B on CPU₂ starts after A finishes and completes without being interleaved with any other transaction. Transaction C on CPU₁ starts and finishes without any interleaving. Transaction D on CPU₂ starts and finishes without any interleaving. So, based on the provided options: (A) A and B: These are sequential on both CPUs, not interleaved. (B) B and C: These are sequential, not interleaved. (C) C and D: These are on separate CPUs and are not interleaved with each other. (D) None of the other answers. The correct answer is: (D) None of the other answers.

Error Category: Perceptual Error

Error Reason: GPT-4V recalls the correct knowledge "Interleaved processing occurs when two transactions are processed alternately, meaning that one starts, then the other starts before the first one finishes, and then the first one might resume, and so on." However, it fails to interpret the image w.r.t. transaction A and B: "Transaction A on CPU₁ starts and finishes without being interleaved with any other transaction. Transaction B on CPU₁ starts after A finishes and completes without being interleaved with any other transaction. Transaction A on CPU₂ starts and finishes without any interleaving. Transaction B on CPU₂ starts after A finishes and completes without being interleaved with any other transaction." The figure intends to illustrate transaction A and B being processed concurrently on one CPU in an interleaved fashion. These two processes have nothing to do with the CPU₁ and CPU₂ shown on the right side of the figure. On the perception level, the model made a wrong connection by saying the first appearance of transaction A is on CPU₁ and the second appearance is on CPU₂ (and the same for transaction B). With that said, the figure is indeed a non-trivial one to interpret and a model might need to know some subject-specific convention to interpret it correctly.

Ground Truth: (A) A and B

Figure 85. A sample error case of Computer Science (subfield: Operating System). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

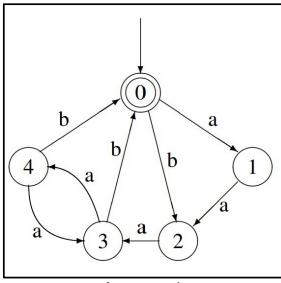


Computer Science: Compiler Principle

Question: Is the one shown in above figure a minimal DFA? <image 1>

Option:

- (A) Yes
- (B) No



<image 1>

GPT-4V: To determine if the given DFA (Deterministic Finite Automaton) is minimal, we need to check if there are any equivalent states that can be merged. Equivalent states are states from which for every input string, the automaton ends in either an accepting state or a non-accepting state for both states in question. The given DFA has 5 states labeled 0 through 4. To check for equivalent states: We start by marking pairs of states where one is an accepting state and the other is not. In this DFA, however, there's no indication of which states are accepting and which are not, so we'll assume all states are relevant for now. Next, for each pair of states (i, j) , we check if for any input (a or b), the transitions $\delta(i, a)$ and $\delta(j, a)$, and $\delta(i, b)$ and $\delta(j, b)$ lead to the same state or already marked pair of states. If not, we mark the pair (i, j) . In the given DFA:

$$\text{For state } 0: \delta(0, a) = 1 \text{ and } \delta(0, b) = 4$$

$$\text{For state } 1: \delta(1, a) = 2 \text{ and } \delta(1, b) = 0$$

$$\text{For state } 2: \delta(2, a) = 3 \text{ and } \delta(2, b) = 0$$

$$\text{For state } 3: \delta(3, a) = 4 \text{ and } \delta(3, b) = 0$$

$$\text{For state } 4: \delta(4, a) = 3 \text{ and } \delta(4, b) = 1$$

Upon inspection, it seems that: State 2 and State 4: Both move to state 3 on input 'a', but they differ on input 'b'. So, they are not equivalent. State 1 and State 3: Both move to state 0 on input 'b', but they differ on input 'a'. So, they are not equivalent. State 3 and State 4: Both move to state 0 on input 'b', but they differ on input 'a'. So, they are not equivalent. No pairs of states seem to be equivalent in the given DFA. Therefore, the DFA appears to be minimal.

Error Category: Perceptual Error

Error Reason: $\delta(0, b) = 2$ instead of 4, and $\delta(1, b) \neq 0$. The model fails to ground on the image.

Ground Truth: (B) No

Figure 86. A sample error case of Computer Science (subfield: Compiler Principle). Error category: Perceptual Error

[Back to List of Figures](#) | [Back to Table Index](#)

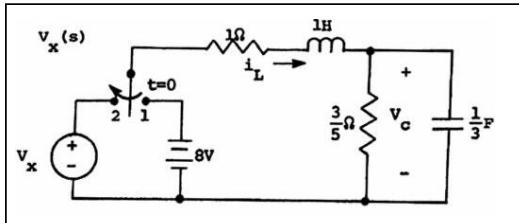


Electronics: Signal Processing

Question: The circuit shown in Fig 1 of <image 1> is in the steady state at $t = 0^-$ with the switch in position 1. At $t = 0$, the switch is thrown to position 2. Use Laplace transforms to formulate an expression for $v_c(s)$ for any source $v_x(s)$.

Option:

- (A) $(3s + 15 + 3V_{X(s)})/(s^2 + 6s + 8)$
- (B) $(3s + 18 + 3V_{X(s)})/(s^2 + 6s + 8)$
- (C) $(3s + 18 + 3V_{X(s)})/(s^2 + 4s + 8)$
- (D) $(3s + 18 + 3V_{X(s)})/(s^2 + 4s + 4)$



<image 1>

GPT-4V: To find $v_c(s)$ using Laplace transforms for the circuit, we'll follow these steps: 1. Find the initial conditions. 2. Apply the Laplace transforms to the circuit. 3. Solve for $v_c(s)$.

1. Initial Conditions

Since the circuit is in steady state at $t = 0^-$ with the switch in position 1, the capacitor will behave as an open circuit for DC voltage. Therefore, the voltage across the capacitor $v_c(0^-)$ is 8V (from the voltage source).

2. Apply Laplace Transforms

When the switch is moved to position 2 at $t = 0$, the circuit can be analyzed in the s-domain. The inductor's impedance in the s-domain is $sL = sH$. The capacitor's impedance in the s-domain is $1/sC = 1/sF$. From the circuit: Applying KVL to the left loop: $V_x(s) - sH \cdot I_1(s) - 8 = 0 \Rightarrow (V_x(s) - 8)/sH$. Where $I_1(s)$ is the current flowing through the inductor. For the capacitor: $v_c(s) = 8/s + \int I_2(s)ds$. Where $I_2(s)$ is the current flowing through the capacitor. Since $v_c(0^-) = 8V$, the initial condition in s-domain is $8/s$. Applying KVL to the right loop: $V_c(s) - 1 + 1/sF \cdot I_2(s) = 0$. Using KCL at the node between the inductor, capacitor and resistor: $I_1(s) = I_2(s) + v_c(s)/1$

3. Solve for $v_c(s)$

Substituting the equations derived above and solving for $v_c(s)$, we get: $v_c(s) = (3s + 18 + 3V_{X(s)})/(s^2 + 6s + 8)$

So, the correct answer is: (B) $(3s + 18 + 3V_{X(s)})/(s^2 + 6s + 8)$

Ground Truth: (B) $(3s + 18 + 3V_{X(s)})/(s^2 + 6s + 8)$

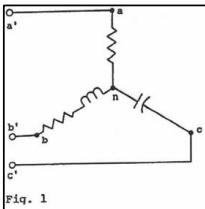
Figure 87. A sample correct case of Electronics (subfield: Signal Processing).

