

---

# YKS<sup>B</sup>ench: Stress-Testing Multimodal Models with Exam-Style Questions

---

Egemen Sert<sup>1</sup>

Şeyda Ertekin<sup>1,2</sup>

<sup>1</sup>Department of Computer Engineering, METU

<sup>2</sup> METU BILTIR CAD/CAM & Robotics Research Center

{egemen.sert, sertekin}@metu.edu.tr

## Abstract

Large Multimodal Models (LMMs) have made remarkable progress in bridging language and vision, yet their performance on visually grounded scientific and exam-style reasoning tasks remains far below human-level ability. To systematically probe these limitations, we introduce YKS<sup>B</sup>ench, a multimodal benchmark of 2,047 university entrance exam questions spanning mathematics, geometry, physics, chemistry, biology, and geography. Unlike prior benchmarks that focus narrowly on mathematics or synthetic tasks, YKS<sup>B</sup>ench captures diverse question formats where visual information is indispensable. Despite the apparent simplicity of many problems to humans, state-of-the-art LMMs show striking deficiencies: the best-performing proprietary model, GPT-5, reaches only 39.34% accuracy, while the strongest open-source model, Gemma 3.27B, achieves 26.82% accuracy. We further provide qualitative analyses and novel scientific figures illustrating failure modes where models misinterpret diagrams, schematics, or spatial layouts. Positioned as a complementary resource to existing benchmarks such as MathVista, MathVision, and MMStar, YKS<sup>B</sup>ench broadens the evaluation landscape and emphasizes the urgent need for improved visual reasoning in LMMs. Dataset is open-sourced at [metu-yks/YKS<sup>B</sup>ench](http://metu-yks/YKS<sup>B</sup>ench).

## 1 Introduction

Large Multimodal Models (LMMs) have advanced rapidly, showing strong performance on vision–language benchmarks. A parallel ecosystem of evaluations probes reasoning in mathematics and science—e.g., MathVista for math-in-vision [1], SEED-Bench for generative comprehension [2], ScienceQA for science VQA [3], and broad capability suites such as MMBench [4]. These benchmarks standardized evaluation and enabled steady progress.

Yet, two issues persist. Many items are visual-unnecessary—solvable without image reasoning, as seen in ScienceQA and MMMU [7]. Others suffer from training leakage, where models recall benchmark items rather than reason. Both inflate scores and obscure failure modes. Domain breadth is another gap. MathVista advanced math reasoning [1], but real exams demand integration of diagrams, maps, and scientific schematics across subjects. Existing datasets capture only part of this diversity.

We introduce YKS<sup>B</sup>ench, 2,047 exam-style questions across Mathematics, Physics, Chemistry, Biology, and Geography. To mitigate leakage, we build from a curriculum-uniform seed, select candidates via cosine similarity, and filter defective or trivial items. Each item has bounding boxes, subject/unit labels, parent IDs, and visual tags. Positioned against MathVista, MMStar,

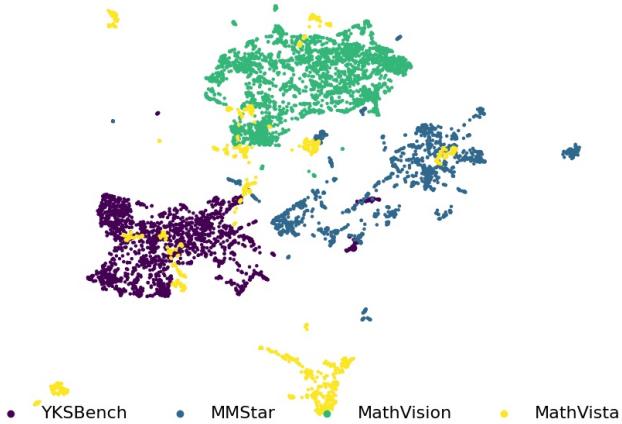


Figure 1: UMAP [5] projection of figure features from MMStar [6], MathVision [7], MathVista [1], and YKSbench. Clusters highlight YKSbench’s unique coverage: maps, atomic diagrams, trees/flowcharts, simple machines, optics, grid arithmetic, isohypsuses, and wave setups. Embeddings from DINOv3 [8] clustered with HDBSCAN [9].

and MathVision, YKSbench covers complementary clusters (Figure 1) with unique visual elements underrepresented elsewhere. These visuals demand indispensable reasoning yet remain challenging for current models. On YKSbench, GPT-5 reaches 39.34%, while GLM-4.5V lags at 26.82%, far below human performance. YKSbench thus complements MathVista [1], SEED-Bench [2], ScienceQA [3], and MMBench [4], isolating where vision still fails and guiding improvements in multimodal training and evaluation.

## 2 Related Work

Reasoning-centric benchmarks such as MMStar [6], MathVista [1], and MathVision [7] have become standard for LMM evaluation and are widely adopted by major labs. MMStar is a target for models like InternLM2 [10] and LLaVA-NeXT [11, 12]. MathVista’s visual math tasks drove evaluation in works like VL-Rethinker [13], R1-VL [14], and large-scale systems such as Kimi [15], Seed1.5-Thinking [16], and Ovis2.5 [17]. These benchmarks shaped analysis of GPT-4 [18], Gemini 2.5 [19], Gemma 3 [20], and GLM-4.5V [21].

Beyond math, MaCBench [22] and LAB-Bench [23] reveal that even state-of-the-art models struggle with domain-specific scientific visuals and multi-step reasoning, suggesting persistent limitations in multimodal integration.

YKSbench builds on this foundation by extending beyond math into broader scientific domains and visual schemas, offering a complementary stress test for multimodal reasoning.

