Almieyar-Oryx-BloomBench: A Bilingual Multimodal Benchmark for Cognitively Informed Evaluation of Vision-Language Models

Mohammad Mahdi Abootorabi[†§‡,1], **Omid Ghahroodi**[†,1], **Anas Madkoor**[†], **Marzia Nouri**[†], **Doratossadat Dastgheib**[†], **Ehsaneddin Asgari**[†,*]

[‡]University of British Columbia [§]Zuse School ELIZA
[†]Qatar Computing Research Institute (QCRI), Hamad Bin Khalifa University

[1]Equal contribution [*]Corresponding author: easgari@hbku.edu.qa

Abstract

Despite the rapid progress of Vision-Language Models (VLMs), the field lacks benchmarks that rigorously diagnose their true reasoning abilities and chart meaningful progress toward human-like multimodal intelligence. Most existing evaluations focus on piecemeal or disconnected tasks, obscuring critical cognitive weaknesses and providing little insight for targeted improvement. To address this gap, we introduce **BloomBench**, part of the *Almieyar* benchmarking series, the first cognitively human-grounded, bilingual (English–Arabic) multimodal benchmark for VLMs. Grounded in Bloom's Taxonomy, BloomBench systematically evaluates six levels of cognition (Remember, Understand, Apply, Analyze, Evaluate, Create) through carefully designed image–question–answer tasks. Built with a semi-automated pipeline and validated through a stratified hybrid quality assurance protocol, it ensures scalability, cultural inclusivity, and linguistic fidelity. Leveraging this framework, we conduct a comprehensive study of state-of-the-art VLMs to diagnose their cognitive profiles. Our analysis reveals a sharp cognitive asymmetry: while state-of-the-art models achieve strong performance ceilings in semantic understanding, they struggle substantially with factual recall and creative synthesis. This demonstrates that current general multimodal proficiency masks deeper limitations in specific cognitive layers. Furthermore, our study highlights a critical performance gap between Arabic and English, exposing limitations in current cross-lingual multimodal reasoning. These findings establish a foundation for developing more cognitively aligned and inclusive VLMs. The benchmark framework and dataset is available at: https://github.com/qcri/Almieyar-Oryx-BloomBench.

1 Introduction

Advances in transformer architectures (Vaswani et al., 2017), increasing computational resources, and the availability of large-scale training corpora (Naveed et al., 2024) have driven rapid progress in language modeling. Foundational Large Language Models (LLMs) (Ouyang et al., 2022; Grattafiori et al., 2024; Touvron et al., 2023; Qwen et al., 2025; Anil et al., 2023) now excel at tasks such as instruction following (Qin et al., 2024), reasoning (Wei et al., 2024), in-context learning (Brown et al., 2020), and multilingual translation (Zhu et al., 2024). Despite these advances, two broad limitations persist: First, the supply of high-quality and diverse text data is finite (Villalobos et al., 2024), which motivates research into data-efficient and extrapolative generalization methods (Li et al., 2024). Second, single-modality architectures are inherently limited when processing real-world information that spans modalities (e.g., text, images, and video) and requires reasoning about cross-modal relationships (Goodwin and Bjørndahl, 2018; Yin et al., 2024).

Vision-Language Models. The pursuit of artificial general intelligence (AGI), combined with the aforementioned limitations, has led to the development of Vision-Language Models (VLMs). By extending LLMs to process visual inputs (e.g., images and videos), VLMs gain a more comprehensive understanding of spatial relationships, objects, scenes, and abstract concepts (Li et al., 2025; Bordes et al., 2024; Liu et al., 2023; Team et al., 2024; Li et al., 2023a). Early work such as CLIP (Radford et al., 2021) catalyzed rapid advances in multimodal tasks that integrate visual and textual understanding, including visual question answering (Song et al., 2022), image captioning (Dai et al., 2023), embodied agents (Ma et al., 2024), and document understanding (Luo et al., 2024). Notably, models like GPT-4 (OpenAI et al., 2024) demonstrate human-level performance by jointly processing text and images, marking a significant milestone in multimodal AI. Nevertheless, important challenges remain: current VLMs still struggle

Findings of the Association for Computational Linguistics: ACL 2026, pages 28404–28436

with complex visual reasoning (Lu et al., 2022), object hallucination (Leng et al., 2024), fine-grained spatial understanding (Daxberger et al., 2025), and compositional reasoning (Sahin et al., 2024). These limitations underscore the need for a comprehensive evaluation that systematically assesses multimodal capabilities.

Multimodal Benchmarks. To systematically measure capabilities, LLMs and VLMs are typically evaluated on benchmarks: curated collections of tasks or questions designed to assess performance in domains such as mathematics (Wang et al., 2024a), programming (Yang et al., 2025), and biology (Justen, 2025; Phan et al., 2025). Significant progress has been made in designing domain-specific and large-scale benchmarks in recent years. However, conventional benchmarks have been criticized for relying on artificial datasets that fail to capture the complexity of human-level tasks (Zhong et al., 2024). As a result, high performance on one benchmark (e.g., reading comprehension) does not necessarily translate to proficiency in other cognitive skills (e.g., arithmetic), and low performance on a benchmark does not straightforwardly reveal general weaknesses. This paradigm encourages the development of models that learn narrow, "shortcut" solutions specific to a benchmark's statistical patterns rather than acquiring robust, generalizable abilities. This makes it exceptionally difficult to diagnose the underlying cognitive strengths and weaknesses of models or to identify concrete directions for improvement, hindering a thorough understanding of their true capabilities.

Bloom's Taxonomy. A promising direction for more informative evaluation is to align benchmarks with cognitive science frameworks, particularly Bloom's Taxonomy (Zhong et al., 2024; Huber and Niklaus, 2025). This taxonomy, originally developed by Bloom and later revised by Anderson and Krathwohl, organizes cognitive processes into a hierarchy ranging from lower- to higher-order skills (Remember, Understand, Apply, Analyze, Evaluate, and Create) (Adams, 2015; Wilson, 2016). By framing diverse multimodal tasks within this hierarchy, assessments can move beyond surface-level accuracy to measure deeper visual reasoning abilities of VLMs. This perspective motivates our design of a benchmark that evaluates VLMs across multiple cognitive levels rather than narrow task-specific metrics. Figure 1 visualizes the taxonomy and our high-level mapping.

Contributions. In this work, **(i)** we introduce ***BloomBench***, the first multimodal, bilingual benchmark for VLMs explicitly grounded in Bloom's Taxonomy, enabling comprehensive evaluation across multiple levels of cognitive complexity. **(ii)** We address the diagnostic limitations of existing benchmarks by designing tasks that measure the depth of VLM reasoning abilities, rather than just its performance on a set of disconnected tasks. **(iii)** We develop a scalable generation pipeline backed by a systematic LLM-as-a-judge and human validation study on a representative subset to ensure high data quality. **(iv)** We benchmark a diverse set of state-of-the-art VLMs, providing a taxonomy-driven analysis of their strengths, weaknesses, and open challenges, and incorporate both answer-based and likelihood-based evaluation methods to reveal discrepancies between model accuracy and confidence, exposing hidden reasoning gaps. **(v)** By incorporating a bilingual English-Arabic evaluation, our work challenges the anglocentric focus of current VLM benchmarks and enables a more inclusive assessment of how cognitive abilities generalize across diverse linguistic and cultural contexts. Together, these contributions establish a cognitively informed framework for assessing VLM progress and guiding future research.

2 Related Works

LLM Benchmarks. With the rapid growth of LLM research, benchmarks have become essential tools for tracking and comparing model capabilities. Widely adopted evaluations such as MMLU (Hendrycks et al., 2021b,a), BIG-Bench Hard (Suzgun et al., 2022), GSM8K (Cobbe et al., 2021), HumanEval (Chen et al., 2021), and GPQA (Rein et al., 2024) provide headline scores across heterogeneous tasks. Due to their scalability and objectivity, many of these benchmarks rely on multiple-choice or short-answer formats (Dua et al., 2019; Rajpurkar et al., 2016; Wang et al., 2018; Sarlin et al., 2020; Yang et al., 2018), which facilitate automated evaluation but limit diagnostic depth.

VLM Benchmarks. VLMs are increasingly applied in domains ranging from generative AI systems (Abootorabi et al., 2025a) and retrieval-augmented generation (RAG) (Abootorabi et al., 2025b), to education (Baral et al., 2025) and healthcare (Hartsock and Rasool, 2024). Despite progress, current models still face challenges in visual arithmetic (Huang et al., 2025), geometric problem-solving (Gao et al., 2023), and spatial

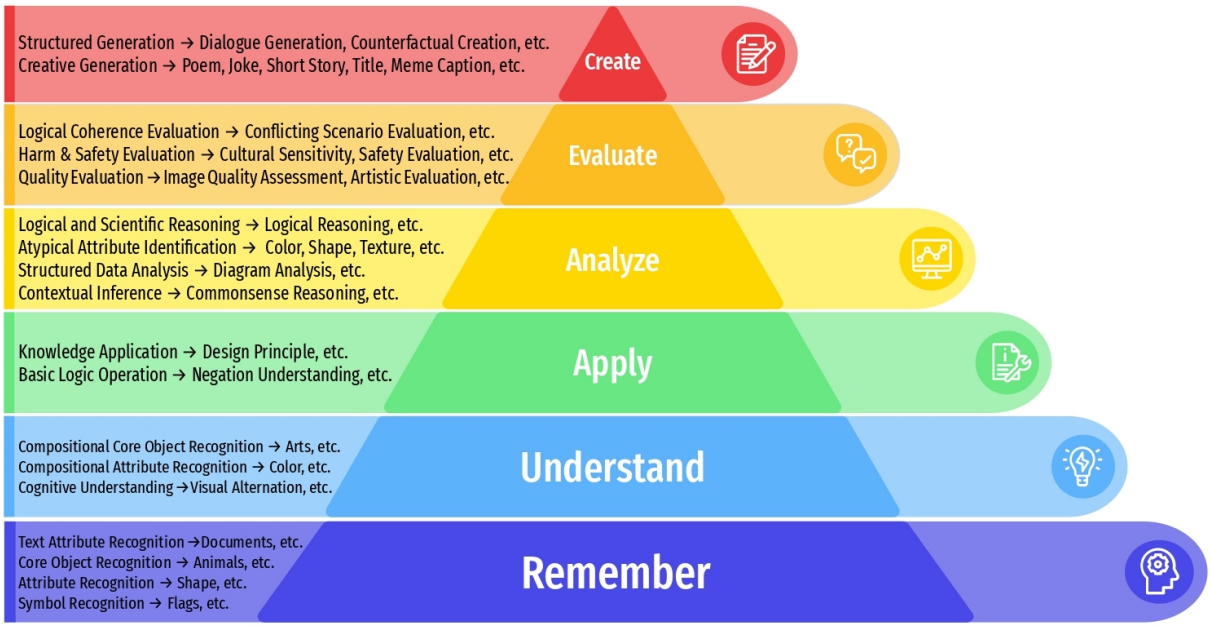

Figure 1: Hierarchical overview of the BloomBench Taxonomy. Grounded in Bloom's cognitive framework, this hierarchy organizes multimodal tasks across six levels of cognitive complexity. Each level is further decomposed into specific task families to enable fine-grained evaluation of VLM reasoning capabilities.

reasoning tasks such as orientation, relations, and navigation (Stogiannidis et al., 2025; Chen et al., 2024a). To evaluate performance in these areas, various benchmarks have been introduced.

Early evaluation efforts were largely task-specific, focusing on specific tasks such as visual question answering (VQA) (Schwenk et al., 2022), image captioning (Liu et al., 2021), and hallucination detection (Li et al., 2023b). More recent work has sought broader coverage and higher complexity. For example, VLM2-Bench (Zhang et al., 2025) evaluates fine-grained cue association across nine subtasks and 3,000 test cases, while revealing persistent weaknesses in visual grounding. MMMU (Yue et al., 2024) combines multiple-choice and open-ended formats to assess perception, knowledge, and reasoning. MMT-Bench (Ying et al., 2024) spans 32 expert-level tasks requiring reasoning and localization. Other efforts focus on spatial reasoning specifically (Stogiannidis et al., 2025; Chen et al., 2024a). Together, these benchmarks highlight important advances but still leave gaps in systematically evaluating higher-order reasoning.

