Jako Tako or Fluent? Presenting PoVisLE: A Polish Vision-Language Evaluation

Anonymous ACL submission

Abstract

Vision-language models (VLMs) have achieved strong performance on tasks such as image captioning, visual question answering, and image-to-text generation. However, they are predominantly trained on English-centric data, which limits their ability to handle culturally grounded visual understanding and leads to failures in interpreting region-specific meanings, symbolic content, and context-dependent visual cues. Existing benchmarks for cultural competence are often template-driven and focused on surface-level recognition, making them insufficient for evaluating deeper linguistic and pragmatic understanding in culturally situated settings. We introduce **PoVisLE**, a monocultural vision-language benchmark for Polish designed to evaluate culturally grounded multimodal understanding under a grounded evaluation paradigm, where language is interpreted in interaction with visual context. The dataset contains 966 images and 2,016 manually annotated VQA pairs. Overall, our dataset provides a controlled and challenging testbed for assessing culturally grounded vision-language understanding beyond surface-level recognition.[1]

1 Introduction

Recent advances in VLMs have enabled high-quality image captioning, visual question answering, and image-to-text generation, accelerating their deployment in applications such as advertising, content creation, and digital assistants.

However, these systems are predominantly trained on English-centric datasets, leading to the underrepresentation of local contexts, cultures, and languages. Consequently, models often struggle with culturally specific meanings, symbolic interpretations, and context-dependent visual cues, particularly in mid- and low-resource languages.

While constructing large-scale, culturally grounded datasets remains challenging, robust evaluation benchmarks are essential (Vintar et al., 2025). Misalignment with regional contexts is not only an ethical concern but also a practical limitation, potentially resulting in inaccurate or inappropriate outputs in real-world applications.

This issue is especially relevant in the European context, where accessibility regulations such as WCAG require alternative text descriptions for visual content. Although multimodal models offer a promising solution (Mähr and Twente, 2025; Zheng et al., 2025; Elisiário and Watanabe, 2025), they must correctly interpret culturally localized content to be reliably deployed.

Current evaluations of cultural competence remain limited. Existing datasets often focus on surface-level recognition tasks, such as identifying objects or landmarks, and rely on template-based formats that restrict linguistic and contextual variability (Yadav et al., 2025). As a result, they fail to capture deeper dimensions of cultural meaning, including symbolic interpretation and pragmatic usage.

To partially address these limitations within a specific cultural setting, we propose **PoVisLE** (**Po**lish **Vis**ion-**L**anguage **E**valuation, see Figure 1), a monocultural benchmark focused on Polish cultural and linguistic competence, including region-specific knowldge. Our dataset adopts a grounded evaluation paradigm, where linguistic understanding is assessed in interaction with visual context, enabling a more comprehensive and realistic evaluation. It extends prior work on culturally grounded text evaluation, notably PLCC (Dadas et al., 2025).

Unlike template-based datasets, PoVisLE is constructed through manual, template-free annotation, prioritizing linguistic richness and cultural specificity. To quantify these properties, we conduct a computational stylometric analysis of the manually authored questions (see Appendix B). Despite

[1] We share our code and data for future research: https://anonymous.4open.science/r/PoVisLE-B856/README.md

this focus on quality, the dataset achieves a moderate scale, comprising 966 images and 2,016 VQA pairs.

Our contributions can be summarized as follows:

- We introduce a culturally grounded vision-language benchmark for Polish, designed to evaluate multimodal models beyond surface-level entity recognition. Nearly 30% of the images in the original image set (before augmentation) are sourced from annotator-provided, private collections that have not been publicly available online, ensuring that they are not part of standard web-scale training corpora and providing a stronger test of model generalization.

- We propose a grounded evaluation paradigm, in which linguistic competence is assessed through its interaction with visual context and design a hierarchical taxonomy covering both cultural knowledge and language phenomena, enabling fine-grained diagnostic analysis. Our results further demonstrate that questions involving dialects and regionalisms within the language category yield lower performance. This underscores the challenges models face in handling linguistic variation and highlights the importance of ensuring fair representation of the Polish language and culture, both at the national and regional levels.

- We demonstrate that high-quality, human-curated annotation can be achieved at moderate scale without relying on template-based generation or synthetic data.

2 Related Work

Early vision-language benchmarks primarily focused on image captioning, visual relations, object recognition and simple reasoning. Representative datasets such as MS COCO (Lin et al., 2014), Visual Question Answering (VQA) v2 (Goyal et al., 2017), and CLEVR (Johnson et al., 2017) introduced tasks involving counting, spatial relations, and compositional reasoning (e.g., "How many objects are in the image?"). Subsequent benchmarks, including OK-VQA (Marino et al., 2019) and KVQA (Shah et al., 2019), extended evaluation toward external and commonsense knowledge, while recent culturally grounded VQA benchmarks investigate whether models can correctly interpret culturally specific symbols, practices, social norms, and geographically localized knowledge.

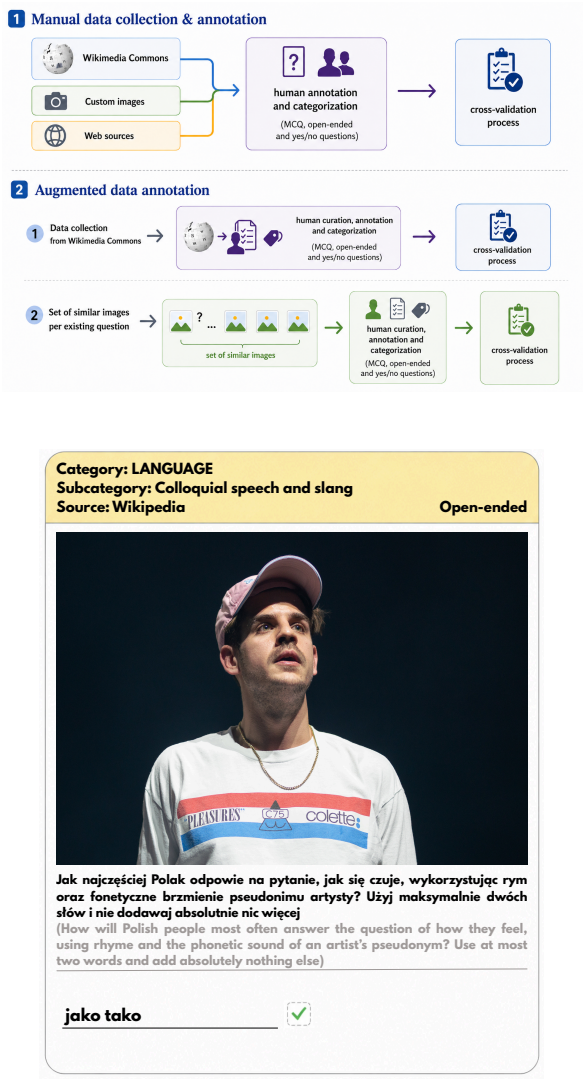

Figure 1: **Above**: Overview of the two-stage PoVisLE dataset construction process. **Below**: a single example from our dataset, more examples in Appendix A.

More recently, a growing body of work has focused on evaluating vision-language models in culturally diverse settings. Large-scale multicultural benchmarks, such as CVQA (Romero et al., 2024), cover dozens of countries and languages, enabling cross-cultural comparison. Similarly, CulturalVQA (Nayak et al., 2024) and BlendVis (Tan et al., 2026) evaluate models across multiple geographic regions and categories, highlighting performance variability across cultures. The WorldCuisines dataset (Winata et al., 2025) further broadens cultural evaluation through food-related visual recognition tasks spanning multiple national cuisines, although Pol-

ish culture is represented only marginally within its coverage. However, these datasets prioritize breadth over depth, limiting their ability to capture fine-grained cultural understanding within a single context. In addition, some benchmarks rely on English-only questions, template-based generation, or synthetically generated images, which may introduce biases and conceptual inaccuracies undermining the value of the evaluation.

Another line of work focuses on region-specific benchmarks, including settings such as India (Maji et al., 2025), China and Taiwan (Wang et al., 2025; Hsieh et al., 2026), and the Arab world (Kadaoui et al., 2026). While more localized, these benchmarks often reflect substantial internal diversity, including multilingual and multidialectal variation, making controlled evaluation more challenging. Additionally, CVLUE, a Chinese vision-language understanding benchmark, focuses on image-level perception while addressing the Western-centric bias present in existing datasets, particularly in concept hierarchies derived from resources such as WordNet (Wang et al., 2025). We also include a comparison between selected visual and culturally-grounded benchmarks in Appendix C.

Despite the growing body of region-specific and multicultural benchmarks, Polish-language resources remain limited and fragmented. The PLCC benchmark (Dadas et al., 2025) represents the first structured effort to evaluate culturally grounded knowledge in Polish, focusing on 600 text-only questions. Similarly, LLMzSzŁ (Jassem et al., 2025) introduces a large-scale evaluation framework for Polish language models based on a collection of national exams, covering nearly 19k closed-ended questions across multiple domains. While comprehensive, it remains restricted to the text modality and does not address multimodal understanding. On the vision-language side, reVISION (Ciesiółka and Graliński, 2025) provides a large-scale Polish benchmark for evaluating VLMs using questions derived from Polish national exams. However, its focus is primarily exam-driven and task-oriented, without explicitly modeling culturally grounded visual-linguistic competence or fine-grained cultural context.

To the best of our knowledge, there is no existing vision-language benchmark specifically designed for Polish cultural and linguistic evaluation beyond exam-based settings. In particular, current resources either focus on text-only evaluation (PLCC, LLMzSzŁ) or exam-centric multimodal reasoning (reVISION), leaving a gap in culturally grounded VLM benchmarks tailored to Polish context and everyday visual semantics.

