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

**Anna Kołos, Grzegorz Statkiewicz, Karolina Seweryn, Katarzyna Kowol, Karolina Piosek, Wojciech Kusa**

NASK National Research Institute, Warsaw, Poland

**Correspondence:** {firstname.lastname}@nask.pl

## 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 1,117 images and 2,366 manually annotated VQA pairs. Overall, our dataset provides a controlled and challenging resource 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 Web Content Accessibility Guidelines (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, especially beyond English. 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).

To address these limitations in a specific cultural setting, we introduce **PoVisLE** (**Po**lish **Vis**ion-**L**anguage **E**valuation; see Figure 1), the first monocultural vision-language benchmark for Polish cultural and linguistic competence, including region-specific knowledge. Whereas existing culturally grounded benchmarks target cultural knowledge alone, PoVisLE also tests linguistic phenomena such as dialect and regional variation through their interaction with visual context, extending text-only Polish evaluation, notably PLCC (Dadas et al., 2025), to the multimodal setting. All questions and answers are created manually and without templates, and designed to require multi-hop reasoning over visual evidence, language, and cultural knowledge. Our contributions are as follows:

- We release PoVisLE, comprising 1,117 images and 2,366 manually authored VQA pairs target-

[1]To facilitate future research, we publicly release the dataset and code: https://huggingface.co/collections/NASK-PIB/PoVisLE, https://github.com/NASK-NLP/PoVisLE

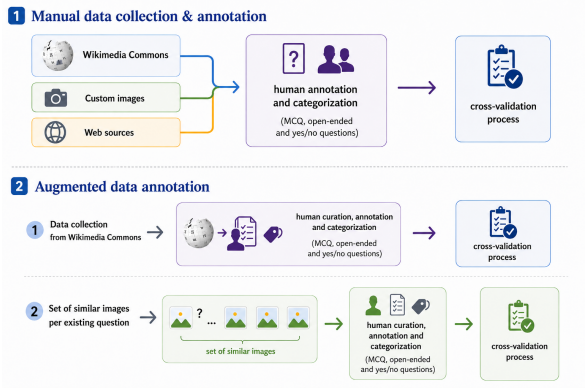



Figure 1: **Above**: Overview of the two-stage PoVisLE dataset construction process. **Below**: a single example from our dataset, with reasoning steps provided for clarity. More examples are included in Appendix A.

ing Polish cultural and linguistic competence. Nearly 30% of the images come from annotators' private collections and are absent from web-scale training corpora, providing a stricter test of generalization.

- We adapt and extend a hierarchical taxonomy of cultural and linguistic phenomena to the Polish multimodal setting, enabling fine-grained diagnostic analysis, and verify the linguistic richness of our questions through a stylometric comparison against template-based benchmarks.
- We benchmark 16 open and proprietary VLMs. The strongest model, Qwen3.5-397B, reaches 71.45% accuracy. Dialect and regionalism questions form the weakest category, indicating limited coverage of intra-language variation. Ablations removing the image or the question confirm that the benchmark cannot be solved from textual priors or answer-option artefacts alone, while evaluation in Polish, English, and German shows that performance also depends on the prompt language.

## 2 PoVisLE Dataset Construction

The primary objective of the dataset construction was to develop a highly diversified, manually annotated dataset designed to evaluate VLMs' understanding of Polish culture and language. Following prior work on culturally situated evaluation, 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 incorporates geographically distributed cultural and linguistic diversity across regions of Poland. This includes region-specific traditions as well as dialectal and lexical variation. 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, preventing models from relying solely on textual associations or memorized knowledge. Critically, this design principle also extends to linguistically oriented questions that are grounded in visual context.

Our taxonomy is described in Section 2.1. The dataset construction process was carried out in two stages, described in Sections 2.2 and 2.3, with the overall workflow illustrated in Figure 1.

### 2.1 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 adjustments 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 (Libovickỳ 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 (168 out of 2,366 VQA 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.

### 2.2 Image collection

The visual data constituted the foundation for the subsequent annotation process. The data collection procedure was carried out in two stages: manual curation and Wikimedia-based augmentation.