Arabic-English VLM Benchmarks. While the VLM landscape has been predominantly English-centric, a growing body of work is developing resources for other languages, particularly Arabic. Initial efforts focused on translating existing English corpora, such as Violet for image captioning (Mohamed et al., 2023) and AraCLIP for Arabic image-text alignment (Al-Barham et al., 2024). The availability of Arabic data was further expanded by large multilingual datasets such as WIT (Srinivasan et al., 2021) and PALI (Ahmadi et al., 2023). More recently, research has advanced towards creating culturally and dialectally aware benchmarks. For instance, the Peacock model family was introduced with Henna, a benchmark assessing understanding of Arabic culture (Alwajih et al., 2024), and CAMEL-Bench (Ghaboura et al., 2024) provides a comprehensive suite for domains ranging from handwritten document understanding to medical imaging. These efforts highlight the importance of inclusive, culturally grounded evaluation resources, particularly as dedicated Arabic-centric platforms such as Fanar continue to emerge (Team et al., 2025a; Abbas et al., 2026). Our benchmark contributes to this line of work by offering bilingual evaluation (Arabic and English), with quality rigorously validated through an LLM-as-a-judge framework (Zheng et al., 2023) using Gemini 3 Pro and human to ensure high linguistic fidelity and cognitive alignment.

Human Cognition-based Benchmarks. To address the diagnostic limitations of task-based evaluations, researchers increasingly turn to cognitive science frameworks. In text-only settings, Bloom's Taxonomy helps analyze LLMs' capabilities: Huber and Niklaus (2025) map popular benchmarks to the taxonomy's six levels, finding that evaluations

concentrate on mid-level skills (Apply, Analyze) while under-representing foundational (Remember) and higher-order (Evaluate, Create) skills. Beyond general analysis, recent works have operationalized the taxonomy for domain-specific evaluation. BloomAPR (Ma et al., 2025) is a dynamic framework for automated program repair that transforms static benchmarks into hierarchical tasks; their findings reveal that while models effectively memorize fixes (Remember), they struggle significantly with higher-order analysis and transfer in real-world coding contexts. Similarly, BLOOMQA (Chen et al., 2026) proposed a framework for generating benchmarks from expert guidelines in practice-based domains (e.g., teaching, dietetics), observing that LLMs occasionally exhibit non-intuitive behavior by outperforming on higher-order reasoning (Analyze) while failing foundational recall tasks. Educators also employ the taxonomy to design assessments that span the full skill spectrum (Elkins et al., 2024). However, these efforts remain restricted to text and code modalities, lacking a comprehensive framework to evaluate cognitive depth in vision-language processing.

Beyond Bloom's Taxonomy, ToMBench (Chen et al., 2024b) evaluates LLMs on theory-of-mind reasoning tasks, highlighting important aspects of social cognition but remaining limited to text-only settings. More recently, Weng et al. (2025) proposed a multimodal framework grounded in psychological faculties such as perception, attention, and memory. While valuable, their approach lacks a cumulative hierarchy, making it difficult to assess reasoning depth in a systematic manner. These limitations highlight that their benchmark and findings are narrow in domain (synthetic objects, controlled settings) and lack linguistic and cultural diversity. In contrast, our work is explicitly grounded in Bloom's Taxonomy, which provides a hierarchical structure for evaluating VLMs and enables a fine-grained diagnosis of cognitive abilities across the full spectrum of complexity.

3 BloomBench Methodology

3.1 Design Principles

Our goal is to design a benchmark capable of evaluating different hierarchical levels of cognitive reasoning in VLMs. Translating an abstract educational theory, such as Bloom's Taxonomy, into a concrete, measurable, and rigorous multimodal evaluation framework requires a principled design process. After a careful review of prior works and an in-depth analysis of Bloom's revised taxonomy, we define a hierarchical taxonomy of multimodal tasks, where each cognitive level (e.g., Remember, Understand, Apply, Analyze, Evaluate, Create) is further decomposed into sub-levels and specific task types. This structure allows for a comprehensive and fine-grained assessment of VLM capabilities. The design of BloomBench is guided by the following principles:

Cognitive Completeness. We aim for a holistic evaluation by providing comprehensive coverage across the entire spectrum of Bloom's Taxonomy. Unlike benchmarks that focus on a narrow range of skills, BloomBench is explicitly designed to assess the full depth of a model's multimodal reasoning, from foundational abilities like remembering and understanding to the highest levels of evaluating and creating. This principle ensures not only comprehensive coverage but also interpretability in understanding how different reasoning abilities contribute to overall multimodal cognition.

Hierarchical Dependence. Bloom's Taxonomy is a cumulative hierarchy in which higher-order reasoning builds on more basic skills. We adopt this structure as an organizing scaffold for BloomBench: advanced tasks are designed to presuppose competence in foundational ones. While this mirrors human cognition, we do not assume VLMs will follow the same progression. Instead, the hierarchy offers a principled way to assess where models align with or diverge from expected cognitive trajectories.

VLM-Specificity. We target core multimodal abilities over text-only reasoning with incidental visual context. Tasks require models to ground language in visual evidence, reason about spatial relations and orientation, understand visual compositionality of objects and attributes, and link abstract concepts to perceptual cues (e.g., numeracy, causality). This design isolates vision–language competence and reduces reliance on textual shortcuts.

Real-World Relevance and Scenario Diversity. A key principle is to ground our evaluation in realistic, context-rich scenarios that reflect genuine human perceptual and reasoning challenges. To achieve this, our tasks draw from diverse, authentic web-sourced images covering everyday, abstract, and domain-specific contexts. Such diversity ensures that models are tested on the variability and ambiguity present in natural visual environments, conditions under which human cognition operates. By moving beyond synthetic or overly controlled

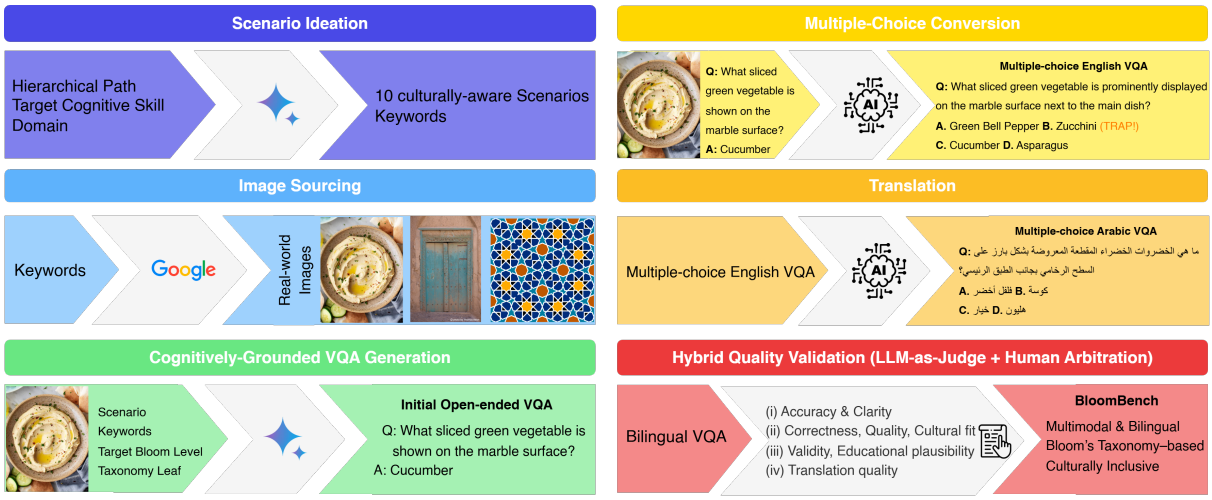

Figure 2: Overview of the BloomBench data generation pipeline. The process combines scenario ideation, cognitively-grounded VQA generation, multiple-choice conversion, translation, and hybrid quality validation of the representative subset to ensure high-quality, culturally relevant benchmark items across all Bloom's levels.

settings, we assess a model's ability to generalize its cognitive skills to complex, real-world visual understanding.

Scalability and Transparency. Another core design principle of BloomBench is to ensure a scalable yet methodologically rigorous construction process. To this end, we adopt a semi-automated pipeline supported by a hybrid validation layer. This approach allows us to leverage the efficiency of automated methods for initial task generation while ensuring reliability through statistically grounded quality assurance. By employing state-of-the-art reasoning models to validate a stratified representative subset, audited by humans, we establish high-confidence quality baselines without the bottleneck of exhaustive manual verification. The result is a benchmark that is not only large-scale but also transparent, statistically validated, and reproducible by design.

3.2 Task Design for Each Cognitive Level

BloomBench operationalizes the cognitive hierarchy through six task families, each corresponding to a level in Bloom's Taxonomy. These families are designed to be VLM-specific, testing the progressive development of multimodal reasoning from perception to creative synthesis. We outline the objectives and representative task types for each level. A summary of the taxonomy is shown in Figure 1, with the complete task list in Appendix (§A).

Remember. This foundational level assesses perceptual recognition and factual recall. Tasks are grouped into several key areas: *Core Object Recognition* (e.g., animals, scenes), *Attribute Recognition* (e.g., color, texture), *Activity Recognition* (e.g., interactions, professions), and the identification of symbolic and textual information via *Symbol Recognition* (e.g., logos, traffic signs) and *Text Attribute Recognition*.

Understand. Moving beyond recognition, this level probes comprehension of relationships and compositional meaning. The main subcategories are *Compositional Recognition*, which tests the understanding of multiple objects and their attributes within a scene, and *Cognitive Understanding*, which requires interpreting more abstract concepts like emotions, semantic knowledge, and visual paraphrasing.

Apply. This level tests the ability to use learned knowledge in novel visual contexts. Tasks fall into two primary groups: *Knowledge Application*, which involves applying external concepts, such as mathematical formulas or scientific principles, to a visual input; and *Basic Logic Operations*, which test the understanding of negation, word order, and coordination in a multimodal context.

Analyze. Analytical tasks require a model to deconstruct a scene to infer relationships and patterns. This family is broken down into four key reasoning types: *Logical and Scientific Reasoning*, *Contextual Inference* (e.g., resolving ambiguity or pronouns), *Structured Data Analysis* (e.g., interpreting charts and tables), and *Atypical Attribute Identification* (e.g., spotting unusual colors or shapes).

Evaluate. This level measures a model's capacity for judgment and critical assessment. Tasks require making and justifying evaluations, organized into three core areas: *Logical Coherence Evalua-*

tion (e.g., detecting hallucinations), *Harm & Safety Evaluation* (e.g., identifying cultural insensitivity or toxicity), and *Quality Evaluation* (e.g., assessing the artistic or technical quality of an image).

Create. At the highest level, tasks assess the ability to synthesize information to generate novel content. We operationalize this in an MCQ format as *discriminative creativity*: requiring the model to evaluate potential syntheses (e.g., poem endings) and identify the one best satisfying complex constraints (rhyme, narrative coherence) over flawed distractors. This measures latent creative judgment, a critical prerequisite for generative ability, across *Creative Generation* (e.g., storytelling) and *Structured Creation* (e.g., designing experiments).

3.3 Data Generation Pipeline

To construct BloomBench, we designed a scalable, semi-automated data generation pipeline that integrates LLMs with rigorous human oversight. This multi-stage process, illustrated in Figure 2, ensures that each benchmark item is cognitively grounded, visually relevant, and bilingually validated. The pipeline leverages prompt engineering techniques (Sahoo et al., 2024) and an agentic design framework (Plaat et al., 2025) to maintain both efficiency and quality throughout the generation process. Prompts used in the data generation pipeline are provided in the Appendix (§F).

Scenario Ideation and Image Sourcing. The pipeline begins by generating a diverse set of visual scenarios for each leaf node in the BloomBench taxonomy. A high-capacity language model (Gemini 2.5 Pro Comanici et al., 2025) is prompted with the complete hierarchical path, target cognitive skill, and domain. For each skill, the model produces ten culturally aware scenarios, including Western, MENA, and Arabic contexts, alongside carefully selected keywords that result in visually concrete images and minimize textual elements. These keywords are then used to source authentic, real-world images from the web, ensuring diversity and contextual relevance throughout the dataset.

Cognitively-Grounded VQA Generation. For each sourced image, an initial open-ended visual question–answer (VQA) pair is generated. We prompt the same LLM used in the previous step with the image, the full scenario context and keywords, and a detailed description of the target Bloom's level and taxonomy leaf. The prompt is carefully crafted to ensure the resulting question is answerable solely through the visual content, reinforcing alignment with the intended cognitive skill and minimizing reliance on external knowledge or textual cues.

Multiple-Choice Conversion and Translation. Each open-ended VQA pair is then converted into a high-quality multiple-choice question (MCQ) with four answer options. A separately instruction-tuned model is prompted to act as a professional test creator, generating three plausible distractors in addition to the correct answer, with one deliberately crafted as a "trap" distractor to probe for deeper understanding. The complete MCQ is then translated into Modern Standard Arabic using the same LLM, with special care to maintain the item's semantic and cognitive integrity across both languages.

Quality Validation. To ensure benchmark robustness, we implemented a *Hybrid Quality Validation* protocol. First, we employed an LLM-as-a-judge phase to filter out invalid or nonsensical samples. Following this, we conducted a statistically grounded validation on a representative subset of 969 samples ($\approx$1/8 of the dataset). This sampling was stratified to include at least four random examples from every one of the 106 taxonomy leaf nodes. We utilized Gemini 3 Pro, leveraging its state-of-the-art visual reasoning capabilities, to audit sample quality. The model identified only 15 samples as potentially incorrect. Subsequent human verification of the selected representative subset confirmed that only these flagged cases were indeed errors (see Appendix §G for annotation protocol). This establishes a 98.45% quality rate, providing strong statistical confidence in the dataset's fidelity.