3 PoVisLE Dataset Construction

The primary objective of the dataset construction was to develop a highly diversified, manually annotated, culturally grounded vision-language dataset designed to evaluate models' understanding of Polish culture and language. Following prior work on culturally situated evaluation of VLMs, as well as existing text-only PLCC benchmark (Dadas et al., 2025), we operationalize culture as a structured combination of tangible cultural artifacts and intangible shared social practices.

In this work, we treat Poland as a proxy for a culturally and linguistically coherent group, assuming a shared baseline of cultural knowledge among annotators and target users, while acknowledging internal regional variation. Importantly, the dataset is not centered on a single urban or institutionalized perspective, which could favor capital-centric culture, but instead deliberately incorporates geographically distributed cultural and linguistic diversity across regions of Poland. This includes region-specific traditions, dialectal and lexical variation, as well as culturally grounded visual concepts tied to particular localities. As a result, the dataset reflects a multi-centered view of Polish culture rather than a homogenized national prototype, ensuring broader coverage of cultural practices and reducing urban or capital-region bias. We assume that a concept is considered culturally relevant if it satisfies at least one of the following requirements: (i) is widely recognized within Poland, either nationally or within specific regional or cultural subgroups, (ii) is taught in primary or secondary education, (iii) appears in shared media discourse, (iv) is necessary for interpreting culturally grounded visual scenes. Therefore, our definition explicitly excludes highly specialized or expert-level knowledge (e.g., university-level domain-specific concepts) that are not part of shared cultural understanding.

Recent work has identified a systematic "grounding gap" in vision-language models: while models can recall factual associations from textual representations, their performance degrades when they must rely on visual inputs referring to the same entities (Ashok et al., 2025). Motivated by this observation, all questions in our dataset are explicitly designed to require visual reference, prevent-

ing models from relying solely on textual associations or memorized knowledge. Critically, this design principle extends also to the *Language* category: even linguistically oriented questions are grounded in visual context, ensuring that the evaluation probes grounded language understanding rather than text-only reasoning.

The dataset construction process was carried out in two stages, described in Sections 3.2 and 3.3, with the overall workflow illustrated in Figure 1.

3.1 Content-based taxonomy

The original six PLCC core categories were *Art and entertainment*, *Culture and tradition*, *Geography and nature*, *History*, *Grammar*, and *Vocabulary*. In the VQA setting, we retain the first four categories and merge *Grammar* and *Vocabulary* into a unified *Language* category. The subcategories were further refined through minor modifications and the introduction of stricter content-matching criteria. An overview of the resulting hierarchical taxonomy, which enables domain-specific error analysis and more precise benchmark diagnostics, is presented in Table 1.

The defined domains cover both fact-based knowledge related to tangible cultural artifacts and elements of cultural reasoning associated with intangible practices, such as symbols, customs, and shared societal references. Unlike some knowledge-driven VQA datasets, e.g., CUS-QA (Libovický et al., 2025), where questions are often constructed around template-based fact retrieval (e.g., "when", "where", "who") and predominantly yield named-entity answers, our dataset prioritizes linguistic and cultural naturalness. Specifically, questions are manually crafted to reflect the values, language patterns, and authentic communicative behaviors of native speakers, emphasizing multi-hop visual understanding, rather than relying on surface-level factual querying.

Furthermore, our approach explicitly incorporates the linguistic dimension, which is often overlooked in multimodal evaluation. The dataset includes questions spanning subcategories such as phraseology, semantics, grammar, dialects and regionalisms, and orthography. This design allows for a more comprehensive assessment of vision-language models, capturing their ability to process culturally grounded language phenomena in multimodal contexts, beyond what is typically evaluated in text-only language benchmarks.

While the benchmark primarily targets culturally grounded reasoning, a small subset (127 out of 2,016 QA pairs) focuses on more general multimodal understanding, categorized as *Image understanding* and *Visual reasoning*. Although these instances do not always require explicit cultural knowledge, they remain embedded in Polish visual and linguistic contexts (e.g., public spaces or Polish-language text), and thus still rely on the model's ability to interpret culturally situated cues.

3.2 Image collection

The visual data constituted the foundation for the subsequent annotation process. The data collection procedure was carried out in two stages.

3.2.1 Manual Curation

In the initial phase, annotators were tasked with the individual selection of images from three primary sources: (i) Wikimedia Commons, (ii) other openly licensed, publicly available datasets or images in the public domain, in compliance with applicable copyright regulations, and (iii) own resources, provided that annotators explicitly consented to waive their copyright prerogatives and contribute the images for project purposes.

The annotators uploaded the selected images to a dedicated GUI-assisted application and were responsible for providing accurate metadata descriptions, including a valid hyperlink and source attribution, along with information on the source type and corresponding licensing conditions.

Due to the more time-consuming nature of sourcing images from personal collections, it was anticipated from the outset that the use of such resources would be limited compared to readily available sources such as Wikimedia Commons. Nevertheless, a targeted proportion of 28.41% of such images was intentionally included. These images are particularly valuable, as they constitute authentic, non-public data that would not have been included in the training pipelines of existing vision-language models. A similar approach has been adopted in culture-specific benchmarks, such as TaiwanVQA (Hsieh et al., 2026).

The majority of the dataset consists of images sourced from Wikimedia Commons (69.18%), while only a small fraction (2.30%) originates from other publicly available resources. In total, 637 images were collected at this stage, including 182 images contributed directly by annotators from their personal collections.

By selecting images themselves, annotators were

Category / Subcategory	Open	Open %	MCQ	MCQ %	Yes/No	Yes/No %	Total	Global %
ART AND ENTERTAINMENT	71	14.8%	233	48.4%	177	36.8%	481	23.9%
Literature	14	15.4%	39	42.9%	38	41.8%	91	4.5%
Sport	12	14.8%	36	44.4%	33	40.7%	81	4.0%
Paintings	16	21.6%	32	43.2%	26	35.1%	74	3.7%
Architecture	8	11.3%	41	57.7%	22	31.0%	71	3.5%
Film	16	25.8%	28	45.2%	18	29.0%	62	3.1%
Music	3	5.4%	31	55.4%	22	39.3%	56	2.8%
Sculpture	2	7.1%	15	53.6%	11	39.3%	28	1.4%
Media	0	0.0%	11	61.1%	7	38.9%	18	0.9%
GEOGRAPHY AND NATURE	30	7.5%	209	52.4%	160	40.1%	399	19.8%
Man-made	15	8.7%	100	57.8%	58	33.5%	173	8.6%
Socio-political	12	10.9%	47	42.7%	51	46.4%	110	5.5%
Inanimate nature	3	4.0%	43	57.3%	29	38.7%	75	3.7%
Animate nature	0	0.0%	19	46.3%	22	53.7%	41	2.0%
HISTORY AND SOCIETY	31	8.9%	161	46.1%	157	45.0%	349	17.3%
Current affairs and society	12	9.7%	57	46.0%	55	44.4%	124	6.2%
Early modern and modern history	8	8.4%	47	49.5%	40	42.1%	95	4.7%
Post-war history	4	8.5%	19	40.4%	24	51.1%	47	2.3%
Middle Ages	3	7.1%	23	54.8%	16	38.1%	42	2.1%
World War II	4	9.8%	15	36.6%	22	53.7%	41	2.0%
CULTURE AND TRADITION	17	5.0%	174	51.5%	147	43.5%	338	16.8%
Religion and tradition	4	4.2%	47	49.5%	44	46.3%	95	4.7%
Cuisine	3	3.2%	44	46.8%	47	50.0%	94	4.7%
Regional and ethnic cultures	4	4.4%	53	58.2%	34	37.4%	91	4.5%
Popculture	6	10.3%	30	51.7%	22	37.9%	58	2.9%
LANGUAGE	29	9.0%	158	49.1%	135	41.9%	322	16.0%
Phraseology	3	3.5%	36	42.4%	46	54.1%	85	4.2%
Semantics	7	10.9%	37	57.8%	20	31.2%	64	3.2%
Grammar	10	19.2%	30	57.7%	12	23.1%	52	2.6%
Dialects and regionalisms	1	2.0%	19	38.8%	29	59.2%	49	2.4%
Colloquial speech and slang	3	13.0%	13	56.5%	7	30.4%	23	1.1%
Language basics and phonetics	0	0.0%	11	61.1%	7	38.9%	18	0.9%
Orthography	5	27.8%	5	27.8%	8	44.4%	18	0.9%
Rhetorical figure	0	0.0%	7	53.8%	6	46.2%	13	0.6%
IMAGE UNDERSTANDING	23	26.7%	43	50.0%	20	23.3%	86	4.3%
VISUAL REASONING	12	29.3%	15	36.6%	14	34.1%	41	2.0%
GLOBAL TOTAL	**213**	**10.6%**	**993**	**49.3%**	**810**	**40.2%**	**2016**	**100.0%**

Table 1: Hierarchical distribution of questions by category and question type (counts and row-wise percentages).

able to draw on their cultural knowledge, experience, and interpretative intuition to construct questions that go beyond surface-level factual queries. In contrast to standard prompts such as "Who is depicted?", the resulting questions often required deeper cultural, historical, or linguistic understanding and reasoning, reflecting aspects of the content that are unlikely to be captured through template-based approaches or fully automated LLM-generated annotation.

However, this approach may also introduce subjective biases, as annotators may favor content aligned with their personal experience or regional background. This risk was taken into account in the subsequent design of the dataset.

3.2.2 Data Augmentation

The second phase of data collection aimed to mitigate selection bias and increase dataset diversity by assigning annotators images from Poland-related categories on Wikimedia Commons, rather than allowing them to select images independently. Annotators first assessed whether an image was suitable for culturally grounded VQA and, if so, formulated corresponding questions.

To account for domain expertise, annotators could mark images as suitable but outside their knowledge, enabling reassignment to more qualified annotators. This ensured both relevance and high-quality, contextually grounded annotations. Additionally, the use of Wikimedia Commons improved efficiency by eliminating the need for manual metadata curation.

