

# EYE-Q: A Multilingual Benchmark for Visual Word Puzzle Solving and Image-to-Phrase Reasoning

Anonymous ACL submission

## Abstract

Vision-Language Models (VLMs) have achieved strong performance on standard vision-language benchmarks, yet often rely on surface-level recognition rather than deeper reasoning. We propose visual word puzzles as a challenging alternative, as they require discovering implicit visual cues, generating and revising hypotheses, and mapping perceptual evidence to non-literal concepts in ways that are difficult to solve via literal grounding, OCR-heavy shortcuts, or simple retrieval-style matching. We introduce EYE-Q, a multilingual benchmark designed to assess this form of complex visual understanding. EYE-Q contains 1,343 puzzles in which a model observes a conceptually dense scene with a brief description and must infer a specific target word or phrase. The puzzles are intentionally unstructured and cue-implicit, with distractors and contextual relationships that demand selective attention, abstraction, and associative inference. The benchmark spans English, Persian, Arabic, and cross-lingual puzzles. We evaluate state-of-the-art VLMs using an open-ended, human-aligned protocol that probes hypothesis formation and revision under lightweight assistance. Results reveal substantial performance gaps, especially on abstract and cross-lingual puzzles, highlighting limitations in current models' ability to construct and search over appropriate conceptual representations for flexible image-to-phrase inference; maximum accuracy reaches only 60.27%.

## 1 Introduction

Vision-language models (VLMs) have rapidly evolved into general-purpose assistants that can describe images, answer questions, and follow complex visual instructions (xAI, 2025; Bai et al., 2023; Gemini Team, 2025; OpenAI, 2025; Touvron et al., 2023). As these systems become more capable and widely deployed, benchmarking them

becomes increasingly important. Yet many existing benchmarks rely on questions with explicit, directly queryable evidence, and a large fraction of their content (or close variants) is available on the web or in widely circulated sources. This can let models succeed via surface recognition, shortcut heuristics, or training-data exposure rather than the reasoning we aim to measure (Agrawal et al., 2018; Si et al., 2022; Deng et al., 2024; Li et al., 2024b,a). A natural next step is to push toward evaluations that are harder to solve by appearance matching or retrieval. In particular, a challenging benchmark should be (i) **multimodal** in how it encodes meaning, (ii) **multi-step** in a way that supports hypothesis testing rather than one-shot guessing, and (iii) **multi/cross-lingual**, measuring performance beyond English scripts, while presenting these criteria in a creative and previously unseen manner.

**Multimodality.** A broad suite of multimodal benchmarks probes higher-level cognition beyond perception, including multi-domain understanding (Liu et al., 2024), compositional reasoning (Zhu et al., 2025; Lei et al., 2025), abstract reasoning (Chia et al., 2024; Wüst et al., 2025), and scientific understanding (Ma et al., 2025; Yin et al., 2025). These benchmarks have been valuable, but many tasks still emphasize literal grounding: recognizing entities, reading text, and answering questions with evidence that is localized and directly queryable. Strong performance therefore does not always imply that a model can discover subtle cues in a complex scene or map perceptual evidence to non-literal concepts. This leaves a gap for evaluations where the signal is implicit.

**Multi-step Reasoning.** Many evaluations reward systematic and rule-following reasoning that narrows toward a single correct solution (Cromptley, 2006). This gap appears in puzzles such as rebus-like riddles and picture-word games (Gritsevskiy et al., 2024; Kraaijveld et al., 2025; Khurdula et al., 2025; Chia et al., 2024). Solving these tasks is

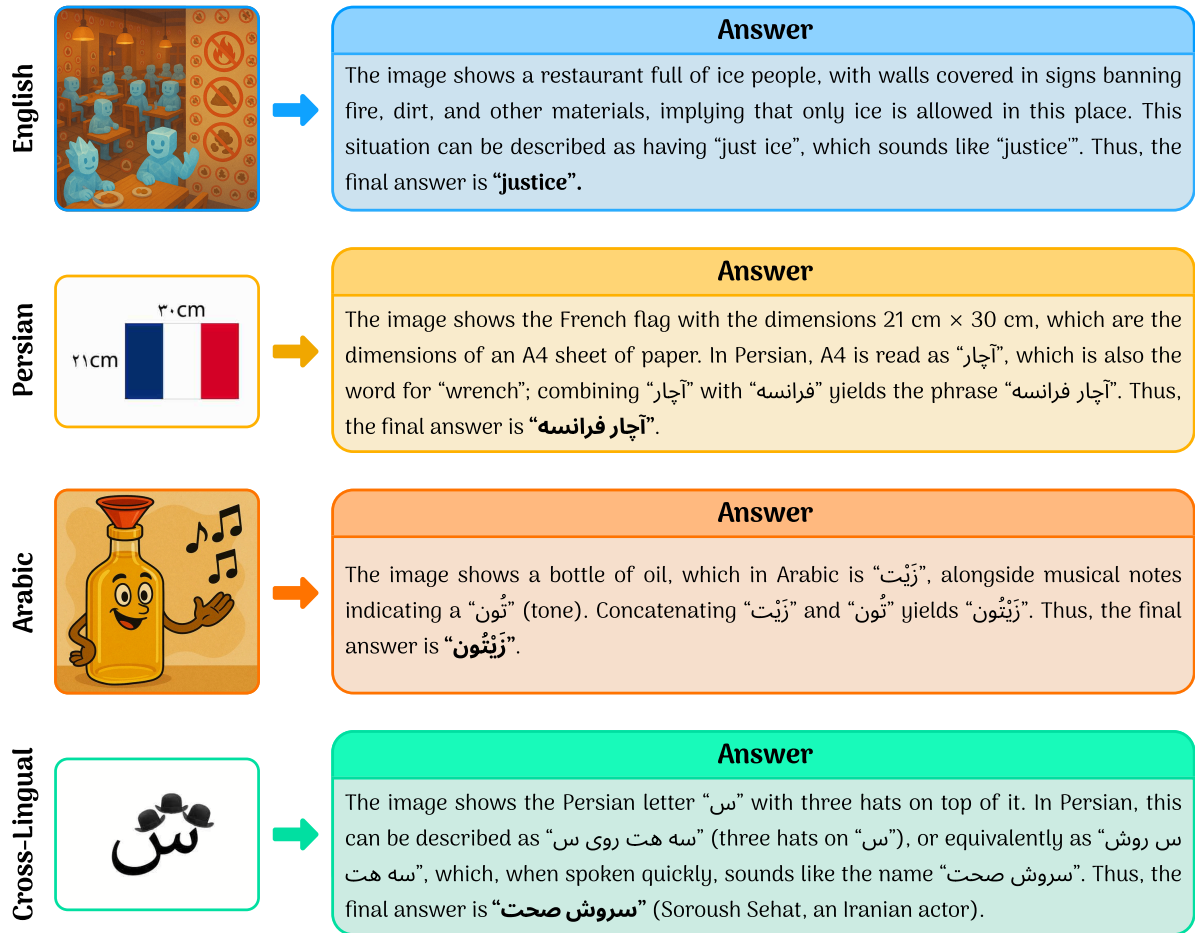


Figure 1: Illustrative EYE-Q examples from the English, Persian, Arabic, and cross-lingual subsets. Each example shows an input image along with a textual derivation that leads to the target word or phrase.

085 abductive and iterative: solvers propose hypothe- 106  
 086 ses about which aspects of the scene are cues, test 107  
 087 them, and revise when they fail. In this setting, 108  
 088 multi-step reasoning alone is insufficient, because 109  
 089 without improving the underlying hypothesis, addi- 110  
 090 tional steps can amplify unproductive search. Re- 111  
 091 cent puzzle-oriented evaluations underscore that 112  
 092 even strong VLMs struggle when success depends 113  
 093 on abstraction, symbolic substitution, and non- 114  
 094 literal image-to-language associations (Khurdula 115  
 095 et al., 2025; Gritsevskiy et al., 2024; Kraaijveld 116  
 096 et al., 2025; Lee et al., 2025; Ren et al., 2025; 117  
 097 Guetta et al., 2024). However, many of these works 118  
 098 rely on visually clean compositions or explicit text, 119  
 099 making success largely a matter of OCR and shal- 120  
 100 low heuristics. Moreover, protocols can under- 121  
 101 measure models when they do not allow hypothesis 122  
 102 revision. This suggests a need for puzzle-style eval- 123  
 103 uations that are visually dense, cue-implicit, and 124  
 104 paired with a human-aligned setup that supports 125  
 105 controlled refinement without changing the task. 126

**Multilinguality.** Multilinguality further com- 106  
 pounds the challenge. While multilingual vision- 107  
 language resources and evaluations have grown 108  
 (Liu et al., 2021; Leng et al., 2025; Fabbri et al., 109  
 2025), puzzle solving remains under-studied be- 110  
 yond English, particularly for non-Latin scripts 111  
 and culturally grounded expressions. Puzzles of- 112  
 ten hinge on phonetic, orthographic, and idiomatic 113  
 mechanisms that do not translate cleanly, thereby 114  
 offering a rigorous test of reasoning across linguis- 115  
 tic contexts. Uniquely, our work introduces the 116  
 first systematic evaluation of visual word puzzles 117  
 with Persian answers and cross-lingual puzzles that 118  
 bridge English visual cues with Persian solutions. 119  
**Our Approach.** We introduce EYE-Q (Figure 1), 120  
 a multilingual benchmark for visual word puzzle 121  
 solving and open-ended image-to-phrase inference. 122  
 The task is designed to require implicit cue dis- 123  
 covery in dense scenes and supports hypothesis re- 124  
 vision via lightweight, human-aligned assistance. 125  
 Even with these supports, performance across pro- 126

proprietary VLMs remains low, especially in cross-lingual settings, highlighting persistent gaps in non-literal vision–language reasoning.

**Contributions.** Our work makes the following contributions:

- **Task Formulation.** We introduce visual word puzzle solving as a vision–language reasoning task requiring multi-modal cue integration, multi-step reasoning and search, and multilingual/cross-lingual generalization.
- **Benchmark Dataset.** We introduce a multilingual benchmark of 1,343 puzzles spanning English, Persian, Arabic, and cross-lingual settings.
- **Cue-Implicit Design.** The puzzles are cue-implicit, unstructured, and conceptually dense, avoiding token-explicit cues (OCR or templated layouts) and requiring abstraction and associative reasoning.
- **Evaluation Protocol.** We propose a human-aligned, open-ended evaluation protocol that probes hypothesis formation and revision through lightweight assistance (answer-length hints, partial character reveals, few-shot examples, iterative refinement).
- **Empirical Findings.** We benchmark state-of-the-art VLMs and reveal substantial performance gaps, especially on abstract and cross-lingual puzzles, highlighting limits in current models’ ability to construct and search over conceptual representations.

## 2 Related Work

### VLM Benchmarks Beyond Literal Grounding.

Vision–language evaluation has long focused on VQA and captioning tasks with explicit, directly grounded evidence (Goyal et al., 2017; Hudson and Manning, 2019; Chen et al., 2015). Newer benchmarks target compositional or multi-step reasoning and more abstract visual understanding (Kil et al., 2024; Zhou et al., 2025; Chia et al., 2024). Yet multiple analyses show that high performance can reflect language priors or dataset shortcuts rather than robust grounding (Yang et al., 2025; Krojer et al., 2025; Chi et al., 2025), motivating evaluations that require implicit cue discovery and non-literal inference.

### Visual Wordplay and Rebus-Style Puzzles.

Puzzle-style benchmarks probe non-literal image-to-phrase mappings. Rebus and visual wordplay datasets encode short phrases via symbolic substitution, layout, and phonetic or semantic composition (Gritsevskiy et al., 2024; Kraaijveld et al., 2025; Khurdula et al., 2025; Das et al., 2025). Relevant work studies visual riddles and puns, including culturally grounded variants (Guetta et al., 2024; Chung et al., 2024; Zhang et al., 2024). However, many settings use clean, token-like cues (text/icons) and analyses report weak hint use and limited hypothesis revision (Lee et al., 2025; Movva, 2025).

### Hinted and Iterative Evaluation Protocols.

A growing literature evaluates reasoning under iterative protocols where models revise answers using feedback or critique. Representative methods include self-refinement and reflection-style loops (Shinn et al., 2023; Madaan et al., 2023) and critique-based verification with tools (Gou et al., 2023). Benchmarks also measure critique-and-correct behavior directly and in multi-turn settings (Wu et al., 2024; Lin et al., 2024; Banatt et al., 2024), while other studies emphasize that self-correction is unreliable without strong verifiers and careful reporting (Huang et al., 2023; Kamoi et al., 2024). Our evaluation uses lightweight assistance to separate cue-discovery failures from output brittleness.

### Multi/Cross-Lingual VLM Evaluation.

Although most benchmarks are English-centric, recent work expands evaluation across languages and cultures through multilingual suites, cross-lingual VQA, and culturally grounded reasoning datasets (Liu et al., 2021; Pfeiffer et al., 2022; Bugliarello et al., 2022; Romero et al., 2024). Complementary efforts broaden coverage for underrepresented regions and non-Latin scripts, including Arabic-focused resources (Barua et al., 2024; Alwajih et al., 2024, 2025). Still, multilingual multimodal puzzle solving remains limited, particularly for culturally grounded wordplay and cross-lingual transformations into Persian.

### Position of EYE-Q.

EYE-Q connects these threads by evaluating open-ended image-to-phrase inference for non-literal puzzles in cue-implicit, distractor-rich scenes, and by extending the setting to English, Persian, Arabic, and cross-lingual puzzles.

Subset	# Puzzles	Source
English	300	Our synthetic puzzles
Persian	671	<i>Aftabe</i> game
Arabic	50	Our synthetic puzzles
Cross-lingual	322	<i>Aftabe</i> game
Total	1,343	–

Table 1: Basic statistics for EYE-Q.

### 3 EYE-Q

This section introduces EYE-Q and the design principles behind its puzzles. We first summarize the benchmark composition and instance format, then describe puzzle design and the data sources used for the monolingual and cross-lingual subsets.

#### 3.1 Benchmark Overview

EYE-Q is a multimodal benchmark designed to evaluate open-ended image-to-phrase reasoning under cue-implicit visual settings and multilingual wordplay. The benchmark contains 1,343 picture puzzles organized into four subsets: English, Persian, Arabic, and cross-lingual (Table 1).

Each puzzle consists of an image and a short textual prompt that specifies the game rule (i.e., the image encodes a hidden target word or short phrase) and requests a single final answer. The task is open-ended: models must generate the target word/phrase rather than select from candidates.

The images are intentionally cue-implicit and conceptually dense: scenes contain multiple objects, attributes, and relationships that may function as semantic cues, often alongside distractors. Cues can involve object orientation, item counts, color, relative size, spatial relations (e.g., containment, overlap, inversion), material properties, posture, and affect. Text may appear in some images, but EYE-Q is designed to be anti-OCR: solving rarely reduces to simply reading visible text.

#### 3.2 Puzzle Design

The core of EYE-Q lies in requiring a semantic leap from literal recognition to a non-literal linguistic target. Unlike standard VQA tasks that ask “What is in the image?”, our puzzles ask “What does this scene imply or represent?”. Solving typically requires: (i) *cue discovery* (identifying which elements are informative versus distractors), (ii) *relational abstraction* (reasoning over relations and transformations rather than isolated objects), and

(iii) *linguistic association* (mapping the inferred concept to an idiom, pun, phonetic resemblance, or conventional phrase in the target language).

Puzzles range from monolingual wordplay to cross-lingual reasoning. In cross-lingual instances, cues naturally suggest words or concepts in one language, while the correct answer is expressed in the other, requiring phonetic, orthographic, or semantic bridging between English and Persian. Figure 1 provides representative examples.