3.3.1 Dataset Statistics

BloomBench contains 7,747 bilingual (English–Arabic) image–question–answer pairs across 106 distinct task types, spanning all six levels of Bloom's Taxonomy and providing hierarchically comprehensive coverage from basic perceptual recall to high-level creative reasoning. Specifically, the dataset includes 2,948 samples for *Remember*, 1,592 for *Understand*, 499 for *Apply*, 1,431 for *Analyze*, 592 for *Evaluate*, and 685 for *Create*. Overall, BloomBench offers a cognitively structured benchmark for evaluating VLM reasoning across multiple levels of complexity. Full statistics are provided in Appendix (§B), and representative examples are shown in Appendix (§E).

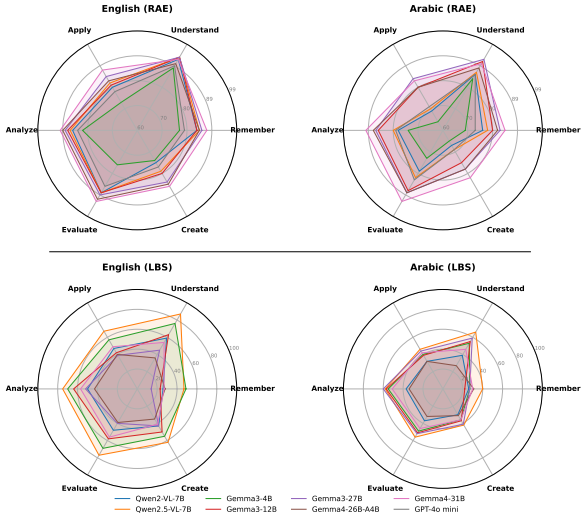

Figure 3: Performance of different Vision-Language Models on the six BloomBench cognitive levels. The charts compare results for English (left column) and Arabic (right column) using two evaluation methods: Regex-based Answer Extraction (RAE) (top row) and Likelihood-based Scoring (LBS) (bottom row).

4 Evaluation

We evaluate the performance of several representative state-of-the-art open- and closed-source VLMs on BloomBench and analyze their capabilities across different tasks and cognitive levels. We describe the evaluation setup and present several quantitative results.

Setup We evaluate a selection of representative open-source models, including Gemma 3 (4B, 12B, and 27B) (Team et al., 2025b), Gemma 4 (26B-A4B and 31B) (Farabet and Lacombe, 2026), Qwen2.5-VL-7B (Qwen et al., 2025), and Qwen2-VL-7B (Wang et al., 2024b). We also include the closed-source model GPT-4o mini (OpenAI et al., 2024) to benchmark against state-of-the-art proprietary systems. By including multiple sizes for Gemma 3 (Team et al., 2025b), we aim to analyze the impact of model scale on cognitive reasoning abilities. All models are prompted with zero-shot instructions. For inference, we set the decoding temperature to 0. Following prior work in robust VLM evaluation (Ghahroodi et al., 2025, 2024), we employ two distinct answer extraction techniques: *Regex-based Answer Extraction* and *Likelihood-based Scoring*. This dual approach allows us to assess both the model's explicitly generated output and its underlying confidence distribution over the answer choices.

Regex-based Answer Extraction (RAE). This method (abbreviated as *RAE*) parses the model's free-form text output to identify the selected answer choice (e.g., "A", "B", "C", or "D"). This approach simulates real-world usage where a user interprets the generated response directly. In cases where the model fails to produce a valid format, we assign a wrong choice to establish a baseline and account for catastrophic failures in instruction following.

Likelihood-based Scoring (LBS). This method (abbreviated as *LBS*) offers a more principled evaluation by directly querying the model's internal probability distribution over the possible answers, removing the dependency on output formatting. For each multiple-choice question, we compute the conditional log-probability of each answer choice given the image I and the question Q. For each choice $C_i \in \{C_A, C_B, C_C, C_D\}$, we compute a score based on the conditional probability of the choice's token sequence $(w_1, ..., w_k)$ given the context:

$$\text{Score}(C_i) = \sum_{j=1}^{k} \log P(w_j | I, Q, w_1, ..., w_{j-1}) \quad (1)$$

To ensure a fair comparison between choices of different lengths, we normalize this score by the number of tokens in the choice:

$$\text{NormalizedScore}(C_i) = \frac{1}{k} \text{Score}(C_i) \quad (2)$$

The model's prediction is then the choice with the highest normalized score. This approach offers a more robust evaluation of the model's internal knowledge, independent of its ability to format the final answer:

$$\text{Answer} = \underset{i \in \{A,B,C,D\}}{\arg\max} \ \text{NormalizedScore}(C_i) \quad (3)$$

This likelihood-based approach offers a more direct measure of the model's confidence in its answers, sidestepping errors from parsing, formatting, or exploiting alternative choices, and exposing reasoning signals that most conventional benchmarks overlook.

Evaluation Metrics. For both methods, we report accuracy as the primary metric, calculated as the fraction of correctly answered questions:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{prediction}_i = \text{ground\_truth}_i) \quad (4)$$

where N represents the total number of test samples, $\mathbb{I}(\cdot)$ denotes the indicator function, and prediction_i and ground_truth_i correspond to the model's prediction and ground truth for the i-th sample, respectively.

Model	English (Accuracy ↑)				Arabic (Accuracy ↑)			
	RAE		LBS		RAE		LBS	
	Micro	*Macro*	*Micro*	*Macro*	*Micro*	*Macro*	*Micro*	*Macro*
Qwen2-VL-7B	0.854	0.845	0.421	0.455	0.773	0.758	0.326	0.335
Qwen2.5-VL-7B	0.869	0.860	**0.654**	**0.692**	0.792	0.777	**0.503**	**0.513**
Gemma3-4B	0.796	0.785	0.609	0.627	0.729	0.715	0.408	0.433
Gemma3-12B	0.867	0.860	0.450	0.500	0.836	0.835	0.398	0.435
Gemma3-27B	0.883	0.881	0.336	0.387	0.859	0.856	0.440	0.472
Gemma4-26B-A4B	0.876	0.877	0.347	0.368	0.846	0.843	0.309	0.312
Gemma4-31B	**0.898**	**0.898**	0.430	0.473	**0.876**	**0.875**	0.397	0.418
GPT-4o mini	0.824	0.823	N/A	N/A	0.769	0.768	N/A	N/A

Table 1: **Overall accuracy (↑) of VLMs on BloomBench.** We report both **Micro** (standard) accuracy and **Macro** (balanced) accuracy to account for class imbalance across cognitive levels. N/A: Closed-source models do not support LBS calculation.

5 Discussion

The overall results of all evaluated models on BloomBench using both evaluation methods are presented in Table 1, while Table 2 and Figure 3 detail their performance across the six cognitive levels of our taxonomy. Task-specific results are provided in Appendix (§H). The key findings are summarized below.

Insights and Challenges from BloomBench. BloomBench presents substantial challenges. While Gemma 4 31B achieves the state-of-the-art performance in Regex-based accuracy (89.8% English / 87.6% Arabic), overtaking Qwen2.5-VL, it notably struggles under the LBS evaluation. Performance on Arabic tasks generally lags behind English; however, the Gemma 3 family demonstrates remarkable cross-lingual consistency, with the 27B model showing a minimal performance drop between languages.

Coverage Comparison with Existing Benchmarks. To concretely illustrate BloomBench's diagnostic value relative to existing resources, we examined MMMU (Yue et al., 2024) as an established example of current evaluation standards. Specifically, we mapped 1,080 of its samples onto the BloomBench taxonomy using Gemini 3 Flash (Google DeepMind, 2025) as a judge. The results highlight a pronounced disparity: the *Analyze* level alone accounts for 66.4% of MMMU's coverage, driven almost entirely by Math Reasoning (344 samples), Table Analysis (85), and Chart Analysis (68), while the *Create* and *Evaluate* levels combined represent under 1.1% of the dataset. Forty-five taxonomy leaf nodes, such as Ambiguity Resolution, Toxicity Detection, and Dialogue Generation, have zero representation in MMMU. This confirms that while MMMU excels at evaluating expert domain knowledge and analytical reasoning, it cannot serve as a proxy for the full spectrum of multimodal cognition. A complete breakdown is provided in Appendix (§D).

Comparison Between RAE and LBS Evaluation. Previous benchmarks typically rely on RAE, but our experiments reveal that LBS exposes deeper weaknesses. While critics might attribute LBS difficulty to scoring artifacts, the disparate impact across models contradicts this. While Qwen2.5-VL maintains relative stability (0.869 RAE $\rightarrow 0.654$ LBS), the Gemma 3 family exhibits an inverse scaling trend in LBS. Most strikingly, Gemma 3 27B achieves the highest RAE accuracy (0.883) yet suffers the most severe drop in LBS (0.336). This divergence suggests that Gemma3 relies more on surface-level pattern recognition, while the Qwen family demonstrates superior internal consistency. Thus, the two metrics provide complementary perspectives: RAE simulates real-world usage, while LBS validates the model's underlying reasoning confidence.

Cross-Linguistic Performance. As expected, performance is consistently higher in English than in Arabic. We note that this gap is amplified by tokenization bias in LBS: Arabic's higher morphological fertility leads to more tokens per word, disproportionately penalizing the length-normalized score (Equation 2). To rigorously disentangle genuine reasoning gaps from metric sensitivity, we conducted a controlled ablation using Spanish, a high-resource language with distinct tokenization properties from English, confirming that LBS scores drop significantly even when RAE remains near-parity with English, validating that the Arabic LBS gap reflects a compound effect of tokenization fertility and lower non-English probability priors (see Appendix (§C)). Despite this, the Gemma-3 family shows the smallest drop when shifting to Arabic, reflecting stronger cross-lingual generalization (Team et al., 2025b), while Qwen models exhibit more pronounced declines.

Model Size Effect on Performance. We observe a distinct decoupling of metrics as model size increases. Under RAE, performance scales predictably with size (27B > 12B > 4B). However, under LBS, we observe an inverse scaling phenomenon within the Gemma 3 family, where the largest 27B model yields the lowest likelihood scores. This suggests that increasing model scale and instruction tuning intensity may improve deterministic answer generation at the cost of the raw probabilistic calibration required for LBS.

Model	English (Accuracy ↑)												Arabic (Accuracy ↑)											
	Remember		Understand		Apply		Analyze		Evaluate		Create		Remember		Understand		Apply		Analyze		Evaluate		Create	
	RAE	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**	**RAE**	**LBS**
Qwen2-VL-7B	0.84	0.26	0.93	0.59	0.80	0.47	0.86	0.50	0.89	0.48	0.75	0.43	0.76	0.26	0.86	0.39	0.69	0.32	0.78	0.37	0.79	0.37	0.67	0.30
Qwen2.5-VL-7B	0.85	0.47	0.94	**0.87**	0.82	**0.67**	0.87	**0.75**	0.89	**0.77**	0.79	**0.62**	0.78	**0.40**	0.87	**0.66**	0.70	**0.46**	0.80	0.58	0.82	**0.56**	0.69	**0.42**
Gemma3-4B	0.77	**0.49**	0.89	0.76	0.73	0.57	0.82	0.70	0.76	0.69	0.74	0.55	0.69	0.27	0.84	0.53	0.64	0.40	0.74	0.55	0.73	0.49	0.65	0.36
Gemma3-12B	0.84	0.23	0.94	0.63	0.81	0.42	0.88	0.64	0.89	0.58	0.80	0.50	0.80	0.22	0.92	0.55	0.80	0.39	0.86	0.57	0.88	0.51	0.75	0.37
Gemma3-27B	0.86	0.14	**0.94**	0.45	0.85	0.39	0.89	0.52	0.90	0.40	0.84	0.43	0.83	0.27	**0.93**	0.59	**0.84**	0.44	0.87	**0.60**	0.89	0.52	0.78	0.41
Gemma4-26B-A4B	0.85	0.28	0.91	0.36	0.83	0.40	0.90	0.43	0.92	0.39	0.85	0.35	0.82	0.31	0.89	0.27	0.80	0.32	0.88	0.34	0.89	0.32	0.78	0.31
Gemma4-31B	**0.88**	0.27	0.93	0.54	**0.88**	0.49	**0.91**	0.57	**0.93**	0.56	**0.86**	0.41	**0.85**	0.30	0.91	0.46	0.83	0.42	**0.91**	0.51	**0.93**	0.46	**0.82**	0.36
GPT-4o mini	0.79	N/A	0.90	N/A	0.78	N/A	0.84	N/A	0.86	N/A	0.77	N/A	0.73	N/A	0.86	N/A	0.71	N/A	0.79	N/A	0.83	N/A	0.69	N/A

Table 2: Accuracy (↑) of VLMs across BloomBench cognitive levels. Results reported for Regex-based Answer Extraction (RAE) and Likelihood-based Scoring (LBS). Bold denotes the best results for each metric–language pair. N/A: Closed-source models do not support LBS calculation.