To further enhance diversity, an augmentation step introduced for questions originally linked to the same image. We searched for alternative images within the same source categories as the original image and ranked the candidates by visual similar-

ity using CLIP embeddings. The highest-ranked non-identical candidates were selected for manual review by annotators. Region-specific or culturally ambiguous replacements were verified during this review before inclusion. This process increased the number of unique images from 637 to 966 while preserving the original set of questions.

3.3 Data annotation

Once an image was selected and its associated metadata had been provided (either manually during the initial phase or automatically in the second phase) the VQA annotation process was carried out. Annotators were tasked with the following:

- coupling each image with between 1 and 10 questions,
- assigning topic coverage by classifying each question into one of the 7 main categories and 29 corresponding subcategories,
- labeling the question type, including *open-ended*, *multiple-choice*, or *yes/no* formats.

The annotation process was governed by the following fundamental rules, aligned with the ultimate goal of reliable and deterministic evaluation of large language models:

- **No ambiguity:** Questions were required to be precise and unambiguous, allowing for a single clearly defined and non-debatable answer.
- **Visual grounding:** Questions had to rely on the visual content of the image and could not be answerable based solely on general knowledge without access to the image.
- **Strict task formulation:** Each question was formulated as a prompt containing explicit instructions regarding the expected answer format, ensuring consistent responses from evaluated models.

More details on annotation guidelines can be found in Appendix D.

The annotation team consisted of two primary annotators with background in linguistics, supported by an expert-level super-annotator responsible for continuous quality control. The super-annotator played a key role in identifying ambiguities, ensuring annotation consistency, and maintaining adherence to the defined guidelines throughout the process.

In addition, a controlled subset of inputs was contributed by seven auxiliary annotators, selected to ensure demographic diversity, particularly with respect to regional representation within Poland. The included regions covered Warsaw and the Masovian Voivodeship, Greater Poland, Lower Silesia, Upper Silesia, and Lesser Poland. This design allowed the dataset to incorporate a broader range of cultural perspectives while maintaining overall annotation consistency through centralized supervision.

3.4 Quality Assurance

To ensure annotation quality and consistency, a multi-stage validation process was implemented throughout dataset construction. First, a cross-checking procedure was applied, whereby annotations were reviewed by a second annotator to identify potential errors, ambiguities, or inconsistencies.

In addition, regular team meetings and discussions were conducted to address emerging issues, clarify annotation guidelines, and provide continuous feedback to annotators. This iterative process allowed for the refinement of annotation practices and improved alignment across contributors.

A central role in quality assurance was fulfilled by an expert-level super-annotator, who systematically reviewed representative samples of the dataset. The super-annotator was responsible for identifying recurring ambiguities, detecting inconsistencies in labeling and question formulation, and ensuring adherence to the established guidelines. Based on this analysis, targeted feedback was provided to annotators, and problematic instances were corrected or revised where necessary.

This combination of cross-validation, collaborative feedback, and centralized supervision ensured high annotation quality, consistency, and reliability across the dataset.

4 Dataset statistics

The dataset comprises 637 manually curated images paired with a total of 2,016 question–answer pairs, resulting in an average of 3.16 questions per image. In terms of image provenance, 69.54% of the data was sourced from Wikimedia Commons, 28.41% consists of annotator-provided (custom) images, and the remaining 2.04% originates from other permissively licensed resources.

Regarding question types, the dataset is dominated by multiple-choice questions (49.3%), followed by binary (yes/no) questions (40.2%) and

a smaller proportion of open-ended questions (10.6%). After the data augmentation phase, which introduced additional images while preserving the original question set, the total number of unique images increased to 966, reflecting improved visual diversity without altering the overall distribution of question types. In the final dataset the percentage of annotators' custom images decreased to 18.74%, while Wikimedia Commons covers 79.91% of the images.

5 Experiment Setup

Models are evaluated on the PoVisLE test split using the described evaluation pipeline. For each instance, the model is given the image and a prompt with a question, which explicitly specifies the expected answer format, length, word order, and, where relevant, grammatical form.

For multiple-choice questions, we use circular evaluation. Answer options are cyclically permuted (e.g., for three options, A, B, C, we evaluate the orders A, B, C, B, C, A, and C, A, B), and an instance is counted as correct only when the model selects the correct answer under all rotations. This reduces the effect of option-position bias while requiring fewer evaluations than checking all possible permutations. Yes/no and open-ended questions are evaluated in a single pass. Yes/no questions always require a binary *yes* or *no* answer in Polish. Open-ended predictions are compared with the gold answer after normalization (including lowercasing, diacritic removal, and alphanumeric/whitespace normalization). A prediction is correct if it exactly matches the normalized gold answer or contains it as a substring.

5.1 Models

We evaluate both open-weight and proprietary VLMs, across several model families and scales. Where available, we also evaluate reasoning variants of the selected models. The evaluated models include Mistral, Qwen, Gemma, GLM, LLaVA-based models, including Polish-oriented LLaVa-PLLuM and LLaVa-Bielik (Statkiewicz et al., 2026), and GPT-family proprietary models. We also report a random baseline.

Open-weight models are served locally using vLLM, through a unified backend that applies the corresponding model processors and chat templates. Proprietary and externally hosted models are evaluated through API backends. Unless a model-specific configuration requires otherwise, we use deterministic decoding with temperature set to zero and top-p set to 1.0. For reasoning variants, we follow the settings recommended by model cards, including the appropriate reasoning mode, decoding parameters and generation budget. All models are evaluated using the same prompts. Further computational details are in Appendix E.

5.2 Metrics

We report accuracy as the main evaluation metric. An example is counted as correct if the parsed model prediction satisfies the task-specific scoring rule described above, including correct answers for all circular variants in the multiple-choice setting.

5.3 Single modality input

To estimate how much of the benchmark can be solved without visual grounding, we also evaluate models in a single-modality textual setting, where the prompt remains unchanged, but the image is removed. The same scoring rules are used as in the full multimodal setting. This comparison shows whether the proposed benchmark requires models to use the image, as intended, or whether the textual input alone is sufficient to answer correctly.

6 Results and Discussion

Table 2 reports our main results. GPT-5.4 obtains the highest overall score, with 75.89% accuracy, and ranks first on all five culture- and language-oriented categories. Among open-weight models, Gemma-4-31B-it (Thinking) obtains the highest overall score (72.02%), followed by Gemma-4-31B-it (67.21%), Qwen3.5-122B-A10B (Thinking) (66.77%), and GLM-4.6V (65.62%). Most evaluated models score above the random baseline, but only GPT-5.4 and Gemma-4-31B-it (Thinking) exceed 70% overall.

Among the main cultural categories, *Culture and tradition* has the highest scores for the top-ranked models. GPT-5.4 scores 82.84% in this category, and Gemma-4-31B-it (Thinking) scores 78.40%. *Art and entertainment* has the lowest scores among the same categories, with 68.81% for GPT-5.4 and 59.25% for the best open-weight models. The remaining main categories fall between these two and do not change the overall model ranking.

Within the evaluated families, larger variants generally score higher. This ordering appears for Qwen, Gemma, and InternVL, although model family remains important. Gemma-4-31B-it (Thinking)

Model	Overall	Art & Entert.	Culture & Trad.	Geogr. & Nature	History & Society	Language	Image Und.	Visual Reas.
PROPRIETARY MODELS								
GPT-5.4	**75.89**	**68.81**	**82.84**	**74.94**	**77.65**	**77.64**	86.05	60.98
GPT-5.4-mini	62.50	56.96	68.93	63.66	63.32	60.87	74.42	43.90
OPEN-WEIGHTS MODELS								
Gemma-4-31B-it (*Thinking*)	72.02	59.25	78.40	73.68	72.78	76.71	89.53	73.17
Gemma-4-31B-it	67.21	56.34	76.04	67.67	67.05	67.39	90.70	68.29
Qwen3.5-122B-A10B (*Thinking*)	66.77	52.81	72.19	63.91	73.64	70.19	**91.86**	**75.61**
GLM-4.6V	65.62	59.25	64.79	68.42	69.63	62.11	88.37	65.85
Qwen3.5-27B (*Thinking*)	61.86	46.36	64.50	58.90	69.05	69.25	88.37	**75.61**
Qwen3.5-122B-A10B (*Instruct*)	57.59	45.95	61.83	62.91	61.03	53.11	86.05	53.66
Mistral-Medium-3.5-128B	56.60	49.48	61.83	57.89	57.59	53.11	83.72	46.34
Qwen3.5-27B (*Instruct*)	55.51	42.41	60.36	58.40	57.88	54.35	90.70	56.10
Qwen3.5-9B (*Thinking*)	54.91	42.00	60.36	53.63	61.60	53.42	83.72	68.29
LLaVA-OneVision-Qwen2-72B-OV-HF	50.79	47.19	60.65	51.88	52.72	45.65	52.33	21.95
InternVL3.5-38B	49.01	44.91	51.78	50.38	48.71	44.72	75.58	41.46
LLaVA-Bielik-11B-v2.6	48.36	43.66	55.62	49.62	51.29	40.68	68.60	24.39
Ministral-3-14B-Instruct-2512	47.67	39.71	50.89	49.12	51.00	44.72	73.26	41.46
Ministral-3-8B-Instruct-2512	46.08	40.75	51.48	46.62	48.14	41.93	61.63	41.46
LLaVA-PLLuM-12B-250715	45.04	43.45	52.07	45.61	47.56	38.20	50.00	21.95
InternVL3.5-14B	44.35	40.96	47.63	44.86	47.56	37.89	59.30	43.90
Qwen3.5-9B (*Instruct*)	44.00	33.89	46.45	45.11	48.42	40.37	80.23	46.34
LLaVA-PLLuM-12B	43.11	41.37	50.00	43.86	44.13	36.65	52.33	21.95
InternVL3.5-8B	41.47	39.29	44.08	42.61	40.97	36.02	61.63	39.02
LLaVA-OneVision-Qwen2-7B-OV-HF	39.83	40.54	47.34	41.10	38.40	32.92	41.86	19.51
LLaVA-v1.6-Mistral-7B-HF	32.14	27.44	37.28	32.83	34.10	26.71	47.67	31.71
SINGLE-MODALITY TEXTUAL INPUT								
Qwen3.5-122B-A10B (*Thinking*)	39.83	34.51	49.41	34.34	47.85	41.93	23.26	26.83
Qwen3.5-27B (*Thinking*)	35.37	29.31	39.94	28.32	43.55	41.30	27.91	36.59
GLM-4.6V	33.53	28.69	39.94	29.32	40.11	34.47	26.74	29.27
Qwen3.5-122B-A10B (*Instruct*)	32.39	26.82	36.39	29.82	37.25	36.34	25.58	31.71
LLaVA-Bielik-11B-v2.6	32.34	30.56	37.57	28.82	37.25	33.54	23.26	12.20
Random	19.89	18.81	21.15	19.67	22.06	20.96	10.47	17.07