#### 3.3 Data Collection and Ground Truth

EYE-Q is built from two sources: (i) the mobile game *Aftabe* (Appendix A), and (ii) English and Arabic puzzles designed by the authors and rendered into images using text-to-image generation models. We have obtained permission from *Aftabe* to use and redistribute the curated puzzles as part of EYE-Q.

**Persian and Cross-Lingual Puzzles.** We curated Persian and cross-lingual puzzles from *Aftabe*, which contains visual puzzles with Persian answers. For each selected instance, we preserve the original image and the game-provided intended solution as ground truth. Because puzzles can admit multiple interpretations, we manually filter out items that are ambiguous or under-specified, retaining puzzles with a clear intended reading and a unique target answer. We then categorize each retained puzzle as Persian or cross-lingual based on whether solving requires an English–Persian bridge.

**English and Arabic Puzzles.** We build the English and Arabic subsets by first selecting a target word or short phrase and writing a concise scene description intended to lead a human solver to that target. We then render the scene using text-to-image models (GPT (OpenAI, 2025) and Nano Banana (Raisinghani, 2025)). All generated images are manually reviewed, and we retain only those that clearly support the intended solution and do not contain rendering artifacts or alternative plausible interpretations. Because the target answer is fixed during design and each image is filtered for clarity, these instances have controlled ground-truth labels that are verified to be consistent with the visual content.

**Evaluation Hints.** Lightweight assistance signals such as answer-length hints and partial character reveal are not separate annotation fields in

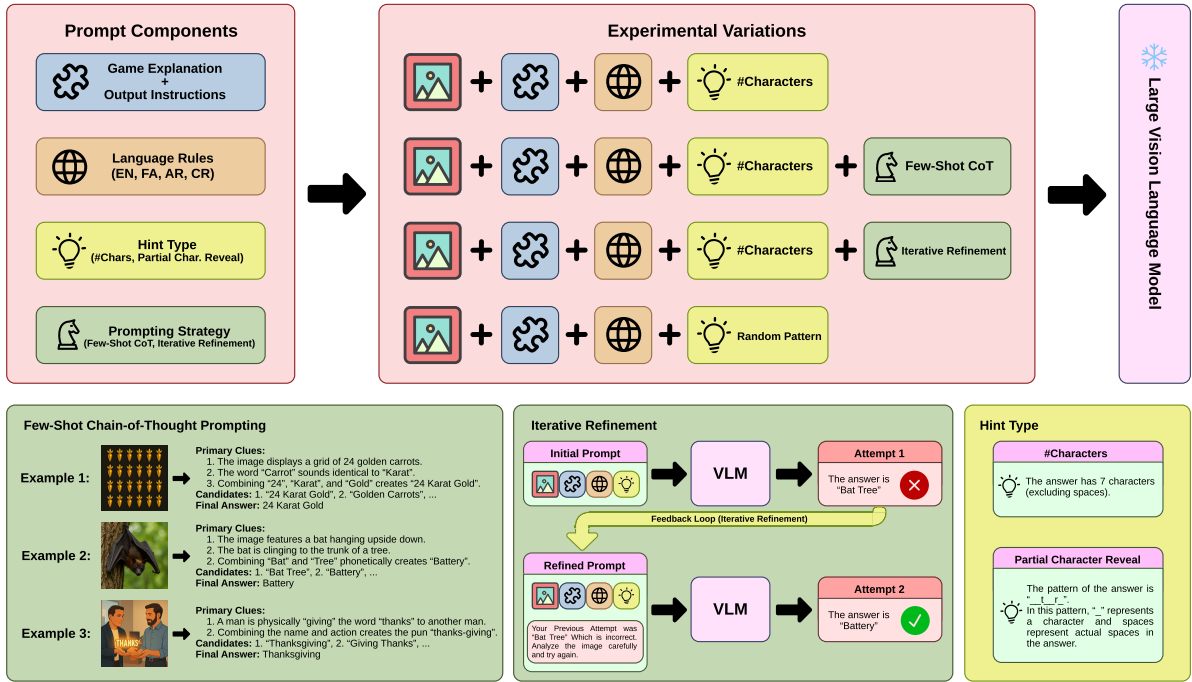


Figure 2: Overview of experimental variations in EYE-Q. We instantiate prompts by combining a shared base template, game explanation and subset-specific language rules, with optional hinting and prompting strategy modules.

EYE-Q. However, they are directly derivable from the ground-truth answer (e.g., character count or revealing a fixed fraction of characters), so any user of the dataset can apply the same hinting schemes. In our experiments (Section 4), we use these derived hints as controlled evaluation variants to provide human-aligned assistance without changing the underlying task.

## 4 Experiments

This section evaluates vision–language models on EYE-Q. We compare multiple models under four prompt variants and report accuracy across the benchmark’s language subsets. We include both proprietary and open-source LVLMs. Our goal is to measure open-ended puzzle solving under a human-aligned setup that supports hypothesis formation and revision, while keeping the underlying task unchanged.

### 4.1 Models

We evaluate a set of proprietary and open-source LVLMs: GPT-5.2 (OpenAI, 2025), Gemini 2.5 Flash (Gemini Team, 2025), Gemini 2.5 Pro (Gemini Team, 2025), Grok 4.1 Fast reasoning (xAI, 2025), Llama 4 Scout (Touvron et al., 2023), and Qwen 3 VL (Bai et al., 2023). All experiments use the default inference configu-

ration provided by each model endpoint or release, including each model’s default decoding strategy and default visual preprocessing.

### 4.2 Prompt Design and Input Format

For each puzzle, we query the model with a single image and a text prompt instantiated from a fixed base template, shown in Figure 2. The base template includes: (i) a brief explanation of the game objective, and (ii) subset-specific instructions for English, Persian, Arabic, or cross-lingual puzzles. Across all experiments, we keep the base instructions and output format consistent and vary only the experiment-specific modules described below. We provide the exact prompts used for the experiments in Appendix D.

**Decoding and scoring.** All main results reported in the paper use each model’s default decoding configuration rather than enforcing a shared decoding strategy such as greedy or fixed-temperature sampling. We report exact-match accuracy against the ground-truth answer, using the model’s final answer string with only minimal cleanup, specifically trimming whitespace and removing Persian and Arabic diacritics (A’rab).

**Experimental Variations.** Using the shared base template, we evaluate four prompting vari-

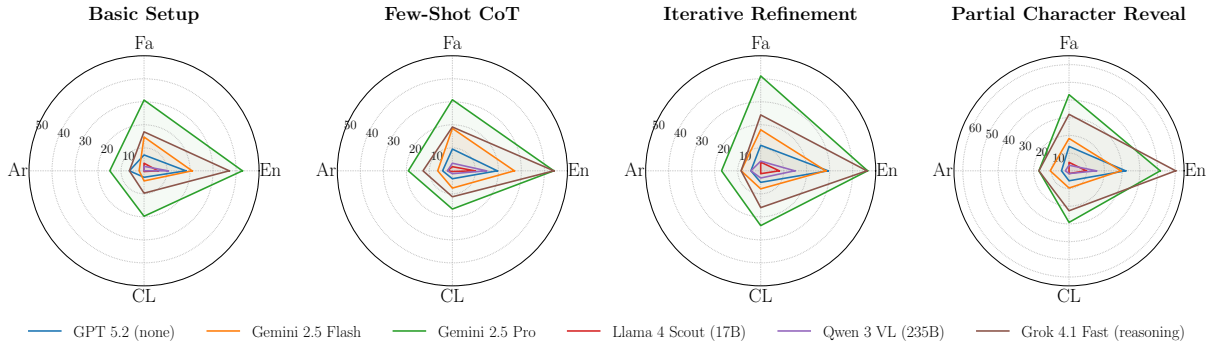


Figure 3: Accuracy(%) of six LVLMs across four prompt variants and four language subsets.

ants:

1. **Basic.** We append an orthographic hint indicating the target answer length in characters to narrow the search space, and ask the model to output a single final answer.
2. **Few-Shot Chain-of-Thought.** We prepend three solved examples to the prompt. Each demonstration includes an image, the correct answer, and a concise description of the intended image–answer relation. Demonstrations are sampled from the same subset pool as the test puzzle, English, Persian, Arabic, or cross-lingual, and exclude the test instance. For fair comparison, the selected demonstration set is fixed across models for each evaluated sample.
3. **Iterative Refinement.** We first query the model using the basic prompt. If the answer is incorrect, we append the model’s previous guess and a short instruction to revise, then requery. We allow up to two such revisions, for a total of three attempts. A puzzle is counted as correct if any attempt matches the ground truth.
4. **Partial Character Reveal.** We reveal a randomly selected 25% subset of the ground-truth answer’s non-space characters and mask the remainder, for example with underscores, then ask the model to output the full answer. The revealed positions are sampled with a fixed random seed to ensure reproducibility.

Across all variants, the input image is unchanged and the core task remains open-ended image-to-phrase inference; only the prompting strategy and lightweight hinting described above are varied.

## 5 Results

Figure 3 reports accuracy for six LVLMs across the four language subsets and four prompt variants described in Section 4.2, following the evaluation protocol in Section 4. Each value is the normalized exact-match accuracy over all puzzles in a subset, where normalization applies minimal surface-form cleanup (e.g., trimming punctuation/whitespace artifacts and lowercasing when applicable). Exact numerical values for all model–subset–prompt combinations are provided in Appendix B.1. We additionally report controlled temperature sweeps in Appendix B.2.

### 5.1 Overall Performance

Across all models and prompt variants, accuracy remains far from high, indicating that EYE-Q poses a challenging open-ended image-to-phrase inference problem. Even the best-performing models struggle substantially on the harder subsets. Considering the best result achieved on each subset across all evaluated models and prompt variants, the peak accuracy is 60.27% on English (Grok 4.1 Fast, Partial Character Reveal), 43.03% on Persian (Gemini 2.5 Pro, Partial Character Reveal), 19.15% on Arabic (Gemini 2.5 Pro, Few-Shot CoT), and 29.15% on cross-lingual puzzles (Gemini 2.5 Pro, Partial Character Reveal). Thus, while English can reach moderate performance under supportive prompting, Arabic and cross-lingual puzzle solving remain particularly difficult even for the strongest models.

Performance differences across model families are also clear. Proprietary models (GPT-5.2, Gemini 2.5, Grok 4.1) dominate overall, while open-source models (Llama 4 Scout, Qwen 3 VL) remain low, including near-zero performance on Arabic and cross-lingual subsets in multiple settings.

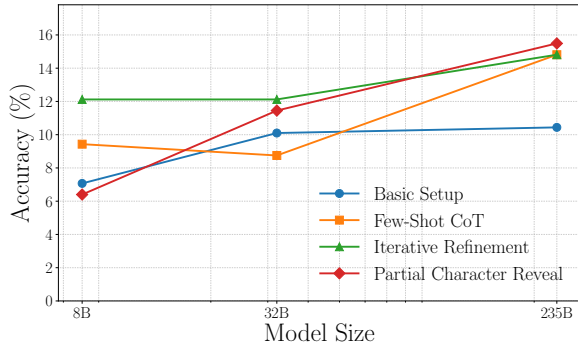


Figure 4: English-subset accuracy versus model size for the Qwen3-VL family (8B, 32B, 235B-A22B) under four prompting variants.

This highlights persistent weaknesses in multilingual cue discovery and non-literal phrase inference.

## 5.2 Assistance Effects

Prompt variants provide lightweight, human-aligned assistance, but they do not close the gap. Averaged across models and language subsets, iterative refinement improves accuracy over the basic setup from 11.53% to 15.63%, suggesting that allowing hypothesis revision can recover from some initial misreadings. Partial character reveal improves further to 18.11% on average.

These assistance variants also directly reduce a common concern in open-ended evaluation: that low exact-match accuracy might be dominated by output-space brittleness (large lexical search space or superficial near-misses). Partial character reveal constrains the output space without supplying answer candidates, and refinement gives models multiple chances to correct small mistakes without turning the task into candidate verification. Despite this, accuracy remains low even under supportive settings: for example, Arabic stays below 10% on average even with partial character reveal, and cross-lingual performance remains around 12% on average. This persistence suggests that the dominant failure mode is upstream of surface form: models often fail to identify which elements in a complex, distractor-rich scene constitute the right cues and to construct the correct abstract interpretation linking the scene to the intended hidden phrase. In this sense, EYE-Q stresses a different bottleneck than OCR-heavy or visually clean rebus-style settings, where reading explicit tokens and combining them compositionally can account for much of the solution.

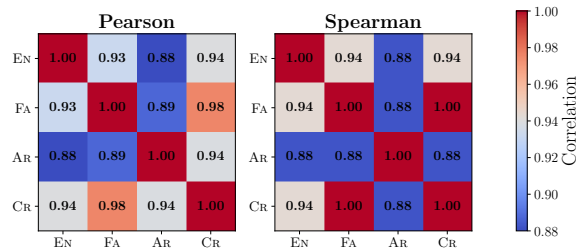


Figure 5: Correlation of model accuracies across language subsets in the Basic setup. Pearson (left) and Spearman (right) correlations are computed across models using subset-level accuracies for English, Persian, Arabic, and cross-lingual puzzles.

## 5.3 Model Size Effect

To probe the effect of scale in a controlled, same-family setting, Figure 4 plots accuracy as a function of model size for Qwen3-VL on the English subset under all four prompt variants (Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct, Qwen3-VL-235B-A22B-Instruct). Larger models consistently perform better, and the gains tend to be most visible under assistance variants that encourage revision or constrain the output space. However, even at the largest size, accuracy remains far from saturated, reinforcing that simply scaling parameters is insufficient to overcome the core challenge posed by EYE-Q: cue discovery and abstraction in complex, cue-implicit scenes.

Results for the same Qwen3-VL scaling analysis on Persian, Arabic, and cross-lingual subsets are provided in Appendix B.3.

## 5.4 Language Coupling

To test whether model strengths transfer across languages, we analyze language coupling by correlating model accuracies between language subsets in the basic setup. Figure 5 reports Pearson and Spearman correlations computed across models using subset-level accuracies. Correlations are high overall, indicating that relative model ranking is largely consistent across languages.

The tightest coupling is between Persian and cross-lingual performance (Pearson 0.98, Spearman 1.00), suggesting that models strong on Persian puzzles tend also to be strong on cross-lingual puzzles that require bridging Persian and English. In contrast, Arabic shows weaker coupling with other subsets (with correlations around 0.88 in multiple pairings), consistent with Arabic being the most challenging setting for current models and

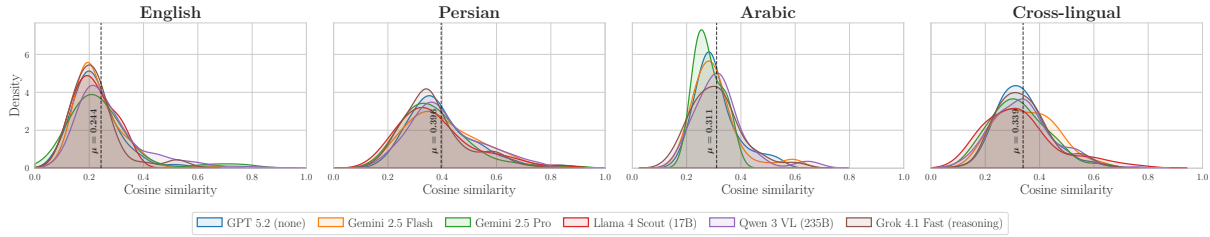


Figure 6: Density estimates of cosine similarity between model outputs and ground-truth answers under iterative refinement. Similarities are computed between each model’s final refined answer and the ground-truth answer using OpenAI `text-embedding-3-large` embeddings (OpenAI, 2024). Curves are estimated over failure cases only and shown across English, Persian, Arabic, and cross-lingual subsets; higher values indicate closer semantic alignment. The concentration of mass at low similarity suggests that many errors are not near-miss paraphrases.