Cognitive-Level Analysis. Performance trends across Bloom's hierarchy reveal non-linear patterns. Models demonstrate near-ceiling proficiency in *Understand* and *Evaluate* (achieving > 0.88 RAE), indicating that discriminative visual reasoning is highly advanced in state-of-the-art VLMs. However, this competence does not transfer to generative tasks, with performance degrading significantly on *Apply* and *Create*. Surprisingly, *Remember* also performs poorly under LBS. This discrepancy reflects current VLM training biases: models are optimized for semantic association rather than factual recall or creative synthesis. Thus, BloomBench's diagnostic value lies in quantifying this cognitive asymmetry, revealing that high discriminative accuracy often masks deeper deficiencies in precise reasoning and generation.

Cognitive-Level Cross-Linguistic Analysis. We observe distinct cross-lingual trends based on the RAE metric. First, the *Understand* level exhibits the strongest alignment across languages, showing the lowest average degradation, indicating that core semantic comprehension is highly transferable. In contrast, across all evaluated model families, we observe a consistent and pronounced performance degradation in the *Create* level when shifting from English to Arabic. This trend persists even in the Gemma 3 family, which is otherwise renowned for its multilingual capabilities, highlighting that while semantic understanding transfers effectively across languages, the high-order generative synthesis required for creation remains a significant cross-lingual bottleneck. Beyond creation, we identify a sharp divergence in procedural reasoning (*Apply*). The Qwen families and GPT-4o mini suffer substantial drops in this category (e.g., 12% degradation for Qwen2.5-VL). This failure in procedural reasoning is further corroborated by Likelihood-based Scoring (LBS): Qwen2.5-VL, which struggled in RAE, shows an even steeper decline in likelihood confidence ($0.67 \rightarrow 0.46$) for *Apply*, confirming that its procedural failures stem from fundamental reasoning gaps rather than formatting errors. Conversely, the larger Gemma models (12B and 27B) maintain remarkable stability in this category, with negligible performance loss. However, this robustness is strictly scale-dependent in these models: while Gemma 3 12B and 27B effectively bridge the cross-lingual gap, the 4B variant suffers significant losses across all cognitive levels (e.g., 9% drop in *Apply* in comparison to 1% drop of 12B and 27B in the same category). This underscores that robust cross-lingual alignment in complex reasoning tasks, specifically procedural application and creative synthesis, is likely an emergent property of model scale.

6 Conclusion

We present BloomBench, a cognitively informed, bilingual benchmark designed to evaluate VLMs through the hierarchical lens of Bloom's Taxonomy. By integrating cognitive theory with a semi-automated, hybrid-verified data pipeline, BloomBench enables systematic and interpretable assessment of multimodal reasoning across six levels of cognition. Our analyses reveal both the strengths and persistent weaknesses of current VLMs, underscoring the value of cognition-driven evaluation for guiding model development. Future work should expand this framework with more challenging and diverse task types, particularly at higher cognitive levels where complex reasoning extends beyond the current MQA format, and explore adaptive difficulty scaling to better capture the evolving multimodal reasoning abilities of next-generation models.

7 Limitations

While this work introduces a comprehensive, cognitively grounded benchmark for evaluating Vision-Language Models, some limitations should be acknowledged. First, computational resource constraints, including limited access to large-scale

GPU infrastructure and the high costs of proprietary model APIs, restricted the breadth of models we could evaluate. Consequently, several noteworthy VLMs from recent literature were not included in our experiments. We attempted to ensure diversity by selecting models spanning different architectural families and scales, but future work should expand the evaluation to encompass a broader range of state-of-the-art systems to provide more comprehensive insights into the current landscape of VLM capabilities. Moreover, because our pipeline relies on agent-driven generation over rich public web data, the benchmark can naturally improve as stronger agents (e.g., Gemini 3 Pro) and more compute become available. Access to newer high-performing models and larger GPU resources would enable future iterations of BloomBench to produce higher-quality and more accurate multimodal items.

Second, all questions in BloomBench are presented in multiple-choice format. While this design choice facilitates standardized evaluation and enables robust automated scoring, it may not fully capture the range of reasoning abilities required in open-ended, real-world scenarios. Incorporating additional question formats, such as fill-in-the-blank, short-answer generation, or multi-step reasoning tasks, could provide richer diagnostic information about model capabilities across cognitive levels. Future iterations of the benchmark could explore these alternative formats to complement the current evaluation framework.

Finally, while we conducted comprehensive validation on a stratified subset ($N{\approx}1k$), we did not manually verify every item in the full dataset. However, the high agreement rate ($> 97\%$) across all 106 taxonomy nodes provides strong statistical confidence in the benchmark's overall reliability.

8 Ethical Considerations.

Data Provenance and Licensing. The images in BloomBench are sourced from publicly available repositories. To strictly adhere to copyright laws and fair use principles, we do not host or redistribute the image files directly. Instead, we release the dataset as a collection of image URLs accompanied by a download script. This ensures that users retrieve content from the original sources, a standard practice in recent vision-language research to respect the intellectual property and distribution rights of content creators (Deitke et al., 2025).

Content Safety. Given that visual data is scraped from the web, we implemented a strict filtering pipeline. We utilized both automated safety classifiers and manual inspection to remove any images containing offensive or violent content. We believe this dataset is safe for research use; however, users should exercise standard caution when utilizing web-crawled data.

Broader Impact. BloomBench aims to advance the cognitive evaluation of LVLMs. We anticipate this will help the community move beyond surface-level pattern matching toward genuine multimodal reasoning. We do not foresee immediate negative societal impacts, but we encourage researchers to use this benchmark to identify and mitigate failures in safety-critical reasoning applications.

Acknowledgments

We would like to express our sincere gratitude to Mohamed Hefeeda, lead of the Fanar Multimodal Team at the Qatar Computing Research Institute, for his invaluable feedback and guidance throughout this work. We also extend our thanks to Hamza Aldaghstany, Mohammad Amin Sadeghi, and Mohamed Eltabakh, as well as the broader Fanar team, for their contributions and support in developing this benchmark. Their collective efforts were instrumental in bringing BloomBench to this stage. The first author was also supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

Ummar Abbas, Mohammad Shahmeer Ahmad, Minhaj Ahmad, Abdulaziz Al-Homaid, Anas Al-Nuaimi, Enes Altinisik, Ehsaneddin Asgari, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, and 1 others. 2026. Fanar 2.0: Arabic generative ai stack. *arXiv preprint arXiv:2603.16397*.

Mohammad Mahdi Abootorabi, Omid Ghahroodi, Pardis Sadat Zahraei, Hossein Behzadasl, Alireza Mirrokni, Mobina Salimipanah, Arash Rasouli, Bahar Behzadipour, Sara Azarnoush, Benyamin Maleki, and 1 others. 2025a. Generative ai for character animation: A comprehensive survey of techniques, applications, and future directions. *arXiv preprint arXiv:2504.19056*.

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025b.

Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Nancy E Adams. 2015. Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, 103(3):152.

Sina Ahmadi, Milind Agarwal, and Antonios Anastasopoulos. 2023. Pali: A language identification benchmark for perso-arabic scripts. *arXiv preprint arXiv:2304.01322*.

Muhammad Al-Barham, Imad Afyouni, Khalid Almubarak, Ashraf Elnagar, Ayad Turky, and Ibrahim Hashem. 2024. Araclip: Cross-lingual learning for effective arabic image retrieval. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 102–110.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, and 34 others. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil Heffernan, and Kyle Lo. 2025. DrawEduMath: Evaluating vision language models with expert-annotated students' hand-drawn math images. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6902–6920, Albuquerque, New Mexico. Association for Computational Linguistics.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, and 1 others. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code.

Si Chen, Le Huy Khiem, Annalisa Szymanski, Ronald Metoyer, Ting Hua, and Nitesh V Chawla. 2026. Automated benchmark generation from domain guidelines informed by bloom's taxonomy. *arXiv preprint arXiv:2601.20253*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024b. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 17 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and 1 others. 2025. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. *arXiv preprint arXiv:2503.13111*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Sabina Elkins, Ekaterina Kochmar, Jackie CK Cheung, and Iulian Serban. 2024. How teachers can use large language models and bloom's taxonomy to create educational quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23084–23091.

Clement Farabet and Olivier Lacombe. 2026. Gemma 4: Byte for byte, the most capable open models. https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/. Google Blog. Accessed: 2026-04-10.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Alharthi, Ines Riahi, Abduljalil Saif, Jorma Laaksonen, Fahad S. Khan, Salman Khan, and Rao M. Anwer. 2024. Camel-bench: A comprehensive arabic lmm benchmark. *arXiv preprint arXiv:2410.18976*.

Omid Ghahroodi, Arshia Hemmat, Marzia Nouri, Seyed Mohammad Hadi Hosseini, Doratossadat Dastgheib, Mohammad Vali Sanian, Alireza Sahebi, Reihaneh Zohrabi, Mohammad Hossein Rohban, Ehsaneddin Asgari, and 1 others. 2025. Meena (persianmmmu): Multimodal-multilingual educational exams for n-level assessment. *arXiv preprint arXiv:2508.17290*.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *arXiv preprint arXiv:2404.06644*.

Charles Goodwin and Johanne Stege Bjørndahl. 2018. Why multimodality? why co-operative action?(transcribed by j. philipsen). *Social Interaction. Video-Based Studies of Human Sociality*, 1(2).

Google DeepMind. 2025. Gemini 3 flash: Frontier intelligence built for speed. https://blog.google/products/gemini/gemini-3-flash/. Google Blog. Accessed: 2026-04-10.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 37 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2025. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. *arXiv preprint arXiv:2502.11492*.

Thomas Huber and Christina Niklaus. 2025. Llms meet bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246.

Lennart Justen. 2025. Llms outperform experts on challenging biology benchmarks. *arXiv preprint arXiv:2505.06108*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in*

Natural Language Processing, pages 292–305, Singapore. Association for Computational Linguistics.

Yijiang Li, Sucheng Ren, Weipeng Deng, Yuzhi Xu, Ying Gao, Edith Ngai, and Haohan Wang. 2024. Beyond finite data: Towards data-free out-of-distribution generalization via extrapolation. *arXiv preprint arXiv:2403.05523*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640.

Yinghang Ma, Jiho Shin, Leuson Da Silva, Zhen Ming, Song Wang, Foutse Khomh, Shin Hwei Tan, and 1 others. 2025. Bloomapr: A bloom's taxonomy-based framework for assessing the capabilities of llm-powered apr solutions. *arXiv preprint arXiv:2509.25465*.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Inciarte, and Muhammad Abdul-Mageed. 2023. Violet: A vision-language model for Arabic image captioning with gemini decoder. In *Proceedings of ArabicNLP 2023*, pages 1–11, Singapore (Hybrid). Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models. *Preprint*, arXiv:2307.06435.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 36 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models, a survey. *arXiv preprint arXiv:2503.23037*.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. 2024. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. 2022. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, Dublin, Ireland. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. 2025. Mind the gap: Benchmarking spatial reasoning in vision-language models. *Preprint*, arXiv:2503.19707.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, and 1 others. 2025a. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 37 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 22 others. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 26 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Lei Wang, Shan Dong, Yuhui Xu, Hanze Dong, Yalu Wang, Amrita Saha, Ee-Peng Lim, Caiming Xiong, and Doyen Sahoo. 2024a. Mathhay: An automated benchmark for long-context mathematical reasoning in llms. *arXiv preprint arXiv:2410.04698*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zihan Weng, Lucas Gomez, Taylor Whittington Webb, and Pouya Bashivan. 2025. Caption this, reason that: Vlms caught in the middle. *arXiv preprint arXiv:2505.21538*.

Leslie Owen Wilson. 2016. Anderson and krathwohl bloom's taxonomy revised understanding the new version of bloom's taxonomy. *The Second Principle*, 1(1):1–8.