**Manual collection** 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, 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, in the initial manually curated set, a targeted proportion of 39.5% of images originated from annotator-provided collections. These images are particularly valuable, as they constitute authentic, non-public data. A similar approach has been adopted in culture-specific benchmarks, such as TaiwanVQA (Hsieh et al., 2026). The dataset metadata includes source-type annotations, enabling analysis of model performance on annotator-provided versus publicly sourced images.

By selecting images themselves, annotators were able to draw on their cultural knowledge, experience, and interpretative intuition to construct questions that go beyond surface-level factual queries. Unlike standard prompts such as "Who is depicted?", the resulting questions often require deeper cultural, historical or linguistic understanding and reasoning, reflecting aspects of the content

| Category / Subcategory | Open | MCQ | Y/N | Total (%) |
|---|---|---|---|---|
| **Art and entertainment** | 150 | 208 | 186 | 544 (23.0) |
| Literature | 25 | 38 | 39 | 102 (4.3) |
| Architecture | 25 | 34 | 24 | 83 (3.5) |
| Sport | 22 | 26 | 34 | 82 (3.5) |
| Paintings | 24 | 24 | 26 | 74 (3.1) |
| Music | 12 | 30 | 27 | 69 (2.9) |
| Film | 21 | 27 | 17 | 65 (2.7) |
| Media | 14 | 16 | 7 | 37 (1.6) |
| Sculpture | 7 | 13 | 12 | 32 (1.4) |
| **Language** | 149 | 159 | 169 | 477 (20.2) |
| Phraseology | 11 | 35 | 47 | 93 (3.9) |
| Semantics | 26 | 34 | 23 | 83 (3.5) |
| Grammar | 33 | 28 | 14 | 75 (3.2) |
| Dialects and regionalisms | 21 | 19 | 30 | 70 (3.0) |
| Orthography | 21 | 5 | 16 | 42 (1.8) |
| Rhetorical figure | 8 | 10 | 23 | 41 (1.7) |
| Language basics and phonetics | 19 | 11 | 8 | 38 (1.6) |
| Colloqual speech and slang | 10 | 17 | 8 | 35 (1.5) |
| **Geography and nature** | 110 | 161 | 164 | 435 (18.4) |
| Man-made | 56 | 76 | 60 | 192 (8.1) |
| Socio-political | 31 | 41 | 53 | 125 (5.3) |
| Inanimate nature | 12 | 30 | 27 | 69 (2.9) |
| Animate nature | 11 | 14 | 24 | 49 (2.1) |
| **History and society** | 86 | 131 | 162 | 379 (16.0) |
| Current affairs and society | 32 | 44 | 56 | 132 (5.6) |
| Early modern and modern history | 24 | 39 | 42 | 105 (4.4) |
| World War II | 12 | 14 | 23 | 49 (2.1) |
| Post-war history | 9 | 14 | 25 | 48 (2.0) |
| Middle Ages | 9 | 20 | 16 | 45 (1.9) |
| **Culture and tradition** | 81 | 133 | 149 | 363 (15.3) |
| Cuisine | 24 | 28 | 50 | 102 (4.3) |
| Religion and tradition | 25 | 32 | 44 | 101 (4.3) |
| Regional and ethnic cultures | 17 | 46 | 33 | 96 (4.1) |
| Pop culture | 15 | 27 | 22 | 64 (2.7) |
| **Image understanding** | 50 | 45 | 23 | 118 (5.0) |
| **Visual reasoning** | 17 | 18 | 15 | 50 (2.1) |
| **Total** | 643 | 855 | 868 | 2366 (100.0) |
| **Total %** | 27.2% | 36.1% | 36.7% | |

Table 1: Distribution of questions by category and question type. Y/N denotes yes/no questions; the figure in parentheses is the (sub)category's percentage share of all 2,366 questions.

that are unlikely to be captured using template-based approaches or 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.

**Wikimedia Data Augmentation** The second phase of data collection aimed to mitigate selection bias and increase dataset diversity by assigning images from Poland-related categories to annotators 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 the corresponding questions.

To account for annotator confidence, images could be marked as suitable for VQA but outside the annotator's confidence scope, indicating that while the image met general suitability criteria, it required more certain or specialized knowledge for question formulation. These cases were reassigned accordingly. Additionally, the use of Wikimedia Commons improved efficiency by eliminating the need for manual metadata curation.