Table 2: Model accuracy by dataset category, with all values reported as percentages. The **best result** in each column is shown in bold, and the best result among open-weight models is underlined.

is the highest-scoring open-weight model overall, despite being smaller than Qwen3.5-122B-A10B (Thinking). The LLaVA-PLLuM variants also fall in the same range as other models of comparable size, with 43.11% and 45.04% overall accuracy, close to InternVL3.5-14B at 44.35%. Reasoning variants score higher for Gemma and Qwen in the evaluated pairs.

The single-modality textual rows in Table 2 report the five highest-scoring variants in this setting. The full single-modality textual results are reported in Table 6 of Appendix F. The random baseline is 19.89% overall, while the best single-modality textual result reaches 39.83%. The same model reaches 66.77% when the image is provided, and the best full-setting result is 75.89%. These differences indicate that textual cues alone do not account for the scores obtained in the full setting and that PoVisLE requires combining the Polish question with visual evidence.

7 Conclusion and Future Work

In this work, we introduced PoVisLE, a culturally grounded vision-language benchmark designed to evaluate multimodal models on Polish cultural and linguistic competence. The dataset combines manual, template-free annotation with a grounded evaluation paradigm, in which correct answers depend on the interaction between linguistic input and visual context. By focusing on a single, well-defined cultural setting, our approach enables a controlled and fine-grained assessment of culturally situated understanding.

Looking forward, our results highlight that open-ended questions constitute the most informative yet challenging evaluation setting, as they require models to generate precise, contextually grounded answers rather than select from predefined options. However, it also exposes limitations of current evaluation protocols. Strict exact-match evaluation, although reliable, can be overly rigid given natural linguistic variation. Future work should therefore focus on developing more robust and interpretable evaluation metrics that can balance flexibility with determinism, enabling reliable assessment of open-ended, multi-clause question answering without sacrificing fairness or reproducibility.

Limitations

We acknowledge that constructing a culturally grounded benchmark of this nature involves inherent trade-offs between annotation quality, dataset size, and thematic coverage. Our decision to avoid template-based generation and synthetic data improves linguistic naturalness and cultural authenticity, but limits the scale of the dataset and may constrain the breadth of covered topics.

A central challenge lies in the evaluation of open-ended answers. While this sort of question offers the richest signal for assessing model capabilities, it is difficult to evaluate automatically using standard metrics. Approaches such as LLM-as-a-judge are particularly problematic in this setting, as they rely on models that may themselves exhibit cultural biases, creating a paradox when evaluating cultural competence. As a result, open-ended evaluation introduces both methodological complexity and potential bias.

Traditional evaluation methods based on string matching (e.g., exact match or token-level overlap) are insufficient for culturally grounded tasks, where multiple valid expressions may exist. To ensure reliable and reproducible evaluation, we therefore rely primarily on multiple-choice and binary formats. While these formats allow for deterministic evaluation, they also introduce their own limitations: models may succeed by eliminating implausible options rather than demonstrating genuine understanding. Although techniques such as circular evaluation help mitigate positional biases, they do not fully address this issue.

An additional challenge arises from the relationship between culturally prototypical expressions, annotator perspective, and conceptual variability. Annotators, as members of the target culture, naturally operate from an insider (emic) perspective and tend to favor culturally salient and widely shared labels. For example, certain architectural forms characteristic of the socialist era in Poland are commonly referred to as *wielka płyta*, which functions as a culturally dominant prototype. However, the same phenomenon admits multiple valid descriptions, including more technical or general formulations (e.g., prefabricated housing structures), which may be preferred by experts or non-local observers.

This creates a perspective asymmetry, where culturally canonical answers are privileged over alternative, semantically correct interpretations. Furthermore, annotators may implicitly treat certain phenomena as culturally specific or unique, even when they are in fact shared across multiple regions. This may lead to over-localization, where questions assume a single culturally grounded interpretation despite the existence of broader or cross-cultural variants. As a result, models may be penalized for producing correct but non-prototypical answers, especially when these reflect alternative cultural, technical, or global perspectives.

Together, these effects highlight a fundamental challenge in culturally grounded evaluation: reconciling the need for precise and verifiable answers with the inherently plural, context-dependent nature of cultural and conceptual knowledge.

In our approach, we mitigate the impact of these biases by either reformulating questions to enforce a strict and unambiguous response format or by converting such instances into closed-ended questions. While this strategy improves evaluation consistency and reduces ambiguity, it introduces an inherent trade-off between linguistic and conceptual flexibility on the one hand and evaluation reliability on the other.

References

Dhananjay Ashok, Ashutosh Chaubey, Hirona J Arai, Jonathan May, and Jesse Thomason. 2025. Can vlms recall factual associations from visual references? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15691–15708.

Michał Ciesiółka and Filip Graliński. 2025. revision: A polish benchmark for evaluating vision-language models on multimodal national exam data. In *2025 20th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 665–673. IEEE.

Sławomir Dadas, Małgorzata Grebowiec, Michał Perełkiewicz, and Rafał Poświata. 2025. Evaluating polish linguistic and cultural competency in large language models. In *International Conference on Artificial Intelligence and Soft Computing*, pages 60–71. Springer.

V. A. Elisiário and W. M. Watanabe. 2025. Multimodal large language models for portuguese alternative text generation for images. In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 493–501. SCITEPRESS – Science and Technology Publications, Lda.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Hsin Yi Hsieh, Shang-Wei Liu, Chang-Chih Meng, Chien-Hua Chen, Shuo-Yueh Lin, Hung-Ju Lin, Hen-Hsen Huang, I Wu, and 1 others. 2026. Taiwanvqa: Benchmarking and enhancing cultural understanding in vision-language models. *Advances in Neural Information Processing Systems*, 38.

Krzysztof Jassem, Michał Ciesiółka, Filip Graliński, Piotr Jabłoński, Jakub Pokrywka, Marek Kubis, Monika Jabłońska, and Ryszard Staruch. 2025. LLMzSzŁ: a comprehensive LLM benchmark for Polish. *arXiv preprint arXiv:2501.02266*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karima Kadaoui, Hanin Atwany, Hamdan Al-Ali, Abdelrahman Mohamed, Ali Mekky, Sergei Tilga, Natalia Fedorova, Ekaterina Artemova, Hanan Aldarmaki, and Yova Kementchedjhieva. 2026. Jeem: Vision-language understanding in four arabic dialects. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 331–354.

Jindřich Libovickỳ, Jindřich Helcl, Andrei Manea, and Gianluca Vico. 2025. Cus-qa: Local-knowledge-oriented open-ended question answering dataset. *arXiv preprint arXiv:2507.22752*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Moritz Mähr and Moritz Twente. 2025. Seeing history unseen: Evaluating vision-language models for wcag-compliant alt-text in digital heritage collections. *Anthology of Computers and the Humanities*, 3:1148–1168.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, Sriparna Saha, and 1 others. 2025. Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790.

Inez Okulska, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głąbińska, and Adam Nowakowski. 2023. Stylometrix: An open-source multilingual tool for representing stylometric vectors. *arXiv preprint arXiv:2309.12810*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Grzegorz Statkiewicz, Alicja Dobrzeniecka, Karolina Seweryn, Aleksandra Krasnodębska, Karolina Piosek, Katarzyna Bogusz, Sebastian Cygert, and Wojciech Kusa. 2026. Annotation-efficient vision-language model adaptation to the polish language using the llava framework. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 569–589.

Bryan Chen Zhengyu Tan, Weihua Zheng, Zhengyuan Liu, Nancy Chen, Hwaran Lee, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2026. Blend-vis: Benchmarking multimodal cultural understanding in vision language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4669.

Špela Vintar, Taja Kuzman Pungeršek, Mojca Brglez, and Nikola Ljubešić. 2025. Charting the european llm benchmarking landscape: A new taxonomy and a set of best practices. *arXiv preprint arXiv:2510.24450*.

Yuxuan Wang, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, Wanxiang Che, and Hongyang Chen. 2025. Cvlue: A new benchmark dataset for chinese vision-language understanding evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8196–8204.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2025. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. 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 3242–3264.

Srishti Yadav, Lauren Tilton, Maria Antoniak, Taylor Arnold, Jiaang Li, Siddhesh Milind Pawar, Antonia Karamolegkou, Stella Frank, Zhaochong An, Negar Rostamzadeh, and 1 others. 2025. Evaluation of cultural competence of vision-language models. *arXiv preprint arXiv:2505.22793*.

Amber Yijia Zheng, Jae Joong Lee, Bedrich Benes, and Raymond A Yeh. 2025. Webaccessvl: Making an accessible web via violation-conditioned vlm. *arXiv preprint arXiv:2602.03850*.

A Dataset examples and statistics visualization

Figure 2 presents representative examples from the dataset, illustrating the diversity of question types and linguistic formulations, while figure 3 visualizes this distribution in a hierarchical form, offering an intuitive overview of the balance between high-level categories and their internal structure.