Lei Yang, Renren Jin, Ling Shi, Jianxiang Peng, Yue Chen, and Deyi Xiong. 2025. Probench: Benchmarking large language models in competitive programming. *arXiv preprint arXiv:2502.20868*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, and 1 others. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R. Fung. 2025. VLM2-bench: A closer look at how well VLMs implicitly link explicit matching visual cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7510–7545, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Full BloomBench Taxonomy

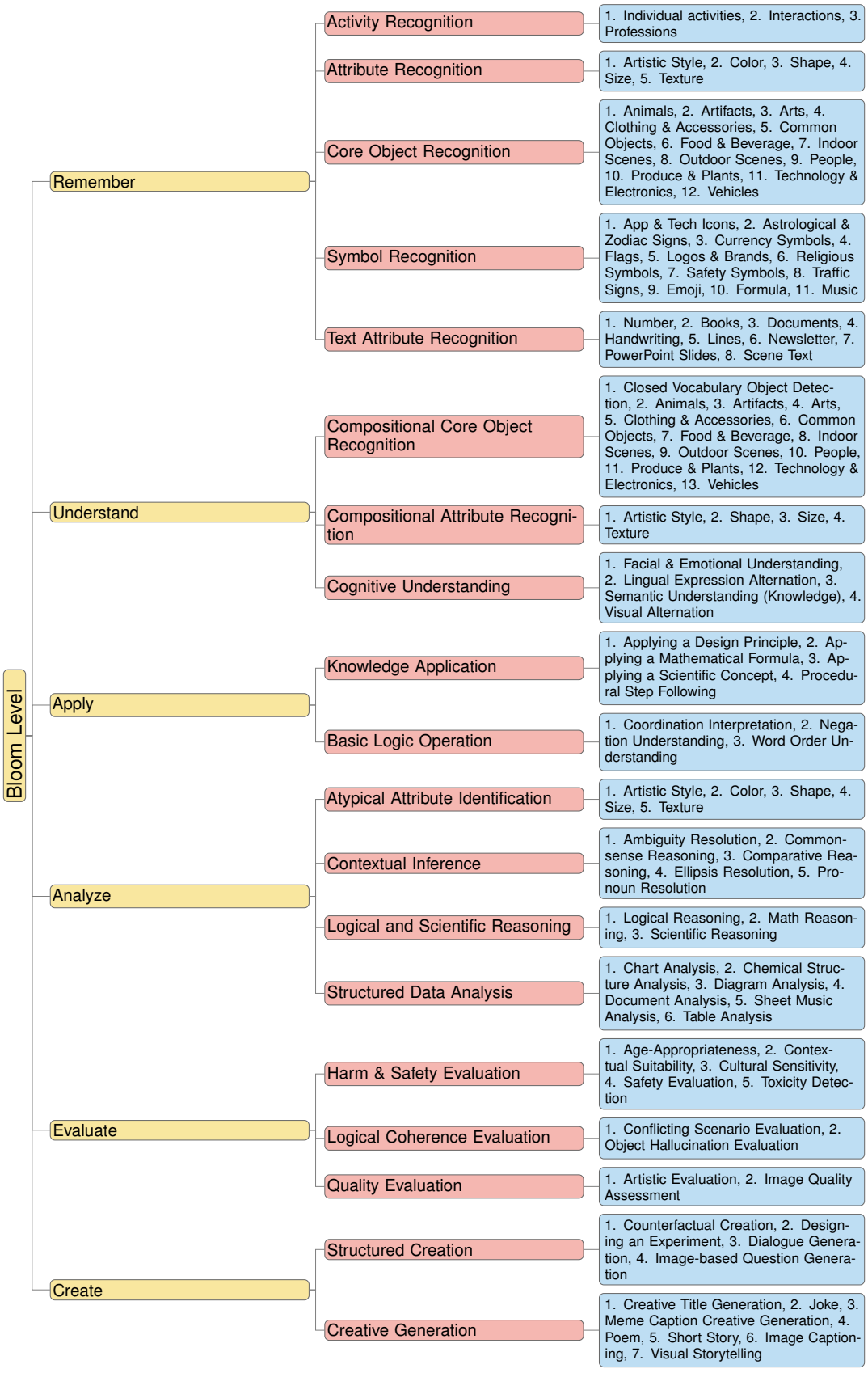

Figure 4: The complete hierarchical structure of the BloomBench taxonomy. The diagram illustrates the decomposition of Bloom's six cognitive levels into specific sub-categories and fine-grained task types, defining the full scope of multimodal capabilities evaluated in the benchmark.

B Statistics

Table 3: Distribution of Items by Bloom's Taxonomy Category

Bloom Category	Count
Remember	2,948
Understand	1,592
Analyze	1,431
Create	685
Evaluate	592
Apply	499
Total	**7,747**

Table 4: Hierarchical Statistics — Bloom's Level: Remember

Subcategory	Sub-subcategory	Count
core object recognition	clothing & accessories	95
attribute recognition	color	95
core object recognition	indoor scenes	93
symbol recognition	traffic signs	93
core object recognition	artifacts	92
core object recognition	people	90
symbol recognition	flags	89
core object recognition	outdoor scenes	88
core object recognition	vehicles	87
core object recognition	produce & plants	87
core object recognition	common objects	85
text attribute recognition	books	85
attribute recognition	shape	83
text attribute recognition	lines	83
attribute recognition	texture	82
core object recognition	arts	81
core object recognition	animals	80
symbol recognition	music	79
core object recognition	food & beverage	79
symbol recognition	safety symbols	78
attribute recognition	artistic style	77
text attribute recognition	handwriting	75
attribute recognition	size	75
core object recognition	technology & electronics	74
text attribute recognition	scene text	73
text attribute recognition	number	73
symbol recognition	emoji	69
symbol recognition	religious symbols	68
text attribute recognition	documents	66
text attribute recognition	newletter	63
symbol recognition	formula	62
symbol recognition	astrological & zodiac signs	62
text attribute recognition	power point slides	61
activity recognition	individual activities	61
activity recognition	interactions	57
activity recognition	professions	56
symbol recognition	logos & brands	53
symbol recognition	currency symbols	51
symbol recognition	app & tech icons	48

Table 5: Hierarchical Statistics — Bloom's Level: Understand

Subcategory	Sub-subcategory	Count
compositional core object recognition	closed vocabulary object detection	89
cognitive understanding	semantic understanding (knowledge)	88
compositional core object recognition	food & beverage	85
compositional core object recognition	clothing & accessories	84
compositional attribute recognition	texture	83
compositional core object recognition	outdoor scenes	83
compositional core object recognition	arts	82
compositional core object recognition	animals	81
compositional attribute recognition	shape	81
compositional attribute recognition	artistic style	79
compositional core object recognition	artifacts	76
compositional core object recognition	vehicles	75
cognitive understanding	lingual expression alternation	74
compositional core object recognition	people	73
cognitive understanding	facial & emotional understanding	71
cognitive understanding	visual alternation	71
compositional attribute recognition	size	70
compositional core object recognition	produce & plants	67
compositional core object recognition	indoor scenes	63
compositional core object recognition	common objects	62
compositional core object recognition	technology & electronics	55

Table 6: Hierarchical Statistics — Bloom's Level: Apply

Subcategory	Sub-subcategory	Count
basic logic operation	negation understanding	84
knowledge application	applying a design principle	77
knowledge application	procedural step following	75
knowledge application	applying a scientific concept	73
basic logic operation	word order understanding	67
basic logic operation	coordination interpretation	62
knowledge application	applying a mathematical formula	61

Table 7: Hierarchical Statistics — Bloom's Level: Analyze

Subcategory	Sub-subcategory	Count
structured data analysis	chart analysis	93
structured data analysis	document analysis	88
logical and scientific reasoning	logic reasoning	86
logical and scientific reasoning	scientific reasoning	86
contextual inference	comparative reasoning	82
logical and scientific reasoning	math reasoning	81
atypical attribute identification	artistic style	81
contextual inference	commonsense reasooning	78
structured data analysis	table analysis	75
structured data analysis	chemical structure analysis	73
atypical attribute identification	shape	73
contextual inference	pronoun resolution	71
contextual inference	ambiguity resolution	69
contextual inference	ellipsis resolution	68
atypical attribute identification	color	68
atypical attribute identification	size	67
structured data analysis	diagram analysis	64
structured data analysis	sheet music analysis	52

Table 8: Hierarchical Statistics — Bloom's Level: Evaluate

Subcategory	Sub-subcategory	Count
quality evaluation	image quality assessment	82
quality evaluation	artistic evaluation	78
logical coherence evaluation	object hallucination evaluation	76
harm & safety evaluation	safety evaluation	75
harm & safety evaluation	contextual suitability	70
harm & safety evaluation	age-appropriateness	63
logical coherence evaluation	conflicting scenario evaluation	59
harm & safety evaluation	toxicity detection	55
harm & safety evaluation	cultural sensitivity	34

Table 9: Hierarchical Statistics — Bloom's Level: Create

Subcategory	Sub-subcategory	Count
creative generation	creative title generation	81
creative generation	image captioning	77
structured creation	counterfactual creation	74
creative generation	visual stroytelling	64
structured creation	designing an experiment	62
structured creation	image-based question generation	60
structured creation	dialogue generation	59
creative generation	meme caption creative generation	58
creative generation	joke	54
creative generation	poem	53
creative generation	short story	43

C Cross-Lingual LBS Ablation Study

To rigorously validate our claim that tokenization fertility amplifies the Likelihood-based Scoring (LBS) gap observed in Arabic, we conducted a controlled ablation study using Spanish as an intermediate language. Spanish is a high-resource language that uses the Latin script and shares many morphological properties with English, but exhibits different tokenization characteristics (i.e., lower token-per-word ratios than Arabic, but higher than English for the models evaluated). We evaluated Qwen2.5-VL-7B (Qwen et al., 2025) on a stratified representative subset of 908 samples, drawn with at least four examples from each of the 106 taxonomy leaf nodes.

Results. As shown in Table 10, the model's deterministic accuracy (RAE) in Spanish remains high at 84.91%, closely comparable to its English performance of 87.22%, a gap of only 2.31 percentage points. However, the LBS score for Spanish suffers a substantially steeper decline (8.26 points below English), despite the near-parity in RAE. This dissociation between RAE and LBS in a language where the model demonstrably understands the content confirms that LBS is sensitive to the model's intrinsic lower probability priors for non-English text, a phenomenon that is further amplified by tokenization fertility.

Interpretation. These findings support the interpretation that the Arabic LBS gap observed in the main experiments (Section 5) is a *compound effect*: (i) a **metric artifact** stemming from higher token-per-word ratios in morphologically rich languages, which disproportionately penalizes the length-normalized score (Equation 2), and (ii) **genuine reasoning gaps** arising from the model's weaker Arabic-language priors. Crucially, this also validates LBS as a measure of *calibrated confidence density* rather than surface-level accuracy, distinguishing it from RAE as a complementary and more diagnostically sensitive evaluation signal.

Language	$\text{RAE}_{\text{micro}}$	$\text{RAE}_{\text{macro}}$	$\text{LBS}_{\text{micro}}$	$\text{LBS}_{\text{macro}}$
English	87.22	85.76	67.07	67.06
Spanish	84.91	83.29	58.81	58.81
Arabic	79.96	79.20	50.77	50.24

Table 10: RAE vs. LBS accuracy (%) of Qwen2.5-VL-7B on a stratified subset of 908 samples across English, Spanish, and Arabic. The near-parity in RAE between English and Spanish, coupled with a significant LBS drop, isolates the effect of tokenization fertility on likelihood-based scoring from genuine reasoning capability.

D MMMU Taxonomy Coverage Analysis

To quantitatively assess the difference in cognitive coverage between BloomBench and an established benchmark, we mapped 1,080 samples from MMMU (Yue et al., 2024) onto the BloomBench taxonomy using Gemini 3 Flash (Google DeepMind, 2025) as a judge model. Table 11 summarizes the distribution of these samples across the six levels of Bloom's Taxonomy. Note that 94 samples were classified as unmatched and are excluded from leaf-level counts.

Distributional Skew. The *Analyze* level alone accounts for 66.4% (717/1,080) of the total samples, reflecting MMMU's strong orientation toward exam-style academic problems in STEM domains. The three most frequent leaf nodes together account for nearly half ($\approx$46%) of the entire dataset:

- **Math Reasoning:** 344 samples
- **Table Analysis:** 85 samples
- **Chart Analysis:** 68 samples

Gaps in Cognitive Coverage. Forty-five taxonomy leaf nodes received zero coverage, revealing structural gaps in MMMU's ability to test higher-order synthesis and judgment. The *Create* and *Evaluate* levels combined represent only $\approx$1.0% (11/1,080) of the dataset. Critical task types completely absent from the MMMU sample include:

- **Contextual Inference:** Ambiguity Resolution, Commonsense Reasoning
- **Harm & Safety Evaluation:** Contextual Suitability, Toxicity Detection
- **Creative Generation:** Dialogue Generation, Designing an Experiment

This comparison confirms that while MMMU serves as a robust benchmark for expert domain knowledge and analytical reasoning, it lacks the breadth to evaluate the full spectrum of multimodal cognition, particularly in creative, evaluative, and socially grounded tasks that BloomBench is designed to assess.

Bloom Level	Covered Leaves	Total Nodes	Coverage (%)	Samples
Remember	25	39	64.1%	77
Understand	12	21	57.1%	134
Analyze	13	19	68.4%	717
Apply	3	7	42.9%	47
Evaluate	4	9	44.4%	5
Create	4	11	36.4%	6
Total	**61**	**106**	**57.5%**	**986**[†]

Table 11: Distribution of 1,080 MMMU samples across BloomBench cognitive levels. [†]94 samples were classified as unmatched and are excluded from leaf-level counts.