To further enhance diversity, we introduced an augmentation step for images associated with multiple questions. For each such image, we retrieved visually similar candidates from the same source categories and ranked them using CLIP embeddings (Radford et al., 2021). Annotators then selected suitable non-identical replacements for individual questions, ensuring that each question remained visually grounded and unequivocal. As a result, each augmented image was individually validated by a human annotator. If no suitable candidate image was found, the sample was not augmented. This process increased the number of unique images from 790 to 1,117 while preserving the original question set.

### 2.3 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), annotators wrote between 1 and 10 questions per image, assigned each to one of 7 main categories and 29 subcategories, and labelled its type as *open-ended*, *multiple-choice*, or *yes/no*.

Detailed annotation guidelines can be found in Appendix D. 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.

The annotation team consisted of two primary annotators with background in linguistics, supported by an expert-level super-annotator responsible for continuous quality control. Prior to annotation, annotators underwent a structured train-

ing phase, including guideline familiarization and example-based calibration, to ensure consistency and adherence to the defined quality criteria. In addition, a controlled subset of inputs was annotated by 12 auxiliary annotators selected to ensure demographic diversity, particularly across Polish regions (see Appendix D.5 for details).

### 2.4 Quality Assurance

To ensure annotation quality and consistency, a structured multi-stage validation process was implemented, with particular emphasis on cross-validation. Each annotated instance was reviewed by a second annotator, whose role was to verify: (i) the logical and linguistic correctness of the question, (ii) the correctness and unequivocal nature of the answer, (iii) whether the answer could be derived solely from the visual content, (iv) compliance with VQA task requirements, and (v) the accuracy of associated metadata.

Regular team discussions were conducted to resolve ambiguities and refine annotation guidelines, which supported consistent decision-making between annotators. In addition, an expert-level super-annotator reviewed representative samples to identify systematic issues, ensure consistency in labeling and formulation, and provide targeted feedback. Problematic instances were revised accordingly.

Additionally, as part of the iterative quality control process, three validation procedures were regularly performed using LLMs, as described below.

First, for MCQs, models were provided with the image and answer options but not the question. This analysis was conducted using a subset of 11 LLMs selected for the evaluation study. Questions for which the models achieved a high success rate (approximately above 60%) were considered too easy and were subsequently reviewed. Revisions included replacing distractors with more plausible alternatives, increasing the number of answer options, adding options such as "none of the above" when appropriate, or converting the question into an open-ended format when a single clear and unambiguous answer could be expected.

Second, visual grounding was assessed using text-only variants. Models were given the question without access to the image. Questions that could be answered correctly without visual information were considered insufficiently grounded in the image content and were reformulated or removed.

Third, open-ended questions were evaluated through iterative manual inspection of LLM-generated answers. This process was used to verify that correct answers were consistently accepted and that incorrect answers were not mistakenly treated as valid. The findings informed the development of the evaluation protocol and additional answer-matching rules. Responses were required to be grammatically and orthographically correct. Consequently, factually correct answers containing orthographic errors were not accepted. For example, *powstanie warszawskie* was considered correct, as adjectives derived from most proper nouns are written in lowercase in standard Polish, whereas *Powstanie Warszawskie* was rejected due to incorrect capitalization. This procedure helped ensure consistent evaluation of open-ended responses while preserving linguistic correctness requirements.

The latter validation step also served as a feedback mechanism for the evaluation process. If an issue cannot be resolved through the addition of alternative accepted answers or through existing matching rules, annotators should report it to the evaluation team. Such cases may indicate that the evaluation protocol requires further modification or refinement.

## 3 Dataset statistics

The dataset consists of 2,366 question-answer pairs associated with 1,117 unique images, with an average of 2.12 questions per image (Table 1). In the final dataset, Wikimedia Commons constitutes the majority of the data (70.28%), followed by annotator-provided images (28.74%) and other sources (0.98%). The distribution of question types is relatively balanced, with yes/no questions accounting for 36.7% of the dataset, closely followed by multiple-choice questions (36.1%), while open-ended questions represent 27.2%. MCQs vary in the number of answer options, ranging from 2 to 8, with an average of 4.29 options per question.

The questions exhibit substantial linguistic diversity and structural complexity, with an average length of 18.7 tokens and frequent multi-clause formulations, supporting context-rich and non-template-based reasoning. A detailed stylometric analysis, which provides further evidence of these characteristics, is presented in Appendix B.