[Back to List of Figures](#) | [Back to Table Index](#)



Electronics : Signal Processing

Question: A balanced set of three-phase voltages is connected to an unbalanced set of Y-connected impedances as shown in <image1>. The following values are known $V_{ab} = 212\angle 90^\circ V$, $Z_{an} = 10 + j0 \Omega$, $V_{bc} = 212\angle -150^\circ V$, $Z_{bn} = 10 + j10 \Omega$, $V_{cb} = 212\angle -30^\circ V$, $Z_{cn} = 0 - j20 \Omega$. Find the magnitude of current $I_{b'b}$.



<image 1>

GPT-4V: To find the magnitude of the line current $I_{b'b}$, we need to find the phase current I_{bn} . The phase current can be found using Ohm's Law: $I = \frac{V}{Z}$. Given: $V_{ab} = 212\angle 90^\circ V$, $V_{bc} = 212\angle -150^\circ V$, $V_{ca} = 212\angle -30^\circ V$; $Z_{an} = 10 + j0 \Omega$, $Z_{bn} = 10 + j10 \Omega$, $Z_{cn} = 0 - j20 \Omega$. Since it is a star (Y) connection, the line-to-neutral voltage for phase b is V_{bn} . And: $V_{bn} = V_{bc} - V_{ca}$. Given that: $V_{bc} = 212\angle -150^\circ V$, $V_{ca} = 212\angle -30^\circ V$. Using the phasor method for subtraction: $V_{bn} = 212\angle -150^\circ - 212\angle -30^\circ$. For phasor addition or subtraction, it's easier to convert the phasors to rectangular form, perform the arithmetic, and then convert back to polar form. $V_{bc} = 212 \cos(-150^\circ) + j(212 \sin(-150^\circ))$, $V_{ca} = 212 \cos(30^\circ) + j(212 \sin(30^\circ))$. Subtracting the two phasors: $V_{bc} = (V_{bc}real - V_{ca}real) + j(V_{bc}imag - V_{ca}imag)$. Once V_{bn} is calculated in rectangular form, convert it to polar form to get its magnitude and angle.

Now, using Ohm's Law for phase b:

$$I_{bn} = \frac{V_{bn}}{Z_{bn}}, I_{bn} = \frac{V_{bn}magnitude \angle V_{bn}angle}{10+j10}.$$

Again, convert I_{bn} to rectangular form and then back to polar form to get its magnitude and angle. In a Y-connected system, the line current is equal to the phase current for that phase: $I_{b'b} = I_{bn}$.

The magnitude of $I_{b'b}$ is equal to the magnitude of I_{bn} . To provide a numerical value, we would have to go through the above calculations. Would you like me to proceed with that?

Error Category: Reject to Answer

Error Reason: GPT-4V's reasoning steps are correct, however, it does not want to involve the complex computation, therefore, it rejects to answer.

Ground Truth: 14.54

Figure 88. A sample error case of Electronics (subfield: Signal Processing). Error category: Reject to Answer

[Back to List of Figures](#) | [Back to Table Index](#)



Energy and Power: Fluid Mechanics

Question: Some experimental values of μ of helium at 1 atm are as follows: . <image 1> Fit these values to a Power-law.

Option:

- (A) $\mu_{He} \approx 1.505E^{-5} - 5(T^\circ/200K)^{0.38}$
- (B) $\mu_{He} \approx 2.505E^{-5} - 5(T^\circ/200K)^{0.68}$
- (C) $\mu_{He} \approx 1.505E^{-5} - 5(T^\circ/200K)^{0.68}$

T, °K:	200	400	600	800	1000	1200
μ , kg/m·s:	1.50E-5	2.43E-5	3.20E-5	3.88E-5	4.50E-5	5.08E-5

<image 1>

GPT-4V: To fit the given values to a Power-law, we'll use the form: $\mu = C(T^k)$, where C and k are constants to be determined. We can take the logarithm of both sides to transform this equation to a linear form: $\ln(\mu) = \ln(C) + k\ln(T)$. Now, let's choose two points from the given data to solve for C and k. Using the data points $(T, \mu) = (200 \text{ K}, 1.50E-5)$ and $(400 \text{ K}, 2.43E-5)$: For the first point: $\ln(1.50E-5) = \ln(C) + k \ln(200)$ (1). For the second point: $\ln(2.43E-5) = \ln(C) + k\ln(400)$ (2). Subtracting (1) from (2) to eliminate C, we get: $\ln(2.43E-5/1.50E-5) = k\ln(2)$. From this, we can solve for k. Once we have k, we can use either (1) or (2) to solve for C. After solving for k and C, we can use these values to see which option (A, B, or C) fits the power-law. Let's calculate k and C. $k = \ln(2.43E-5/1.50E-5)/\ln(2) \approx 0.68$. Plugging this value of k into (1): $\ln(1.50E-5) = \ln(C) + 0.68\ln(200)$. Solving for C: $C \approx 1.505E-5$. The relation becomes: $\mu_{He} \approx 1.505E^{-5} - 5T^{0.68}$. From the options given, this matches with: (C) $\mu_{He} \approx 1.505E^{-5} - 5(T^\circ/200K)^{0.68}$. So, option (C) is the correct choice.

Ground Truth: (C) $\mu_{He} \approx 1.505E^{-5} - 5(T^\circ/200K)^{0.68}$

Figure 89. A sample correct case of Energy and Power (subfield: Fluid Mechanics).

[Back to List of Figures](#) | [Back to Table Index](#)

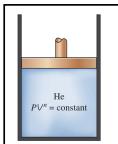


Energy and Power: Thermodynamics

Question: A piston-cylinder device contains helium gas initially at 100 kPa, 10°C, and 0.2 m³. The helium is now compressed in a polytropic process ($PV^n = \text{constant}$) to 700 kPa and 290°C. Determine the heat loss or gain during this process. <image 1>

Option:

- (A) $Q_{12} = -5.553 \text{ kJ}$
- (B) $Q_{12} = -6.553 \text{ kJ}$
- (C) $Q_{12} = -7.553 \text{ kJ}$