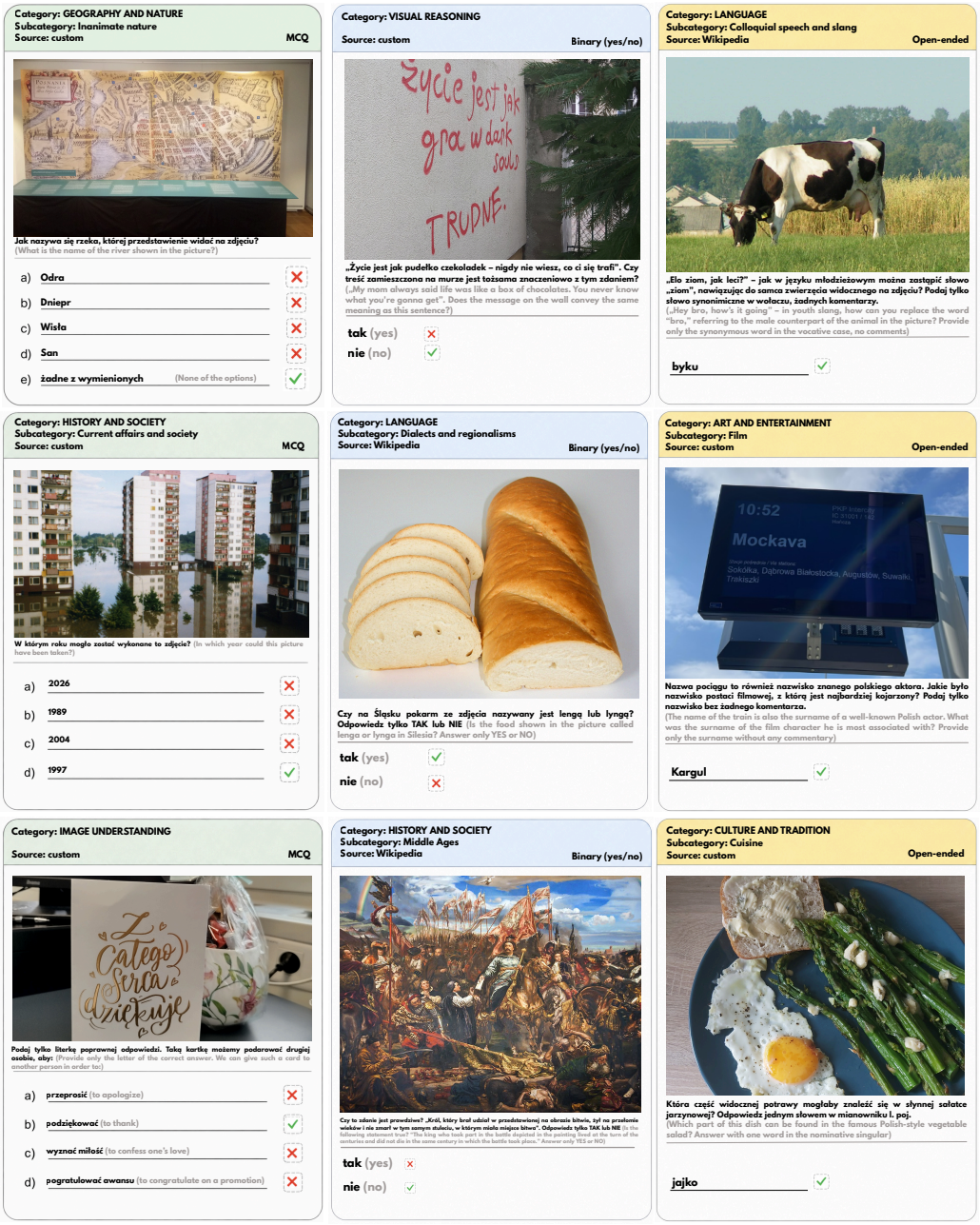

Figure 2: Examples from the dataset illustrating multiple-choice, binary (yes/no), and open-ended questions across all seven main categories. Image sources include Wikimedia Commons and annotators' personal collections contributed to the project. English translations are omitted for Polish-specific named entities and for open-ended answers requiring Polish vocabulary.

B Linguistic Analysis of Questions

To quantify the linguistic properties of the dataset, we focus exclusively on questions, as answers are typically very short (one to two tokens) or provided as multiple-choice options, limiting their usefulness for stylometric analysis.

The average question length is 22.4 tokens (see Figure 4), indicating that questions are relatively long and often involve more complex structures.The distribution is right-skewed, with most questions concentrated around the mean (22.4 tokens), and a long tail extending to much longer, multi-clause formulations.

B.1 Stylometric features

To further investigate the linguistic nature of the dataset, we employed StyloMetrix (Okulska et al., 2023), a library designed for Polish, which represents texts as vectors of interpretable linguistic features normalized to the range $[0, 1]$. We focused in particular on metrics capturing lexical diversity and syntactic structure, including indicators associated with the presence of subordinate clauses, which may signal multi-hop reasoning. Selected metrics are reported in Table 3.

Interestingly, we observe a minimal difference between the surface-form type–token ratio (L_TTR_IA) and its lemmatized counterpart (L_TTR_LA) (0.758 vs. 0.749), despite the fact that many questions follow similar, though not identical, task formulations designed to ensure unambiguous evaluation of model behavior.This suggests that lexical diversity is not primarily driven by morphological variation, which is characteristic of richly inflected Slavic languages, but instead reflects genuinely diverse vocabulary usage.

Furthermore, the combination of a high named entity ratio (0.13) and a high type–token ratio indicates that entity mentions are not dominated by a small set of frequently repeated references, but instead span a broad range of distinct entities. This contributes to the dataset's topical and regional diversity. This observation is reinforced by the small gap between content word incidence and content word types (0.511 vs. 0.499), suggesting that many content words occur only once and thus contribute directly to lexical diversity.

The proportion of pronouns in the dataset is comparable to that of adjectives (0.07 vs. 0.079), indicating that reference to entities is frequently realized through pronominal forms rather than descriptive modification. Relative and interrogative pronouns further contribute to the prevalence of subordinate structures, while the presence of negative pronouns reflects the inclusion of adversarial question formulations.

Despite the expectation that a question dataset should consist predominantly of interrogative sentences, our dataset contains a slightly higher proportion of tokens associated with declarative constructions (0.523 vs. 0.498). This can be partly attributed to task formulations, which are often expressed in short declarative phrases.

However, these formulations alone do not account for the observed distribution. The predominance of declarative tokens is further driven by the frequent use of contextual or narrative framing preceding the actual question. While the interrogative component itself may be relatively short, it is often embedded within a longer descriptive context, resulting in a higher overall proportion of declarative structures. A smaller, yet non-negligible proportion of tokens corresponds to negative constructions (0.048), indicating the presence of adversarial or contrastive question formulations.

Finally, high values for words within modifiers (0.236) and words in nominal phrases (0.598) indicate a strongly periphrastic style of question formulation. This style likely supports the construction of indirect, image-grounded references and enables the formulation of questions that require interpretation beyond explicitly depicted visual elements.

B.2 Question-level syntactic complexity

To further illustrate the distribution of complexity across the questions and having thoroughly examined stylometric features extracted with StyloMetrix, we aimed to quantify syntactic complexity, which may indicate the need for multi-hop reasoning and contribute to the overall linguistic difficulty of the questions. To this end, we introduce three custom metrics based on rule-based detection of subordinate structures: (i) Relative Clause Proportion (RCP), (ii) Subordinate Conjunction Proportion (SCP), and (iii) Subordination Proportion (SP), which captures the presence of either structure.

Each metric is computed at the question level as the proportion of questions containing at least one instance of the corresponding structure. For all metrics, sentence-initial tokens (including capitalized pronouns and conjunctions) are excluded, as they typically correspond to interrogative openings

rather than embedded subordinate constructions.

To further characterize the distribution of structural complexity across the questions, and building on the stylometric analysis performed with StyloMetrix, we aimed to quantify syntactic complexity, which may indicate the need for multi-hop reasoning and contribute to the overall linguistic difficulty of the dataset.

To this end, we introduce three custom metrics based on rule-based detection of subordinate structures: (i) Relative Clause Proportion (RCP), (ii) Subordinate Conjunction Proportion (SCP), and (iii) Subordination Proportion (SP), which captures the presence of either structure.

Each metric is computed at the question level as the proportion of questions containing at least one instance of the corresponding structure. For all metrics, sentence-initial tokens (including capitalized pronouns and conjunctions) are excluded, as they typically correspond to interrogative openings rather than embedded subordinate constructions.

The resulting values are as follows: RCP = 0.318, SCP = 0.101, and SP = 0.381. The resulting values are as follows: RCP = 0.318, SCP = 0.101, and SP = 0.381. This also reveals an overlap between the two indicators, corresponding to questions that contain both a relative clause and a subordinate conjunction (3.7% of the dataset), indicating the presence of more complex, multi-layered clause structures. Importantly, these metrics are computed as sentence-level incidence proportions rather than token-level ratios. This design choice is motivated by the presence of task formulations, which could otherwise inflate token-based measurements and obscure the true distribution of structurally complex questions.

Overall, the dataset exhibits high lexical diversity, rich semantic content, and structurally complex, multi-clause question formulations, indicating that it is not dominated by simple templates but supports more demanding, context-driven reasoning.