508 potentially requiring additional robustness beyond  
509 what drives performance in English and Persian.

### 5.5 Semantic Near-Miss Analysis

511 Exact-match evaluation can in principle underes-  
512 timate performance if models frequently produce  
513 answers that are semantically close to the ground  
514 truth but differ in surface form. To test this, we  
515 measure semantic proximity between model out-  
516 puts and ground-truth answers using cosine sim-  
517 ilarity in an embedding space. Specifically, we  
518 compute cosine similarity between embeddings of  
519 the model’s final answer under iterative refinement  
520 and the ground-truth answer. We compute this sim-  
521 ilarity over failure cases only and visualize den-  
522 sity estimates across language subsets in Figure 6. Ad-  
523 ditional plots for other prompting variants show a  
524 similar trend and are provided in Appendix B.5.

525 The density estimates are concentrated at low  
526 similarity values across all four language settings,  
527 indicating that incorrect outputs are typically not  
528 semantically close to the target phrase. This sug-  
529 gests that low accuracy on EYE-Q is not primarily  
530 driven by minor linguistic differences or near-miss  
531 paraphrases penalized by strict string matching. In-  
532 stead, the dominant failure mode appears to arise  
533 earlier: models often fail to identify the right cues  
534 amid distractors and to form the correct abstract  
535 interpretation linking the complex scene to the in-  
536 tended hidden phrase.

## 6 Conclusion

538 We introduced EYE-Q, a multilingual benchmark  
539 for visual word puzzle solving and open-ended  
540 image-to-phrase inference. EYE-Q contains 1,343  
541 cue-implicit puzzles spanning English, Persian,  
542 Arabic, and cross-lingual Persian–English settings.  
543 Unlike prior puzzle-style evaluations that often

544 center on visually clean rebus compositions or ex-  
545 plicit text cues, EYE-Q emphasizes visually dense  
546 scenes with distractors, where success requires dis-  
547 covering which elements function as clues and  
548 composing a non-literal conceptual interpretation.

549 Across six state-of-the-art LVLMs, performance  
550 remains low. Even with human-aligned assistance  
551 such as few-shot CoT, iterative refinement, and par-  
552 tial character reveal, models continue to struggle,  
553 particularly on Arabic and cross-lingual subsets. A  
554 semantic similarity analysis of failures further sug-  
555 gests that errors are typically not near-miss para-  
556 phrases, reinforcing that the bottleneck is not only  
557 output-space brittleness but also cue selection and  
558 abstraction. We hope EYE-Q will provide a useful  
559 testbed for developing and measuring multimodal  
560 systems that better support hypothesis generation,  
561 revision, and multilingual reasoning beyond literal  
562 grounding.

## Limitations

563  
564 First, on rare occasions where alternative inter-  
565 pretations exist, EYE-Q prioritizes the intended  
566 ground truth to maintain rigorous evaluation met-  
567 rics.

568 Second, dataset generation relies on contribu-  
569 tors who are fluent in the target languages. This  
570 can make it harder to scale the benchmark to ad-  
571 ditional languages, dialects, or low-resource vari-  
572 eties while keeping difficulty and style consistent.

573 Third, reliable human evaluation is also non-  
574 trivial. To assess correctness fairly across lan-  
575 guages, each language subset should ideally be  
576 evaluated by native speakers or annotators with  
577 comparable proficiency. This increases coordina-  
578 tion effort and cost, and it can introduce variability  
579 between annotator groups, especially when border-  
580 line cases depend on subtle linguistic judgments.

581	If these challenges are addressed, the benchmark’s coverage is likely to improve, and its evaluations to become more reliable and reproducible.	
582		
583		
584	<b>Acknowledgments</b>	
585	We thank the developers of the <i>Aftabe</i> mobile game for creating an engaging collection of visual word puzzles that inspired and enabled the Persian and cross-lingual portions of EYE-Q.	
586		
587		
588		
589	<b>References</b>	
590	Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. <a href="#">Don’t just assume; look and answer: Overcoming priors for visual question answering</a> . In <i>2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018</i> , pages 4971–4980. Computer Vision Foundation / IEEE Computer Society.	
591		
592		
593		
594		
595		
596		
597		
598	Fakhraddin Alwajih, S. Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Abdelfadil, Abdulfattah Mohammed Yahya, H. Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmeh, Houdaifa Atou, Serry Sibae, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged Saeed Al-shaibani, Aya El aatar, Nour Qandos, Rahaf Alhamouri, Samar Ahmad, Razan Khassib, Lina Hamad, Mohammed Anwar Al-Ghrawi, Fatimah Alshamari, C. Malainine, D. Qawasmeh, Aminetou Yacoub, Tfeil moilid, Ruwa AbuHweidi, Ahmed Aboeitta, Vatimetou Mohamed Lemin, Reem Abdel-Salam, Ahlam Bashiti, A. Ammar, Aisha Alansari, Ahmed Ashraf, N. Alturayef, Sara Shatnawi, Alcides Alcoba Inciarte, AbdelRahim Elmadany, Mohamedou Cheikh Tourad, Ismail Berrada, Mustafa Jarar, Shady Shehata, and M. Abdul-Mageed. 2025. <a href="#">Pearl: A multimodal culturally-aware arabic instruction dataset</a> . <i>ArXiv</i> , abs/2505.21979.	
599		
600		
601		
602		
603		
604		
605		
606		
607		
608		
609		
610		
611		
612		
613		
614		
615		
616		
617	Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and M. Abdul-Mageed. 2024. <a href="#">Peacock: A family of arabic multimodal large language models and benchmarks</a> . In <i>Annual Meeting of the Association for Computational Linguistics</i> .	
618		
619		
620		
621		
622		
623	Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. <a href="#">Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond</a> . <i>Preprint</i> , arXiv:2308.12966.	
624		
625		
626		
627		
628		
629	Eryk Banatt, Jonathan Cheng, Skanda Vaidyanath, and Tiffany Hwu. 2024. <a href="#">Wilt: A multi-turn, memorization-robust inductive logic benchmark for llms</a> . <i>ArXiv</i> , abs/2410.10998.	
630		
631		
632		
	Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Farhan Ishmam, Fabiha Haider, Faraha Tanjim Shifat, Md Fahim, and Md Farhad Alam. 2024. <a href="#">Chitrojera: A regionally relevant visual question answering dataset for bangla</a> . <i>ArXiv</i> , abs/2410.14991.	633 634 635 636 637 638
	Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, E. Ponti, and Ivan Vulic. 2022. <a href="#">Iglue: A benchmark for transfer learning across modalities, tasks, and languages</a> . <i>ArXiv</i> , abs/2201.11732.	639 640 641 642 643
	Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. In <i>arXiv preprint arXiv:1504.00325</i> .	644 645 646 647 648
	Ziheng Chi, Yifan Hou, Chenxi Pang, Shaobo Cui, Mubashara Akhtar, and Mrinmaya Sachan. 2025. <a href="#">Chimera: Diagnosing shortcut learning in visual-language understanding</a> . <i>ArXiv</i> , abs/2509.22437.	649 650 651 652
	Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. 2024. <a href="#">PuzzleVQA: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns</a> . In <i>Findings of the Association for Computational Linguistics: ACL 2024</i> , pages 16259–16273, Bangkok, Thailand. Association for Computational Linguistics.	653 654 655 656 657 658 659
	Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, and Youngjae Yu. 2024. <a href="#">Can visual language models resolve textual ambiguity with visual cues? let visual puns tell you!</a> In <i>Conference on Empirical Methods in Natural Language Processing</i> .	660 661 662 663 664
	Arthur Cropley. 2006. <a href="#">In praise of convergent thinking</a> . <i>Creativity Research Journal - CREATIVITY RES J</i> , 18:391–404.	665 666 667
	Trishanu Das, Abhilash Nandy, Khush Bajaj, and S. Deepiha. 2025. <a href="#">Rebus bus: A large and diverse multimodal benchmark for evaluating the ability of vision-language models to understand rebus puzzles</a> . <i>ArXiv preprint</i> .	668 669 670 671 672
	Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. <a href="#">Investigating data contamination in modern benchmarks for large language models</a> . In <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.	673 674 675 676 677 678 679 680 681
	Alexander R. Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee, Bing Liu, and Chen Xing. 2025. <a href="#">Multinrc: A challenging and native multilingual reasoning evaluation benchmark for llms</a> . <i>ArXiv</i> , abs/2507.17476.	682 683 684 685 686
	Gemini Team. 2025. <a href="#">Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long</a>	687 688

689	context, and next generation agentic capabilities.	Benno Krojer, Mojtaba Komeili, Candace Ross,	743
690	<i>Preprint</i> , arXiv:2507.06261.	Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and	744
691	Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen,	Mahmoud Assran. 2025. <a href="#">A shortcut-aware video-qa</a>	745
692	Yujiu Yang, Nan Duan, and Weizhu Chen. 2023.	<a href="#">benchmark for physical understanding via minimal</a>	746
693	<a href="#">Critic: Large language models can self-correct with</a>	<a href="#">video pairs</a> . <i>ArXiv</i> , abs/2506.09987.	747
694	<a href="#">tool-interactive critiquing</a> . <i>ArXiv</i> , abs/2305.11738.		
695	Yash Goyal, Tushar Khot, Douglas Summers-Stay,	Heekyung Lee, Jiaxin Ge, Tsung-Han Wu, Minwoo	748
696	Dhruv Batra, and Devi Parikh. 2017. Making the	Kang, Trevor Darrell, and David M. Chan. 2025.	749
697	v in VQA matter: Elevating the role of image under-	<a href="#">Puzzled by puzzles: When vision-language models</a>	750
698	standing in visual question answering. In <i>CVPR</i> .	<a href="#">can't take a hint</a> . In <i>Proceedings of the 2025 Con-</i>	751
699	Andrew Gritsevskiy, Arjun Panickssery, Aaron Kirt-	<i>ference on Empirical Methods in Natural Language</i>	752
700	land, Derik Kauffman, Hans Gundlach, Irina Grit-	<i>Processing</i> , pages 21735–21748, Suzhou, China. As-	753
701	sevskaya, Joe Cavanagh, Jonathan Chiang, Lydia La	sociation for Computational Linguistics.	754
702	Roux, and Michelle Hung. 2024. <a href="#">Rebus: A ro-</a>		
703	<a href="#">bust evaluation benchmark of understanding sym-</a>	Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li,	755
704	<a href="#">bols</a> . <i>Preprint</i> , arXiv:2401.05604.	and Yang Liu. 2025. <a href="#">Scaffolding coordinates to</a>	756
705	Nitzan Bitton Guetta, Aviv Slobodkin, Aviya Maimon,	<a href="#">promote vision-language coordination in large multi-</a>	757
706	Eliya Habba, Royi Rassin, Yonatan Bitton, Idan	<a href="#">modal models</a> . In <i>Proceedings of the 31st Inter-</i>	758
707	Szpektor, Amir Globerson, and Yuval Elovici. 2024.	<i>national Conference on Computational Linguistics</i> ,	759
708	<a href="#">Visual riddles: a commonsense and world knowl-</a>	pages 2886–2903, Abu Dhabi, UAE. Association for	760
709	<a href="#">edge challenge for large vision and language models</a> .	Computational Linguistics.	761
710	In <i>The Thirty-eight Conference on Neural Informa-</i>		
711	<i>tion Processing Systems Datasets and Benchmarks</i>	Jixuan Leng, Chengsong Huang, Langlin Huang,	762
712	<i>Track</i> .	Bill Yuchen Lin, William W. Cohen, Haohan	763
713	Jie Huang, Xinyun Chen, Swaroop Mishra,	Wang, and Jiaxin Huang. 2025. <a href="#">Crosswordbench:</a>	764
714	Huaixiu Steven Zheng, Adams Wei Yu, Xiny-	<a href="#">Evaluating the reasoning capabilities of llms and</a>	765
715	ing Song, and Denny Zhou. 2023. <a href="#">Large language</a>	<a href="#">l1lms with controllable puzzle generation</a> . <i>ArXiv</i> ,	766
716	<a href="#">models cannot self-correct reasoning yet</a> . <i>ArXiv</i> ,	abs/2504.00043.	767
717	abs/2310.01798.		
718	Drew Hudson and Christopher D. Manning. 2019. Gqa:	Yucheng Li, Frank Guerin, and Chenghua Lin. 2024a.	768
719	A new dataset for real-world visual reasoning and	<a href="#">Latesteval: Addressing data contamination in lan-</a>	769
720	compositional question answering. In <i>CVPR</i> .	<a href="#">guage model evaluation through dynamic and time-</a>	770
721	Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han,	<a href="#">sensitive test construction</a> . In <i>Thirty-Eighth AAAI</i>	771
722	and Rui Zhang. 2024. <a href="#">When can llms actually cor-</a>	<i>Conference on Artificial Intelligence, AAAI 2024,</i>	772
723	<a href="#">rect their own mistakes? a critical survey of self-</a>	<i>Thirty-Sixth Conference on Innovative Applications</i>	773
724	<a href="#">correction of llms</a> . <i>Transactions of the Association</i>	<i>of Artificial Intelligence, IAAI 2024, Fourteenth Sym-</i>	774
725	<a href="#">for Computational Linguistics</a> , 12:1417–1440.	<i>posium on Educational Advances in Artificial Intelli-</i>	775
726	Harsha Vardhan Khurdula, Basem Rizk, and Indus	<i>gence, EAAI 2014, February 20-27, 2024, Vancou-</i>	776
727	Khaitan. 2025. <a href="#">Beyond visual understanding intro-</a>	<i>ver, Canada, pages 18600–18607</i> . AAAI Press.	777
728	<a href="#">ducing PARROT-360V for vision language model</a>		
729	<a href="#">benchmarking</a> . In <i>Proceedings of the 31st Interna-</i>	Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua	778
730	<i>national Conference on Computational Linguistics: In-</i>	Lin. 2024b. <a href="#">An open-source data contamination re-</a>	779
731	<i>dustry Track</i> , pages 68–75, Abu Dhabi, UAE. Asso-	<a href="#">port for large language models</a> . In <i>Findings of the</i>	780
732	ciation for Computational Linguistics.	<i>Association for Computational Linguistics: EMNLP</i>	781
733	Jihyung Kil, Farideh Tavazoei, Dongyeop Kang, and	2024, pages 528–541, Miami, Florida, USA. Associ-	782
734	Joo-Kyung Kim. 2024. <a href="#">Ii-mm: Identifying and im-</a>	ation for Computational Linguistics.	783
735	<a href="#">proving multi-modal multi-hop reasoning in visual</a>		
736	<a href="#">question answering</a> . In <i>Annual Meeting of the Asso-</i>	Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo,	784
737	<a href="#">ciation for Computational Linguistics</a> .	Haowei Liu, and Yujiu Yang. 2024. <a href="#">Criticbench:</a>	785
738	Koen Kraaijveld, Yifan Jiang, Kaixin Ma, and Filip	<a href="#">Benchmarking llms for critique-correct reasoning</a> .	786
739	Ilievski. 2025. <a href="#">Columbus: Evaluating cognitive lat-</a>	<i>ArXiv</i> , abs/2402.14809.	787
740	<a href="#">eral understanding through multiple-choice rebuses</a> .	Alisa Liu, Gabriele Bugliarello, Edoardo M. Ponti, and	788
741	<i>Proceedings of the AAAI Conference on Artificial In-</i>	et al. 2021. <a href="#">Marvl: A multilingual and multicultural</a>	789
742	<i>telligence</i> , 39(4):4410–4418.	<a href="#">visual reasoning dataset</a> . In <i>NeurIPS</i> .	790
		Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li,	791
		Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi	792
		Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua	793
		Lin. 2024. <a href="#">Mmbench: Is your multi-modal model an</a>	794
		<a href="#">all-around player?</a> <i>Preprint</i> , arXiv:2307.06281.	795
		Jizheng Ma, Xiaofei Zhou, Yanlong Song, and Han	796
		Yan. 2025. <a href="#">Cocova: Chain of continuous vision-</a>	797
		<a href="#">language thought for latent space reasoning</a> . <i>ArXiv</i> ,	798
		abs/2511.02360.	799



914 need for video generalization: A counterfactual  
915 benchmark with sub-question evaluation. *ArXiv*,  
916 abs/2503.10691.