## 3 Experiments

To assess the multimodal reasoning capabilities of contemporary models on complex scientific and mathematical problems, we conduct a series of experiments on our proposed YKSbench dataset. We evaluate overall performance, analyze results broken down by visual modality, and perform a qualitative error analysis to identify common failure modes.

### 3.1 Experimental Setup

**Models** We evaluate four leading multimodal models, comprising two proprietary models, GPT-5 and Gemini 2.5 Pro, and two powerful open-weight models, Gemma 3 27B and GLM 4.5V. This selection allows us to compare the capabilities of both closed and open-source systems.

**Evaluation Protocol** All models are evaluated in a zero-shot setting to test their intrinsic problem-solving abilities without task-specific prompt optimization. We use a deterministic sampling strategy by setting the temperature to 0 for all models. A notable exception is GPT-5, for which the API requires a fixed temperature of 1. The maximum number of completion tokens is set to 4,096 for all evaluations. Same prompt is used for each model and can be found in code provided at Section F.

### 3.2 Overall Performance

The overall accuracy of each model on YKS<sup>B</sup>ench is presented in Table 1. The results underscore the significant difficulty of our benchmark. The top-performing model, GPT-5, achieves an accuracy of only 39.47%. Given that the questions are 5-choice multiple-choice, the random chance baseline is 20%. The best model’s performance is thus less than 20 percentage points above random guessing.

Furthermore, we observe a clear performance gap between proprietary and open-weight models. Gemini 2.5 Pro follows GPT-5 with 35.27% accuracy, while Gemma 3 27B and GLM 4.5V score 26.82% and 19.44%, respectively. Strikingly, GLM 4.5V’s performance is 0.56 points below the random chance baseline.

Table 1: **Accuracy score of varying models on YKS<sup>B</sup>ench.** Each accuracy is indicated as percentage. Row labeled ALL overall accuracy over the whole dataset. Column labeled AVG represents average performance for the row across models. Rest of the rows indicate accuracy per figure tag. Each tag is represented as tag’s first letters. Tag names are shared in Section 3.3

Model	GPT 5* [18]	Gemini 2.5 Pro [19]	Gemma 3 27B [20]	GLM 4.5V [21]	AVG
ALL	39.47	35.27	26.82	19.44	30.25
OD	15.23	16.56	19.87	10.60	15.56
G	18.04	21.57	22.75	13.73	19.02
WD	31.25	21.88	20.31	18.75	23.05
HD	22.92	31.25	27.08	14.58	23.96
VF	35.71	19.64	21.43	19.64	24.11
EG	36.40	18.61	27.20	14.93	24.28
CS	34.52	29.76	26.19	13.10	25.89
SM	30.53	32.06	29.01	19.08	27.67
ES	43.50	34.96	28.86	21.54	32.22
SI	41.97	42.17	28.05	22.24	33.61
T	35.00	40.00	35.00	25.00	33.75
CF	46.84	45.57	22.78	25.32	35.13
FP	48.87	48.87	26.24	25.79	37.44
GI	62.50	50.00	12.50	25.00	37.50
C	51.97	53.95	25.66	26.32	39.47
F	57.69	61.54	26.92	15.38	40.38
S	53.57	60.71	25.00	32.14	42.86
M	52.34	70.09	35.51	18.69	44.16

\* Each model uses  $temperature = 0$  for completions except GPT-5 where the  $temperature = 1$  is unchangeable.

### 3.3 Performance by Input Modality

To investigate how different visual schemas affect model performance, we first categorized each figure in the dataset. We used Gemini 2.5 Flash to programmatically assign tags from a predefined set of 18 categories (listed in Section 3.3). To ensure high-quality labels, we generated tags three times for each image and kept those present in at least two runs, followed by a final manual review.

The performance breakdown in Table 1 reveals that certain visual modalities pose a much greater challenge than others. Specifically, models perform below their overall average on questions involving Optics Diagrams (OD), Grids (G), Wave Diagrams (WD), Hydrostatics Diagrams (HD), Vector Fields (VF), Euclidean Geometry (EG), Circuit Schematics (CS), and Simple Machines (SM).

While the literature on solving Geometry questions with language models is rich, our dataset demonstrates that this is not an isolated issue. We identify five other diagrammatic modalities where

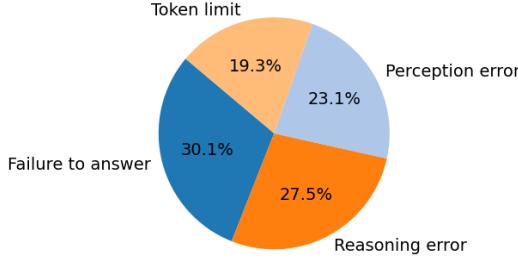


Figure 2: Error distribution of random 500 incorrect solutions generated by Gemini 2.5 Pro. More than half of the errors were due to an error in perception or reasoning.

models perform worse than on EG questions: Optics Diagrams, Grids, Wave Diagrams, Hydrostatics Diagrams, and Vector Fields. This highlights a broader weakness in precise spatial and abstract reasoning that extends beyond pure geometry.

### 3.4 Qualitative Error Analysis

We analyzed 500 incorrect answers from Gemini 2.5 Pro, chosen for its deterministic outputs ( $T = 0$ ) and reasoning tokens. Errors fall into four categories:

1. **Perception:** Misreading values or shapes.
2. **Reasoning:** Correct perception but faulty logical or mathematical steps.
3. **Failure to Answer:** Halting without a final response.
4. **Token Limit:** Reasoning cut off mid-generation.

Reasoning (27.5%) and Perception (23.1%) together account for over half of failures, while 30.1% are refusals and 19.3% truncations (Figure 3.3). By modality, Perception errors cluster in Grid spaces (G), Euclidean Geometry (EG), and Scientific Illustrations (SI). Reasoning failures dominate SI, EG, and Electromagnetism Schematics (ES). Failures to Answer are most common in EG. These patterns show weaknesses in both perception and multi-step scientific reasoning.