Metric	Description	Score
Grammatical Forms		
G_N	Nouns	0.262
G_V	Verbs	0.094
G_ADJ	Adjectives	0.079
G_ADV	Adverbs	0.021
G_PRO	Pronouns	0.070
G_PRO_NEG	Negative pronouns	0.004
G_PRO_REL	Relative pronouns	0.006
G_PRO_INT	Interrogative pronouns	0.019
G_CONJ	Conjunctions	0.031
G_CCONJ	Coordinating conjunctions	0.028
G_SCONJ	Subordinating conjunctions	0.003
Punctuation		
PUNCT_TOTAL	Total punctuation	0.163
Syntactic		
SY_MOD	Words within modifiers	0.236
SY_NPHR	Words in nominal phrases	0.598
SY_S_DE	Words in declarative sentences	0.523
SY_S_IN	Words in interrogative sentences	0.498
SY_S_NEG	Words in negative sentences	0.048
SY_QUOT	Words in quotation marks	0.002
Lexical		
L_TTR_IA	Type-token ratio for non-lemmatized tokens	0.758
L_TTR_LA	Type-token ratio for lemmatized tokens	0.749
L_CONT_A	Incidence of content words	0.511
L_CONT_T	Content word types	0.499
L_FUNC_A	Incidence of function words	0.258
L_FUNC_T	Function words types	0.243
L_NAME_ENT	Named entities	0.130
L_NAME	Proper names	0.051
L_PERSN	Person names	0.084
L_PLACEN_GEOG	Place and geographical names	0.016

Table 3: Stylometric feature distribution of the dataset across grammatical, syntactic, and lexical categories.

C Comparison to Other Datasets

Table 4 compares PoVisLE with existing culturally grounded VQA benchmarks across scale, geographic coverage, and key dataset properties. While some prior datasets achieve larger scale by covering multiple countries or regions, PoVisLE adopts a monocultural design focused on a single, well-defined context. This allows for a more controlled and fine-grained evaluation of cultural and linguistic competence. In particular, we emphasize fully manual, creative question construction, the use of a local (non-English) language, and the inclusion of non-synthetic visual data, including images sourced from annotators' private collections that are not publicly available online.

We additionally include CUS-QA in the comparison. While it is primarily knowledge-oriented, it is embedded in the Czech, Slovak, and Ukrainian context, which is relevant due to the geographic and cultural proximity of these regions.

When reporting dataset statistics, we provide numbers to the best of our knowledge, focusing specifically on the visual question answering setting. In cases where datasets support multiple evaluation formats, only the visual components are considered to ensure comparability. The only exception is PLCC, which we include as a baseline

Figure 3: Distribution of 1,889 culturally grounded questions across categories and subcategories in the dataset. Numbers in parentheses indicate the number of samples. The figure excludes the additional categories *Image Understanding* and *Visual Reasoning* which are not divided into subcategories.

for Polish cultural understanding; notably, it is a text-only benchmark.

We decided not to include the Polish Cultural Vision Benchmark (PCVB) v2 in this comparison.[2] While it is described as a benchmark for Polish cultural understanding in vision-language models, no detailed documentation regarding its construction, annotation methodology, or evaluation protocol is publicly available, which prevents a meaningful and reproducible comparison.

[2] https://huggingface.co/spaces/speakleash/Polish_Cultural_Vision_Benchmark

D Annotation Guidelines

D.1 Manual data collection and annotation

The aim of the dataset annotation is to manually craft questions regarding visual content defined by the following categories:

- **Art & Entertainment**
  - Architecture
  - Film
  - Literature
  - Media
  - Music
  - Paintings

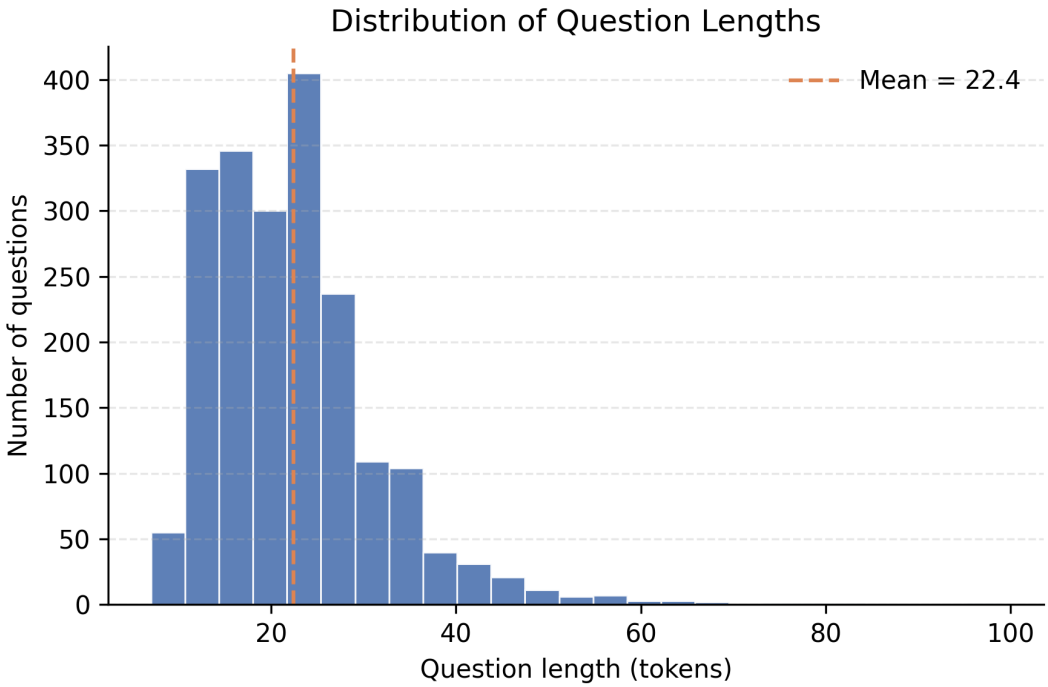

Figure 4: Distribution of question lengths (in tokens) across the dataset. The dashed vertical line indicates the mean question length (22.4 tokens).

  - Sculpture
  - Sport

- **Geography & Nature**
  - Animate nature
  - Inanimate nature
  - Man-made
  - Socio-political

- **Culture & Tradition**
  - Cuisine
  - Pop culture
  - Regional and ethnic cultures
  - Religion and tradition

- **History & Society**
  - Middle Ages
  - Early modern and modern history
  - World War II
  - Post-war history
  - Current affairs and society

- **Language**
  - Colloquial speech and slang
  - Dialects and regionalisms
  - Grammar
  - Language basics and phonetics
  - Orthography
  - Phraseology
  - Rhetorical figures
  - Semantics

- **Image Understanding**

- **Visual Reasoning**

The annotator's task consists of the following steps:

1. **Image selection:** Choose an image from one of the following sources:
   - Wikimedia Commons
   - Publicly available sources (after careful verification of the license terms)
   - Personal image collections
2. **Image upload and metadata annotation:** Upload the selected image to the annotation interface and provide the required metadata, including:
   - A source hyperlink, or
   - A declaration confirming ownership of the image and agreement to share it for

Dataset	Regions	#Img	#Q	Lang.	Manual	MCQ	Open	Own Img.	Synth Img.
CULTURALVQA	11 (global)	2,328	2,378	EN	✓	✗	✓	✗	✗
BLEND-VIS	16 (global)	4,916	21,782	EN	✗	✓	✗	✗	✓
DRISHTIKON	1 (IN)	2,126	64,288	EN + dialects	✗	✓	✗	✗	✗
JEEM	4 (MENA)	2,178	10,890	AR	✓	✗	✓	✓	✗
CVLUE	1 (CN)	30,009 (?)	72,306 (?)	ZH	✓	✗	✓	✗	✗
TaiwanVQA	1 (TW)	2,736	5,472	ZH	✓	✓	✓	✓	✗
CUS-QA	3 (CZ, SK, UA)	(?)	1,097	CZ, SK, UA	✓	✗	✓	✗	✗
PLCC	1 (PL)	—	600	PL	✓	✓	✓	—	—
PoVisLe	**1 (PL)**	**966**	**2,016**	**PL**	✓	✓	✓	✓	✗

Table 4: Comparison of culturally grounded VQA datasets and culture-aware benchmarks. While most listed resources are visual question answering (VQA) datasets, we additionally include PLCC as a text-only baseline for Polish culture-aware evaluation. Regions are reported as the number of covered countries or regions with indicative geographic scope. Checkmarks indicate key dataset properties, including annotation type, use of annotator-provided (non-web) images, and the presence of synthetic images.

research purposes under the CC BY-SA license

3. **VQA pair creation:** Create from one up to ten question–answer (VQA) pairs based on the image content.

4. **Annotation labels:**

   - **Task type:** multiple-choice, binary (yes/no), open-ended questions
   - **Category:** category and subcategory

Follow the rules and guidelines for annotation:

- **Formulation of questions:** Construct natural-sounding questions that reflect authentic language use. Ensure linguistic correctness and fluency. Avoid repetitive or overly uniform question structures; instead, aim for diversity in formulation and richness of vocabulary.

- **Answer unambiguity:** Ensure that all answers are unambiguous and allow for an objective and verifiable evaluation. In the case of multiple-choice questions (MCQ), the distinction between correct and incorrect options must be clear and indisputable. Avoid answer choices that could be interpreted as correct depending on regional variation, interpretation, or context (e.g., cases where answer A may be valid in one region while answer B is accepted elsewhere). For open-ended questions, aim to formulate them in such a way that only one answer is clearly plausible. Where appropriate, include subtle guidance in the question (e.g., expected format or level of specificity) to reduce variability in answers. If in doubt regarding the ambiguity of a VQA pair, consult with the super-annotator or discuss it with the team to ensure consistency and agreement across annotators.

- **Coverage and difficulty:** Ensure that the dataset covers a broad range of thematic categories and includes varying levels of difficulty. Questions may rely on cultural and region-specific knowledge that is broadly accessible through public education, media, and cultural institutions. However, they should not require specialized academic knowledge or expert-level training.

- **Dependence on visual content:** All questions must require image understanding. The answer should not be obtainable from textual knowledge alone without reference to the image.

- **Image-referential language:** Use image-grounded expressions (e.g., "the man in the picture") rather than overly specific or leading descriptions (e.g., "the famous musician in the picture"). Avoid including clues that reveal the answer directly.