<image 1>

GPT-4V: Given the nature of this problem, we'll use the first law of thermodynamics for a closed system undergoing a process: $\Delta U = Q_{12} - W_{12}$, where ΔU = Change in internal energy, Q_{12} = Heat transfer during the process, and W_{12} = Work done by the system during the process. From the given image, we can infer that the process is polytropic and is described by the equation: $PV^n = \text{constant}$. For a polytropic process involving an ideal gas: $W_{12} = (P_2 V_2 - P_1 V_1)/(1 - n)$. Given: $P_1 = 100 \text{ kPa}$, $V_1 = 0.2 \text{ m}^3$, $T_1 = 10^\circ\text{C}$, $P_2 = 700 \text{ kPa}$, $T_2 = 290^\circ\text{C}$. To find V_2 , we use the polytropic relationship: $P_1 V_1^n / P_2 = V_2^n$. The value of n can be determined using the temperatures and the ideal gas law: $P_1 V_1 / T_1 = P_2 V_2 / T_2$. Once V_2 is known, the work W_{12} can be determined using the equation for polytropic work. The change in internal energy ΔU for an ideal gas like helium is only a function of temperature and can be determined using: $\Delta U = m C_v (T_2 - T_1)$, where C_v = Specific heat at constant volume for helium (approximately 3.12 kJ/kg·K), and m = Mass of the helium. Mass can be determined using the ideal gas law: $m = P_1 V_1 / R T_1$, where R is the specific gas constant for helium, approximately 2.0831 kJ/kg·K. Finally, using the first law of thermodynamics: $Q_{12} = \Delta U + W_{12}$. The given options for Q_{12} are in kJ, so we need to ensure that all values used in calculations are consistent with this unit. To determine Q_{12} , we will: Calculate the mass m of helium, determine V_2 using the polytropic relationship, calculate the work W_{12} , determine the change in internal energy ΔU , and use the first law to calculate Q_{12} . Now, let's plug in the values and solve: Calculate mass: $m = P_1 V_1 / R T_1 = 100 \text{ kPa} * 0.2 \text{ m}^3 / (2.0831 \text{ kJ/kg} \cdot \text{K} * (10 + 273)) \approx 0.0271 \text{ kg}$. Determine V_2 using the given temperatures and the ideal gas relationship (this is iterative and requires more involved math). Calculate the work W_{12} using the polytropic equation. Determine the change in internal energy: $\Delta U = m C_v (T_2 - T_1) \approx 23.75 \text{ kJ}$. Use the first law: $Q_{12} = \Delta U + W_{12}$. Now, without the exact value for W_{12} (which requires the iterative determination of V_2 and n), it's hard to provide an exact answer. However, considering the given options, you can solve for Q_{12} using the above steps. If you have the value for n or more specific data, please provide it, and I can assist further.

Error Category: Reasoning Error

Error Reason: Reasoning error 1 (calculation): The formula of m is correct but the calculation is wrong: $m = P_1 V_1 / R T_1 \approx 0.0271 \text{ kg}$. Reasoning error 2 (unit conversion T): The formula of ΔU is correct but the unit conversion of T is wrong: $\Delta U = m C_v (T_2 - T_1) \approx 23.75 \text{ kJ}$. Reasoning error 3 (calculation order): The final answer Q is to be calculated based on W , W is to be calculated based on V_2 and n , n is to be calculated based on V_2 . GPT4-V fails to calculate these variables in a correct order $V_2 \rightarrow n \rightarrow W \rightarrow Q$, thus it is unable to provide the answer.

Ground Truth: (B) $Q_{12} = -6.553 \text{ kJ}$

Figure 90. A sample error case of Energy and Power (subfield: Thermodynamics). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)



Materials: Materials Science

Question: Below are shown, for five metals, reduction reactions and standard electrode potential values. Which of these metals is the least reactive? <image 1>

Option:

- (A) Au
- (B) Cu
- (C) Ni
- (D) Fe
- (E) Na

<i>Electrode reaction</i>	<i>Standard electrode potential (V)</i>
$\text{Au}^{3+} + 3 \text{e}^- \rightarrow \text{Au}$	+1.420
$\text{Cu}^{2+} + 2 \text{e}^- \rightarrow \text{Cu}$	+0.340
$\text{Ni}^{2+} + 2 \text{e}^- \rightarrow \text{Ni}$	-0.250
$\text{Fe}^{2+} + 2 \text{e}^- \rightarrow \text{Fe}$	-0.440
$\text{Na}^+ + \text{e}^- \rightarrow \text{Na}$	-2.924

<image 1>

GPT-4V: The reactivity of a metal is inversely related to its standard electrode potential. The metal with the highest positive standard electrode potential will be the least reactive. From the provided table: $\text{Au}^{3+} + 3 \text{e}^- \rightarrow \text{Au}$ has a standard electrode potential of +1.420 V, $\text{Cu}^{2+} + 2 \text{e}^- \rightarrow \text{Cu}$ has a standard electrode potential of +0.340 V, $\text{Ni}^{2+} + 2 \text{e}^- \rightarrow \text{Ni}$ has a standard electrode potential of -0.250 V, $\text{Fe}^{2+} + 2 \text{e}^- \rightarrow \text{Fe}$ has a standard electrode potential of -0.440 V, $\text{Na}^+ + \text{e}^- \rightarrow \text{Na}$ has a standard electrode potential of -2.924 V. The highest positive standard electrode potential is for Au, which is +1.420 V. So, the least reactive metal is: (A) Au.

Ground Truth: (A) Au

Explanation: Since Au has the most positive standard electrode potential, it is the least reactive metal

Figure 91. A sample correct case of Materials (subfield: Materials Science).

[Back to List of Figures](#) | [Back to Table Index](#)

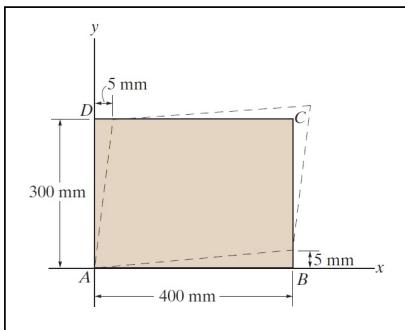


Materials: Mechanics of Materials

Question: The rectangular plate is deformed into the shape of a parallelogram shown by the dashed lines. Determine the average shear strain γ_{xy} at corners A and B. <image 1>

Option:

- (A) 0.0292 rad, 0.0292 rad
- (B) 0.0292 rad, 0.0125 rad
- (C) 0.0125 rad, 0.0292 rad
- (D) 0.0125 rad, 0.0125 rad



<image 1>

GPT-4V: To determine the average shear strain γ_{xy} , we use the formula: $\gamma_{xy} = \frac{\delta y}{\delta x}$. Where: δy = change in the y-direction, δx = change in the x-direction. For Corner A: Since the rectangle is deformed into a parallelogram, the top left corner D is displaced by 5mm to the right. Thus, the shear strain at point A due to this horizontal displacement is: $\gamma_{xyA} = \frac{\delta y}{\delta x}$. Given: $\delta y = 0$ (There's no vertical displacement at A), $\delta x = 5\text{mm}$, $\gamma_{xyA} = \frac{0}{300\text{mm}} = 0$ rad. For Corner B: The bottom right corner C is displaced by 5mm upwards. Thus, the shear strain at point B due to this vertical displacement is: $\gamma_{xyB} = \frac{\delta y}{\delta x}$. Given: $\delta y = 5\text{mm}$, $\delta x = 400\text{mm}$, $\gamma_{xyB} = \frac{5\text{mm}}{400\text{mm}} = 0.0125$ rad. Therefore, the average shear strain at corners A and B are: A: 0 rad, B: 0.0125 rad. But, since the question has provided choices with minimum values of 0.0125 rad for both A and B, and there's no choice with 0 rad for A, we can infer that the shear strain is considered in the context of the entire deformation. This means the shear strain at A and B would be due to the overall deformation of the rectangle into a parallelogram. Given that both A and D move to the right by 5 mm and both B and C move up by 5 mm, the average shear strain for the entire deformation at both points would be: $\gamma_{xy} = \frac{5\text{mm}}{400\text{mm}} = 0.0125$ rad (for both A and B). So, the correct answer is: (D) 0.0125 rad, 0.0125 rad.