917 Yinglun Zhu, Jiancheng Zhang, and Fuzhi Tang.  
918 2025. Test-time matching: Unlocking composi-  
919 tional reasoning in multimodal models. *Preprint*,  
920 arXiv:2510.07632.

## A Aftabe Mobile Game

The Persian and cross-lingual subsets of EYE-Q are sourced from *Aftabe*, a popular Iranian puzzle game released in 2014. The game achieved viral success in Iran due to its creative use of Persian wordplay and abstract visual riddles. It is available on Google Play at: <https://play.google.com/store/apps/details?id=com.freeapks.aftabe>.

## B Additional Results

This appendix reports additional results that complement the main experiments and provide further context for the observed trends.

### B.1 Overall accuracy

For completeness, we report the exact accuracies corresponding to Figure 3 in Table 2.

### B.2 Temperature Ablation

Table 3 shows that performance is strongly and consistently sensitive to decoding temperature, with clear differences across models. For the most capable model, Gemini 2.5 Pro, accuracy is relatively stable at low temperature but declines steadily as  $T$  increases. This monotonic degradation matches the intuition that, in a constrained exact-match evaluation, additional sampling noise mainly perturbs reasoning traces that would otherwise remain correct. Grok 4.1 Fast follows the same overall trend from  $T=0.01$  to  $T=1.0$ , but drops sharply at  $T=2.0$ , suggesting a distinct failure regime once stochasticity becomes large.

By contrast, several weaker or more generation-sensitive models exhibit an inverted-U pattern: moderate randomness improves results, while higher temperature is harmful. Gemini 2.5 Flash and Qwen 3 VL reach their best performance at  $T=1.0$ , and GPT-5.2 displays a milder version of the same effect. Llama 4 Scout gains slightly at low temperature but breaks down completely at very high  $T=5.0$ . Overall, these findings suggest that modest stochasticity can help less reliable models recover from early incorrect decisions, whereas stronger reasoning-focused models are best evaluated under near-deterministic greedy decoding.

### B.3 Model Size Effect

**Attempt Scaling with Model Size.** Beyond final accuracy, we analyze how model size affects the process of iterative reasoning by measuring the

number of refinement attempts used under a five-attempt iterative refinement setting on the English subset only. Figure 7 reports two complementary statistics as a function of model size: (i) the average number of attempts conditioned on solved puzzles only, and (ii) the average number of attempts over all puzzles, where unsolved instances are assigned the maximum of five attempts.

When conditioning on solved puzzles, the average number of attempts exhibits a non-monotonic trend, it first decreases from 8B to 32B models, then increases for the largest 235B model. The initial decrease suggests that moderately larger models can identify the correct interpretation more quickly, requiring fewer revision steps to reach a valid solution. The subsequent increase at the largest scale reflects a different regime where larger models appear able to sustain longer chains of productive reasoning, using additional refinement steps to recover from partial hypotheses and eventually converge on the correct answer. In contrast, when averaging over all puzzles, the mean number of attempts decreases monotonically with model size. This trend is driven by the sharp reduction in unsolved puzzles at larger scales, which otherwise dominate the mean by exhausting all five attempts.

**Accuracy Scaling with Model Size.** Figure 8 extends the controlled scaling analysis of the Qwen3-VL family to all language subsets. We plot accuracy versus model size for Qwen3-VL-8B, 32B, and 235B under the same four prompt variants.

Across subsets, scaling generally improves performance, but the magnitude and stability of gains depend on both language and prompting. The English subset shows the most consistent and monotonic improvement with size across variants, with particularly clear gains under iterative refinement and partial character reveal, suggesting that larger models benefit more from settings that enable hypothesis revision or restrict the output space. Persian and cross-lingual subsets also trend upward with size, though improvements are smaller and more variant-dependent, indicating that cross-lingual bridging and non-Latin orthography remain challenging even for larger models. Arabic exhibits the weakest and least consistent scaling behavior. The accuracy remains low overall and improvements with size are modest, indicating that the dominant bottleneck is not simply capacity.

Overall, these results suggest that increasing pa-

Model	Basic setup				Few-Shot CoT				Iterative Refinement				Partial Character Reveal			
	En	Fa	Ar	CL	En	Fa	Ar	CL	En	Fa	Ar	CL	En	Fa	Ar	CL
GPT 5.2 (none)	18.18	6.90	<u>6.38</u>	2.82	19.53	9.45	4.26	3.45	29.29	11.09	4.26	5.02	31.99	13.79	4.26	5.64
Gemini 2.5 Flash	20.88	14.69	2.13	4.39	<u>26.94</u>	18.44	6.38	7.52	28.28	17.84	<u>8.51</u>	7.84	29.29	18.29	<u>10.64</u>	9.72
Gemini 2.5 Pro	<b>42.76</b>	<b>30.73</b>	<b>14.89</b>	<b>19.75</b>	<b>44.11</b>	<b>30.88</b>	<b>19.15</b>	<b>16.61</b>	<b>46.46</b>	<b>41.23</b>	<b>17.02</b>	<b>23.82</b>	<u>51.52</u>	<b>43.03</b>	<b>17.02</b>	<b>29.15</b>
Llama 4 Scout (17B)	6.06	3.30	0.00	0.31	9.76	1.80	2.13	0.31	8.08	3.77	0.00	1.25	9.76	4.80	0.00	1.57
Qwen 3 VL (235B)	10.44	1.95	0.00	0.00	14.81	3.30	2.13	1.25	14.81	4.20	4.26	3.13	15.49	3.15	2.13	1.57
Grok 4.1 Fast (reasoning)	<u>37.04</u>	<u>16.94</u>	<u>6.38</u>	<u>9.72</u>	<b>44.11</b>	<u>19.04</u>	<u>12.77</u>	<u>11.29</u>	<u>46.13</u>	<u>24.29</u>	<u>8.51</u>	<u>15.99</u>	<b>60.27</b>	<u>31.93</u>	<b>17.02</b>	<u>22.57</u>

Table 2: Accuracy(%) of models for four prompt types across languages. The best performance in each column is in **bold** and the second best is underlined.

Model	$T$	Acc (%)	$T$	Acc (%)	$T$	Acc (%)
GPT 5.2 (none)	0.01	18.86	1.0	19.53	2.0	17.17
Gemini 2.5 Flash	0.01	23.91	1.0	<u>26.94</u>	2.0	<u>23.91</u>
Gemini 2.5 Pro	0.01	<b>47.14</b>	1.0	<b>44.11</b>	2.0	<b>42.76</b>
Llama 4 Scout (17B)	0.01	7.74	0.15	9.76	5.0	0.00
Qwen 3 VL (235B)	0.01	13.80	1.0	14.81	2.0	8.75
Grok 4.1 Fast	0.01	<u>46.13</u>	1.0	<b>44.11</b>	2.0	0.34

Table 3: English task accuracy in the Few-Shot CoT setup. The table presents results for different temperature ( $T$ ) values, indicating the specific  $T$  used alongside the corresponding accuracy. The best performance in each column is in **bold** and the second best is underlined.

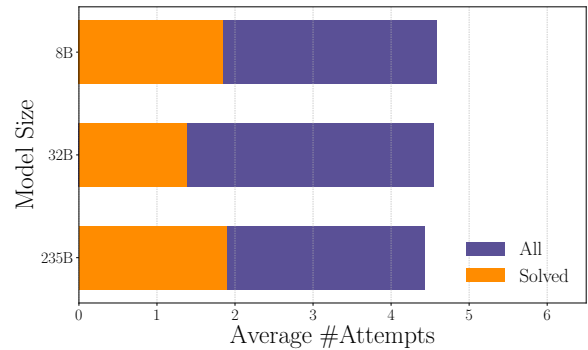


Figure 7: Average number of refinement attempts vs. model size for Qwen3-VL on the English subset. The plot reports the mean number of attempts under a five-attempt iterative refinement setting as a function of model size (Qwen3-VL-8B, 32B, 235B), shown for two aggregation regimes: averaging over solved puzzles only, and averaging over all puzzles, where unsolved instances are assigned the maximum of five attempts.

rameters helps within a same-family comparison, yet performance remains far from saturated across all subsets, highlighting that scale alone does not resolve the core reasoning challenges posed by EYE-Q.

#### B.4 Language Coupling

Figure 9 extends the language-coupling analysis from the basic setup in Section 5.4 to all prompting variants. For each variant, we compute Pearson and Spearman correlations across models using subset-level accuracies, producing a  $4 \times 4$  correlation matrix over {En, Fa, Ar, Cr}. Across all variants, correlations remain high, indicating that model strengths are largely consistent across languages even when the evaluation protocol changes. In particular, Persian and cross-lingual accuracies remain tightly coupled across variants, suggesting that the abilities driving performance on Persian puzzles also transfer to cross-lingual puzzles. Arabic shows comparatively weaker coupling in several variants, consistent with Arabic being the hardest subset and potentially introducing more variance across models.

#### B.5 Failure-Case Similarity Density

Figure 10 extends the near-miss analysis from Figure 6 to all evaluation variants (Basic, Few-shot

CoT, Iterative Refinement, and Partial Character Reveal). As in the main paper, we compute cosine similarity between model outputs and ground-truth answers in an embedding space and restrict the analysis to failure cases.

Across all variants, the qualitative pattern remains consistent: similarity distributions place most mass at relatively low values, with only small high-similarity tails. This indicates that incorrect outputs are generally not close paraphrases of the ground truth, and that the main conclusions from iterative refinement hold broadly rather than being an artifact of a particular prompting protocol.

Notably, even partial character reveal, which substantially narrows the lexical search space, does not produce a large shift toward high semantic similarity among failures, suggesting that many errors reflect misinterpretations of the underlying visual cues rather than surface-form variation. This indicates that when models fail, they typically fail by selecting the wrong latent concept rather than producing a near-miss wording of the correct one.

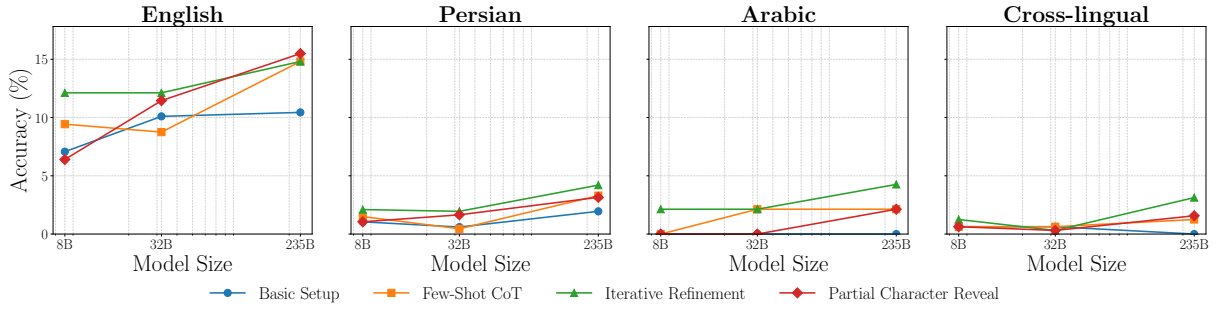


Figure 8: Accuracy vs. model size for Qwen3-VL across language subsets. Each panel reports accuracy on one subset as a function of model size (Qwen3-VL-8B, 32B, 235B) under four prompt variants: Basic, Few-shot CoT, Iterative Refinement, and Partial Character Reveal.

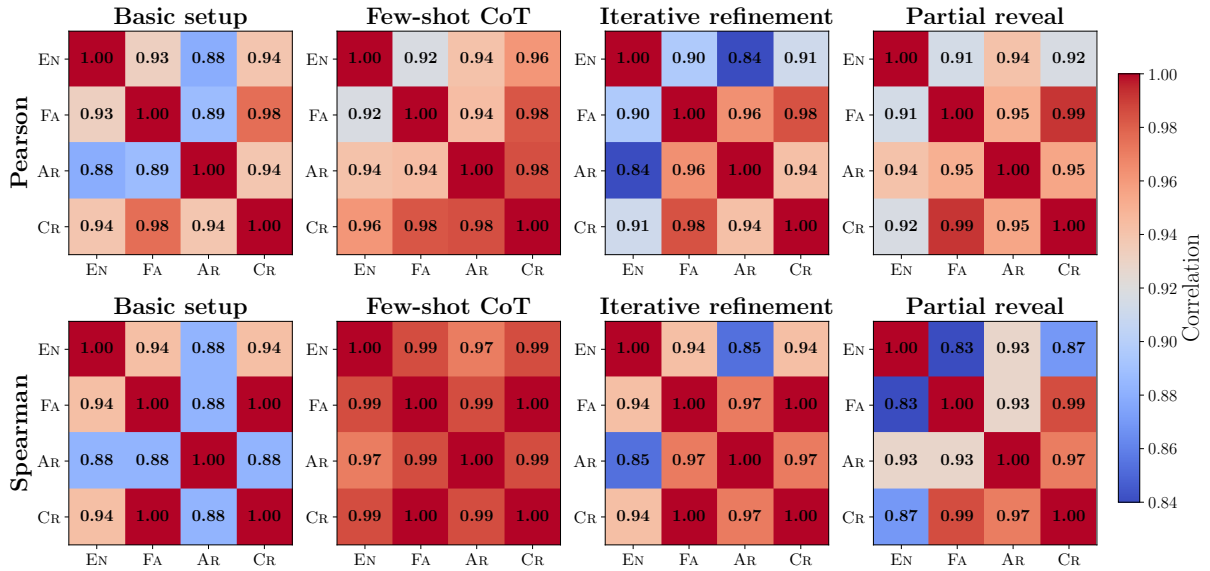
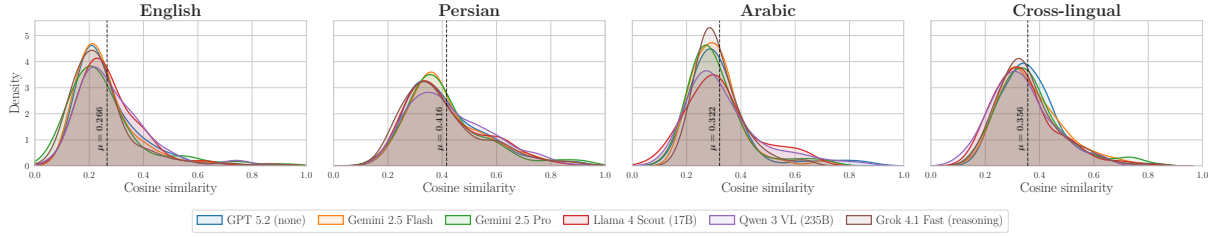
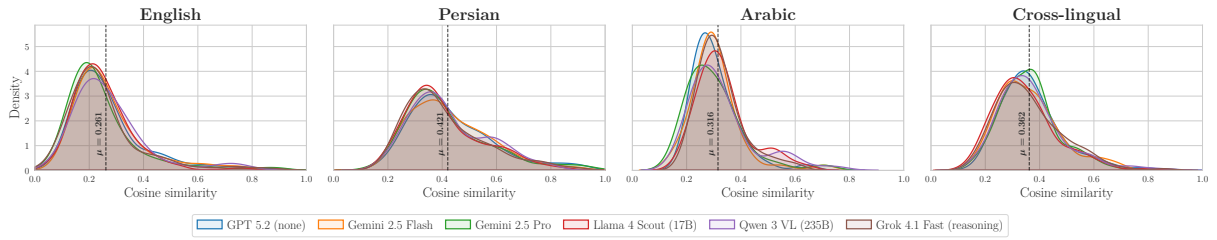


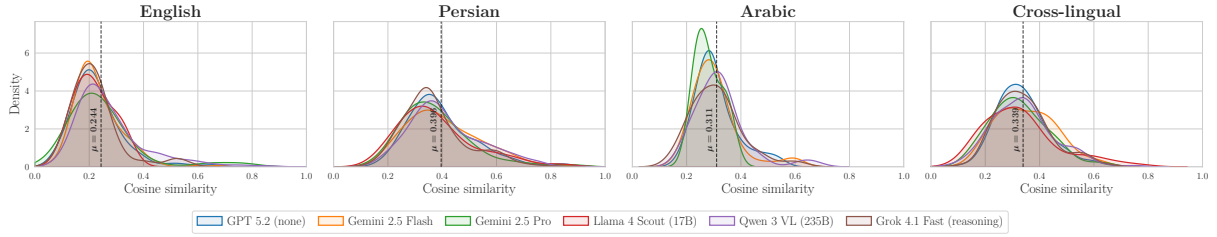
Figure 9: Cross-language coupling across prompt variants. Each matrix shows Pearson (top row) or Spearman (bottom row) correlations computed across models using subset-level accuracies for English (En), Persian (Fa), Arabic (Ar), and cross-lingual (Cr) puzzles, separately for the Basic setup, Few-shot CoT, Iterative Refinement, and Partial Character Reveal. Higher values indicate more consistent model performance across languages under a given evaluation variant.