- **Reasoning and multi-hop questions:** Whenever possible, formulate questions that require multi-step reasoning or combining visual information with general knowledge.

- **Use of associations:** Treat the image as a visual anchor that can connect to broader cultural or knowledge contexts. Questions do

not need to refer strictly to what is explicitly shown; they may involve associated entities or concepts, as long as the connection is clearly grounded in the visual content and can be reasonably inferred.

- **Adversarial questions:** Include a subset of carefully designed adversarial questions to evaluate models' robustness. These questions should be intentionally challenging, e.g., involving misleading visual cues, while still remaining answerable based on the image.

- **Design of incorrect answer options:** In multiple-choice questions (MCQ), ensure sufficient variation and plausibility among incorrect answers. For lower-difficulty questions, distractors (incorrect options) may be more similar to the correct answer, requiring careful distinction. For higher-difficulty questions, distractors may be less subtle; however, they should remain relatively plausible. Avoid overly absurd or clearly incorrect options, as these allow the correct answer to be identified through simple elimination without reference to the image.

D.2 Augmented Data Annotation

D.2.1 Image Filtering and Annotation

At the stage of augmented data annotation the task is to decide whether the pre-collected set of images can be annotated in consistence with the established guidelines and insights from the previous stage of manual annotation. The set of images contains different Wikimedia-based images retrieved from the Poland-related categories.

The annotator's task is to:

- Annotate the VQA pair and assign the appropriate task type and category, following the procedures defined in the first stage.

- If the image is rejected, assign one of the following labels:
  - Low quality
  - Not suitable for VQA
  - Potentially relevant but outside my domain of expertise

D.2.2 Question-Guided Image Matching

At the final stage of the augmented data annotation process, annotators are tasked with identifying images that match previously created questions. Since multiple questions may be associated with a single source image, this step aims to diversify the dataset by retrieving visually similar images from Wikimedia Commons for further human inspection. The goal is to distribute questions across a broader set of non-identical images whenever possible.

The matching procedure is based on the previously constructed annotations (image–question pairs) and follows these guidelines:

- Verify whether a given question can be answered using any of the automatically retrieved candidate images other than the original one. If so, mark the image as a valid match.

- Note that some candidate images may be near-duplicates (e.g., slightly different crops or framing). Such cases should not be treated as distinct images.

- For artworks such as paintings, partial views of the original canvas may be labeled as valid matches, provided that they contain sufficient information to answer the question.

- If none of the candidate images are applicable, submit the task and proceed to the next sample.

D.3 Cross-validation

At this stage, the task is to review annotations created by other team members through careful manual inspection.

- Verify that all required metadata is present and that licensing terms have been correctly applied.

- Check that the question is linguistically correct, natural in phrasing, and that the answer is unambiguous and factually accurate.

- Ensure that the question is properly grounded in the image and cannot be answered without reference to it.

- Consider that annotators may come from different cultural or regional backgrounds. Use this diversity to identify potential biases or assumptions embedded in the question. In particular, verify whether the association or interpretation made by the original annotator is universally understandable or strongly tied to a specific cultural or regional context. If a

question relies on knowledge or associations that may not be broadly shared, reformulate it to ensure wider accessibility or make the required regional context explicit.

- If issues are identified, take appropriate corrective actions:
  - Reformulate the question for clarity or correctness
  - Modify the question type if necessary
  - Remove the question and create a new one if it does not meet the guidelines
- If any uncertainty arises, consult with other annotators or a super-annotator to ensure consistency and quality.

E Computational Details

Open-weight models were evaluated on GPU clusters. Most experiments used NVIDIA A100 GPUs with 40 GB of memory. For larger models, we used NVIDIA GH200 GPUs with 96 GB of memory. The number of GPUs used in each experiment depended on the size and memory requirements of the model. As a rough estimate of computational cost, a full evaluation run required around 1 GPU-hour for models up to 12B parameters, around 2 GPU-hours for models up to 38B parameters, and around 3–4 GPU-hours for larger models. Reasoning variants were more computationally expensive and required around 8 GPU-hours per run.

E.1 Prompts

For open-ended and yes/no examples, for the prompt we provide the raw question. For multiple-choice examples, the answer options are appended below the question with letter labels. Circular evaluation changes only the order and labels of the options.

Images are provided as separate multimodal inputs. Qwen and LLaVA-based vLLM configurations prepend `<image>` and a newline to the text prompt before applying the chat template. Other open-weight configurations use the model processor chat template without this prefix. For Qwen and Gemma variants, we set `enable_thinking=true` for thinking runs and `enable_thinking=false` for non-thinking runs in `chat_template_kwargs`. No few-shot examples or culture-specific hints are added. All model-specific evaluation configurations are provided in the repository.

F Detailed Results

Below we report additional tables with detailed results. Table 5 gives full multimodal accuracy by question type. Table 6 reports the single-modality textual ablation using the same prompts and scoring rules. Table 7 provides the full breakdown by category and subcategory.

Model	Overall	MCQ	Yes/No	Open
PROPRIETARY MODELS				
GPT-5.4	**75.89**	**75.53**	**80.49**	**60.09**
GPT-5.4-mini	62.50	60.22	70.49	42.72
OPEN-WEIGHTS MODELS				
Gemma-4-31B-it (*Thinking*)	<u>72.02</u>	<u>70.59</u>	<u>79.26</u>	51.17
Gemma-4-31B-it	67.21	65.66	75.43	43.19
Qwen3.5-122B-A10B (*Thinking*)	66.77	58.81	78.77	<u>58.22</u>
GLM-4.6V	65.62	59.21	77.41	50.70
Qwen3.5-27B (*Thinking*)	61.86	52.97	77.16	45.07
Qwen3.5-122B-A10B (*Instruct*)	57.59	53.47	67.78	38.03
Mistral-Medium-3.5-128B	56.60	51.96	66.54	40.38
Qwen3.5-27B (*Instruct*)	55.51	51.96	65.31	34.74
Qwen3.5-9B (*Thinking*)	54.91	44.21	72.96	36.15
LLaVA-OneVision-Qwen2-72B-OV-HF	50.79	53.68	60.62	0.00
InternVL3.5-38B	49.01	44.11	62.84	19.25
LLaVA-Bielik-11B-v2.6	48.36	45.12	57.78	27.70
Ministral-3-14B-Instruct-2512	47.67	36.86	66.67	25.82
Ministral-3-8B-Instruct-2512	46.08	34.54	66.91	20.66
LLaVA-PLLuM-12B-250715	45.04	39.88	55.43	29.58
InternVL3.5-14B	44.35	38.97	58.40	15.96
Qwen3.5-9B (*Instruct*)	44.00	35.65	59.75	23.00
LLaVA-PLLuM-12B	43.11	38.07	54.81	22.07
InternVL3.5-8B	41.47	31.42	61.36	12.68
LLaVA-OneVision-Qwen2-7B-OV-HF	39.83	37.36	53.33	0.00
LLaVA-v1.6-Mistral-7B-HF	32.14	24.87	47.78	6.57
Random	19.89	0.76	48.58	0.00

Table 5: Model accuracy by question type, with all values reported as percentages. The **best result** in each column is shown in bold, and the best result among open-weight models is <u>underlined</u>.

Model	Overall	MCQ	Yes/No	Open
PROPRIETARY MODELS				
GPT-5.4	30.00	20.00	50.00	20.00
OPEN-WEIGHTS MODELS				
Qwen3.5-122B-A10B (*Thinking*)	**39.83**	**26.69**	**59.88**	**24.88**
Qwen3.5-27B (*Thinking*)	35.37	20.04	57.78	21.60
GLM-4.6V	33.53	19.34	56.67	11.74
Qwen3.5-122B-A10B (*Instruct*)	32.39	21.05	53.58	4.69
LLaVA-Bielik-11B-v2.6	32.34	17.52	57.04	7.51
Qwen3.5-27B (*Instruct*)	31.30	20.54	50.25	9.39
Qwen3.5-9B (*Thinking*)	30.11	12.29	54.69	19.72
LLaVA-PLLuM-12B	29.81	15.81	52.84	7.51
Ministral-3-14B-Instruct-2512	29.46	11.08	57.53	8.45
InternVL3.5-38B	29.41	15.51	53.21	3.76
LLaVA-PLLuM-12B-250715	29.17	12.79	54.20	10.33
InternVL3.5-14B	28.57	13.70	53.33	3.76
Gemma-4-31B-it	28.57	15.91	50.49	4.23
InternVL3.5-8B	27.83	10.78	54.69	5.16
Ministral-3-8B-Instruct-2512	25.15	7.15	52.10	6.57
Qwen3.5-9B (*Instruct*)	23.81	8.16	48.02	4.69
Mistral-Medium-3.5-128B	23.07	6.75	48.15	3.76
LLaVA-OneVision-Qwen2-7B-OV-HF	0.10	0.00	0.00	0.94
LLaVA-OneVision-Qwen2-72B-OV-HF	0.05	0.00	0.00	0.47
LLaVA-v1.6-Mistral-7B-HF	0.05	0.00	0.00	0.47
Random	19.89	0.76	48.58	0.00

Table 6: Model accuracy by question type in the single-modality textual input setting, with all values reported as percentages. The **best result** in each column is shown in bold, and the best result among open-weight models is underlined.