## 4 Conclusion

We presented **YKS<sup>B</sup>ench**, 2,047 exam-style multimodal questions across math, science, and geography. Even GPT-5 achieves only 39.47%, far below human performance. Error analysis shows most failures arise from perception and reasoning breakdowns, especially in optics, grids, and machines. YKS<sup>B</sup>ench surfaces these weaknesses, offering a benchmark for advancing multimodal reasoning.

## 5 Limitations

**Monolingual Scope** All items are in Turkish, potentially disadvantaging models with limited Turkish ability. A multilingual version would isolate visual reasoning more effectively.

**Evaluation Parameters** GPT-5 requires  $T = 1$ , introducing stochasticity absent from the  $T = 0$  runs of other models.

**Metrics** We report only final-answer accuracy. This omits reasoning coherence, factuality, or overall helpfulness, which remain important future metrics.

## 6 Social Impact

YKS<sup>B</sup>ench is derived from practice questions for the Turkish university entrance exam (YKS), a high-stakes assessment. Strong model performance raises risks of misuse (e.g., automated cheating).

We recommend suspending systems with high YKSBench accuracy during exam hours to preserve fairness, ensuring AI supports education without undermining competition.

## 7 Acknowledgements

This research received funding from the Research Universities Support Program (YOK-ADEP) with project number ADEP-312-2024-11490.

## References

- [1] Pan Lu et al. *MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts*. 2024. arXiv: 2310.02255 [cs.CV]. URL: <https://arxiv.org/abs/2310.02255>.
- [2] Bohao Li et al. *SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension*. 2023. arXiv: 2307.16125 [cs.CL]. URL: <https://arxiv.org/abs/2307.16125>.
- [3] Pan Lu et al. *Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering*. 2022. arXiv: 2209.09513 [cs.CL]. URL: <https://arxiv.org/abs/2209.09513>.
- [4] Yuan Liu et al. *MMBench: Is Your Multi-modal Model an All-around Player?* 2024. arXiv: 2307.06281 [cs.CV]. URL: <https://arxiv.org/abs/2307.06281>.
- [5] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML]. URL: <https://arxiv.org/abs/1802.03426>.
- [6] Lin Chen et al. *Are We on the Right Way for Evaluating Large Vision-Language Models?* 2024. arXiv: 2403.20330 [cs.CV]. URL: <https://arxiv.org/abs/2403.20330>.
- [7] Ke Wang et al. *Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset*. 2024. arXiv: 2402.14804 [cs.CV]. URL: <https://arxiv.org/abs/2402.14804>.
- [8] Oriane Siméoni et al. *DINOv3*. 2025. arXiv: 2508.10104 [cs.CV]. URL: <https://arxiv.org/abs/2508.10104>.
- [9] Claudia Malzer and Marcus Baum. “A Hybrid Approach To Hierarchical Density-based Cluster Selection”. In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, Sept. 2020, pp. 223–228. DOI: 10.1109/mfi49285.2020.9235263. URL: <http://dx.doi.org/10.1109/MFI49285.2020.9235263>.
- [10] Zheng Cai et al. *InternLM2 Technical Report*. 2024. arXiv: 2403.17297 [cs.CL]. URL: <https://arxiv.org/abs/2403.17297>.
- [11] Feng Li et al. *LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models*. 2024. arXiv: 2407.07895 [cs.CV]. URL: <https://arxiv.org/abs/2407.07895>.
- [12] Zhe Chen et al. *Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling*. 2025. arXiv: 2412.05271 [cs.CV]. URL: <https://arxiv.org/abs/2412.05271>.
- [13] Haozhe Wang et al. *VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning*. 2025. arXiv: 2504.08837 [cs.LG]. URL: <https://arxiv.org/abs/2504.08837>.
- [14] Jingyi Zhang et al. *RI-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization*. 2025. arXiv: 2503.12937 [cs.AI]. URL: <https://arxiv.org/abs/2503.12937>.
- [15] Kimi Team et al. *Kimi k1.5: Scaling Reinforcement Learning with LLMs*. 2025. arXiv: 2501.12599 [cs.AI]. URL: <https://arxiv.org/abs/2501.12599>.
- [16] ByteDance Seed et al. *Seed1.5-Thinking: Advancing Superb Reasoning Models with Reinforcement Learning*. 2025. arXiv: 2504.13914 [cs.CL]. URL: <https://arxiv.org/abs/2504.13914>.
- [17] Shiyin Lu et al. *Ovis2.5 Technical Report*. 2025. arXiv: 2508.11737 [cs.CV]. URL: <https://arxiv.org/abs/2508.11737>.
- [18] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: <https://arxiv.org/abs/2303.08774>.
- [19] Gheorghe Comanici et al. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. 2025. arXiv: 2507.06261 [cs.CL]. URL: <https://arxiv.org/abs/2507.06261>.
- [20] Gemma Team et al. *Gemma 3 Technical Report*. 2025. arXiv: 2503.19786 [cs.CL]. URL: <https://arxiv.org/abs/2503.19786>.

- [21] V Team et al. *GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning*. 2025. arXiv: 2507.01006 [cs.CV]. URL: <https://arxiv.org/abs/2507.01006>.
- [22] Nawaf Alampara et al. “Probing the limitations of multimodal language models for chemistry and materials research”. In: *Nature computational science* (2025), pp. 1–10.
- [23] Jon M. Laurent et al. *LAB-Bench: Measuring Capabilities of Language Models for Biology Research*. 2024. arXiv: 2407.10362 [cs.AI]. URL: <https://arxiv.org/abs/2407.10362>.