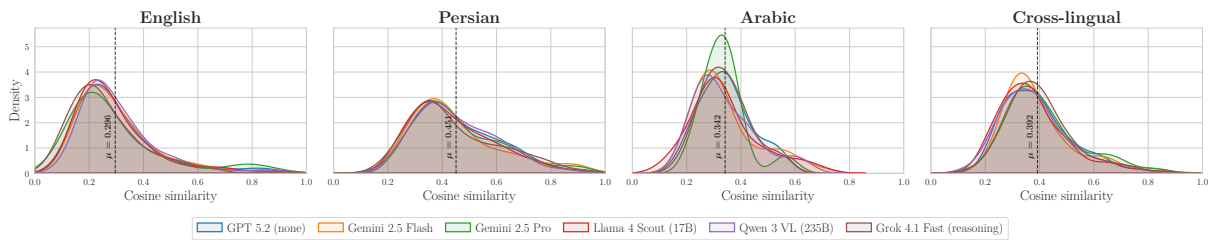
(a) Basic Setup



(b) Few-Shot CoT



(c) Iterative Refinement



(d) Partial Character Reveal

Figure 10: Semantic similarity of failure cases across evaluation variants. Each row corresponds to a prompting variant and each column to a language subset. Curves show kernel density estimates of cosine similarity between embedding representations of model outputs and ground-truth answers, computed over incorrect predictions only. Dashed lines indicate mean similarity for each plot.

1067  
1068  
1069

**C More EYE-Q examples**

We provide additional qualitative examples of EYE-Q in [Table 4](#) and [Table 5](#).

Additional EYE-Q examples			
Image	Derivation	Image	Derivation
<b>English</b>			
	The image shows a green landmass character representing “land” running from the cops through an outdoor scene or “scape”. This suggests combining the two parts as “land” + “scape”. Thus the final answer is <b>“landscape”</b> .		The image shows a “home” (a house) decorated with “sweets” like candy canes and lollipops, all placed inside an outline shaped like the U.S. state of “Alabama”. This can be read as “sweet home” in Alabama, pointing to the familiar phrase. Thus the final answer is <b>“Sweet Home Alabama”</b> .
	The image shows a “class” (students in a classroom) where everyone is “sick” (holding tissues and blowing their noses). This can be described as “class sick”, which sounds like “classic”. Thus the final answer is <b>“classic”</b> .		The image shows a chessboard scene in which certain pieces are engaged in a ritual act. This can be read as “rite chess”, which sounds like “righteous”. Thus the final answer is <b>“righteous”</b> .
	The image shows a cup of “tea” placed on a “chair”. This suggests reading the two elements together as “tea chair”, which sounds like “teacher”. Thus the final answer is <b>“teacher”</b> .		The image shows an “add” symbol indicating addition together with a muddy or swampy scene representing a “mire”. Read together as “add mire”, this sounds like “admire”. Thus the final answer is <b>“admire”</b> .
	The image depicts a “shell” alongside a blackboard containing a mathematical “proof”. Reading these two concepts together gives “shell proof”, leading to the word “shellproof”. Thus the final answer is <b>“shellproof”</b> .		The image shows falling “rain” together with bow-shaped ribbons resembling a “bow”. Interpreting these elements together gives “rain bow”, which forms the word “rainbow”. Thus the final answer is <b>“rainbow”</b> .
<b>Persian</b>			
	The image shows a head, which in Persian is written “سر”, placed on a bench/seat, suggesting the idea of sitting (“نشین” which relates to the verb “sit” in Persian). Combining “سر” with “نشین” yields “سرنشین”; thus, the final answer is <b>“سرنشین”</b> .		The image shows a hand, which in Persian is written “دست”, together with an action of throwing/putting, which cues “انداز”. Combining “دست” with “انداز” yields “دست انداز”; thus, the final answer is <b>“دست انداز”</b> .
	The image shows the moon, which in Persian is written “ماه”, sitting on a swing. In Persian, swinging/rocking is associated with “تاب”. Combining “ماه” with “تاب” yields “مهتاب”; thus, the final answer is <b>“مهتاب”</b> .		The image shows a straw, which in Persian is written “نی”, together with a king, which cues “شاه”. The king’s blond hair cues “بور”. Combining “نی” with “شاه” and “بور” yields “نیشابور”; thus, the final answer is <b>“نیشابور”</b> .
	The image shows water, which in Persian is written “آب”, placed next to a warm-up pant, which cues “گرمکن”. Reading these elements together gives “آب گرمکن”; thus, the final answer is <b>“آب گرمکن”</b> .		The image shows a photograph of a surgical operation, which can be read as “عکس عمل”. Interpreting this as “عکس العمل” leads to the intended word; thus, the final answer is <b>“عکس العمل”</b> .
	The image shows a horse, which in Persian is written “اسب”. The horse is eating the Persian letter “ب”, and since “ب” with “خوار” reads as “بخوار”, this can be interpreted as “بخار”. Reading these elements together yields “اسب بخار”; thus, the final answer is <b>“اسب بخار”</b> .		The image shows a hand, which in Persian is written “دست”, holding money or property, which cues “مال”. Interpreting these elements together gives “دستمال”; thus, the final answer is <b>“دستمال”</b> .

Table 4: Additional illustrative EYE-Q examples across English and Persian settings. Each puzzle consists of an input image and a short derivation leading to the target word or phrase.

Additional EYE-Q examples			
Image	Derivation	Image	Derivation
<b>Arabic</b>			
	The image shows the Arabic letter "ق" with wings, indicating that it is flying, which cues the verb "طَارَ". Combining "ق" with "طَارَ" yields "قَطَارَ". Thus the final answer is "قَطَارَ".		The image shows a house, which in Arabic is "بَيْت", constructed out of mulberries, which cue the word "نُوت". Concatenating "بَيْت" and "نُوت" yields "بَيْتُونَة". Thus the final answer is "بَيْتُونَة".
	The image shows a museum artwork, which cues the Arabic word "فَن", depicting a loaf of bread, which cues "نَان". Combining "فَن" with "نَان" yields "فَنَان". Thus the final answer is "فَنَان".		The image shows a bright crescent moon, which cues the Arabic word "قَمَر", above a mosque interior with worshippers, suggesting "دِين" which means religion. Combining "قَمَر" and "دِين" yields "قَمَرَالدِين". Thus, the final answer is "قَمَرَالدِين".
	The image shows a scene related to logic, which cues the Arabic word "مَنْطِق". The presence of a feminine marker suggests adding "ة". Reading these together gives "مَنْطِقَة". Thus the final answer is "مَنْطِقَة".		The image shows a market, which in Arabic is "سُوق", filled with fresh plants, which cue "طَرِي". Interpreting these attachively gives "سُوقَطَرِي". Thus the final answer is "سُوقَطَرِي".
	The image shows a coin, which cues the Arabic word "فَلْس", placed in soil or mud, which cues "طِين". Reading these elements together gives "فَلْسَطِين". Thus the final answer is "فَلْسَطِين".		The image shows a ball, which cues "كُرَة", together with a head belonging to a person, which can be read as "رَأْسُه" (his head). Interpreting these elements together gives "كُرَأْسَة". Thus the final answer is "كُرَأْسَة".
<b>Cross-lingual</b>			
	The image shows a USB port, which directly cues the Persian loanword "پورت". Beneath the port, the label reads GHAL, which cues "غال". Combining "پورت" with "غال" yields "پرتغال". Thus, the final answer is "پرتغال".		The image shows a Share button, which cues the sound "شر" (from "share"), together with a cat, which cues "كت". Combining "شر" with "كت" yields "شركت". Thus, the final answer is "شركت".
	The image shows a koala, which cues "كوالا", a lamp, which cues "لامپ", and a poor beggar, which cues "پور" as a sound cue. Combining "كوالا" with "لامپ" and "پور" yields "كوالالامپور". Thus, the final answer is "كوالالامپور".		The image shows a car, which cues "كار" as a sound cue, and a house, which cues "خانه". Combining "كار" with "خانه" yields "كارخانه". Thus, the final answer is "كارخانه".
	The image shows a mouse ("موش") next to a bag of coffee, cueing the English word "coffee". Reading these together yields "موش" + "coffee", forming "موشكافي". Thus, the final answer is "موشكافي".		The image shows "پای گربه" (a cat's leg). By replacing "گربه" with the English sound cue "cat" (read as "كت"), we obtain "پای كت", which contracts to "پاكت". Thus, the final answer is "پاكت".
	The image shows the English letter "A" placed on denim material, which cues the Persian word "لی". Reading these elements together gives "A" + "لی", forming "عالی". Thus, the final answer is "عالی".		The image shows two snakes, and in Persian a snake is "مار". Reading this as "two" + "مار" yields "طومار". Thus, the final answer is "طومار".

Table 5: Additional illustrative EYE-Q examples across Arabic and cross-lingual settings. Each puzzle consists of an input image and a short derivation leading to the target word or phrase.

## D Prompt Options

This appendix describes the prompt building blocks used in our experiments. Each query includes one image and a text prompt formed by concatenating a fixed set of reusable modules, as outlined below.

### D.1 Language and subset rules

We use four rule modules, one for each subset: English, Persian, Arabic, and cross-lingual. Each module specifies the required answer language and any subset-specific constraints. For instance, in the cross-lingual subset, the target answer language is Persian (Farsi), but solving the puzzle may require English knowledge because the image can contain English words, letters, numbers, or concepts. The intended answer may therefore include English elements directly (e.g., via transliteration) or combine English and Persian components.

### D.2 Shared task and output instructions

All experimental settings include the same core instruction module. This block explains the game setup, states what the model should infer from the image and the accompanying text, and clarifies the expected form of the response. Crucially, it also enforces a strict output constraint to ensure consistent evaluation across models and prompt variants.

### D.3 Hint modules

To provide additional guidance without changing the underlying task, we use three lightweight hint modules that can be optionally appended to the base prompt. These hints are designed to constrain the search space or support multi-round solving:

- **Answer-length hint:** provides the answer length in characters, excluding spaces.
- **Partial-reveal hint:** provides a masked pattern where a subset of characters is revealed and all remaining positions are replaced by underscores.
- **Iterative-refinement hint:** provides feedback that includes the model’s previous answer, enabling a revision step in subsequent rounds.

### D.4 Few-shot CoT demonstrations

For the few-shot CoT condition, we prepend three solved demonstrations from the same subset as the target puzzle. We prepare one demonstration block

per subset (English, Persian, Arabic, and cross-lingual). Each demonstration includes an image and a concise, structured rationale with two fields: primary clues and candidate answers. The full set of demonstrations used in our experiments is provided below.

1115  
1116  
1117  
1118  
1119  
1120

### English Language Rules

#### LANGUAGE RULES:

- The target answer language is English.

### Persian Language Rules

#### LANGUAGE RULES:

- The target answer language is Persian (Farsi).
- **CULTURAL LENS:** Do not simply translate English concepts. You must interpret the visual elements through the lens of Persian culture, literature, and common daily idioms.
- **WORDPLAY:** If the image suggests wordplay, prioritize phonetic/semantic connections natural in Persian.

### Arabic Language Rules

#### LANGUAGE RULES:

- The target answer language is Arabic.
- **CULTURAL LENS:** Do not simply translate English concepts. You must interpret the visual elements through the lens of Arabic culture, literature, and common daily idioms.
- **WORDPLAY:** If the image suggests wordplay, prioritize phonetic/semantic connections natural in Arabic.

### Cross-Lingual Rules

#### LANGUAGE RULES:

- The target answer language is Persian (Farsi).
- **ENGLISH KNOWLEDGE REQUIRED:** The puzzle may rely on English words, concepts, letters, or numbers depicted in the image.
- You may need to use English elements directly in the Persian answer (transliteration) or combine them with Persian to form the intended phrase.

## Game Explanation

You are an expert multi-modal puzzle solver. You solve picture word puzzles.

### GAME DESCRIPTION:

- You will see exactly ONE image per puzzle.
- The image may depict objects, people, scenes, text, icons, or abstract compositions.
- The goal is to infer a SINGLE intended answer: one word or a short phrase.
- The image is a deliberately constructed clue for a linguistic target, NOT a request to describe the scene.
- The intended answer may be:
  - a literal word,
  - an idiom or proverb,
  - a pun or wordplay,
  - a common expression,
  - a culturally meaningful phrase,
  - or a proper noun / named entity (person, place, title, brand, named item).

### GENERAL SOLVING PROCEDURE (follow in order):

1. Identify candidate clue units in the image:
  - the most salient objects/entities
  - any text, letters, numbers, symbols, or icons
  - any repeated motif/pattern
2. Select ONLY 2–4 PRIMARY clue units:
  - prefer central/emphasized/repeated units
  - compress repeated motifs into one unit
  - ignore minor background details unless they clearly change a primary unit
3. Hypothesize a simple composition:
  - the answer is usually formed by combining or transforming the primary units
  - prefer the simplest coherent interpretation with the fewest assumptions
4. Choose the best final answer:
  - it should be natural/common in the target language
  - it should explain the primary units as a single intended construction
  - prioritize global coherence over matching every local detail

### OUTPUT REQUIREMENT:

- Provide exactly ONE final answer (single word or short phrase).
- If uncertain, choose the most plausible candidate under the simplest coherent interpretation.

## Output Instructions

**OUTPUT FORMAT:** Return ONLY a single valid JSON object. Do not output markdown blocks or conversational text.

```
{
  "primary_clues": ["...", "..."],
  "candidates": ["...", "...", "..."],
  "final_answer": "..."
}
```

### #Characters Hint

The answer has <#Characters> characters (excluding spaces).

### Partial Character Reveal Hint

The pattern of the answer is “\_\_<character>\_\_<character>\_”.

In this pattern, “\_” represents a character and spaces represent actual spaces in the answer.

### Iterative Refinement

Your previous attempt was <model\_previous\_answer> which is incorrect. Analyze the image carefully and try again.