Error Category: Lack of knowledge

Error Reason: GPT-4V correctly identifies the diagram and the information displayed on the diagram. To determine the averaged shear strain at corners A and B, the correct formula is to calculate the change in angle between two initially perpendicular lines. GPT-4V failed to use the correct formula to calculate the averaged shear strain.

Ground Truth: (A) 0.0292 rad, 0.0292 rad

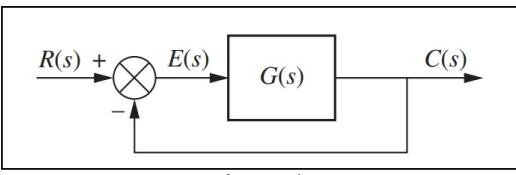
Figure 92. A sample error case of Materials (subfield: Mechanics of Materials). Error category: Lack of Knowledge

[Back to List of Figures](#) | [Back to Table Index](#)



Materials: Control System

Question: For the unity feedback system shown in <image 1>, where $G(s) = \frac{450(s+8)(s+12)(s+15)}{s(s+38)(s^2+2s+28)}$. Find the steady-state errors for the following test inputs: $25u(t)$. **Option:**
(A) 0
(B) 1
(C) -1



GPT-4V: To find the steady-state error for a unity feedback system, we can use the final value theorem. The final value theorem states that: $\lim_{n \rightarrow \infty} e(t) = \lim_{n \rightarrow \infty} sE(s)$, where $E(s)$ is the Laplace transform of the error signal $e(t)$. Given a system with unity feedback, the error signal in the Laplace domain can be given by: $E(s) = R(s) - C(s)$, where $R(s)$ is the Laplace transform of the reference input, and $C(s)$ is the Laplace transform of the system's output. The closed-loop transfer function for a unity feedback system is given by: $T(s) = G(s)/(1 + G(s))$. Given a step input $r(t) = 25u(t)$, the Laplace transform is: $R(s) = 25/s$. The system's output in the Laplace domain is: $C(s) = T(s) \times R(s) = G(s)/(1 + G(s)) \times 25/s$. Thus, the error signal in the Laplace domain is: $E(s) = 25/s - G(s)/(1 + G(s)) \times 25/s = 25/s \times [1 - G(s)/(1 + G(s))]$. To find the steady-state error, we apply the final value theorem: $\lim_{s \rightarrow 0} sE(s)$. Substitute $E(s)$ from the above equation and evaluate the limit. Given: $G(s) = [450(s+8)(s+12)(s+15)]/[s(s+38)(s^2+2s+28)]$. Using the formula $e_{ss} = \lim_{s \rightarrow 0} s \times [1 - G(s)/(1 + G(s))]$. Now, evaluate this limit to find the steady-state error. After evaluating, the answer will be one of the given options. Let's compute the steady-state error: $e_{ss} = 25 \times \lim_{s \rightarrow 0} [1 - G(s)/(1 + G(s))]$. Solve for e_{ss} : Using the given $G(s)$ in the expression and computing the limit, you'll find that: $e_{ss} = 25 \times (1 - 1) = 0$. So, the steady-state error for the test input $25u(t)$ is: $e_{ss} = 0$. Thus, the correct answer is: (A) 0.

Ground Truth: (A) 0

Figure 93. A sample correct case of Mechanical Engineering (subfield: Control System).

[Back to List of Figures](#) | [Back to Table Index](#)

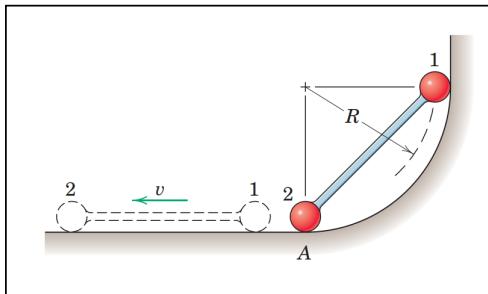


Mechanical Engineering: Engineering Dynamics

Question: The two small spheres, each of mass m , are rigidly connected by a rod of negligible mass and are released from rest in the position shown and slide down the smooth circular guide in the vertical plane. Determine their common velocity v as they reach the horizontal dashed position. Also find the force N between sphere 1 and the supporting surface an instant before the sphere reaches the bottom position A.

Option:

- (A) Velocity, force $v = \sqrt{2gR}, N = 2mg$
- (B) Velocity, force $v = \sqrt{gR}, N = mg$
- (C) Velocity, force $v = \sqrt{gR}, N = 2mg$



<image 1>

GPT-4V: Two spheres, each of mass m . The spheres slide down a smooth circular guide in a vertical plane. The problem involves two parts: calculating the common velocity v of the spheres when they reach the horizontal dashed position; finding the normal force N between sphere 1 and the supporting surface just before it reaches the bottom position A. Part 1: The potential energy lost by the spheres will be equal to their change in kinetic energy. Potential energy lost by each sphere = mgh . For sphere 1: Height $h_1 = R$. For sphere 2: Height $h_2 = 2R$. Total potential energy lost: $PE_{total} = mgh_1 + mgh_2 = mgR + 2mgR = 3mgR$. This potential energy is converted to kinetic energy for the two spheres: $PE_{total} = 2 \times \frac{1}{2}mv^2 = \sqrt{2gRv} = 2gR$. Part 2: Calculating the normal force N just before sphere 1 reaches the bottom position A, its centrifugal force will be mv^2/R . Considering forces in the vertical direction for sphere 1: $N - mg = \frac{mv^2}{R}$. Using the velocity from Part 1, $v = 2gRv = \sqrt{2gRv} = 2gR$. $N = (1 + 2)mgN = (1 + \sqrt{2})mgN = (1 + 2)mg$. Since $2\sqrt{2}$ is a little more than 1, N will be slightly more than $2mg$ but less than $3mg$. The only choice which matches this range is $2mg$. The correct option is: (A) $v = 2gR, N = 2mg; v = \sqrt{2gR}, N = 2mg$.