UMAP Projection of Dataset Features

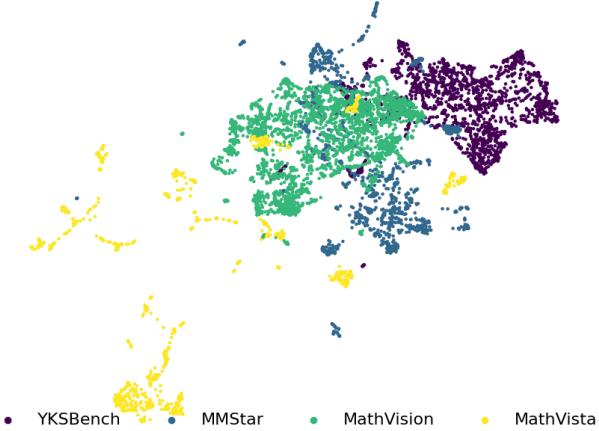


Figure 3: Pipeline of Figure 1 without using local feature embeddings (CLS token only).

## A Measuring Dataset Similarity

**Embedding questions** We measure the similarity across datasets by embedding each sample first. To focus exclusively on figure similarity, we first annotated bounding box for each figure in YKSBench dataset. If there are more than one bounding boxes per image, we take the union of boxes.

Next, we resize each image to 512 pixels on its longest edge. Then propagate the images to DINOv3 [8] (ViT-H/16+). We featurize each sample by concatenating its CLS token embedding with the mean of local patches.

Then we reduce dimensionality of features to 2D using UMAP [5] via its Python package where  $n\_neighbors=15$ ,  $min\_dist=0.1$ . We used cosine as distance metric and set  $random\_state=42$  for reproducibility. Figure 1 shows the resulting 2D manifold. We also applied the same pipeline using only the CLS token embeddings, as shown in Figure 3. You can find code for each UMAP pipeline among the code provided in Section F.

**Clustering samples across datasets** To analyze similarities across datasets, we used each dataset sample’s (CLS + Local) embeddings. Then applied HDBSCAN [9] using cosine as distance metric. We set  $min\_dist=2$  and  $min\_samples=2$  so that if any two samples are within 90% similar to each other, they form a cluster. You can find code for each UMAP pipeline among the code provided in Section F.

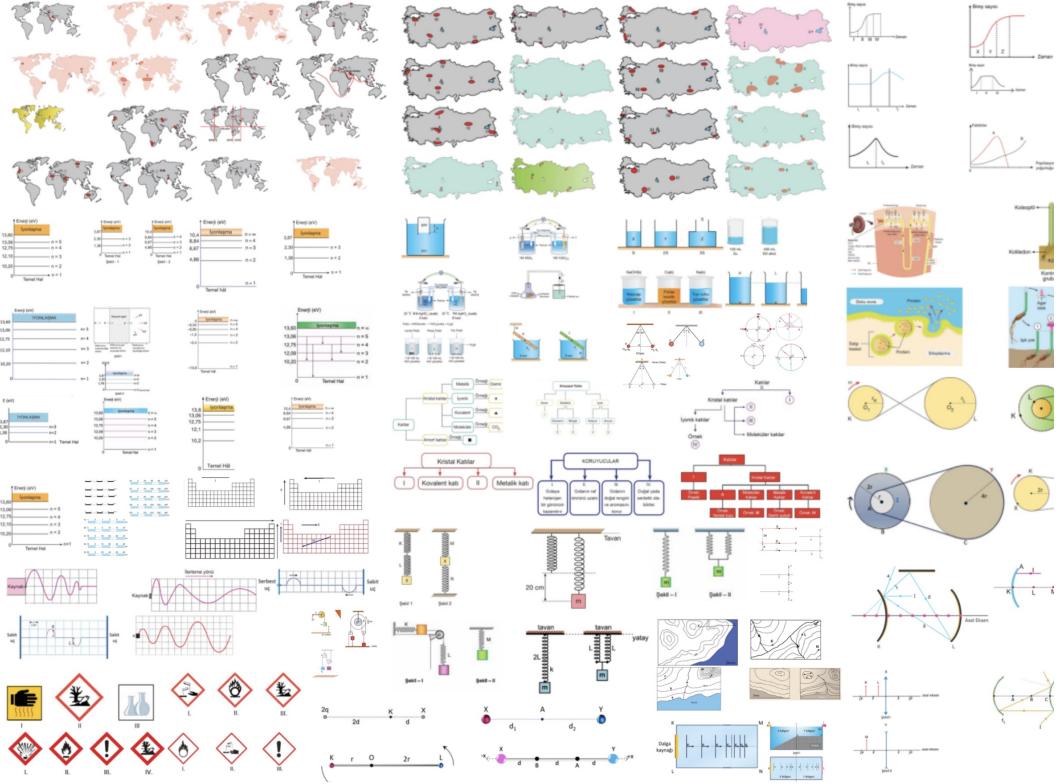


Figure 4: Unique clusters of YKSbench where no other figure is present in the same cluster computed by HDBSCAN algorithm [9]. Clusters highlight YKSbench’s unique coverage: maps, atomic diagrams, trees/flowcharts, simple machines, optics, grid arithmetic, isohypsies, and wave setups. Embeddings from DINOv3 [8] clustered with HDBSCAN.

## B Ablations

### B.1 Commonalities between YKSbench and Other Datasets

Figure 4 shows figure modes unique to YKSbench. Although UMAP analysis on Figure 1 indicates there might be some overlap between datasets. Here we share those overlaps in Figure 5 and 6.

## C Error Analysis

Here we showcase an example per failure mode. Note that each error case is accessible through the error analysis notebook. Readers are encouraged to investigate failure modes.