E Examples

Example

Levels

1: Analyze
2: Atypical Attribute Identification
3: Artistic Style

Question (en): What specific object in this image introduces a stylistic anachronism, contrasting with the classical sculpture style of the figure?

Answer: The modern smartphone held by the figure introduces a stylistic anachronism, as it is a 21st-century device juxtaposed with a Greco-Roman or Renaissance artistic style.

MCQ (en)
Which element in the image creates a striking stylistic anachronism when contrasted with the classical appearance of the sculpture?

(A) The intricate details of the figure's curly hair.

(B) The figure's intense and dramatic facial expression.

(C) The modern smartphone held by the figure.

(D) The flowing, stylized drapery worn by the figure.

Example

Levels

1: Analyze
2: Contextual Inference
3: Comparative Reasoning

Question (en): Contrast the storage approach of the pink-doored pantry with that of the tall wooden cabinets. What does this suggest about their intended contents?

Answer: The pantry uses visible storage with glass doors and open shelves, suggesting it's for frequently accessed or decorative items like glassware and dry goods. The opaque wooden cabinets offer concealed storage, likely for appliances or less aesthetic bulk supplies.

MCQ (en)

Which of the following best contrasts the storage approach of the pink-doored pantry with the tall wooden cabinets, and suggests their intended contents?

(A) Both storage units emphasize concealed storage; however, the pantry is for non-perishable goods, and the wooden cabinets are for cleaning supplies.

(B) The pink-doored pantry is solely for aesthetic display, using its color to enhance the room, and the wooden cabinets are for storing large, industrial kitchen equipment.

(C) The pantry, with its glass doors and open shelves, offers visible storage for frequently accessed or decorative items like glassware and dry goods, while the opaque wooden cabinets provide concealed storage for appliances or less aesthetic bulk supplies.

(D) The pantry's transparency is to showcase rarely used, valuable china, contrasting with the wooden cabinets which offer practical, hidden storage for essential everyday tools.

F Prompts

Scenario Prompt Template

You are an assistant tasked with generating image search scenario queries to support visual learning grounded in Bloom's Taxonomy. Your goal is to create scenarios that describe visually rich scenes with minimal or no text, leading to concrete images that can be used for evaluating a model's visual understanding.

Context

We are creating visual learning prompts based on Bloom's Taxonomy. Each query should describe a clear, specific visual scene that reflects a cognitive ability (Bloom level) applied to a key concept (leaf). The primary goal is to generate scenarios and keywords that produce high-quality, language-independent images from a standard image search engine.

Key Concepts

- **Leaf**: The core concept or topic the scenario should visually represent. This is the **target subject** of the image query. (*Example*: "Pattern Recognition")
- **Bloom Level**: The type of thinking or cognitive ability that the learner is expected to apply to the leaf concept. (*Example*: "Analyzing" → discriminating between different parts)
- **Path**: The full hierarchical classification. (*Example*: "Analyzing → Differentiating → Discriminating → Pattern Recognition")

Bloom's Taxonomy Levels & Abilities

1. **Remembering** – Recall facts and basic concepts.
2. **Understanding** – Demonstrate comprehension of meaning.
3. **Applying** – Use knowledge in new situations.
4. **Analyzing** – Break information into parts.
5. **Evaluating** – Make judgments based on criteria.
6. **Creating** – Generate new ideas or artifacts.

Expected Output (JSON)

```
{
   "scenarios": ["Man crossing street while looking at smartphone.", ...],
    "keywords": ["man crossing street looking at phone", ...]
}
```

VQA Prompt Template

You are an expert assistant for creating Visual Question Answering (VQA) benchmarks grounded in Bloom's Taxonomy. Your task is to generate high-quality, insightful question-answer pairs

answerable *only* by analyzing the visual content of the image.

Guidelines

- **Image-Grounded**: Both the question and answer MUST be directly derivable from the visual information.
- **Bloom Alignment**: The question must genuinely require the cognitive skill of `{{bloom_node}}`.
- **Deterministic**: Answers should be objective and verifiable.

Example Output

```
[
  {
   "question_en": "Based on the dog's attire, what event is taking place?",
    "answer_en": "A birthday party, indicated by the birthday hat."
  }
]
```

Make Multiple Choice Arabic QA Prompt

You are a professional test creator. Your task is to generate a high-quality multiple-choice question in English, and then provide a translation in Modern Standard Arabic.

Important Guidelines

- One of the distractors must be a **trap answer**—a tempting but incorrect option.
- Strong trap answers involve confusing similar regional or cultural features (e.g., Qatari vs. Emirati attire).
- Provide the correct answer as Choice A in the JSON structure.

G Human Annotation and Quality Control Guidelines

Annotators evaluated a stratified subset of the dataset using a custom annotation interface. For each sample, they were provided with the image, the corresponding multiple-choice question in both English and Arabic, the correct answer, and the target taxonomy path.

Annotators were instructed to verify that all components were accurate, coherent, and meaningful. They were required to either approve the sample as fully valid or flag it under one of the following four specific error categories:

- **Image Quality and Alignment:** Verifying that the image is clear, relevant, and correctly aligns with the targeted Bloom's Taxonomy node.
- **Question Integrity:** Ensuring the English question is natural, unambiguous, grammatically correct, and strictly grounded in both the visual content and the assigned cognitive level.
- **Choice Validity:** Confirming that the multiple-choice options are clear, properly structured, and not misleading, with exactly one unequivocally correct answer.
- **Translation Fidelity:** Evaluating the Arabic translation to ensure it accurately, naturally, and faithfully reflects the semantic meaning of the original English text.

Samples that passed all four criteria without issue were marked as valid. If a sample was flagged for any of the above issues, it was recorded for removal or revision. The annotation platform also included concurrency controls to ensure each question was reviewed exactly once by the annotator pool.

H Detailed Results of Benchmarking

Sub-category	Sub-sub-category	Gemma-3-4B-en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27B-it_en	Gemma-3-12B-it_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
atypical attribute identification	artistic style	0.84	0.52	0.75	0.64	0.77	0.54	0.63	0.4	0.52	0.57
atypical attribute identification	color	0.75	0.56	0.88	0.18	0.57	0.53	0.59	0.38	0.65	0.47
atypical attribute identification	shape	0.62	0.44	0.64	0.48	0.53	0.52	0.55	0.34	0.48	0.41
atypical attribute identification	size	0.63	0.4	0.57	0.45	0.64	0.52	0.48	0.45	0.49	0.49
atypical attribute identification	texture	0.79	0.51	0.82	0.64	0.75	0.63	0.62	0.32	0.53	0.54
contextual inference	ambiguity resolution	0.65	0.43	0.81	0.48	0.59	0.52	0.58	0.33	0.77	0.46
contextual inference	commonsense reasooning	0.76	0.47	0.82	0.5	0.68	0.6	0.59	0.49	0.72	0.67
contextual inference	comparative reasoning	0.77	0.59	0.88	0.63	0.78	0.63	0.62	0.34	0.6	0.66
contextual inference	ellipsis resolution	0.56	0.5	0.72	0.41	0.5	0.44	0.51	0.32	0.41	0.41
contextual inference	pronoun resolution	0.51	0.41	0.72	0.32	0.46	0.49	0.56	0.31	0.49	0.45
logical and scientific reasoning	logic reasoning	0.68	0.53	0.71	0.51	0.6	0.62	0.6	0.34	0.59	0.62
logical and scientific reasoning	math reasoning	0.6	0.46	0.59	0.51	0.52	0.43	0.48	0.33	0.48	0.42
logical and scientific reasoning	scientific reasoning	0.74	0.56	0.81	0.47	0.62	0.6	0.62	0.42	0.63	0.56
structured data analysis	chart analysis	0.74	0.57	0.76	0.58	0.68	0.65	0.62	0.39	0.54	0.65
structured data analysis	chemical structure analysis	0.74	0.48	0.68	0.6	0.75	0.53	0.63	0.37	0.59	0.56
structured data analysis	diagram analysis	0.7	0.53	0.8	0.66	0.73	0.66	0.67	0.48	0.72	0.7
structured data analysis	document analysis	0.78	0.51	0.8	0.62	0.68	0.65	0.67	0.36	0.64	0.59
structured data analysis	sheet music analysis	0.71	0.52	0.73	0.54	0.56	0.58	0.73	0.37	0.58	0.54
structured data analysis	table analysis	0.71	0.53	0.75	0.59	0.67	0.63	0.65	0.39	0.56	0.55

Table 12: Likelihood-based Scoring (LBS) performance comparison across models for category: *analyze*

Sub-category	Sub-sub-category	Gemma-3-4B-en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27B-it_en	Gemma-3-12B-it_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
basic logic operation	coordination interpretation	0.71	0.55	0.84	0.6	0.53	0.5	0.56	0.31	0.47	0.56
basic logic operation	negation understanding	0.4	0.38	0.57	0.27	0.32	0.24	0.29	0.27	0.32	0.24
basic logic operation	word order understanding	0.34	0.24	0.48	0.13	0.24	0.21	0.24	0.21	0.37	0.24
knowledge application	applying a design principle	0.64	0.58	0.74	0.43	0.53	0.45	0.52	0.39	0.55	0.49
knowledge application	applying a mathematical formula	0.39	0.38	0.51	0.28	0.26	0.36	0.26	0.28	0.41	0.26
knowledge application	applying a scientific concept	0.82	0.64	0.79	0.6	0.6	0.55	0.63	0.41	0.58	0.59
knowledge application	procedural step following	0.65	0.53	0.72	0.43	0.44	0.44	0.56	0.39	0.52	0.45

Table 13: Likelihood-based Scoring (LBS) performance comparison across models for category: *apply*

Sub-category	Sub-sub-category	Gemma-3-4B-en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27B-it_en	Gemma-3-12B-it_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
creative generation	creative title generation	0.41	0.38	0.56	0.31	0.35	0.25	0.31	0.26	0.43	0.28
creative generation	image captioning	0.64	0.45	0.74	0.38	0.49	0.39	0.49	0.38	0.52	0.34
creative generation	joke	0.31	0.33	0.39	0.39	0.3	0.2	0.3	0.15	0.2	0.17
creative generation	meme caption creative generation	0.47	0.33	0.47	0.48	0.5	0.36	0.43	0.21	0.31	0.45
creative generation	poem	0.45	0.38	0.62	0.28	0.42	0.19	0.25	0.21	0.26	0.23
creative generation	short story	0.7	0.47	0.63	0.51	0.7	0.6	0.63	0.33	0.65	0.51
creative generation	visual storytelling	0.7	0.48	0.72	0.56	0.64	0.5	0.52	0.27	0.41	0.41
structured creation	counterfactual creation	0.66	0.51	0.7	0.51	0.59	0.46	0.51	0.36	0.42	0.39
structured creation	designing an experiment	0.69	0.55	0.61	0.65	0.71	0.58	0.55	0.44	0.53	0.6
structured creation	dialogue generation	0.37	0.39	0.54	0.34	0.46	0.22	0.27	0.27	0.34	0.24
structured creation	image-based question generation	0.63	0.45	0.73	0.3	0.4	0.3	0.25	0.42	0.48	0.33

Table 14: Likelihood-based Scoring (LBS) performance comparison across models for category: *create*

Sub-category	Sub-sub-category	Gemma-3-4B-en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27B-it_en	Gemma-3-12B-it_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
harm-safety evaluation	age-appropriateness	0.75	0.49	0.79	0.38	0.59	0.49	0.52	0.44	0.71	0.44
harm-safety evaluation	contextual suitability	0.72	0.54	0.86	0.41	0.63	0.51	0.5	0.37	0.54	0.44
harm-safety evaluation	cultural sensitivity	0.71	0.47	0.88	0.32	0.59	0.56	0.47	0.35	0.59	0.65
harm-safety evaluation	safety evaluation	0.75	0.44	0.8	0.4	0.56	0.49	0.53	0.39	0.47	0.47
harm-safety evaluation	toxicity detection	0.68	0.39	0.7	0.42	0.58	0.65	0.61	0.4	0.54	0.64
logical coherence evaluation	conflicting scenario evaluation	0.66	0.46	0.56	0.32	0.58	0.46	0.51	0.37	0.53	0.46
logical coherence evaluation	object hallucination evaluation	0.61	0.5	0.76	0.33	0.49	0.41	0.51	0.42	0.57	0.41
quality evaluation	artistic evaluation	0.76	0.5	0.82	0.53	0.59	0.51	0.47	0.28	0.55	0.46
quality evaluation	image quality assessment	0.63	0.5	0.78	0.39	0.66	0.52	0.54	0.28	0.57	0.52

Table 15: Likelihood-based Scoring (LBS) performance comparison across models for category: *evaluate*