### Persian Language Few-shot CoT



#### Example 1:

”primary\_clues”: [

”The image displays a human brain, which translates to 'Mokh' (مخ) in Persian.”,

”The visual style of the brain is blurred or faded, which translates to the adjective 'Tar' (تار).”,

”Combining the noun 'Mokh' with the adjective 'Tar' creates the construction 'Mokh-Tar'.”,

”This forms the common Persian name 'Mokhtar' (مختار).” ]

”candidates”: [ "مخ مات", "تارمخ", "مختار" ]

#### Example 2:

”primary\_clues”: [ ”The image shows a courtroom setting with lawyers, representing the legal concept of 'Vekalat' (وکالت) or power of attorney.”,

”The cartoon character is Tom from 'Tom and Jerry', whose name in Persian is 'Tam' (تام).”,

”Combining the legal term 'Vekalat' with the character's name 'Tam' creates the phrase 'Vekalat-e Tam' (وکالت تام).”,

”This phrase is a common legal term meaning 'Full power of attorney'.” ]

”candidates”: [ "وکالت تام", "وکیل تامی", "تام وکیل" ]

#### Example 3:

”primary\_clues”: [ ”The image displays a grilling skewer (seekh) typically used for making kebabs.”,

”Instead of meat, the skewer holds multiple instances of the Persian letter 'Che' (چ).”,

”In Persian, 'multiple Che's' or 'several Che's' translates to 'Chand Che' (چند چ).”,

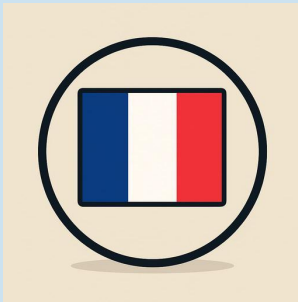
”The phrase 'Chand Che' sounds phonetically very similar to 'Chenjeh' (چنجه).”,

”This creates a pun on 'Kebab Chenjeh' (Lamb Chop Kebab), replacing the meat with letters.” ]

”candidates”: [ "کباب چنجه", "کباب چندچ", "چهار کباب" ]

## English Language Few-shot CoT

### Example 1:



"primary\_clues": [ "The central image is the flag of France.", "The flag is enclosed or surrounded by a circle.", "The prefix for 'around' or 'surrounding' is 'Circum-'.", "Combining the prefix with the country name creates 'Circum-France'.", "Phonetically, this sounds like the geometric term 'Circumference'." ]

"candidates": [ "France Circles", "Rounded France", "Circumference" ]

### Example 2:

"primary\_clues": [ "A man sits in a corner with tattered clothes and a bowl containing broken scraps, symbolizing having lost everything.", "His desolate condition and the empty/broken contents of his begging bowl indicate he is 'in need'.", "A thought bubble above his head displays the letter 'L'.", "Combining his condition ('Need') with the letter ('L') creates the construction 'Need-L'.", "Phonetically, this sounds like the word 'Needle'." ]

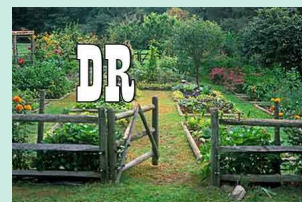
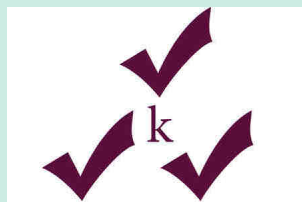
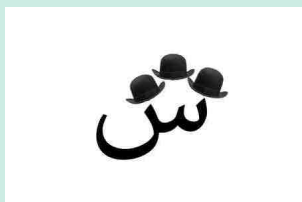
"candidates": [ "Needs L", "L Wants", "Needle" ]

### Example 3:

"primary\_clues": [ "The image depicts a personified church building.", "The church appears sick, with a thermometer in its mouth and a tissue in its hand, indicating it is 'ill'.", "This combines to form 'Church' + 'ill', sounding like 'Churchill'.", "The church is also holding a pack of cigarettes clearly labeled 'Winston'.", "Combining the brand name with the other elements creates 'Winston Churchill'." ]

"candidates": [ "Sick Church Smokes", "Winstons Sick Home", "Winston Churchill" ]

## Cross-Lingual Few-shot CoT



### Example 1:

"primary\_clues": [ "The image shows the Persian letter 'Seen' (س).", "Placed on top ('Roosh' - ( of the 'Seen' are three ('Se' - سه) hats ('Hat' - هت).", "Combining 'Seen' + 'Roosh' creates the name 'Soroush' (سروش).", "Combining 'Se' (Three) + 'Hat' (English word for hat) sounds like 'Sehhat' (صحت).", "Together, they form the name of the famous Iranian director 'Soroush Sehhat' (سروش صحت)." ]

"candidates": [ "سه کلاه س", "کلاه سه س", "سروش صحت" ]

### Example 2:

"primary\_clues": [ "The image shows the letter 'k' surrounded by three checkmarks (ticks).", "The letter 'k' is in the middle, which translates to 'Mian' (میان) in Persian.", "Combining 'K' + 'Mian' phonetically sounds like 'Kamion' (کامیون), meaning Truck.", "The 'k' is placed among ('La' - لا) the three ('Se' - سه) ticks ('Tick' - تیک).", "The phrase 'La-Se-Tick' (لا سه تیک) sounds like 'Lastik' (لاستیک), meaning Tire.", "Together, the visual pun creates 'Lastik Kamion' (لاستیک کامیون)." ]

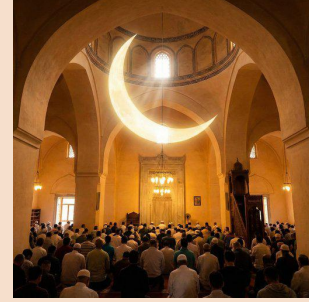
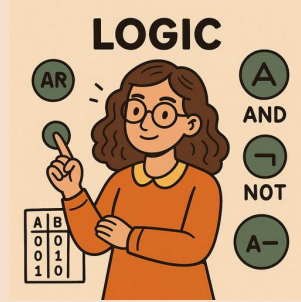
"candidates": [ "لاستیک اتوبوس", "لاستیک دوچرخه", "لاستیک کامیون" ]

### Example 3:

"primary\_clues": [ "The image features a garden, which translates to 'Bagh' (باغ) in Persian.", "The letters 'DR' are superimposed on the scene.", "Phonetically, the English letters 'DR' (Dee-Ar) sound identical to the Persian word 'Diar' (دیار), meaning 'Land' or 'Realm'.", "Combining the two components yields 'Diar' + 'Bagh'.", "This creates a pun on the common phrase 'Diar-e Baghi' (دیار باقی), which means 'The Eternal Realm' or 'The Afterlife'.", "The pun works by substituting the word 'Bagh' (Garden) for the phonetically similar 'Baghi' (Eternal)." ]

"candidates": [ "دیار باقی", "باغ دیوار", "دکتر باقی" ]

## Arabic Language Few-shot CoT



### Example 1:

"primary\_clues": [ "The image features the Arabic letter 'Qaf' (ق).", "The letter has wings attached to it, indicating the action of flying.", "The Arabic verb for 'he flew' is 'Tara' (طار).", "Combining the letter 'Q' (ق) with the verb 'Tara' (طار) creates the sequence 'Q-Tara'.", "This forms the word 'Qitar' (قطار).", "The word 'Qitar' (قطار) means 'Train' in Arabic." ]

"candidates": [ "قطار", "طائر", "طارق" ]

### Example 2:

"primary\_clues": [ "The image features the word 'LOGIC' along with truth tables and logic gates.", "The concept of 'Logic' translates to 'Mantiq' (منطق) in Arabic.", "The central character is female, indicating the presence of a 'feminine marker'.", "In Arabic morphology, the feminine marker is added as 'Ta Marbuta' (ة).", "Combining the word 'Mantiq' (منطق) with the feminine suffix (ة) yields 'Mintaqah' (منطقة).", "While the base word means Logic, the resulting word 'Mintaqah' (منطقة) means 'Area' or 'Zone'." ]

"candidates": [ "منطقة", "منطقي", "منظمة" ]

### Example 3:

"primary\_clues": [ "The image features a large, bright crescent moon inside a structure.", "The Arabic word for 'Moon' is 'Qamar' (قمر).", "The setting is a mosque filled with worshippers, symbolizing 'Religion' or 'Faith', which translates to 'Din' (دين) in Arabic.", "Combining the word for Moon 'Qamar' and Religion 'Din' yields the compound 'Qamar al-Din' (قمر الدين).", "This forms the term 'Qamar al-Din', which refers to sheets of dried apricot paste." ]

"candidates": [ "قمر الدين", "شمس الدين", "نور الدين" ]

## E Model Output Examples

This section presents qualitative examples of model behavior on our multilingual visual word-puzzle benchmark for two systems, **Gemini 2.5 Pro** and **Grok 4.1 Fast**. We evaluated the models in four languages and under four experimental variations, which results in **16 different setups** for each model. For every setup, we report two examples. The examples are reported by language and then by experimental variation. They are ordered as English, Persian, Arabic, and cross-lingual. The examples are not selected uniformly at random. Instead, we focus on edge cases that are particularly useful for analyzing model capabilities and failure modes. These include borderline correct cases that require non-trivial compositional reasoning, as well as near-miss failures where the model produces a plausible but incorrect answer. Each example includes the input image, the corresponding variation configuration such as hint type and context settings, the ground-truth answer, and the model output. This structure allows direct comparison across languages and experimental conditions.

### Basic setup

Task: Solve the picture word puzzle from the image.  
Target language: English.  
Hint type: char\_count.

#### Ground Truth

roadtrip



### Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "A character that is a piece of a road", "The action of tripping over a rock"
  ],
  "candidates": [
    "Road trip", "Roadblock", "Hit rock bottom", "A bump in the road"
  ],
  "final_answer": "Road trip"
}
```

### Model: grok-4.1-fast

```
{
  "primary_clues": [
    "road", "tripping", "rock"
  ],
  "candidates": [
    "rocky road", "roadblock", "roadtrip"
  ],
  "final_answer": "roadtrip"
}
```

### Basic setup

Task: Solve the picture word puzzle from the image.  
Target language: English.  
Hint type: char\_count.

#### Ground Truth

harry potter



Model: gemini-2.5-pro

```
{  
  "primary_clues": [  
    "A hairy man resembling Hagrid", "Making a pot on a potter's wheel"  
  ],  
  "candidates": [  
    "Hairy Potter", "Hagrid's Pottery", "The Potter's Wheel"  
  ],  
  "final_answer": "Hairy Potter"  
}
```

Model: grok-4.1-fast

```
{  
  "primary_clues": [  
    "hairy man", "potter (pottery wheel)"  
  ],  
  "candidates": [  
    "harry potter", "hairy potter", "potter harry"  
  ],  
  "final_answer": "harry potter"  
}
```

### Example: Few-shot chain-of-thought prompting

Task: Solve the picture word puzzle from the image.

Target language: English.

Hint type: char\_count.

Ground Truth

Fishtank



**Model: gemini-2.5-pro**

```
{
  "primary_clues": [
    "The turret of the army tank is a glass container holding the fish, which is a fish tank.",
    "Two fish wearing helmets.", "An army tank."
  ],
  "candidates": [
    "Aquarium", "Tank Fish", "Fish Tank"
  ],
  "final_answer": "Fish Tank"
}
```

**Model: grok-4.1-fast**

```
{
  "primary_clues": [
    "military tank with transparent body and bubbles",
    "two orange fish inside wearing green helmets"
  ],
  "candidates": [
    "army goldfish", "wet soldier fish", "Fishtank"
  ],
  "final_answer": "Fishtank"
}
```

**Example: Few-shot chain-of-thought prompting**

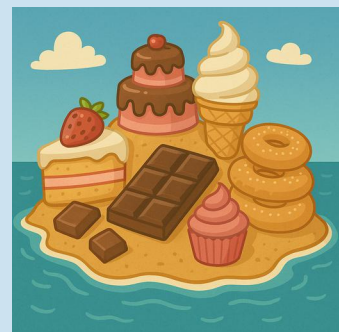
Task: Solve the picture word puzzle from the image.

Target language: English.

Hint type: char\_count.

**Ground Truth**

Deserted island



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "An island in the ocean.",
    "The island is uninhabited, which makes it a 'deserted' island.",
    "The island is covered entirely with sweets, also known as 'desserts'.",
    "A pun is formed by combining 'dessert' with '-ed' to create 'desserted', which sounds like 'deserted'."
  ],
  "candidates": [
    "Dessert island", "Sweet island", "Desserted island"
  ],
  "final_answer": "Desserted island"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "a small island surrounded by ocean",
    "the island is made of desserts (cakes, donuts, ice cream, chocolate)",
    "no people present (empty/abandoned)"
  ],
  "candidates": [
    "dessert island", "desert island", "Deserted island"
  ],
  "final_answer": "Deserted island"
}
```

### Iterative refinement

Task: Solve the picture word puzzle from the image.

Target language: English.

Hint type: char\_count.

Hint: The answer has 20 characters (excluding spaces).

Ground Truth

Scholar of the first sin



Model: gemini-2.5-pro

```
"attempts":
{
  "attempt_idx": 0,
  "model_output": {
    "primary_clues": [
      "A devil figure", "A person wearing a graduation cap and gown",
      "Symbols of knowledge and temptation (apple, snake)"
    ],
    "candidates": [
      "Devil's Advocate", "Sympathy for the Devil", "Devil's Graduate School"
    ],
    "final_answer": "Devil's Graduate School"
  },
  "feedback": "The answer 'Devil's Graduate School' is incorrect. Please Step-by-Step."
},
{
  "attempt_idx": 1,
  "model_output": {
    "primary_clues": [
      "A devil figure in academic regalia", "A sapling representing an 'origin'",
      "Biblical symbols of evil (apple, snake)"
    ],
    "candidates": [
      "On the Origins of Evil", "Knowledge is the root of all evil", "Devil's Advocate"
    ],
    "final_answer": "On the Origins of Evil"
  },
  "feedback": "The answer '' is incorrect. Please Step-by-Step."
},
{
  "attempt_idx": 2,
  "model_output": {
    "primary_clues": [
      "A devil dressed as a scholar or professor", "Symbols of temptation (apple and snake)",
      "A known work of literature with a 20-character title"
    ],
    "candidates": [
      "The Screwtape Letters", "On the Origins of Evil", "Sympathy for the Devil"
    ],
    "final_answer": "The Screwtape Letters"
  }
}
```

**Model: grok-4.1-fast**

```
"attempts": [  
  {  
    "attempt_idx": 0,  
    "primary_clues": [  
      "horned graduate", "bitten apple", "snake on branch"  
    ],  
    "candidates": [  
      "tree of good and evil", "the tree of good and evil", "an apple for the teacher"  
    ],  
    "final_answer": "an apple for the teacher"  
  },  
  {  
    "attempt_idx": 1,  
    "primary_clues": [  
      "devil graduate", "bitten apple", "snake on branch"  
    ],  
    "candidates": [  
      "forbidden fruit", "tree of good and evil", "the tree of good and evil"  
    ],  
    "final_answer": "the tree of good and evil"  
  },  
  {  
    "attempt_idx": 2,  
    "primary_clues": [  
      "a devil figure wearing academic cap and gown (a scholar/teacher)",  
      "a bitten apple", "a snake on a branch (Eden / the first sin)"  
    ],  
    "candidates": [  
      "Scholar of the first sin", "The first sin scholar", "Professor of forbidden fruit"  
    ],  
    "final_answer": "Scholar of the first sin"  
  }  
]
```

**Iterative refinement**

Task: Solve the picture word puzzle from the image.

Target language: English.

Hint type: char\_count.

Hint: The answer has 9 characters (excluding spaces).