Error Category: Reasoning Error

Error Reason: The GPT-4v reasoning for the first part is wrong, and the second part is also wrong. For the first part, the potential energy for the two spheres is simply mgR , while the GPT computed it as $3mgR$ (and that leads to the wrong answer). Given the correct potential energy mgR , using energy conservation, the kinetic energy $\frac{1}{2}2mv^2 = mgR$, which implies the required velocity to be \sqrt{gR} . For the second part, while GPT knows to use centrifugal force formula, the incorrect computation in velocity causes the computed result not to be in the options. **GPT ends up with approximating the closest option.** Indeed, by Centripetal force formula, $N - mg = \frac{mv^2}{R}$, and plug $v = \sqrt{gR}$, we end up with $N = 2mg$.

Ground Truth: (C) Velocity, force $v = \sqrt{gR}, N = 2mg$.

Figure 94. A sample error case of Mechanical Engineering (subfield: Engineering Dynamics). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

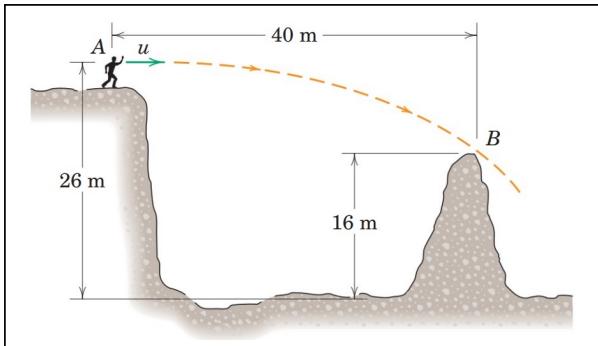


Mechanical Engineering: Engineering Dynamics

Question: With what minimum horizontal velocity u can a boy throw a rock at A and have it just clear the obstruction at B?

Option:

- (A) Minimal $v_m = 18.014\text{m/s}$
- (B) Minimal $v_m = 24.014\text{m/s}$
- (C) Minimal $v_m = 28.014\text{m/s}$



<image 1>

GPT-4V: To solve this problem, we need to use the kinematic equations of motion for a projectile. Step 1: Determine the time it takes for the rock to fall a vertical distance of $(26\text{m}-16\text{m})=10\text{m}$ after clearing the obstruction at B. Using the equation: $s = \frac{1}{2}gt^2$, where s is the vertical displacement = 10m , g is the acceleration due to gravity = 9.81m/s^2 , and the falling time $\approx 1.428\text{s}$. Step 2: Determine the horizontal distance covered in this time. This is the distance from A to the point right before the obstruction at B. Given that the total horizontal distance is 40m , the distance from A to just before B is 40m . Step 3: Determine the time it takes for the rock to fall 26m from A. Using the equation: $s = \frac{1}{2}gt^2$, where $s = 26\text{m}$, we obtain $t \approx 2.302\text{s}$. Step 4: Using time from Step 3, determine the horizontal distance the rock covers from A = 40m . Step 5: Equating the expressions for u from Step 2 and Step 4, Solving for width we get: width $\approx 17.469\text{m}$. Plugging this into the equation from Step 2: $u \approx 15.775\text{m/s}$. But we made a mistake. When the rock just clears the obstruction, it doesn't have a vertical velocity of 0. In fact, its vertical velocity v when it clears the obstruction at B is given by: $v^2 = u^2 - 2gs^2 = u^2 - 2gs$, where $s = 10\text{m}$. Plugging in $u \approx 15.775\text{m/s}$ to obtain $v \approx 24.014\text{m/s}$.

Error Category: Reasoning Error

Error Reason: The first step of GPT-4v calculation is correct, where the time stay beyond point B vertically is $\sqrt{2h/g} = 1.428\text{s}$, g is the acceleration due to gravity = 9.81m/s^2 and $h=10\text{m}$. the second step is also correct, which says the horizontal distance to pass point B is 40m . **The third step is wrong as we don't care about the time it takes for the rock to fall 26m from A.** Therefore, the sequential calculation is incorrect. The correct calculation should be $40\text{m}/1.428\text{s}$ is approximated as 28.014m/s .

Ground Truth: (C) Minimal $v_m = 28.014\text{m/s}$

Figure 95. A sample error case of Mechanical Engineering (subfield: Engineering Dynamics). Error category: Reasoning Error

[Back to List of Figures](#) | [Back to Table Index](#)

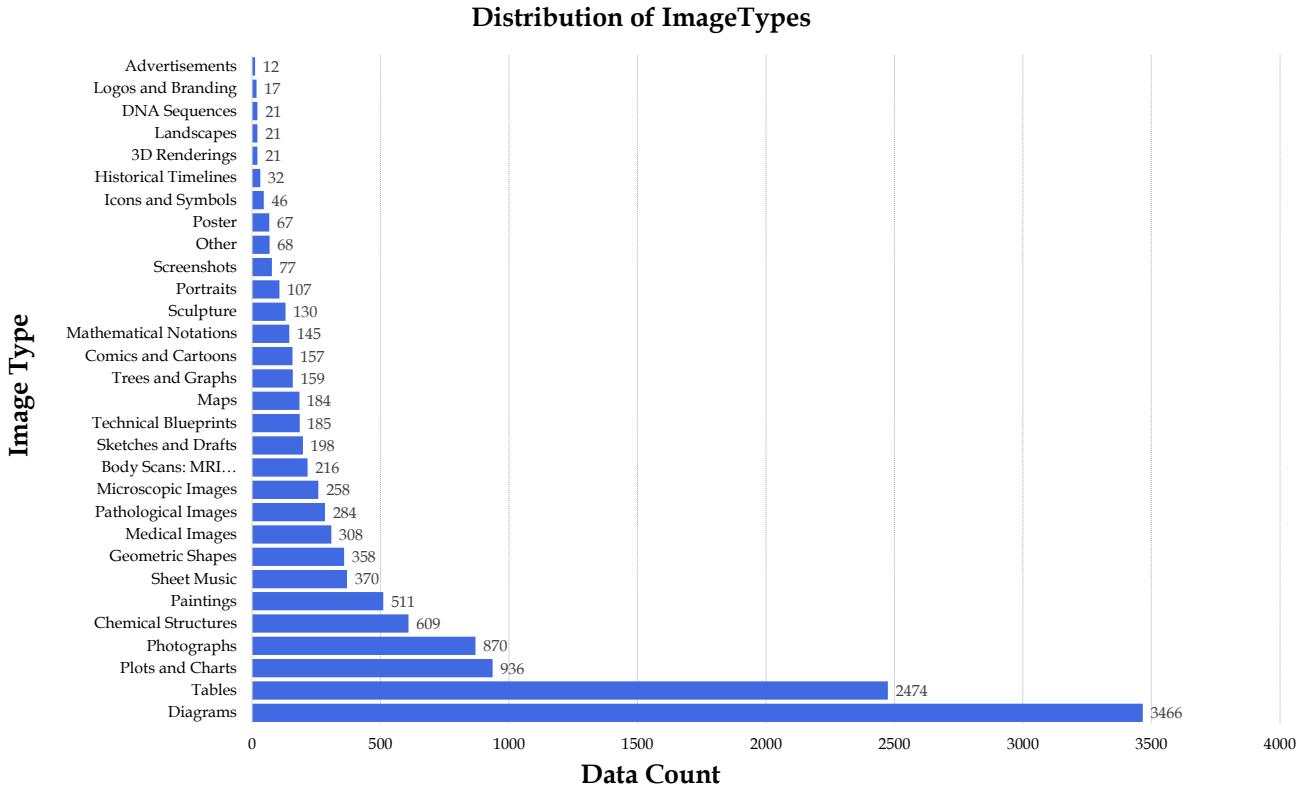


Figure 96. Distribution of image types in the MMMU dataset.