Figure 8 shows a "failure to answer" case. Figure 9 shows a case where model makes a "reasoning error". Figure 10 displays a case where model fails to perceive the visual content accurately - resulting in a "perception error". Finally, Figure 11 displays a case where model's correct reasoning is interrupted due to "token limit".

### C.1 Failure to answer

You can find an example case in Figure 8.

### C.2 Reasoning error

You can find an example case in Figure 9.

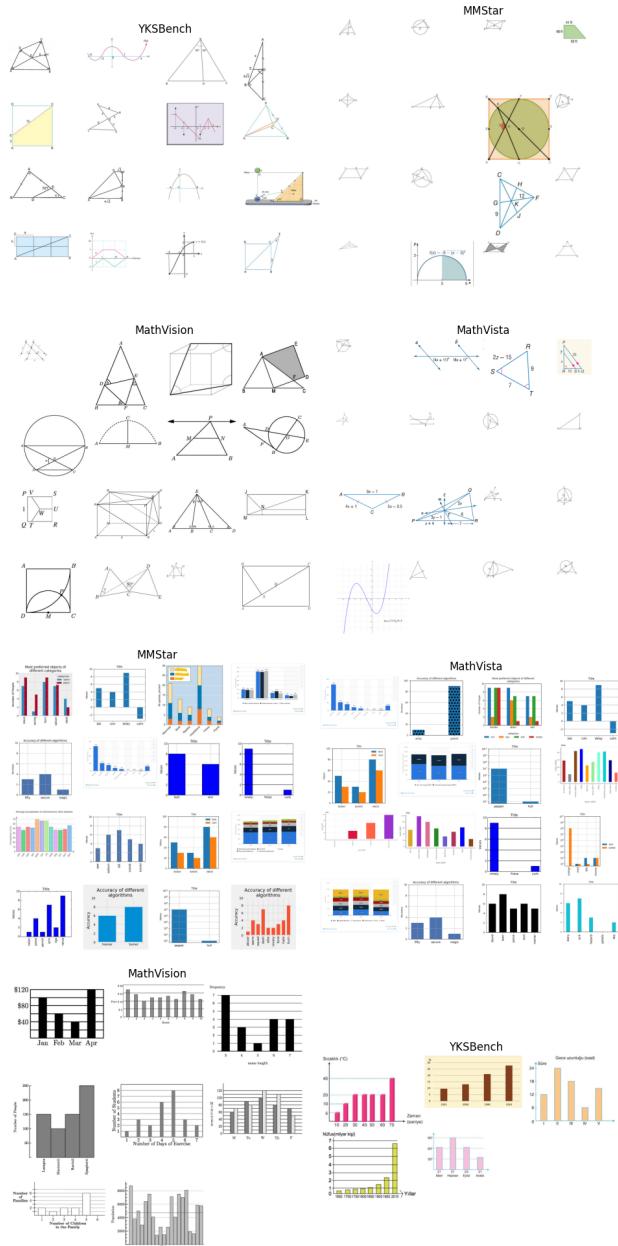


Figure 5: Common figures identified through HDBSCAN clustering, where clusters contain images from each dataset. Geometric shapes and charts are consistently shared figure modes across datasets.

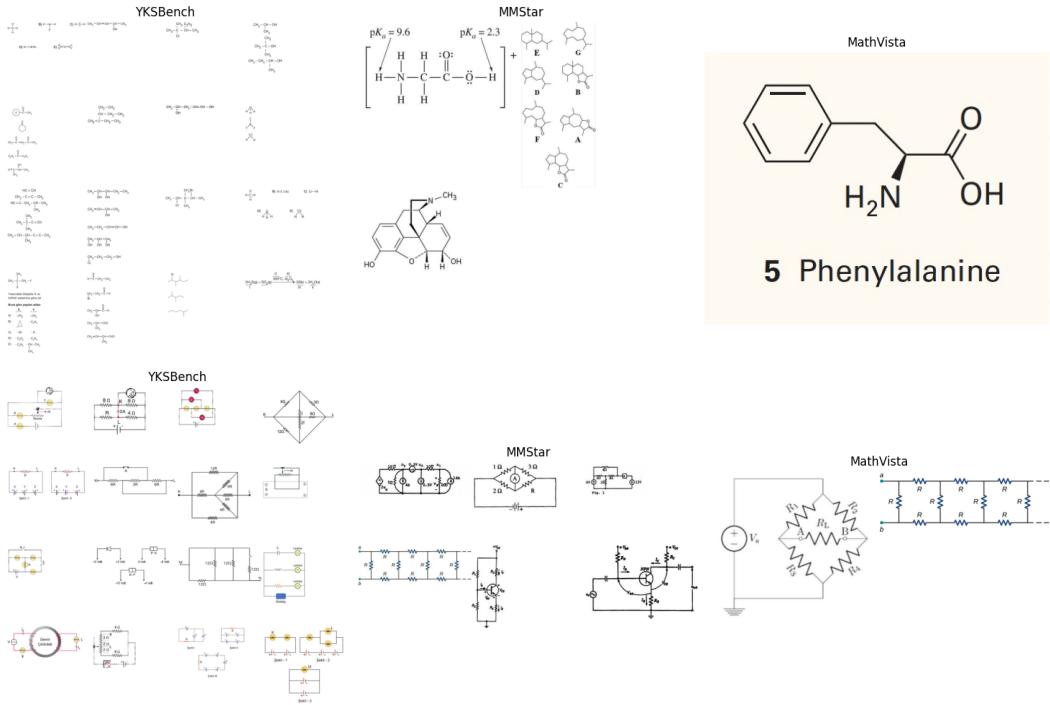


Figure 6: Although less pronounced than in Figure 5, circuits and compounds also exhibit overlap with existing datasets.