Sub-category	Sub-sub-category	Gemma-3-4B-en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27B-it_en	Gemma-3-12B-it_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
activity recognition	individual activities	0.62	0.48	0.75	0.25	0.3	0.34	0.46	0.18	0.62	0.34
activity recognition	interactions	0.6	0.46	0.65	0.3	0.4	0.37	0.44	0.37	0.51	0.3
activity recognition	professions	0.55	0.39	0.73	0.27	0.36	0.23	0.32	0.27	0.68	0.23
attribute recognition	artistic style	0.48	0.47	0.66	0.19	0.34	0.3	0.26	0.32	0.51	0.21
attribute recognition	color	0.56	0.25	0.33	0.07	0.21	0.25	0.24	0.25	0.19	0.37
attribute recognition	shape	0.48	0.24	0.31	0.18	0.27	0.17	0.17	0.27	0.29	0.19
attribute recognition	size	0.35	0.27	0.53	0.15	0.11	0.16	0.2	0.21	0.35	0.15
attribute recognition	texture	0.4	0.38	0.59	0.17	0.21	0.3	0.34	0.3	0.41	0.3
core object recognition	animals	0.52	0.21	0.34	0.06	0.18	0.28	0.31	0.21	0.26	0.29
core object recognition	artifacts	0.46	0.29	0.48	0.17	0.28	0.32	0.31	0.37	0.49	0.35
core object recognition	arts	0.41	0.36	0.57	0.2	0.28	0.2	0.29	0.3	0.47	0.22
core object recognition	clothing-accessories	0.46	0.27	0.35	0.09	0.2	0.27	0.35	0.22	0.34	0.35
core object recognition	common objects	0.47	0.11	0.36	0.14	0.19	0.19	0.2	0.18	0.34	0.19
core object recognition	food-beverage	0.58	0.16	0.44	0.11	0.23	0.18	0.25	0.23	0.41	0.16
core object recognition	indoor scenes	0.49	0.25	0.59	0.17	0.28	0.22	0.26	0.25	0.46	0.28
core object recognition	outdoor scenes	0.49	0.11	0.55	0.11	0.19	0.22	0.24	0.24	0.45	0.17
core object recognition	people	0.44	0.19	0.34	0.1	0.13	0.12	0.27	0.22	0.31	0.22
core object recognition	produce-plants	0.64	0.22	0.37	0.2	0.23	0.18	0.2	0.23	0.34	0.15
core object recognition	technology-electronics	0.43	0.26	0.45	0.11	0.26	0.22	0.32	0.3	0.46	0.26
core object recognition	vehicles	0.51	0.18	0.45	0.16	0.22	0.21	0.29	0.26	0.49	0.26
symbol recognition	app-tech icons	0.52	0.23	0.62	0.1	0.23	0.33	0.33	0.27	0.44	0.38
symbol recognition	astrological-zodiac signs	0.45	0.26	0.29	0.03	0.11	0.16	0.26	0.16	0.18	0.24
symbol recognition	currency symbols	0.67	0.29	0.61	0.14	0.31	0.24	0.29	0.25	0.47	0.39
symbol recognition	emoji	0.51	0.23	0.54	0.17	0.19	0.23	0.35	0.35	0.48	0.33
symbol recognition	flags	0.49	0.27	0.45	0.12	0.27	0.2	0.2	0.24	0.28	0.3
symbol recognition	formula	0.34	0.34	0.5	0.16	0.15	0.18	0.27	0.23	0.32	0.35
symbol recognition	logos-brands	0.58	0.28	0.42	0.13	0.21	0.25	0.32	0.26	0.53	0.32
symbol recognition	music	0.47	0.24	0.48	0.15	0.27	0.18	0.23	0.28	0.52	0.18
symbol recognition	religious symbols	0.47	0.32	0.51	0.16	0.21	0.18	0.26	0.19	0.35	0.21
symbol recognition	safety symbols	0.58	0.27	0.59	0.04	0.19	0.15	0.23	0.21	0.45	0.29
symbol recognition	traffic signs	0.54	0.27	0.42	0.12	0.19	0.16	0.22	0.3	0.32	0.34
text attribute recognition	books	0.41	0.21	0.34	0.19	0.2	0.29	0.26	0.25	0.4	0.29
text attribute recognition	documents	0.5	0.36	0.61	0.2	0.21	0.2	0.26	0.32	0.45	0.2
text attribute recognition	handwriting	0.41	0.27	0.33	0.16	0.33	0.28	0.29	0.27	0.35	0.32
text attribute recognition	lines	0.39	0.2	0.43	0.17	0.17	0.23	0.35	0.29	0.33	0.27
text attribute recognition	newsletter	0.37	0.25	0.6	0.16	0.17	0.21	0.17	0.38	0.57	0.3
text attribute recognition	number	0.48	0.18	0.38	0.03	0.14	0.14	0.19	0.22	0.29	0.34
text attribute recognition	power point slides	0.49	0.32	0.49	0.14	0.25	0.16	0.22	0.35	0.46	0.21
text attribute recognition	scene text	0.45	0.18	0.38	0.08	0.26	0.19	0.25	0.36	0.41	0.3

Table 16: Likelihood-based Scoring (LBS) performance comparison across models for category: *remember*

Sub-category	Sub-sub-category	Gemma-3-4B-en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27B-it_en	Gemma-3-12B-it_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
cognitive understanding	facial-emotional understanding	0.65	0.52	0.85	0.41	0.55	0.52	0.63	0.37	0.65	0.48
cognitive understanding	lingual expression alternation	0.55	0.32	0.45	0.19	0.23	0.31	0.31	0.34	0.45	0.3
cognitive understanding	semantic understanding (knowledge)	0.77	0.59	0.93	0.38	0.55	0.57	0.65	0.45	0.8	0.64
cognitive understanding	visual alternation	0.75	0.56	0.87	0.46	0.62	0.61	0.48	0.44	0.73	0.48
compositional attribute recognition	artistic style	0.8	0.65	0.91	0.53	0.73	0.65	0.68	0.41	0.71	0.61
compositional attribute recognition	shape	0.67	0.51	0.73	0.49	0.62	0.49	0.49	0.4	0.54	0.49
compositional attribute recognition	size	0.66	0.59	0.86	0.3	0.61	0.47	0.53	0.44	0.46	0.4
compositional attribute recognition	texture	0.88	0.58	0.87	0.64	0.76	0.7	0.65	0.23	0.52	0.52
compositional core object recognition	animals	0.78	0.68	0.94	0.41	0.63	0.53	0.62	0.4	0.59	0.54
compositional core object recognition	artifacts	0.82	0.64	0.88	0.55	0.72	0.61	0.61	0.47	0.59	0.54
compositional core object recognition	arts	0.83	0.61	0.93	0.46	0.67	0.49	0.57	0.4	0.7	0.52
compositional core object recognition	closed vocabulary object detection	0.71	0.48	0.76	0.45	0.64	0.57	0.66	0.37	0.62	0.55
compositional core object recognition	clothing-accessories	0.83	0.63	0.94	0.39	0.65	0.54	0.54	0.42	0.7	0.6
compositional core object recognition	common objects	0.81	0.61	0.9	0.5	0.6	0.6	0.65	0.31	0.58	0.56
compositional core object recognition	food-beverage	0.82	0.72	0.93	0.59	0.69	0.53	0.58	0.31	0.66	0.52
compositional core object recognition	indoor scenes	0.76	0.67	0.98	0.46	0.68	0.62	0.75	0.44	0.84	0.57
compositional core object recognition	outdoor scenes	0.82	0.69	0.96	0.46	0.8	0.64	0.72	0.52	0.76	0.63
compositional core object recognition	people	0.77	0.63	0.96	0.47	0.62	0.51	0.59	0.45	0.84	0.62
compositional core object recognition	produce-plants	0.69	0.6	0.87	0.4	0.6	0.57	0.57	0.28	0.57	0.51
compositional core object recognition	technology-electronics	0.76	0.55	0.85	0.44	0.69	0.53	0.62	0.38	0.75	0.62
compositional core object recognition	vehicles	0.76	0.6	0.92	0.37	0.56	0.48	0.57	0.39	0.73	0.52

Table 17: Likelihood-based Scoring (LBS) performance comparison across models for category: *understand*

Sub-category	Sub-sub-category	Gemma-3-4B-it_en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27b-it_en	Gemma-3-12b-it_en	GPT4omini_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	GPT4omini_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
atypical attribute identification	artistic style	0.91	0.93	0.94	0.98	0.94	0.86	0.93	0.95	0.85	0.83	0.8	0.85
atypical attribute identification	color	0.9	0.94	0.96	0.97	0.94	0.94	0.87	0.96	0.87	0.91	0.88	0.81
atypical attribute identification	shape	0.71	0.79	0.82	0.9	0.82	0.81	0.74	0.79	0.66	0.67	0.64	0.6
atypical attribute identification	size	0.79	0.82	0.81	0.84	0.79	0.78	0.84	0.82	0.69	0.7	0.7	0.7
atypical attribute identification	texture	0.86	0.8	0.84	0.89	0.91	0.82	0.83	0.84	0.7	0.72	0.76	0.78
contextual inference	ambiguity resolution	0.9	0.94	0.97	0.96	0.99	0.9	0.93	0.99	0.88	0.78	0.87	0.83
contextual inference	commonsense reasooning	0.82	0.95	0.86	0.91	0.92	0.87	0.9	0.91	0.83	0.88	0.88	0.79
contextual inference	comparative reasoning	0.89	0.93	0.98	0.89	0.91	0.88	0.89	0.88	0.73	0.84	0.89	0.83
contextual inference	ellipsis resolution	0.84	0.9	0.91	0.94	0.96	0.91	0.93	0.9	0.93	0.84	0.9	0.76
contextual inference	pronoun resolution	0.92	0.92	0.96	0.94	0.9	0.89	0.9	0.89	0.83	0.8	0.76	0.82
logical and scientific reasoning	logic reasoning	0.75	0.77	0.77	0.9	0.79	0.83	0.81	0.87	0.79	0.69	0.71	0.67
logical and scientific reasoning	math reasoning	0.62	0.73	0.73	0.64	0.68	0.67	0.65	0.65	0.65	0.72	0.63	0.56
logical and scientific reasoning	scientific reasoning	0.84	0.9	0.85	0.87	0.91	0.9	0.85	0.93	0.84	0.81	0.83	0.74
structured data analysis	chart analysis	0.81	0.84	0.87	0.86	0.82	0.76	0.84	0.83	0.72	0.8	0.82	0.68
structured data analysis	chemical structure analysis	0.79	0.71	0.77	0.88	0.82	0.74	0.86	0.88	0.7	0.68	0.75	0.73
structured data analysis	diagram analysis	0.83	0.89	0.86	0.89	0.92	0.84	0.91	0.89	0.81	0.78	0.84	0.72
structured data analysis	document analysis	0.86	0.9	0.97	0.92	0.97	0.91	0.9	0.92	0.9	0.88	0.91	0.83
structured data analysis	sheet music analysis	0.71	0.83	0.85	0.9	0.87	0.77	0.85	0.87	0.81	0.77	0.81	0.63
structured data analysis	table analysis	0.8	0.87	0.85	0.89	0.93	0.87	0.88	0.88	0.76	0.77	0.76	0.72

Table 18: Regex Based Extraction (RAE) Performance comparison across models for category: *analyze*

Sub-category	Sub-sub-category	Gemma-3-4B-it_en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27b-it_en	Gemma-3-12b-it_en	GPT4omini_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	GPT4omini_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
basic logic operation	coordination interpretation	0.87	0.89	0.94	0.89	0.92	0.87	0.89	0.9	0.74	0.79	0.85	0.77
basic logic operation	negation understanding	0.7	0.75	0.77	0.88	0.79	0.74	0.79	0.82	0.69	0.64	0.65	0.58
basic logic operation	word order understanding	0.66	0.67	0.78	0.76	0.75	0.52	0.66	0.79	0.49	0.61	0.52	0.58
knowledge application	applying a design principle	0.79	0.9	0.9	0.86	0.88	0.88	0.87	0.83	0.81	0.75	0.75	0.75
knowledge application	applying a mathematical formula	0.51	0.66	0.75	0.8	0.72	0.75	0.64	0.8	0.67	0.59	0.69	0.44
knowledge application	applying a scientific concept	0.74	0.89	0.89	0.92	0.85	0.86	0.92	0.9	0.81	0.75	0.74	0.75
knowledge application	procedural step following	0.79	0.81	0.73	0.85	0.77	0.79	0.81	0.83	0.76	0.71	0.71	0.6

Table 19: Regex Based Extraction (RAE) Performance comparison across models for category: *apply*