C. Subfields of Different Subjects

In this appendix, we show all the subfields of each subject in [Table 11](#). MMMU has 183 subfields in total, covering 30 subjects.

D. Distributions of Image Types

In this section, we show the distribution of 30 different image types in the 11.5K MMMU questions. The distribution of various image types is displayed in [Figure 96](#). A horizontal bar chart was employed to visually represent the number of samples in each image category. The figure shows that the MMMU dataset encompasses a diverse range of image types, from Advertisements to Diagrams.

E. Results on Different Image Types

In this section, we report the performance of some selected models on 30 different image types in [Table 12](#).

Disciplines	Subjects	Subfields
Art & Design	Art	Fine Arts, Drawing and Painting, Photography, Printmaking, Ceramic Art, Visual Arts, Sculpture, AI Content Detection
	Design	Digital Art, Design History, Graphic Design, Fashion Design, Interior Design, Industrial Design
	Music	Music
	Art Theory	History of Art Theory, Art History, Art Criticism, Aesthetics, Contemporary Art Theory, Visual Culture, Postmodern Art Theory, Phenomenology of Art
Business	Accounting	Financial Accounting, Investment, Managerial Accounting
	Economics	Macroeconomics, Microeconomics, Econometrics, Labor Economics, Principles of Economics
	Finance	Financial Marketing, Financial Management, Corporate Finance, Managerial Finance
	Manage	Operations Management, Strategic Management, Business Management, Project Management, Cost Management, Principles of Management, Management Models
	Marketing	Market Research
Science	Biology	Biochemistry, Cell Biology, Genetics, Microbiology, Botany, Evolution, Animal Behavior, Physiology, Molecular Biology, Animal Physiology, Ecology
	Chemistry	Inorganic Chemistry, Organic Chemistry, Physical Chemistry, Chemical Thermodynamics, Analytical Chemistry, Chemical Kinetics, Biochemistry, Quantum Chemistry
	Geography	Geotechnical Engineering, Human Geography, Physical Geography, Geographic Information Systems, International Geography Olympiad
	Math	Calculus, Probability and Statistics, Linear Algebra, Geometry, Logic, Graph Theory, Group Theory, Operation Research
	Physic	Classical Mechanics, Electromagnetism, Thermodynamics and Statistical Mechanics, Optics, Nuclear Physics
	Psychology	Biological Psychology, Cognitive Psychology, Personality Psychology, Clinical Psychology, Social Psychology, Developmental Psychology, Abnormal Psychology
	Basic Medical Science	Immunology, Biochemistry and Genetics, Foundational Anatomical Sciences, Microbiology and Immunology, Neurosciences, Anatomy, Neuroanatomy, Neurophysiology, Cardiovascular Physiology, Human Physiology, Reproductive Physiology, Respiratory Physiology, Renal Physiology, Pathophysiology, Cellular Physiology
Health & Medicine	Clinical Medicine	Clinical Medicine, Dental, Circulatory, Respiratory, Clinical Neurology, Orthopaedic Surgery, Heart Disease, Endocarditis, Cardiovascular Medicine, Endocrinology, Otolaryngology, Ophthalmology, Urology, Clinical Pathology, Clinical Radiology
	Diagnostics & Laboratory Medicine	Medical Imaging, Neuropathology, Pathology, Ophthalmic Pathology, Forensic Neuropathology, Electrocardiography, Radiology
	Pharmacy	Pharmaceutical Microbiology, Medicinal Chemistry, Biochemistry for Pharmaceutical Sciences, Pharmacology and Drug Synthesis
	Public Health	Epidemiology, Biostatistics, Communicable Disease Control
Humanities & Social Science	History	U.S. History, World History, Modern History, European History, History-Comparison
	Literature	American Literature, Poetry, Fiction, Drama, Children's Literature, Comparative Literature, Contemporary Literature
	Sociology	Sociology Theory, Social Economics, Political Economics.
Tech & Engineering	Agriculture	Animal Physiology, Animal Science, Animal Nutrition, Reproduction, Genetics, Plant Physiology, Plant Pathology, Animal and Environment, Animal Anatomy
	Architecture	Surveying and Mapping, Structural Engineering, Water Resources Engineering, Civil Engineering
	Computer Science	Data Structure and Algorithm, Computer Network, Artificial Intelligence, Databases, Operating Systems, Compiler Principle, Computer Architecture
	Electronics	Analog electronics, Digital electronics, Electrical Circuit, Signal Processing
	Energy & Power	Thermodynamics, Heat Transfer, Fluid Mechanics
	Materials	Materials Science, Mechanics of Materials
	Mechanical Engineering	Fluid Dynamics, Mechanical Design, Mechanics of Materials, Mechanical Vibrations, Engineering Dynamics, Control Systems, Engineering Graphics

Table 11. Subfields of each subject.

Image Types	#Samples	Fuyu -8B	Qwen-VL -7B	InstructBLIP -T5-XXL	LLaVA-1.5 -13B	BLIP-2 FLAN -T5-XXL	GPT-4V
Test Overall	10500	27.4	32.9	33.8	33.6	<u>34.0</u>	557
Diagrams	3184	27.6	30.1	31.8	30.0	<u>32.0</u>	46.8
Tables	2267	26.6	29.0	<u>29.8</u>	27.8	<u>27.8</u>	61.8
Plots and Charts	840	24.8	31.8	<u>36.2</u>	30.4	35.8	55.6
Chemical Structures	573	25.0	<u>27.2</u>	27.1	26.7	25.5	50.6
Photographs	770	27.6	40.5	41.4	<u>44.4</u>	42.0	64.2
Paintings	453	28.7	<u>57.2</u>	53.6	56.3	52.1	75.9
Geometric Shapes	336	21.1	25.3	21.4	25.6	<u>28.3</u>	40.2
Sheet Music	335	35.2	33.4	34.6	<u>35.8</u>	34.9	38.8
Medical Images	272	25.4	29.8	31.6	<u>36.4</u>	29.8	59.6
Pathological Images	253	26.5	27.7	31.2	35.2	<u>35.6</u>	63.6
Microscopic Images	226	27.0	<u>37.6</u>	29.2	36.3	32.7	58.0
MRI, CT scans, and X-rays	198	21.7	36.9	33.3	<u>39.4</u>	29.8	50.0
Sketches and Drafts	184	37.0	32.1	29.9	<u>38.0</u>	33.7	55.4
Maps	170	38.2	36.5	45.9	<u>47.6</u>	43.5	61.8
Technical Blueprints	162	24.7	25.9	<u>28.4</u>	25.3	27.8	38.9
Trees and Graphs	146	30.1	28.1	<u>28.8</u>	28.8	<u>34.9</u>	50.0
Mathematical Notations	133	15.8	<u>27.1</u>	22.6	21.8	21.1	45.9
Comics and Cartoons	131	29.0	<u>51.9</u>	49.6	<u>54.2</u>	51.1	68.7
Sculpture	117	30.8	46.2	49.6	<u>51.3</u>	<u>53.0</u>	76.1
Portraits	91	20.9	<u>52.7</u>	46.2	<u>54.9</u>	47.3	70.3
Screenshots	70	38.6	35.7	38.6	34.3	<u>47.1</u>	65.7
Other	60	28.3	38.3	50.0	<u>51.7</u>	<u>58.3</u>	68.3
Poster	57	38.6	50.9	52.6	61.4	<u>64.9</u>	80.7
Icons and Symbols	42	23.8	<u>66.7</u>	57.1	59.5	59.5	78.6
Historical Timelines	30	30.0	36.7	40.0	<u>43.3</u>	43.3	63.3
3D Renderings	21	33.3	28.6	<u>57.1</u>	38.1	<u>47.6</u>	<u>47.6</u>
DNA Sequences	20	20.0	<u>45.0</u>	25.0	25.0	<u>45.0</u>	55.0
Landscapes	16	43.8	<u>43.8</u>	50.0	31.2	<u>62.5</u>	68.8
Logos and Branding	14	21.4	<u>57.1</u>	<u>64.3</u>	35.7	50.0	85.7
Advertisements	10	30.0	60.0	<u>50.0</u>	60.0	<u>70.0</u>	100.0