**Ground Truth**

Hat-trick



Model: gemini-2.5-pro

```
"attempts": [
{
  "attempt_idx": 0,
  "model_output": {
    "primary_clues": [
      "A plurality of hats", "A tree", "A tree trunk shaped like the letter 'K'"
    ],
    "candidates": [
      "Hat trick", "Hat-trick", "Hat tricks", "Millinery"
    ],
    "final_answer": "Hat tricks"
  },
  "feedback": "The answer 'Hat tricks' is incorrect. Please Step-by-Step:"
},
{
  "attempt_idx": 1,
  "model_output": {
    "primary_clues": [
      "An impossible scene of hats on a tree, suggesting magic",
      "A tree trunk shaped like the letter 'K'"
    ],
    "candidates": [
      "Hat tricks", "Family tree", "Magic trick"
    ],
    "final_answer": "Magic trick"
  },
  "feedback": "The answer 'Magic trick' is incorrect. Please Step-by-Step:"
},
{
  "attempt_idx": 2,
  "model_output": {
    "primary_clues": [
      "A plurality of hats", "A tree", "A tree trunk shaped like the letter 'K'"
    ],
    "candidates": [
      "Hat tricks", "Magic trick", "Hattricks"
    ],
    "final_answer": "Hattricks"
  }
}
]
```

Model: grok-4.1-fast

```
"attempts": [  
  {  
    "attempt_idx": 0,  
    "primary_clues": [  
      "tree", "hats"  
    ],  
    "candidates": [  
      "hat tree", "mad hatter", "three hats"  
    ],  
    "final_answer": "three hats"  
  },  
  {  
    "attempt_idx": 1,  
    "primary_clues": [  
      "many hats", "K-shaped tree"  
    ],  
    "candidates": [  
      "hat tree", "hat trick", "mad hatter"  
    ],  
    "final_answer": "mad hatter"  
  },  
  {  
    "attempt_idx": 2,  
    "primary_clues": [  
      "three hats", "a tree shaped like the letter K (hinting 'K' in 'trick')"  
    ],  
    "candidates": [  
      "Hat-trick", "hat trick", "hat tricks"  
    ],  
    "final_answer": "Hat-trick"  
  }  
]
```

### Partial character reveal

Task: Solve the picture word puzzle from the image.

Target language: English.

Hint type: shuffle\_chars.

Hint pattern: \_\_t\_\_r\_\_.

Ground Truth

Hat-trick



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "A collection of various hats", "A tree",
    "The tree's trunk and branches are shaped like the letter 'K'"
  ],
  "candidates": [
    "Hat trick", "Hat tree", "Patronage"
  ],
  "final_answer": "Hat trick"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "multiple hats", "a tree that looks like a rack/stand",
    "a clear hint toward the word 'trick' (visual pun setup)"
  ],
  "candidates": [
    "Hat-trick", "hat trick", "hattrick"
  ],
  "final_answer": "Hat-trick"
}
```

### Partial character reveal

Task: Solve the picture word puzzle from the image.

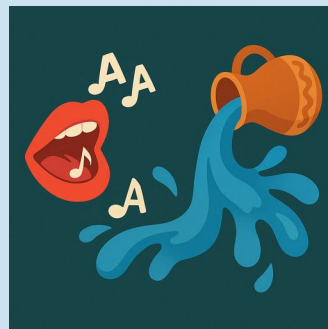
Target language: English.

Hint type: shuffle\_chars.

Hint pattern: \_\_n\_\_o\_\_.

Ground Truth

singapore



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "A mouth singing a sustained note ('A')", "A pitcher pouring a continuous stream of water",
    "The letter pattern is __n__o__"
  ],
  "candidates": [
    "Tenacious", "Aquarius", "Monotonous"
  ],
  "final_answer": "Tenacious"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "the wordplay cue 'sing'", "a 'pore' (visible skin pore / pore-like opening)",
    "the combined sound suggests a place name"
  ],
  "candidates": [
    "singapore", "sing a pore", "singa pore"
  ],
  "final_answer": "singapore"
}
```

### Basic setup

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Hint: The answer has 7 characters (excluding spaces).



Ground Truth

گلاب پاش

Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "گل (flower)", "آپاش (watering can)", "آب (water)"
  ],
  "candidates": [
    "گلاب", "آپاش", "گلابپاش"
  ],
  "final_answer": "گلابپاش"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "آب پاش", "گل"
  ],
  "candidates": [
    "گلاب پاش", "آبشار گل", "باران گل", "گل افشان"
  ],
  "final_answer": "گلاب پاش"
}
```

### Basic setup

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Hint: The answer has 14 characters (excluding spaces).



#### Ground Truth

دسته گل به آب دادن

Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "آب", "دست و گل"
  ],
  "candidates": [
    "آب دادن پای گل", "گلاب", "دست گل به آب دادن"
  ],
  "final_answer": "دست گل به آب دادن"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "آب", "دسته گل"
  ],
  "candidates": [
    "آب دادن به گل", "دسته گل به آب دادن", "گل به آب دادن"
  ],
  "final_answer": "دسته گل به آب دادن"
}
```

### Few-shot chain-of-thought prompting

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Context: Few-shot examples are provided in the prompt.



#### Ground Truth

توپ خانه

### Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "A house, which is 'Khaneh' (خانه) in Persian.",
    "A basketball, which is a type of ball, 'Toop'(توپ) in Persian."
  ],
  "candidates": [
    "توپخانه", "منزل بسکتبال", "خانه توپ"
  ],
  "final_answer": "توپخانه"
}
```

### Model: grok-4.1-fast

```
{
  "primary_clues": [
    "ترکیب «توپ» + «خانه»، «خانه‌ای که از توپ (بسکتبال) ساخته شده»
  ],
  "candidates": [
    "توپخانه", "توپ خانه", "خانه توپ"
  ],
  "final_answer": "توپخانه"
}
```

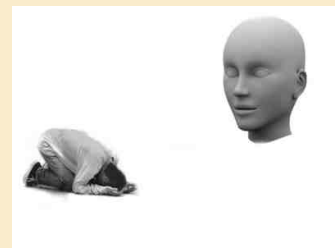
### Few-shot chain-of-thought prompting

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Context: Few-shot examples are provided in the prompt.



#### Ground Truth

سرپرست

Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "A person in the act of prostration or worship, which implies 'Parastesh' (پرستش).",
    "A generic, statue-like head, which represents an 'idol' or 'Bot' (بت) in Persian."
  ],
  "candidates": [
    "بت پرست", "طاغوت", "سر به سجده"
  ],
  "final_answer": "بت پرست"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "فرد در حال سجده/پرستش (پرست)", "سر انسان (سر)"
  ],
  "candidates": [
    "پرستار", "سرپرستی", "سرپرست"
  ],
  "final_answer": "سرپرست"
}
```

### Iterative refinement

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Hint: The answer has 8 characters (excluding spaces).

Ground Truth

لیست سیاه



Model: gemini-2.5-pro

```
"attempts": [
  {
    "attempt_idx": 0,
    "model_output": {
      "primary_clues": [
        "A checklist being marked", "A red pencil", "The action of reviewing or checking"
      ],
      "candidates": [
        "چک لیست", "بررسی کردن", "خط قرمز"
      ],
      "final_answer": "بررسی کردن"
    },
    "feedback": "The answer 'بررسی کردن' is incorrect. Please Step-by-Step:"
  },
  {
    "attempt_idx": 1,
    "model_output": {
      "primary_clues": [
        "A pencil/pen, which is 'قلم' in Persian",
        "The action of crossing an item out, which signifies finding fault, as in 'ایراد گرفتن'"
      ],
      "candidates": [
        "خط قرمز", "سرشماری", "قلم گرفتن"
      ],
      "final_answer": "قلم گرفتن"
    },
    "feedback": "The answer 'قلم گرفتن' is incorrect. Please Step-by-Step:"
  },
  {
    "attempt_idx": 2,
    "model_output": {
      "primary_clues": [
        "A list or form, which can be called 'صورت' in Persian",
        "The action of checking or tallying, related to 'حساب' (accounting)"
      ],
      "candidates": [
        "بررسی کردن", "قلم گرفتن", "صورت حساب"
      ],
      "final_answer": "صورت حساب"
    }
  }
]
```

Model: grok-4.1-fast

```
"attempts": [  
  {  
    "attempt_idx": 0,  
    "primary_clues": [  
      "علامت‌گذاری روی آیتم‌ها", "یک لیست/چک‌لیست روی کلیپ‌برد"  
    ],  
    "candidates": [  
      "لیست سیاه", "لیست کامل", "چک لیست"  
    ],  
    "final_answer": "چک لیست"  
  },  
  {  
    "attempt_idx": 1,  
    "primary_clues": [  
      "علامت‌گذاری به شکل «منفی/رد کردن» (ضربدر/خط زدن)", "لیست آیتم‌ها"  
    ],  
    "candidates": [  
      "خط قرمز", "لیست سیاه", "لیست ممنوع"  
    ],  
    "final_answer": "لیست ممنوع"  
  },  
  {  
    "attempt_idx": 2,  
    "primary_clues": [  
      "تأکید تصویری روی «سیاه/منفی بودن» لیست (رد شدن/علامت منفی)", "یک «لیست» واضح"  
    ],  
    "candidates": [  
      "لیست ممنوعه", "لیست قرمز", "لیست سیاه"  
    ],  
    "final_answer": "لیست سیاه"  
  }  
]
```

### Iterative refinement

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Ground Truth

گرگ باران زده



Model: gemini-2.5-pro

```
"attempts": [  
  {  
    "attempt_idx": 0,  
    "model_output": {  
      "primary_clues": [  
        "گرگ", "باران"  
      ],  
      "candidates": [  
        "گرگ در باران", "گرگ بارش دیده", "گرگ باران دیده"  
      ],  
      "final_answer": "گرگ بارش دیده"  
    },  
    "feedback": "The answer 'گرگ بارش دیده' is incorrect. Please Step-by-Step."  
  },  
  {  
    "attempt_idx": 1,  
    "model_output": {  
      "primary_clues": [  
        "گرگ", "باران"  
      ],  
      "candidates": [  
        "گرگ تگرگ دیده", "گرگ باران دیده", "گرگ باران خور"  
      ],  
      "final_answer": "گرگ باران خور"  
    },  
    "feedback": "The answer 'گرگ باران خور' is incorrect. Please Step-by-Step."  
  },  
  {  
    "attempt_idx": 2,  
    "model_output": {  
      "primary_clues": [  
        "گرگ", "باران"  
      ],  
      "candidates": [  
        "گرگینه", "گرگ تگرگ دیده", "گرگ باران دیده"  
      ],  
      "final_answer": "گرگ باران دیده"  
    }  
  }  
]
```

Model: grok-4.1-fast

```
"attempts": [  
  {  
    "attempt_idx": 0,  
    "primary_clues": [  
      "عضله/رگ", "باران", "گرگ"  
    ],  
    "candidates": [  
      "باران رگباری", "باران گربه سگ", "باران سگ گربه"  
    ],  
    "final_answer": "باران رگباری"  
  },  
  {  
    "attempt_idx": 1,  
    "primary_clues": [  
      "ابر", "باران", "گرگ/سگ"  
    ],  
    "candidates": [  
      "گرگ بارانی", "رگبار گرگی", "باران سگ گربه"  
    ],  
    "final_answer": "باران سگ گربه"  
  },  
  {  
    "attempt_idx": 2,  
    "primary_clues": [  
      "باران شدید و ممتد", "گرگ خیس/باران خورده",  
      "تأکید تصویری روی «زده/خورده» شدن (خیس و بارانی بودن)"  
    ],  
    "candidates": [  
      "گرگ خیس", "گرگ زیر باران", "گرگ باران زده"  
    ],  
    "final_answer": "گرگ باران زده"  
  }  
]
```

### Partial character reveal

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: shuffle\_chars.

Hint pattern: ی\_ک\_

Ground Truth

سیاتیک



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "A check mark, which in Persian is called 'تیک' (tik) and symbolizes correctness or understanding.",
    "A word that fits the letter pattern (a five-letter word where the second letter is 'ی' and the last is 'ک')."
  ],
  "candidates": [
    "بی‌درک", "کیوسک", "فیزیک"
  ],
  "final_answer": "بی‌درک"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "الگوی حروف", "علامت تیک"
  ],
  "candidates": [
    "فیزیک", "تیک‌تاک", "سیاتیک"
  ],
  "final_answer": "سیاتیک"
}
```

### Partial character reveal

Task: Solve the picture word puzzle from the image.

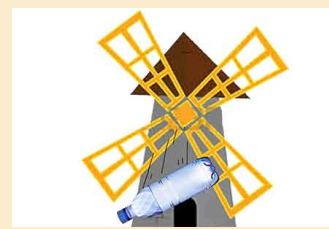
Target language: Persian (Farsi).

Hint type: shuffle\_chars.

Hint pattern: \_ \_ آس \_ ف \_ \_

#### Ground Truth

آب از آسیاب افتادن



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "آسیاب بادی (windmill)", "بطری آب (water bottle)", "افتاده (fallen)"
  ],
  "candidates": [
    "آب از آسیاب افتاده", "آب‌ها از آسیاب افتادن", "از آسیاب افتادن"
  ],
  "final_answer": "آب از آسیاب افتاده"
}
```

Model: grok-4.1-fast

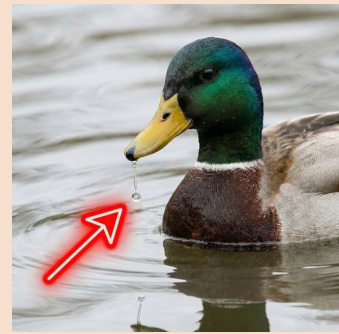
```
{
  "primary_clues": [
    "افتادن", "آسیاب", "آب"
  ],
  "candidates": [
    "آب در آسیاب افتادن", "آب به آسیاب ریختن", "آب از آسیاب افتادن"
  ],
  "final_answer": "آب از آسیاب افتادن"
}
```

### Basic setup

Task: Solve the picture word puzzle from the image.  
Target language: Arabic.  
Hint type: char\_count.

#### Ground Truth

بطريق



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "An image of a duck, which is 'بط' (baTT) in Arabic.",
    "A droplet from the beak interpreted as saliva/drool, which is 'ريق' (reeq) in Arabic."
  ],
  "candidates": [
    "تنقيط", "بطوطة", "بطريق"
  ],
  "final_answer": "بطريق"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "صورة لبط (duck).", "إشارة إلى ريق كجزء لفظي مكمل", " (saliva/drool)."
  ],
  "candidates": [
    "ماء بط", "بط + ريق", "بطريق"
  ],
  "final_answer": "بطريق"
}
```

### Basic setup

Task: Solve the picture word puzzle from the image.

Target language: Arabic.

Hint type: char\_count.

Ground Truth

مرمر



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "قهوة مُرّة (bitter coffee).",
    "خيار مُر (bitter cucumber).",
    "The clue 'اُمُر' (bitter) appears twice, suggesting repetition."
  ],
  "candidates": [
    "مرمر", "مرارة", "مرير"
  ],
  "final_answer": "مرارة"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "قهوة مُرّة", "خيار مُر", "تكرار معنى مُرّ مرتين"
  ],
  "candidates": [
    "مرمر", "مرير", "مرار"
  ],
  "final_answer": "مرمر"
}
```

### Few-shot chain-of-thought prompting

Task: Solve the picture word puzzle from the image.

Target language: Arabic.

Hint type: char\_count.