**Test/validation split** The benchmark is divided into a test split used for final evaluation and a validation split released publicly. The validation split contains 406 question-answer pairs (212 multiple-

choice, 154 yes/no, and 40 open-ended), while the test split contains 1,960 question-answer pairs (643 multiple-choice, 714 yes/no, and 603 open-ended). The validation split is intended primarily to publicly illustrate the range of question types and is therefore not sampled from the same distribution as the test set. In particular, it contains fewer open-ended questions, which represent a more challenging category and are largely retained in the held-out test set.

## 4 Experiment Setup

Models are evaluated on the PoVisLE test and validation splits 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.

### 4.1 Scoring

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 (Zheng et al., 2024) 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 against the gold answers, with correct diacritics required in all cases and correct capitalization required where relevant. For selected questions, multiple answer variants are accepted through predefined inclusion patterns. The answer-matching rules and inclusion patterns were developed through an iterative validation process involving human annotators, as described in Section 2.4.

### 4.2 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), as well as GPT and Claude 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. We use deterministic decoding in zero-shot setting with temperature set to zero and top-p set to 1.0. All models are evaluated using the same prompts. Further details, including specific model versions, prompts, parameters and the random baseline are presented in Appendix E.

### 4.3 Metrics

We report macro-averaged accuracy over question type as the main evaluation metric. An example is counted as correct if the parsed model prediction satisfies the task-specific scoring rule described in Section 4.1, including correct answers for all circular variants in the multiple-choice setting.

### 4.4 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.

### 4.5 Question-free input

Recent work has shown that language models can achieve considerable accuracy on benchmark tasks without the question itself, especially in multiple-choice settings, where models can rely on the answer choices alone (Balepur et al., 2024). We therefore include a question-free setting in which the image remains available, but the question is removed. For multiple-choice items, the model still receives the answer options. For yes/no questions, it is only prompted to answer *yes* or *no*. For open-ended questions, the model receives only the image.

### 4.6 Language impact

Model performance may depend not only on visual and cultural understanding, but also on the language used to formulate the task (Shen et al., 2024). To study the impact of the prompt language on model performance, we evaluate the same questions in Polish, English, and German to measure the effect of prompt language. First, we randomly

| Model | Overall | Art & Entert. | Culture & Trad. | Geogr. & Nature | History & Society | Language | Image Und. | Visual Reas. |
|---|---|---|---|---|---|---|---|---|
| **PROPRIETARY MODELS** | | | | | | | | |
| GPT-5.4 | 65.93 | 57.51 | **75.23** | 66.75 | 65.65 | 65.89 | 78.63 | 53.59 |
| Claude Sonnet 5 | 65.57 | 51.43 | 70.98 | 70.35 | 70.40 | 65.18 | 84.54 | 55.81 |
| **OPEN-WEIGHTS MODELS** | | | | | | | | |
| Qwen3.5-397B-A17B (*Thinking*) | **71.45** | **60.40** | 74.99 | **71.10** | **79.57** | 71.83 | **86.58** | 74.53 |
| Gemma-4-31B-it (*Thinking*) | 66.78 | 54.08 | 74.15 | 65.97 | 66.40 | **72.86** | 82.13 | 74.96 |
| Gemma-4-31B-it | 58.60 | 49.81 | 66.71 | 54.92 | 59.23 | 59.86 | 80.94 | 59.83 |
| Qwen3.5-397B-A17B | 58.25 | 50.65 | 64.84 | 61.27 | 60.23 | 52.18 | 81.86 | 47.69 |
| GLM-4.6V | 58.06 | 51.89 | 57.00 | 66.39 | 60.31 | 49.54 | 82.13 | 62.05 |
| Qwen3.5-27B (*Thinking*) | 55.65 | 41.12 | 57.24 | 52.73 | 60.23 | 61.65 | 82.97 | 69.40 |
| Qwen3.5-27B | 47.19 | 34.70 | 46.77 | 47.83 | 51.64 | 46.81 | 83.06 | 56.50 |
| Qwen3.5-9B (*Thinking*) | 47.00 | 35.99 | 43.32 | 44.86 | 51.98 | 48.25 | 79.55 | **75.73** |
| Qwen3.5-9B | 37.59 | 26.67 | 34.76 | 40.50 | 39.89 | 35.15 | 78.81 | 42.14 |
| Ministral-3-14B-2512 | 37.41 | 31.49 | 37.74 | 38.94 | 38.14 | 33.84 | 62.06 | 42.48 |
| LLaVA-Bielik-11B-v2.6 | 36.89 | 35.35 | 39.96 | 37.04 | 38.55 | 30.03 | 59.93 | 22.22 |
| InternVL3.5-38B | 35.76 | 31.65 | 32.19 | 34.96 | 36.89 | 33.68 | 62.44 | 39.49 |
| Ministral-3-14B-2512 (*Thinking*) | 33.51 | 24.57 | 35.49 | 34.95 | 30.55 | 35.84 | 56.05 | 34.02 |
| LLaVA-PLLuM-12B | 30.85 | 27.15 | 38.50 | 29.72 | 30.96 | 25.32 | 51.79 | 19.23 |
| Random | 16.79 | 16.78 | 16.85 | 16.77 | 16.79 | 16.77 | 16.80 | 16.91 |