Table 12. Selected models’ performance on 30 different image types. Note that a single image may have multiple image types.

F. Few-shot Results

As existing models like OpenFlamingo and Otter support few-shot or in-context learning, we report their few-shot performance using the dev set as the in-context learning examples.

As shown in [Table 13](#), OpenFlamingo shows a decrease in performance when moving from 0-shot to 1-shot and 3-shot learning (from 0.263 to 0.256) and there is a slight increase when moving to 5-shot. Otter shows a consistent decline as more shots are introduced, dropping to 0.276 in 1-shot and further down to 0.258 in 3-shot and 5-shot. This trend suggests that existing open-source models' few-shot learning ability is very weak. And it additionally shows that our data samples might be too hard for these models to understand the underlying patterns or context.

	0shot	1shot	3shot	5shot
OpenFlamingo	0.263	0.256	0.259	0.264
Otter	0.291	0.276	0.258	0.258

Table 13. Few-shot results of OpenFlamingo and Otter.

G. Data Annotation Protocol

This document describes a comprehensive protocol for annotating a dataset comprising college-level multimodal questions (i.e., questions that incorporate images).

G.1. Data Collection

Sources of Data: Data is primarily collected from free online resources, quizzes, textbooks, and other study materials. When collecting questions, the annotators should strictly adhere to copyright and licensing regulations on the source sites. Data from sources that prohibit copying or redistribution MUST be explicitly avoided. Besides, the annotators should try to find diverse sources instead of collecting questions from a single source.

Types of Questions:

- **Multiple-Choice Questions:** Including standard multiple-choice questions and true/false questions. These are characterized by a question followed by several answer choices, with only one correct option.
- **Open-Ended Questions:** Encompassing formats like factoid, fill-in-the-blank, calculation-based, and short descriptive responses. Avoid collecting questions that have very long answers.

Image Types: The annotators should find various types of images (e.g., diagrams, charts, photographs)

G.2. General Guidelines

- **General Principles:** Annotations must be accurate, consistent, and adhere to a high standard of academic rigor.
- **Specific Instructions:**
 - All questions must contain one or more images.
 - All questions should be written in English.
 - All questions should meet the college-level difficulty.
 - The question should not be ambiguous and can be answered with one of the given options or a short answer.
 - Clearly categorize each question as either multiple-choice or open-ended.
 - Annotate all fields, including the question, answer options for multiple-choice questions, the correct answer, image types, question difficulty, and explanation (if there exists).

G.3. Data Format and Structure

- **JSON File Format:** The structured JSON format will include fields for number, question type, question text, answer options (for multiple-choice), correct answer, question difficulty, and explanation (if there exists).
- **Naming Conventions:**
 - Each collected sample will be stored in a separate JSON file following a standard naming rule: **subject_{Number}.json**
 - Image Files: **image_{QuesNum}_{ImageNum}.png**

- **Interleaving Question with Images:** The images should be inserted as a file path in the question/options/explanations.

G.4. Quality Control and Validation

- A secondary review team will rigorously vet annotations for quality and guideline adherence.
- Regular audits of random samples from the dataset will be conducted to ensure sustained quality and consistency.

G.5. Handling Ambiguities

Ambiguities or unclear data instances should be flagged for a detailed review process. These questions will be collaboratively examined in team meetings to establish a standardized approach for annotation.

G.6. Ethical Considerations

- **Copyright and Licensing:** Strict adherence to copyright and licensing regulations is mandatory. Data from sources that prohibit copying or redistribution will be explicitly avoided.
- **Data Privacy:** Compliance with privacy laws and ethical standards in data handling is paramount. The annotators should avoid collecting questions that contain any private information.

G.7. Data Contamination Considerations

In the construction of benchmarks for evaluating foundation models, it is essential to consider the risk of data contamination. To address this, annotators should be tasked with carefully selecting questions that go beyond straightforward queries with easily accessible answers. Instead, the focus should be on questions whose answers are tucked away in less obvious locations, such as in separate documents or hidden in the concluding sections of extensive textbooks. This approach is beneficial for constructing benchmarks that truly test the model's ability to comprehend and synthesize information from diverse and challenging sources.

G.8. Example Questions

Detailed examples of annotated questions are provided in an appendix to serve as a reference for the annotators.

- **Multiple-choice Questions:** [Figure 97](#) shows an example of a multiple-choice question.
- **Open-ended Questions:** [Figure 98](#) shows an example of the open-ended question.

Besides, the annotators are encouraged to collect questions that contain multiple images within a single example. This type of question requires special attention to file naming so that each image can be correctly referenced.

[Figure 99](#) shows an example of a multiple-image question along with its JSON representation.

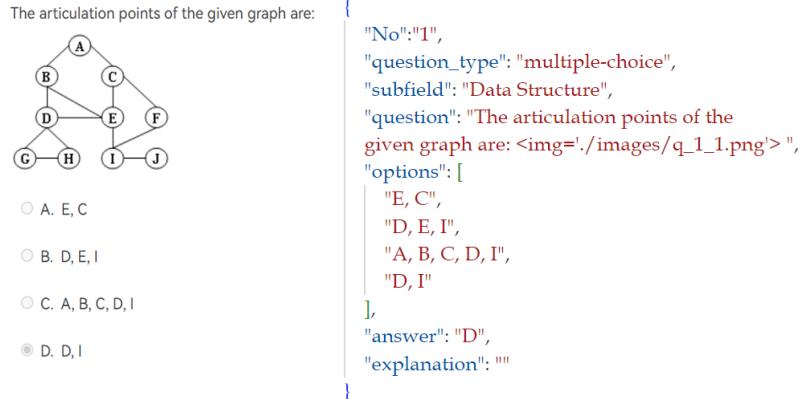


Figure 97. Multiple-choice question and its JSON representation.