Ground Truth

سمسم



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "The image shows two bottles, each marked with a skull and crossbones, indicating they contain poison.",
    "The Arabic word for poison is 'سم' (samm).",
    "The presence of two poison bottles suggests a repetition of the word 'سم'.",
    "Repeating 'سم' twice creates the word 'سمسم' (simsim), which means sesame."
  ],
  "candidates": [
    "سمان", "سموم", "سمسم"
  ],
  "final_answer": "سمسم"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "وجود زجاجتين يعني تكرار كلمة سم", "زجاجتان عليهما علامة خطر وتدلان على السم (سم). مرتين."
  ],
  "candidates": [
    "سمين", "سموم", "سمسم"
  ],
  "final_answer": "سمسم"
}
```

### Few-shot chain-of-thought prompting

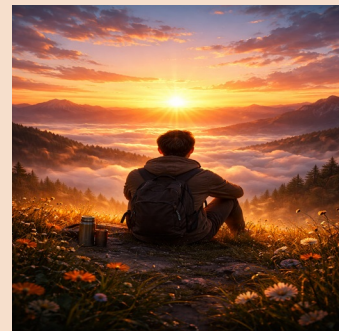
Task: Solve the picture word puzzle from the image.

Target language: Arabic.

Hint type: char\_count.

Ground Truth

منفجر



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "The image shows a person sitting still and looking towards the horizon.",
    "A sunrise is in progress, which is an event one typically waits for and watches.",
    "The person's posture suggests a state of waiting and anticipation.",
    "The Arabic word for 'one who waits' or 'expectant' is 'muntazir' (منتظر).",
  ],
  "candidates": [
    "شروقي", "مسافر", "منتظر"
  ],
  "final_answer": "منتظر"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "A person/man in the scene, suggesting the Arabic component من.",
    "The scene is at dawn/sunrise over clouds, suggesting فجر.",
    "Combining من + فجر yields the 5-letter word منفجر."
  ],
  "candidates": [
    "مسافر", "مغامر", "منفجر"
  ],
  "final_answer": "منفجر"
}
```

### Iterative refinement

Task: Solve the picture word puzzle from the image.

Target language: Arabic.

Hint type: char\_count.

Ground Truth

دبابة



Model: gemini-2.5-pro

```
"attempts": [  
  {  
    "attempt_idx": 0,  
    "model_output": {  
      "primary_clues": [  
        "(bear) دب", "female as personified (door) باب"  
      ],  
      "candidates": [  
        "دبابة", "دبّوب", "باب الدب"  
      ],  
      "final_answer": "دبابة"  
    }  
  }  
]
```

Model: grok-4.1-fast

```
"attempts":  
{  
  "attempt_idx": 0,  
  "model_output": {  
    "primary_clues": [  
      "دب", "باب"  
    ],  
    "candidates": [  
      "دب باب", "دبّاب", "دبابة"  
    ],  
    "final_answer": "دبابة"  
  }  
}
```

### Iterative refinement

Task: Solve the picture word puzzle from the image.

Target language: Arabic.

Hint type: char\_count.

Ground Truth

بيتوتة



Model: gemini-2.5-pro

```
"attempts":
{
  "attempt_idx": 0,
  "model_output": {
    "primary_clues": [
      "توت", "بيت"
    ],
    "candidates": [
      "دار التوت", "بيروت", "بيت توت"
    ],
    "final_answer": "بيت توت"
  },
  "feedback": "The answer 'بيت توت' is incorrect."
},
{
  "attempt_idx": 1,
  "model_output": {
    "primary_clues": [
      "توت", "بيت"
    ],
    "candidates": [
      "دار التوت", "بيت توت", "بيروت"
    ],
    "final_answer": "بيروت"
  },
  "feedback": "The answer '' is incorrect."
},
{
  "attempt_idx": 2,
  "model_output": {
    "primary_clues": [
      "توت", "كوخ"
    ],
    "candidates": [
      "بيروت", "بيت توت", "كوخ توت"
    ],
    "final_answer": "كوخ توت"
  }
}
```

Model: grok-4.1-fast

```
"attempts":
{
  "attempt_idx": 0,
  "model_output": {
    "primary_clues": [
      "عنب", "بيت"
    ],
    "candidates": [
      "بيت العنب", "عنب بيت", "بيت عنب"
    ],
    "final_answer": "بيت عنب"
  },
  "feedback": "The answer 'بيت عنب' is incorrect."
},
{
  "attempt_idx": 1,
  "model_output": {
    "primary_clues": [
      "عنب", "كوخ"
    ],
    "candidates": [
      "قصر عنب", "دار عنب", "قبة عنب", "كوخ عنب"
    ],
    "final_answer": "كوخ عنب"
  },
  "feedback": "The answer 'كوخ عنب' is incorrect."
},
{
  "attempt_idx": 2,
  "model_output": {
    "primary_clues": [
      "عنب", "دار"
    ],
    "candidates": [
      "دار العنب", "عنب دار", "دار عنب"
    ],
    "final_answer": "دار عنب"
  }
}
```

### Partial character reveal

Task: Solve the picture word puzzle from the image.

Target language: Arabic.

Hint type: shuffle\_chars.

Hint pattern: \_ر\_

#### Ground Truth

مرمر



### Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "A bitter substance (e.g., coffee), suggesting the Arabic word root 'مر'.",
    "Another bitter item (e.g., cucumber), reinforcing 'مر'.",
    "Repetition of 'مر' (bitter) indicated by the two bitter clues."
  ],
  "candidates": [
    "مرمر", "مرارة", "مرير"
  ],
  "final_answer": "مرمر"
}
```

### Model: grok-4.1-fast

```
{
  "primary_clues": [
    "علامة + بينهما", "خيار مر", "قهوة مر"
  ],
  "candidates": [
    "مرارة", "مرمر", "مر مر"
  ],
  "final_answer": "مرمر"
}
```

### Partial character reveal

Task: Solve the picture word puzzle from the image.

Target language: Arabic.

Hint type: shuffle\_chars.

Hint pattern: \_\_\_و\_

#### Ground Truth

سودان



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "Two containers/cans: علبتان", "Black color: أسود",
    "The hint pattern matches a 5-letter word with و as the second letter."
  ],
  "candidates": [
    "سودان", "سوداء", "سواد"
  ],
  "final_answer": "سوداء"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "اثنان", "لون أسود", "علب طلاء"
  ],
  "candidates": [
    "سواد", "سودان", "سوداء"
  ],
  "final_answer": "سودان"
}
```

### Basic setup

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Hint: The answer has 9 characters (excluding spaces).

Ground Truth

کلم بروکلی



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "Bruce Lee", "Cabbage"
  ],
  "candidates": [
    "کلم لی", "بروکلی", "کلم بروسلی"
  ],
  "final_answer": "کلم بروسلی"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "کلم", "بروس لی"
  ],
  "candidates": [
    "کلم بروکلی", "بروکلی", "کلم بروس"
  ],
  "final_answer": "کلم بروکلی"
}
```

Example: Basic setup

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Ground Truth

استحکام



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "دهان", "حرف S"
  ],
  "candidates": [
    "سخندان", "دهن سرویس", "اصفهان"
  ],
  "final_answer": "اصفهان"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "دهان", "حرف S"
  ],
  "candidates": [
    "استفهام", "استحکام", "استحصال"
  ],
  "final_answer": "استحکام"
}
```

**Example: Few-shot chain-of-thought prompting**

Task: Solve the picture word puzzle from the image.  
Target language: Persian (Farsi).  
Hint type: char\_count.



**Ground Truth**

دوربین

**Model: gemini-2.5-pro**

```
{
  "primary_clues": [
    "The image shows the English word 'Bin', which sounds like the Persian root 'بین' (bin), meaning 'to see'.",
    "The word 'Bin' is shown receding into the distance, with an arrow pointing to the furthest one, representing 'far' or 'دور' (dur) in Persian."
  ],
  "candidates": [
    "دوربین", "ریزبین", "بین راه"
  ],
  "final_answer": "دوربین"
}
```

**Model: grok-4.1-fast**

```
{
  "primary_clues": [
    "جاده/راه طولانی (دور)", "کلمه BIN"
  ],
  "candidates": [
    "دوربین", "دورراه", "بین راه"
  ],
  "final_answer": "دوربین"
}
```

**Example: Few-shot chain-of-thought prompting**

Task: Solve the picture word puzzle from the image.  
Target language: Persian (Farsi).  
Hint type: char\_count.



**Ground Truth**

دیباچه

Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "The image consists of 12 repeated pictures of a laughing face.",
    "The number 12 in Persian is 'Davazdah' (دوازده).",
    "The pictures are arranged in the shape of the letter 'D', which is the first letter of 'Davazdah'.",

    "The word 'Davazdah' (دوازده) has 6 characters, which matches the hint provided."
  ],
  "candidates": [
    "شادی", "بهروزها", "دوازده"
  ],
  "final_answer": "دوازده"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "صورت‌های خندان زیاد (بچه‌ها)", "حرف D (خوانش فارسی: «دی»)"
  ],
  "candidates": [
    "دیوانه", "دیباچه", "دیپچه"
  ],
  "final_answer": "دیباچه"
}
```

### Iterative refinement

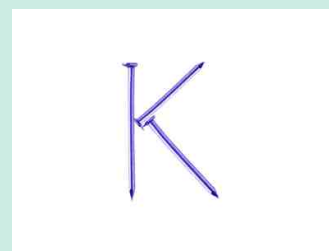
Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Ground Truth

کی می خوابی



Model: gemini-2.5-pro

```
"attempts":
{
  "attempt_idx": 0,
  "model_output": {
    "primary_clues": [
      "Nails, which in Persian are 'میخ', arranged as the letter 'K' / 'ک'.",
      "A violet/purple color, which in Persian is 'بنفش'."
    ],
    "candidates": [
      "میخک بنفشه", "میخک بنفش", "میخک"
    ],
    "final_answer": "میخک بنفشه"
  },
  "feedback": "The answer 'میخک بنفشه' is incorrect."
},
{
  "attempt_idx": 1,
  "model_output": {
    "primary_clues": [
      "The letter 'K'", "Nails 'میخ'", "Purple color 'بنفش'."
    ],
    "candidates": [
      "میخک ارغوانی", "کاوه آهنگر", "میخک بنفشه"
    ],
    "final_answer": "کاوه آهنگر"
  },
  "feedback": "The answer '' is incorrect."
},
{
  "attempt_idx": 2,
  "model_output": {
    "primary_clues": [
      "Nails 'میخ' arranged as the letter 'K' / 'ک'.", "A distinct purple tone associated with 'ارغوانی'."
    ],
    "candidates": [
      "میخک ارغوان", "کاوه آهنگر", "میخک بنفشه"
    ],
    "final_answer": "میخک ارغوان"
  }
}
}
```

Model: grok-4.1-fast

```
"attempts": [  
  {  
    "attempt_idx": 0,  
    "primary_clues": [  
      "میخ", "حرف K"  
    ],  
    "candidates": [  
      "الواح میخی", "نوشته میخی", "کتیبه میخی"  
    ],  
    "final_answer": "کتیبه میخی"  
  },  
  {  
    "attempt_idx": 1,  
    "primary_clues": [  
      "میخ = K", "کی"  
    ],  
    "candidates": [  
      "کی می خوابی", "کی می خ", "کی میخ"  
    ],  
    "final_answer": "کی میخ"  
  },  
  {  
    "attempt_idx": 2,  
    "primary_clues": [  
      "میخ (میخ) = «میخ...» شروع فعل «خوابی», "کی = K"  
    ],  
    "candidates": [  
      "کی می خوابن", "کی می خوابید", "کی می خوابی"  
    ],  
    "final_answer": "کی می خوابی"  
  }  
]
```

### Iterative refinement

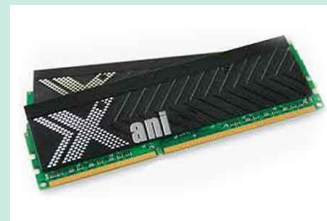
Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: char\_count.

Ground Truth

رمضانی



Model: gemini-2.5-pro

```
”attempts”:  
{  
  ”attempt_idx”: 0,  
  ”model_output”: {  
    ”primary_clues”: [  
      ”A RAM memory stick, which is associated with Persian 'رمل' (raml).”,  
      ”The text 'Xani', interpreted as a pun on Persian 'زنی'.”  
    ],  
    ”candidates”: [  
      ”رمل زنی”, ”حافظه خوانی”, ”رمالی”  
    ],  
    ”final_answer”: ”رمل زنی”  
  },  
  ”feedback”: ”The answer 'رمل زنی' is incorrect. Please Step-by-Step.”  
},  
{  
  ”attempt_idx”: 1,  
  ”model_output”: {  
    ”primary_clues”: [  
      ”A RAM stick (computer memory), i.e., Persian 'حافظه'.”,  
      ”The text 'Xani' as a pun on Persian 'خوانی' (reading), with 'X' used for the 'خ' sound.”  
    ],  
    ”candidates”: [  
      ”رمل زنی”, ”رم خوانی”, ”حافظه خوانی”  
    ],  
    ”final_answer”: ”حافظه خوانی”  
  },  
  ”feedback”: ”The answer 'حافظه خوانی' is incorrect. Please Step-by-Step.”  
},  
{  
  ”attempt_idx”: 2,  
  ”model_output”: {  
    ”primary_clues”: [  
      ”A RAM memory stick, sounding like Persian 'رمل'.”,  
      ”The text 'Xani' as a pun for Persian 'خوانی' (khani), using 'X' for 'خ'.”  
    ],  
    ”candidates”: [  
      ”رمل زنی”, ”حافظه خوانی”, ”رم خوانی”, ”رمل خوانی”  
    ],  
    ”final_answer”: ”رمل خوانی”  
  }  
}
```

**Model: grok-4.1-fast**

```
"attempts": [  
  {  
    "attempt_idx": 0,  
    "primary_clues": [  
      "(RAM) رم", "X"  
    ],  
    "candidates": [  
      "رمضانی", "رم ایکس", "ایکس رم"  
    ],  
    "final_answer": "ایکس رم"  
  },  
  {  
    "attempt_idx": 1,  
    "primary_clues": [  
      "الگوی آج/ طرح تایر = «ضانی»", "رم (RAM)"  
    ],  
    "candidates": [  
      "رمزانی", "رمزنی", "رمضانی"  
    ],  
    "final_answer": "رمضانی"  
  },  
  {  
    "attempt_idx": 2,  
    "primary_clues": [  
      "ضانی (از طرح/ آج)", "رم"  
    ],  
    "candidates": [  
      "رمبانی", "رمزانی", "رمضانی"  
    ],  
    "final_answer": "رمضانی"  
  }  
]
```

**Partial character reveal**

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: shuffle\_chars.

Hint pattern: \_ت\_ا\_

Ground Truth

استحکام



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "The letter 'S'", "An open mouth"
  ],
  "candidates": [
    "استدعا", "استعفا", "حرف اس"
  ],
  "final_answer": "استدعا"
}
```

Model: grok-4.1-fast

```
{
  "primary_clues": [
    "تصویر دهان/کام", "حرف S که در فارسی «اس» خوانده می‌شود",
    "ترکیب «اس» + «تحکام» برای ساخت واژه «استحکام»"
  ],
  "candidates": [
    "استوار", "استحکام‌دار", "استحکام"
  ],
  "final_answer": "استحکام"
}
```

### Partial character reveal

Task: Solve the picture word puzzle from the image.

Target language: Persian (Farsi).

Hint type: shuffle\_chars.

Hint pattern: \_م\_ف\_

Ground Truth

میکروفن



Model: gemini-2.5-pro

```
{
  "primary_clues": [
    "Mickey Mouse", "Telephone"
  ],
  "candidates": [
    "میکروفن", "معروف", "تلفن میکی"
  ],
  "final_answer": "میکروفن"
}
```

**Model: grok-4.1-fast**

```
{
  "primary_clues": [
    "ترکیب «میکی» + «فون» برای رسیدن به «میکروفن»، "فون", "میکی"
  ],
  "candidates": [
    "تلفن میکی", "میکروفون", "میکروفن"
  ],
  "final_answer": "میکروفن"
}
```