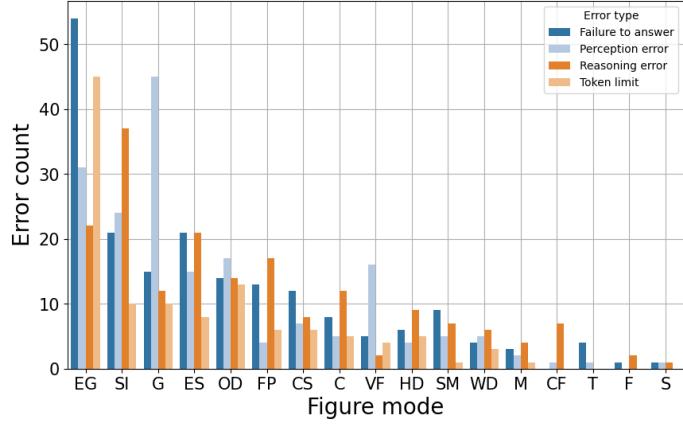
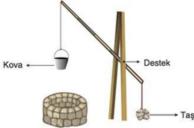


Figure 7: Error count by figure mode. Most amount of perception errors occur on grid spaces (G), questions that contain euclidean geometry (EG) or scientific illustrations (SI). Reasoning errors are mainly focused on scientific illustrations (SI), and questions containing elements of euclidean geometry (EG) or electromagnetism schematics (ES). We see that the model fails to provide an answer mainly on Euclidean Geometry (EG) based questions.

2. Bir öğrencinin su dolu kovayı kuyudan yukarı çıkarabilmek için tasarladığı düzeneğin maketi şekilde verilmiştir.



Bu düzenekte boş kova ipinden çekilecek kuyuya daldırılıyor. Kova suyla dolunca ip serbest bırakılır ve su dolu kova yukarı çıkarıyor.

Buna göre tasarlanan düzeneğin ilgili aşağıdakilerden hangisi doğru olabilir?

- A) Taşın ağırlığı su dolu kovanın ağırlığından küçütür.
- B) Kova ile taşın ağırlığı eşitlenendinde düzenek dengede kalır.
- C) Boş kovayı suya daldıran kuvvet taşın ağırlığından daha azdır.
- D) İçi dolu kovayı kuyudan çıkartan kuvvet taşın ağırlığına eşittir.
- E) Sistemde yoldan kazanç sağlanmıştır.

#### The Physics of a Shaduf

Okay, so I'm looking at this shaduf setup. The problem is clear: we need to figure out which statement is \*correct\* about how this works. First, it's a lever system. I've got a bucket (the load), a counterweight (the stone, or effort), and the pivot (the fulcrum). The distances are crucial here: the distance from the pivot to the bucket is \*greater\* than the distance to the stone. My initial thought is torque must be key here.

I'm thinking, for the bucket to \*rise\* when full, the stone's torque has to be larger than the full bucket's torque. This means the stone needs to be \*heavier\* than the bucket when full. So, option A is wrong (stone's lighter).

If the weights were equal, B is out because the lever wouldn't move! The stone's longer arm will cause it to win.

Now, lowering the empty bucket.. This is where it gets interesting. To lower it, you're \*fighting\* the stone's torque, which \*wants\* to lift it. I need to apply a force. The option C says I'm applying a force less than the weight of the stone. Logically, this makes sense for a simple machine design. I'm applying a force because the stone side is down, making it the effort for the user. With the long lever arm on the bucket side, the small, manual force overcomes the stone on the shorter lever arm.

Let's examine D. D is incorrect. The force is always less than the stone's, because the lever makes lifting the full bucket easier than its actual weight. The rope's tension is less because the stone helps the lift.

This leaves E: distance gain. So, for a small change of the empty bucket being pushed down, the water-filled bucket moves up.

#### Conclusion:

C seems correct. The user applied force must be less than the stone's weight when pushing down on the empty bucket. So, the student is lifting the bucket via a gain in torque. <option>C</option>

Figure 8: Example question where the vision model found the solution; however, failed to report it in correct format. Brown part shows the missing content.

10. Elektrik yükü iki cisimden oluşan sistem şekildeki gibidir.



K noktasında elektriksel alan sıfır olduğuna göre K noktasındaki elektriksel potansiyel kaç  $\frac{kq}{d}$  'dir?

A) 0      B)  $\frac{1}{2}$       C) 1      D)  $\frac{3}{2}$       E) 2

#### <thought>Initial Assessment and Problem Breakdown

Okay, this problem is about electrostatics: specifically, electric fields and potentials. The core is using the knowledge of how fields and potentials change around point charges, and applying superposition (adding individual contributions).

The setup is two point charges: one positive '+2q' at some location, and another charge 'q\_X' at a location we're calling 'X'. We know these facts:

- \* The electric field is zero at a point 'K' situated between them.
- \* The distances: 'K' is '2d' away from the '+2q' charge, and 'd' from the charge at 'X'.

#### ...

#### Step-by-Step Approach

1. Analyze the Zero Field Condition: I need to find 'q\_X' first.

- \* I'll calculate the field from '+2q' at 'K' (let's call it 'E\_1').
- \* I'll then determine what 'E\_X' (the field from 'q\_X') needs to be to cancel 'E\_1'.
- \* This will give me the sign and magnitude of 'q\_X'.

#### 2. Electric Potential Calculation:

- \* The electric potential is a scalar. It just adds up arithmetically.
- \* I'll calculate the potential at 'K' due to '+2q' (let's call it 'V\_1').
- \* I'll calculate the potential at 'K' due to 'q\_X' (let's call it 'V\_X').
- \* I'll add the two potentials 'V\_1 + V\_X' to find 'V\_K', the total potential at 'K'.