Category / Subcategory	GPT-5.4	GPT-5.4-mini	Gemma-4-31B-it (Thinking)	Gemma-4-31B-it	Qwen3.5-122B-A10B (Thinking)	GLM-4.6V	Qwen3.5-27B (Thinking)	Qwen3.5-122B-A10B (Instruct)	Mistral-Medium-3.5-128B	Qwen3.5-27B (Instruct)	Qwen3.5-9B (Thinking)	LLaVA-OneVision-Qwen2-72B-OV-HF	InternVL3.5-38B	LLaVA-Bielik-11B-v2.6	Ministral-3-14B-Instruct-2512	Ministral-3-8B-Instruct-2512	LLaVA-PLLuM-12B-250715	InternVL3.5-14B	Qwen3.5-9B (Instruct)	LLaVA-PLLuM-12B	InternVL3.5-8B	LLaVA-OneVision-Qwen2-7B-OV-HF	LLaVA-v1.6-Mistral-7B-HF	Random
ART AND ENTERTAINMENT	**68.81**	56.96	59.25	56.34	52.81	59.25	46.36	45.95	49.48	42.41	42.00	47.19	44.91	43.66	39.71	40.75	43.45	40.96	33.89	41.37	39.29	40.54	27.44	18.81
Architecture	**66.20**	53.52	61.97	57.75	47.89	60.56	45.07	53.52	52.11	47.89	45.07	45.07	43.66	42.25	46.48	38.03	40.85	45.07	43.66	38.03	38.03	32.39	21.13	19.01
Film	**72.58**	59.68	46.77	41.94	45.16	58.06	35.48	27.42	48.39	32.26	22.58	45.16	40.32	40.32	27.42	35.48	43.55	40.32	22.58	37.10	29.03	38.71	14.52	13.71
Literature	**65.93**	54.95	64.84	56.04	53.85	59.34	47.25	37.36	47.25	40.66	38.46	38.46	37.36	40.66	35.16	38.46	32.97	36.26	26.37	39.56	36.26	37.36	31.87	20.33
Media	**77.78**	72.22	61.11	61.11	50.00	50.00	38.89	38.89	33.33	33.33	38.89	50.00	44.44	22.22	27.78	38.89	33.33	38.89	33.33	27.78	38.89	38.89	44.44	25.00
Music	**55.36**	44.64	50.00	46.43	42.86	50.00	37.50	46.43	46.43	39.29	37.50	46.43	39.29	37.50	37.50	39.29	37.50	30.36	30.36	44.64	37.50	42.86	26.79	20.54
Paintings	**74.32**	64.86	64.86	66.22	55.41	66.22	54.05	62.16	54.05	55.41	45.95	48.65	55.41	56.76	44.59	48.65	52.70	37.84	43.24	51.35	43.24	39.19	35.14	20.95
Sculpture	**85.71**	53.57	64.29	67.86	64.29	57.14	42.86	60.71	53.57	50.00	50.00	64.29	53.57	50.00	50.00	42.86	50.00	50.00	35.71	50.00	42.86	64.29	39.29	17.86
Sport	**67.90**	59.26	59.26	59.26	62.96	61.73	56.79	44.44	50.62	37.04	55.56	53.09	49.38	45.68	44.44	43.21	53.09	50.62	35.80	38.27	48.15	44.44	23.46	16.67
GEOGRAPHY AND NATURE	**74.94**	63.66	73.68	67.67	63.91	68.42	58.90	62.91	57.89	58.40	53.63	51.88	50.38	49.62	49.12	46.62	45.61	44.86	45.11	43.86	42.61	41.10	32.83	19.67
Animate nature	**87.80**	80.49	82.93	78.05	82.93	85.37	73.17	75.61	70.73	75.61	70.73	78.05	68.29	70.73	68.29	65.85	58.54	68.29	53.66	51.22	56.10	60.98	34.15	34.15
Inanimate nature	**84.00**	74.67	82.67	76.00	77.33	76.00	74.67	64.00	66.67	64.00	62.67	62.67	60.00	56.00	52.00	48.00	58.67	56.00	45.33	53.33	46.67	48.00	26.67	14.67
Man-made	**75.72**	60.12	72.25	67.05	55.49	66.47	54.34	60.12	57.23	57.23	46.82	47.40	45.09	46.24	42.77	42.77	46.24	38.15	41.04	41.62	39.31	39.31	31.79	15.61
Socio-political	62.73	55.45	**66.36**	59.09	60.91	60.00	50.00	61.82	48.18	50.00	51.82	41.82	45.45	42.73	50.00	44.55	30.91	39.09	48.18	38.18	40.00	31.82	38.18	24.09
HISTORY AND SOCIETY	**77.65**	63.32	72.78	67.05	73.64	69.63	69.05	61.03	57.59	57.88	61.60	52.72	48.71	51.29	51.00	48.14	47.56	47.56	48.42	44.13	40.97	38.40	34.10	22.06
Current affairs and society	**79.84**	64.52	72.58	67.74	75.00	76.61	70.97	67.74	63.71	58.06	67.74	55.65	52.42	48.39	54.84	47.58	49.19	49.19	52.42	43.55	41.94	38.71	34.68	21.77
Early modern and modern history	**73.68**	61.05	69.47	63.16	70.53	63.16	60.00	55.79	46.32	54.74	57.89	50.53	42.11	49.47	46.32	43.16	36.84	46.32	44.21	43.16	40.00	35.79	32.63	18.95
Middle Ages	73.81	66.67	**78.57**	69.05	71.43	57.14	69.05	57.14	57.14	52.38	61.90	52.38	47.62	47.62	50.00	52.38	50.00	50.00	45.24	38.10	40.48	33.33	21.43	17.86
Post-war history	**76.60**	59.57	**76.60**	70.21	68.09	72.34	72.34	55.32	57.45	57.45	65.96	44.68	48.94	53.19	48.94	48.94	53.19	38.30	53.19	40.43	40.43	38.30	34.04	29.79
World War II	**85.37**	65.85	70.73	68.29	**85.37**	73.17	80.49	63.41	65.85	70.73	46.34	58.54	53.66	65.85	53.66	56.10	58.54	53.66	43.90	58.54	41.46	48.78	48.78	25.61
CULTURE AND TRADITION	**82.84**	68.93	78.40	76.04	72.19	64.79	64.50	61.83	61.83	60.36	60.36	60.65	51.78	55.62	50.89	51.48	52.07	47.63	46.45	50.00	44.08	47.34	37.28	21.15
Cuisine	**90.43**	77.66	87.23	84.04	84.04	77.66	80.85	70.21	69.15	69.15	74.47	67.02	58.51	59.57	57.45	56.38	48.94	44.68	63.83	55.32	53.19	50.00	54.26	27.66
Pop culture	**68.97**	58.62	60.34	63.79	50.00	43.10	48.28	34.48	50.00	37.93	41.38	44.83	39.66	46.55	43.10	39.66	41.38	39.66	25.86	36.21	29.31	36.21	25.86	16.38
Regional and ethnic cultures	**81.32**	73.63	79.12	75.82	72.53	60.44	54.95	62.64	65.93	61.54	57.14	63.74	51.65	56.04	46.15	47.25	56.04	47.25	34.07	46.15	43.96	51.65	25.27	15.38
Religion and tradition	**85.26**	62.11	80.00	75.79	73.68	69.47	67.37	69.47	57.89	64.21	61.05	61.05	52.63	56.84	53.68	57.89	57.89	55.79	53.68	56.84	44.21	47.37	38.95	23.16
LANGUAGE	**77.64**	60.87	76.71	67.39	70.19	62.11	69.25	53.11	53.11	54.35	53.42	45.65	44.72	40.68	44.72	41.93	38.20	37.89	40.37	36.65	36.02	32.92	26.71	20.96
Colloqual speech and slang	**78.26**	43.48	60.87	56.52	47.83	47.83	56.52	43.48	34.78	43.48	21.74	47.83	21.74	30.43	26.09	21.74	39.13	26.09	17.39	30.43	21.74	21.74	21.74	8.70
Dialects and regionalisms	**69.39**	65.31	**69.39**	61.22	55.10	46.94	51.02	55.10	48.98	44.90	40.82	36.73	30.61	40.82	34.69	28.57	42.86	32.65	44.90	38.78	30.61	28.57	36.73	28.57
Grammar	76.92	55.77	**86.54**	71.15	84.62	75.00	80.77	48.08	42.31	59.62	69.23	38.46	46.15	25.00	50.00	46.15	23.08	34.62	46.15	25.00	26.92	23.08	13.46	14.42
Language basics and phonetics	**72.22**	55.56	66.67	50.00	61.11	61.11	66.67	38.89	44.44	38.89	38.89	38.89	38.89	22.22	38.89	22.22	27.78	44.44	33.33	38.89	44.44	33.33	22.22	11.11
Ortography	**88.89**	66.67	72.22	50.00	61.11	55.56	55.56	44.44	55.56	55.56	61.11	27.78	27.78	50.00	55.56	55.56	27.78	33.33	50.00	33.33	33.33	27.78	33.33	27.78
Phraseology	**84.71**	68.24	**84.71**	76.47	77.65	64.71	82.35	58.82	64.71	63.53	60.00	56.47	49.41	48.24	49.41	47.06	43.53	38.82	43.53	42.35	42.35	40.00	31.76	31.18
Rhetorical figure	76.92	46.15	**92.31**	84.62	84.62	76.92	76.92	84.62	69.23	**92.31**	76.92	46.15	84.62	53.85	61.54	53.85	30.77	53.85	61.54	53.85	61.54	30.77	30.77	11.54
Semantics	**73.44**	60.94	70.31	67.19	70.31	64.06	64.06	51.56	54.69	45.31	50.00	50.00	54.69	46.88	43.75	48.44	46.88	43.75	31.25	35.94	37.50	40.62	23.44	14.06
IMAGE UNDERSTANDING	86.05	74.42	89.53	90.70	**91.86**	88.37	88.37	86.05	83.72	90.70	83.72	52.33	75.58	68.60	73.26	61.63	50.00	59.30	80.23	52.33	61.63	41.86	47.67	10.47
VISUAL REASONING	60.98	43.90	73.17	68.29	**75.61**	65.85	**75.61**	53.66	46.34	56.10	68.29	21.95	41.46	24.39	41.46	41.46	21.95	43.90	46.34	21.95	39.02	19.51	31.71	17.07

Table 7: Model accuracy by dataset category and subcategory, with all values reported as percentages. The **best result** in each row is shown in bold, and the best result among open-weight models is underlined.