Sub-category	Sub-sub-category	Gemma-3-4B-it_en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27b-it_en	Gemma-3-12b-it_en	GPT4omini_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	GPT4omini_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
creative generation	creative title generation	0.64	0.77	0.84	0.77	0.74	0.78	0.68	0.64	0.64	0.58	0.62	0.62
creative generation	image captioning	0.77	0.75	0.79	0.83	0.81	0.73	0.78	0.83	0.74	0.69	0.71	0.7
creative generation	joke	0.54	0.5	0.43	0.69	0.56	0.59	0.56	0.63	0.52	0.39	0.48	0.44
creative generation	meme caption creative generation	0.53	0.53	0.6	0.74	0.62	0.6	0.55	0.69	0.5	0.45	0.53	0.47
creative generation	poem	0.85	0.79	0.91	0.91	0.92	0.72	0.83	0.81	0.55	0.64	0.74	0.7
creative generation	short story	0.79	0.74	0.91	0.91	0.86	0.88	0.88	0.86	0.88	0.72	0.72	0.72
creative generation	visual storytelling	0.86	0.81	0.86	0.84	0.91	0.75	0.84	0.78	0.78	0.81	0.88	0.77
structured creation	counterfactual creation	0.72	0.73	0.72	0.84	0.81	0.73	0.72	0.73	0.66	0.68	0.58	0.58
structured creation	designing an experiment	0.84	0.89	0.94	0.95	0.87	0.94	0.84	0.92	0.89	0.82	0.84	0.76
structured creation	dialogue generation	0.76	0.8	0.85	0.86	0.86	0.88	0.78	0.83	0.78	0.75	0.71	0.66
structured creation	image-based question generation	0.82	0.87	0.88	0.9	0.88	0.85	0.83	0.87	0.72	0.8	0.82	0.77

Table 20: Regex Based Extraction (RAE) Performance comparison across models for category: *create*

Sub-category	Sub-sub-category	Gemma-3-4B-it_en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27b-it_en	Gemma-3-12b-it_en	GPT4omini_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	GPT4omini_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
harm-safety evaluation	age-appropriateness	0.89	0.95	0.92	0.95	0.97	0.92	0.94	0.92	0.87	0.87	0.86	0.84
harm-safety evaluation	contextual suitability	0.79	0.91	0.89	0.99	0.93	0.97	0.97	0.97	0.87	0.73	0.81	0.71
harm-safety evaluation	cultural sensitivity	0.82	1.0	0.97	0.94	1.0	1.0	0.94	0.91	0.91	0.91	0.91	0.79
harm-safety evaluation	safety evaluation	0.83	0.92	0.96	0.93	0.91	0.92	0.92	0.93	0.92	0.84	0.88	0.8
harm-safety evaluation	toxicity detection	0.82	0.82	0.88	0.88	0.93	0.79	0.86	0.89	0.74	0.77	0.82	0.75
logical coherence evaluation	conflicting scenario evaluation	0.63	0.75	0.73	0.8	0.78	0.73	0.68	0.76	0.66	0.49	0.58	0.47
logical coherence evaluation	object hallucination evaluation	0.62	0.82	0.84	0.86	0.87	0.74	0.88	0.83	0.79	0.74	0.71	0.59
quality evaluation	artistic evaluation	0.81	0.95	0.94	0.95	0.92	0.91	0.95	0.91	0.9	0.88	0.88	0.86
quality evaluation	image quality assessment	0.67	0.88	0.9	0.84	0.8	0.79	0.82	0.87	0.78	0.87	0.89	0.74

Table 21: Regex Based Extraction (RAE) Performance comparison across models for category: *evaluate*

Sub-category	Sub-sub-category	Gemma-3-4B-it_en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27b-it_en	Gemma-3-12b-it_en	GPT4omini_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	GPT4omini_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
activity recognition	individual activities	0.87	0.87	0.95	0.93	0.92	0.89	0.95	0.98	0.9	0.89	0.92	0.89
activity recognition	interactions	0.84	0.86	0.91	0.84	0.86	0.86	0.88	0.88	0.82	0.84	0.88	0.79
activity recognition	professions	0.95	0.96	0.95	0.95	0.96	0.96	0.91	0.93	0.93	0.95	0.95	0.89
attribute recognition	artistic style	0.83	0.88	0.88	0.92	0.94	0.9	0.9	0.92	0.82	0.83	0.82	0.77
attribute recognition	color	0.59	0.78	0.77	0.78	0.73	0.69	0.65	0.75	0.59	0.66	0.69	0.52
attribute recognition	shape	0.64	0.8	0.71	0.77	0.71	0.75	0.72	0.75	0.65	0.63	0.6	0.54
attribute recognition	size	0.72	0.77	0.76	0.83	0.8	0.64	0.8	0.81	0.56	0.63	0.68	0.68
attribute recognition	texture	0.8	0.77	0.78	0.82	0.8	0.79	0.71	0.77	0.65	0.63	0.66	0.67
core object recognition	animals	0.79	0.86	0.85	0.94	0.89	0.89	0.86	0.92	0.74	0.68	0.7	0.6
core object recognition	artifacts	0.8	0.83	0.8	0.85	0.81	0.76	0.78	0.78	0.72	0.77	0.83	0.72
core object recognition	arts	0.81	0.89	0.92	0.95	0.92	0.87	0.87	0.92	0.82	0.8	0.81	0.72
core object recognition	clothing- accessories	0.74	0.81	0.88	0.86	0.84	0.71	0.74	0.83	0.69	0.68	0.69	0.67
core object recognition	common objects	0.8	0.82	0.8	0.87	0.87	0.82	0.84	0.84	0.73	0.75	0.82	0.72
core object recognition	food-beverage	0.89	0.94	0.94	0.94	0.92	0.9	0.89	0.94	0.84	0.72	0.81	0.76
core object recognition	indoor scenes	0.75	0.89	0.88	0.87	0.88	0.83	0.88	0.85	0.78	0.85	0.88	0.72
core object recognition	outdoor scenes	0.88	0.88	0.94	0.91	0.91	0.84	0.89	0.89	0.77	0.85	0.9	0.82
core object recognition	people	0.76	0.87	0.87	0.86	0.83	0.78	0.77	0.82	0.69	0.81	0.72	0.69
core object recognition	produce-plants	0.9	0.93	0.9	0.94	0.92	0.85	0.86	0.95	0.87	0.79	0.76	0.87
core object recognition	technology-electronics	0.82	0.89	0.88	0.85	0.82	0.88	0.84	0.81	0.8	0.85	0.89	0.68
core object recognition	vehicles	0.68	0.85	0.9	0.86	0.78	0.75	0.77	0.82	0.74	0.86	0.89	0.64
symbol recognition	app-tech icons	0.77	0.81	0.83	0.83	0.83	0.85	0.83	0.81	0.79	0.73	0.77	0.73
symbol recognition	astrological-zodiac signs	0.74	0.77	0.85	0.89	0.89	0.98	0.82	0.82	0.92	0.56	0.69	0.61
symbol recognition	currency symbols	0.92	0.9	0.9	0.88	0.9	0.86	0.9	0.86	0.82	0.88	0.9	0.84
symbol recognition	emoji	0.7	0.78	0.78	0.78	0.74	0.67	0.78	0.81	0.62	0.74	0.7	0.67
symbol recognition	flags	0.64	0.75	0.83	0.8	0.79	0.71	0.75	0.8	0.64	0.58	0.73	0.57
symbol recognition	formula	0.66	0.65	0.73	0.68	0.76	0.52	0.71	0.68	0.53	0.63	0.68	0.53
symbol recognition	logos-brands	0.79	0.87	0.92	0.89	0.87	0.81	0.87	0.87	0.75	0.85	0.96	0.77
symbol recognition	music	0.84	0.85	0.91	0.89	0.89	0.89	0.82	0.82	0.81	0.84	0.8	0.81
symbol recognition	religious symbols	0.9	0.96	0.96	0.94	0.9	0.96	0.85	0.93	0.84	0.91	0.87	0.82
symbol recognition	safety symbols	0.85	0.96	0.95	0.92	0.94	0.88	0.9	0.94	0.83	0.92	0.92	0.81
symbol recognition	traffic signs	0.78	0.86	0.89	0.82	0.82	0.77	0.7	0.74	0.64	0.78	0.8	0.68
text attribute recognition	books	0.64	0.75	0.69	0.75	0.66	0.62	0.64	0.67	0.68	0.69	0.68	0.55
text attribute recognition	documents	0.78	0.83	0.86	0.91	0.89	0.8	0.85	0.88	0.76	0.73	0.73	0.68
text attribute recognition	handwriting	0.65	0.77	0.8	0.76	0.75	0.57	0.65	0.65	0.6	0.67	0.72	0.56
text attribute recognition	lines	0.66	0.67	0.72	0.75	0.71	0.78	0.67	0.76	0.65	0.57	0.61	0.61
text attribute recognition	newsletter	0.76	0.92	0.89	0.84	0.81	0.68	0.75	0.86	0.58	0.87	0.89	0.76
text attribute recognition	number	0.68	0.79	0.82	0.84	0.84	0.68	0.78	0.75	0.59	0.71	0.81	0.66
text attribute recognition	power point slides	0.78	0.81	0.81	0.86	0.84	0.71	0.75	0.83	0.65	0.7	0.75	0.64
text attribute recognition	scene text	0.73	0.85	0.89	0.93	0.9	0.79	0.75	0.86	0.66	0.78	0.77	0.63

Table 22: Regex Based Extraction (RAE) Performance comparison across models for category: *remember*

Sub-category	Sub-sub-category	Gemma-3-4B-it_en	Qwen2-VL-7B-Instruct_en	Qwen2.5-VL-7B-Instruct_en	Gemma-3-27b-it_en	Gemma-3-12b-it_en	GPT4omini_en	Gemma-3-12B-it_ar	Gemma-3-27B-it_ar	GPT4omini_ar	Qwen2-VL-7B-Instruct_ar	Qwen2.5-VL-7B-Instruct_ar	Gemma-3-4B-it_ar
cognitive understanding	facial-emotional understanding	0.93	0.99	0.93	0.93	0.89	0.89	0.85	0.89	0.85	0.86	0.86	0.89
cognitive understanding	lingual expression alternation	0.85	0.85	0.88	0.91	0.92	0.8	0.84	0.88	0.82	0.73	0.68	0.69
cognitive understanding	semantic understanding (knowledge)	0.95	0.98	0.99	0.94	0.97	0.97	0.98	0.97	0.93	0.93	0.97	0.92
cognitive understanding	visual alternation	0.9	0.93	0.94	0.97	0.94	0.96	0.96	0.97	0.93	0.87	0.9	0.85
compositional attribute recognition	artistic style	0.91	0.96	0.97	0.95	0.95	0.92	0.96	0.94	0.94	0.89	0.94	0.87
compositional attribute recognition	shape	0.83	0.84	0.8	0.84	0.88	0.78	0.85	0.78	0.68	0.74	0.72	0.72
compositional attribute recognition	size	0.67	0.79	0.81	0.86	0.91	0.74	0.83	0.84	0.73	0.61	0.7	0.59
compositional attribute recognition	texture	0.8	0.87	0.9	0.88	0.87	0.82	0.8	0.81	0.75	0.8	0.76	0.78
compositional core object recognition	animals	0.96	0.98	0.98	1.0	0.98	0.94	0.96	0.96	0.88	0.89	0.89	0.84
compositional core object recognition	artifacts	0.91	0.87	0.93	0.97	0.95	0.92	0.96	0.96	0.82	0.83	0.82	0.87
compositional core object recognition	arts	0.94	0.98	0.95	0.93	0.95	0.93	0.95	0.95	0.89	0.9	0.91	0.89
compositional core object recognition	closed vocabulary object detection	0.81	0.87	0.93	0.96	0.92	0.85	0.89	0.92	0.76	0.79	0.82	0.78
compositional core object recognition	clothing-accessories	0.89	0.95	0.99	1.0	0.94	0.96	0.92	0.94	0.88	0.85	0.9	0.79
compositional core object recognition	common objects	0.89	0.94	0.89	0.97	0.95	0.89	0.9	0.9	0.84	0.89	0.9	0.82
compositional core object recognition	food-beverage	0.94	0.98	0.99	0.96	0.94	0.91	0.96	0.95	0.88	0.92	0.89	0.92
compositional core object recognition	indoor scenes	0.95	1.0	0.98	1.0	1.0	0.95	0.98	1.0	0.92	0.97	0.98	0.92
compositional core object recognition	outdoor scenes	0.95	0.96	0.98	0.96	0.98	0.98	0.99	0.98	1.0	0.96	0.94	0.94
compositional core object recognition	people	0.93	0.99	0.97	0.96	0.97	0.96	0.93	0.99	0.93	0.95	0.95	0.95
compositional core object recognition	produce-plants	0.88	0.94	0.93	0.93	0.96	0.97	0.93	0.94	0.88	0.84	0.91	0.88
compositional core object recognition	technology-electronics	0.95	0.96	0.95	0.96	0.98	0.87	0.95	0.96	0.89	0.91	0.93	0.87
compositional core object recognition	vehicles	0.91	0.96	0.97	0.96	1.0	0.96	0.96	0.96	0.89	0.95	0.96	0.89

Table 23: Regex Based Extraction (RAE) Performance comparison across models for category: *understand*