#### 3. Units & Result: I'll express the final answer in terms of 'kq/d', as the problem requires.

#### Detailed Calculations

##### \* Finding $q_X$

- \* The field from '+2q' at 'K' is ' $E_1 = k(2q)/(2d)^2 = kq/(2d^2)$ '. It's pointing to the right (away from the positive charge).
- \* For the total field to be zero, ' $E_X$  (from the unknown charge ' $q_X$ ) must point to the left and have the same magnitude. This implies that ' $q_X$  is negative, since ' $E_X$  needs to point towards it.'
- \* ' $E_X = kq_X/d^2$ '. Setting the magnitudes equal: ' $k(2q)/(4d^2) = k|q_X|/d^2$ '. This simplifies to ' $q/2 = |q_X|$ ', and since it is negative, ' $q_X = -q/2$ '.

...

<option>B</option>

Figure 9: Example question where the vision model had an error in its reasoning. Even though the model correctly positioned charged particles, it reasoned the polarity of X to be negative whereas it should be positive. Errors are marked in brown.

### C.3 Perception error

You can find an example case in Figure 10.

1. Eşit karelere bölünmüş düzlemede  $\vec{K}$ ,  $\vec{L}$ ,  $\vec{M}$ ,  $\vec{N}$  vektörleri şekilde verilmiştir.

Buna göre ;

- I.  $|\vec{K}| = |\vec{M}|$
- II.  $|\vec{L}| = |\vec{N}|$
- III.  $|\vec{M}| = |\vec{L}|$

İfadelerinden hangileri doğrudur?

A) Yalnız I	B) Yalnız II	C) Yalnız III
D) I ve II	E) I ve III	

<thought>My Thought Process on the Vector Problem

Okay, so I'm looking at a physics problem involving vectors. Four of them, K, L, M, and N, are presented on a grid. My task is to figure out which of the three statements about their magnitudes and directions are true. Let's break this down systematically.

First, I need to define each vector using components. Each square on the grid has a side length of one unit. I'll set up my coordinate system with the origin at the bottom-left corner of the grid. That makes the calculations easier.

- \* Vector K: Looks like it goes 2 units to the right and 1 unit down. So,  $K = (2, -1)$ . Its magnitude  $|K|$  is the square root of  $(2^2 + (-1)^2)$ , which simplifies to  $\sqrt{5}$ .
- \* Vector L: Moves 2 units right and 2 units up. So,  $L = (2, 2)$ . Its magnitude  $|L|$  is the square root of  $(2^2 + 2^2)$ , which comes out to  $\sqrt{8}$ , or  $2\sqrt{2}$ .
- \* Vector M: Seems to go 3 units right and 1 unit up. So,  $M = (3, 1)$ . Its magnitude  $|M|$  is the square root of  $(3^2 + 1^2)$ , or  $\sqrt{10}$ .
- \* Vector N: Moves 2 units right and 2 units down. So,  $N = (2, -2)$ . Its magnitude  $|N|$  is also  $\sqrt{8}$ , or  $2\sqrt{2}$ .

...

Wait, that's odd. All three statements are false? I should double-check my work. Okay, I'll go through the components again. Let me just confirm everything one more time to be absolutely sure.

I re-examined the vector drawings and the components of each vector.

- \* K: (2, -1) Magnitude is  $\sqrt{5}$  - Checked.
- \* L: (2, 2) Magnitude is  $\sqrt{8}$  - Checked.
- \* M: (3, 1) Magnitude is  $\sqrt{10}$  - Checked.
- \* N: (2, -2) Magnitude is  $\sqrt{8}$  - Checked.

...

Let's test this.

I'm checking the directions again. So, with careful examination, I confirm the grid is drawn and marked clearly.

If, after multiple checks I am wrong... There must be another thing.

If Statement  $I = |L| = |N|$ , then I am correct.

After several checks, the question itself may be incorrect. There might be a typo.

Figure 10: Example question where the vision model had an error in its perception. The model fails to capture direction of K and L vectors. It fails at solving question, then does introspection; however, cannot notice the error in vector positions. Errors are marked in brown.

### C.4 Token limit

You can find an example case in Figure 11.

## D Dataset Description

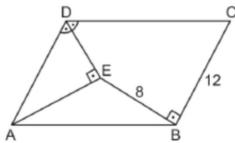
YKSbench consists of 2,047 standardized exam questions. The dataset is constructed on 114 seed level questions. Then remaining questions are selected from a pool of 65K questions based on textual solution similarity to seed level questions. Each question is manually reviewed to ensure quality. Dataset can be found [here](#).

Below we share columns of the dataset.

- **id:** Unique identifier of each sample in the dataset.
- **image:** Question screenshot normalized to 512 pixels on its longest side by preserving original aspect ratio.
- **semantic\_parent:** Question id of the screenshot derived from. If seed item, then this column is equal to an empty string.
- **answer:** Correct option for the question. One of A, B, C, D, or E.
- **topic:** Topic of the question. If the question is seed then topic and subject are from the test the question is from. Otherwise, derived question inherits topic and subject of the parent.
- **subject:** Subject of the question, similar to *topic* column.
- **is\_root:** True if given sample is a seed one, false otherwise.
- **tags:** One of 18 image features that are extracted using Gemini 2.5 Flash. While extracting tags, we used self-consistency of 3 where tags present at 2 out of 3 generations are kept. Then each tag is manually reviewed.
- **annotations:** Manually annotated bounding boxes of figures in the question.

10. ABCD parallelenarında  $[AE] \perp [DE]$ ,  $[EB] \perp [BC]$  ve

$$m(\widehat{ADE}) = m(\widehat{EDC}) \text{ tir.}$$



$|BC| = 12 \text{ cm}$  ve  $|BE| = 8 \text{ cm}$  olduğuna göre  $|AB|$  kaç santimetredir?