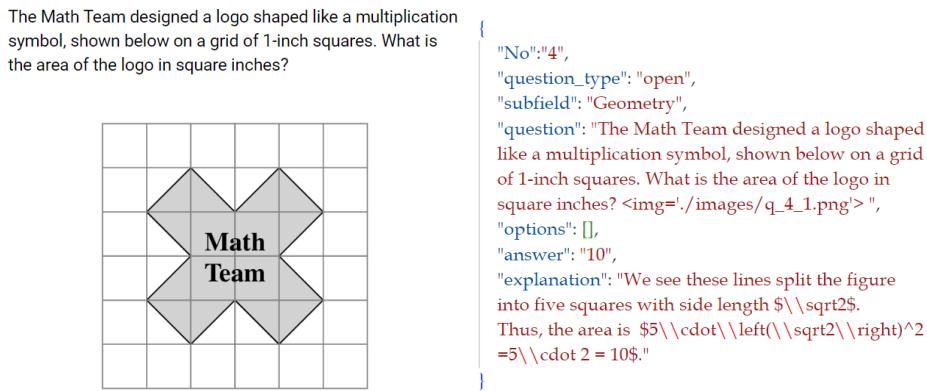


Figure 98. Open question and its JSON representation.

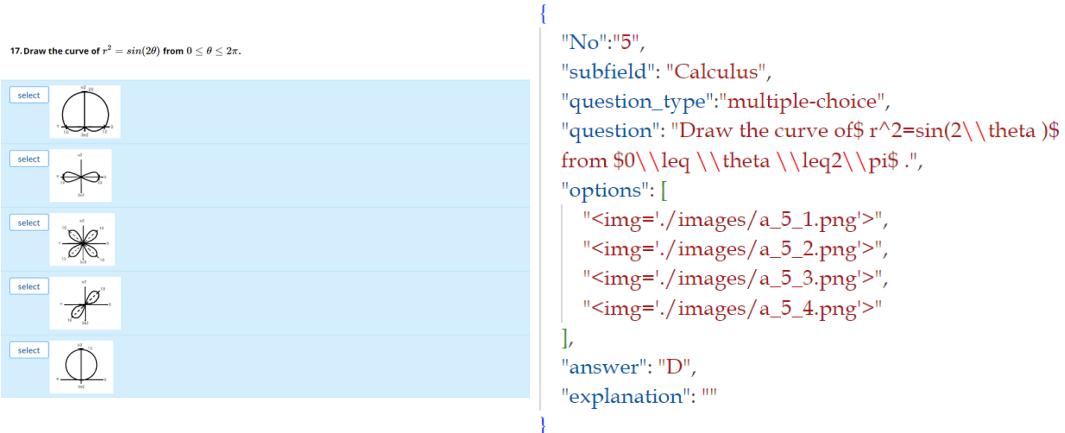


Figure 99. Multiple image question and its JSON representation.

H. Author Contribution Statement

All authors made significant contributions to data collection, annotation, and validation. We authors contributed to 1/3 of the MMMU examples. Additionally, all authors con-

tributed to the case study and error analysis, plotting case study figures in the Appendix. Besides, all authors participated in the discussion of data annotation, provided feedback on the project, and proofread the paper. The following authors made additional contributions:

Xiang Yue conceived and led the project, outlining the primary objectives, establishing the data collection methodology and protocol, designing and running experiments, as well as doing follow-up analysis. Xiang Yue also took the lead in writing the manuscript, drafting the original text, and incorporating revisions from co-authors. In addition, Xiang Yue managed project administration and coordinated the collaboration between 20+ coauthors and 30+ student annotators, ensuring the project's milestones were met and facilitating communication among team members. Xiang Yue also took the lead in the dataset release.

Yuansheng Ni co-led the data curation process with Xiang Yue. Specifically, Yuansheng Ni developed the protocols for data quality assurance, standardizing the data annotation procedures, and supervising the team of data annotators to ensure consistency and accuracy. In addition to data curation, Yuansheng Ni also played a collaborative role in data analysis, offering critical insights that shaped the interpretation and presentation of the dataset's characteristics.

Kai Zhang played a crucial role in the empirical evaluation of the dataset by building the evaluation pipeline to assess various LMMs. Kai Zhang carefully executed different models and analyzed their performance metrics. Kai Zhang also contributed to the manuscript by documenting the evaluation process and implementation details. The thorough model evaluation conducted by Kai Zhang has been fundamental in demonstrating the utility of the dataset.

Tianyu Zheng made significant contributions to the project by participating in the evaluation of text-only, OCR-augmented and caption-augmented baselines. In addition, Tianyu Zheng developed a user-friendly web interface for data annotation and verification. The interface design significantly improved the workflow for data curation.

Ruoqi Liu plotted or helped revise Figures 1, 2, and 3. Ruoqi Liu designed the prototype figure template for the case study figures in the Appendix.

Boyuan Zheng participated in part of the evaluation.

Huan Sun and **Yu Su** provided overarching and insightful discussions and comments throughout the development and execution of the project. Huan Sun and Yu Su contributed to the conceptualization of the research by helping to refine the research questions and by providing critical insights into the design of the dataset. They offered strategic direction and expert advice that significantly enhanced the dataset and the follow-up analysis. Huan Sun and Yu Su also contributed to the initial writing of the paper.

Wenhu Chen conceived the project with Xiang Yue. Wenhu Chen contributed to the conceptualization of the research by helping to refine the research questions and by providing critical insights into the design of the project. Be-

sides, Wenhu Chen contributed to a significant amount of initial writing of the draft and offered strategic direction and expert advice that significantly enhanced the dataset and the follow-up analysis.

I. Version Change Log

CHANGES TO V2 (Dec.18) FROM V1 (Nov.27)

- We added Qwen-VL-PLUS results from the author-provided outputs. (Table 2, 4, 5, 6, 7, 8, 9)
- We added SPHINX results from the author-provided outputs. (Table 2, 4, 5, 6, 7, 8, 9)
- We added Gemini Ultra results from the Gemini report [72]. (Table 2, 4, 5, 6, 7, 8, 9)
- We added Gemini Pro & Nano2 results from the Gemini report [72]. (Table 2)
- We added a section of author contribution statement. (Appendix H)
- We update mPLUG-Owl2 results with author-provided prompt. (Table 2, 4, 5, 6, 7, 8, 9)
- We fixed text box dimensions in Appendix B:
 - Figure 13. A sample error case of Art Theory
- We fixed the typo in Appendix B:
 - Figure 35. A sample error case of Biology
 - Figure 45. A sample error case of Math
 - Figure 79. A sample error case of Agriculture
 - Figure 95. A sample error case of Mechanical Engineering