A) 12      B) 14      C) 15      D) 16      E) 17

<thought>Alright, here's my approach to this geometry problem. I'm focusing on breaking down the given information and finding relationships within the parallelogram ABCD.

First, I identified the key information: the parallelogram properties, the right angles formed by AE and BE, the angle bisector DE, and the given side lengths. My primary goal is to determine the length of AB.

I initially considered using angle properties and extending lines to form triangles, like extending DE to meet AB at a point and considering parallel lines and transversals. However, this didn't immediately lead to a direct solution. I then moved to the angle bisector property within a parallelogram, extending the bisector to AB to form an isosceles triangle and determine if I could find AF.

Next, I considered extending lines to form different shapes. Extending AE and BC to meet, or drawing perpendiculars from D to BC. These did not yield immediate solution.

Then I considered trigonometry with the triangle ADE, but I got stuck with the equations. The law of cosines led me down a complex path, and I wasn't able to get there. I tried drawing a line parallel to AD and EB through E, intersecting DC and AB at G and H.

Then, I focused on the given perpendiculars, but I still couldn't make progress. The next focus was the assumption that AE is the angle bisector of  $\angle DAB$ . This was a key insight because it confirmed a relationship and simplified the problem. I validated this with the property that adjacent angle bisectors in a parallelogram are perpendicular to each other.

With this new information, I extended AE to meet DC, and this gave me two isosceles triangles with properties of equal sides and equal angles. Then I revisited extending DE, but I didn't see a clear path.

I also tried various constructions such as lines parallel to other sides. After getting stuck, I then went back to focusing on the properties of a parallelogram and the relationship between the angle bisector. Extending AE to meet BC at F was my first move. This created an isosceles triangle, but it didn't help. I moved to the alternative to extend AE to G on DC and tried this direction.

Then, I came up with an alternative plan of construction - I drew a line from E to the midpoints of the base. Since E is the right angle with respect to BE, the problem shifted to focusing on triangles and properties of parallelograms to solve the problem. I tried different ways to find a triangle with  $|AB|$  as a leg.

Finally, I drew a line through E parallel to DC and AB and proved that this line goes through the midpoints of the legs AD and BC. Then I created right-angled triangle EBQ to find  $|EQ|$ . Then, by summing, I got my answer! The critical point was identifying that AE is also an angle bisector of  $\angle DAB$  and the use of the midline of the trapezoid. I used properties of parallelograms, parallel lines and alternate interior angles, and the fact that adjacent angle bisectors in a parallelogram are perpendicular.

After careful analysis and considering the properties of parallelograms and right-angled triangles, I have a solution! [TOKEN\_LIMIT]

Figure 11: Example question where model followed correct reasoning; however, it depleted its token budget before arriving at a conclusion.

## E Citing GPT-5

Since technical report of GPT-5 is not available at the time of writing the paper, we cited GPT 4's technical report [18].

## F Data, Code and Prompts

Data, code and prompts can be found at following links:

- **Data:** <https://huggingface.co/datasets/metu-yks/yksbench>
- **Dataset similarity analysis:** <https://colab.research.google.com/drive/1jsh0gpGC5LFMj9v0Wz5YmaWFEup1dPlP?usp=sharing>
- **Solver models along with their prompts:** [https://colab.research.google.com/drive/1wtHukR00xSSCDiHm3E6n\\_P9VHyLEEibZ?usp=sharing](https://colab.research.google.com/drive/1wtHukR00xSSCDiHm3E6n_P9VHyLEEibZ?usp=sharing)

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our main claim is that we include new, challenging, figure modes for multi-modal reasoning. We show the novelty of our samples through UMAP analysis and showcase samples unique to YKSbench using HDBSCAN. Also we showcase that challenging nature of our dataset by benchmarking on top proprietary and open-weight datasets.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Please refer to Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theoretical result available.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Here we share links to reproduce our results and claims. Link to YKSbench, Colab notebook to analyze dataset similarity, Colab notebook to generate/analyze VLM completions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Here we share links to reproduce our results and claims. Link to YKSbench, Colab notebook to analyze dataset similarity, Colab notebook to generate/analyze VLM completions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Code and data are fully open sourced and key parameters are shared.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[NA\]](#)

Justification: We set each VLMs temperature as zero for reproducibility. Only exception is GPT-5 where we could not generate error bars due to budgetary reasons. To ensure reproducibility we relied our error analysis on Gemini 2.5 Pro a more reproducible counterpart of GPT-5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: **[No]**

Justification: We mainly used API services for inference. With any Colab notebook, you can reproduce our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: **[Yes]**

Justification: Code and dataset is anonymized so meet NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: **[Yes]**

Justification: Please refer to Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: To uphold the highest ethical standards, every image sourced for this benchmark underwent a rigorous manual inspection to proactively identify and redact any potentially private information.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The questions comprising YKSbench are sourced from numerous publicly available Turkish University Entrance Exam (YKS) practice books and preparatory materials. These educational materials are intended for widespread public distribution and use for exam preparation; however, they are not typically released under explicit, standardized digital licenses (e.g., Creative Commons). Our compilation of these questions constitutes a new, derivative work created under the principles of fair use for transformative, non-commercial academic research. The purpose of our benchmark is not to redistribute the original content for its primary purpose (exam preparation), but to provide a novel tool for the scientific evaluation of AI models' reasoning capabilities. The YKSbench dataset is being released for academic, non-commercial research purposes only.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](http://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Documentation of code and data can be found on individual links.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: **[Yes]**

Justification: We have used Gemini 2.5 Flash to generate solutions on 60K pool questions. Then used Gemini Embeddings 001 model to get embeddings. Also used Gemini 2.5 Flash to detect image tags. Prompts used for each step is shared in Appendix.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.