Table 2: Model accuracy by dataset category on the test split, 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.

sample 15 questions from each subcategory (or top-level category without subcategories), resulting in 465 samples in total. Then, the sample was automatically translated from Polish into English and German using `DeepSeek-V4-Pro`[2]. The translations were then manually reviewed and corrected.

For multiple-choice and yes/no questions, the model may answer in the language of the prompt. For open-ended questions, the prompt includes an instruction to answer only in Polish. The reference answers remain in Polish and are not translated.

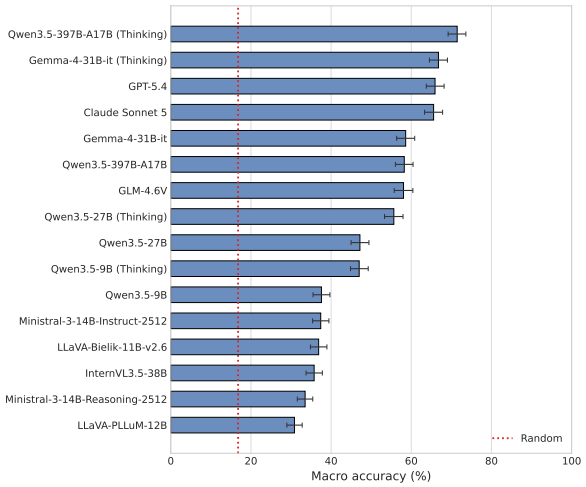


Figure 2: Overall macro accuracy on the test split with 95% image-cluster bootstrap confidence intervals. The red dotted line denotes the random baseline.

[2]https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro

## 5 Results and Discussion

Figure 2 present model performance on the PoVisLE test split with 95% image-cluster bootstrap confidence intervals. Qwen3.5-397B-A17B *Thinking* achieves the highest overall macro accuracy of 71.45%. Its result is higher than those of GPT-5.4 and Claude Sonnet 5 models, which obtain 65.93% and 65.57%, respectively. Additional test and validation results are reported in Appendices G and H.

**Performance across question types and categories.** Table 2 presents results by each category. Image Understanding is the highest-scoring category for every evaluated model. This suggests that direct recognition and interpretation of visual content is more reliable than answering questions that require additional cultural or linguistic knowledge. Among the main culturally grounded categories, Art and Entertainment often produces the lowest results. The ordering of the remaining categories varies across models.

As shown in Table 13 (Appendix H), yes/no questions produce the highest raw accuracy for all evaluated models. Open-ended questions are not consistently the lowest-scoring format, but they require models to produce an answer in the expected language, grammatical form, length, and format. The error analysis presented in Appendix F shows that hallucination is the most frequent error label, followed by instruction non-adherence mostly in weaker performing models.

**Impact of reasoning.** The *Thinking* configuration improves performance for all evaluated Qwen variants and for Gemma-4-31B-it. Qwen3.5-397B-A17B gains +13.20 pp, while Qwen3.5-27B, Qwen3.5-9B, and Gemma-4-31B-it gain +8.46, +9.41, and +8.18 pp, respectively. For Qwen3.5-397B-A17B, the largest gains occur in Visual Reasoning, Language, and History and Society, while Image Understanding improves only slightly. Ministral-3-14B-2512 is the only exception, decreasing by -3.90 pp in the *Thinking* configuration.

| Model | Overall | | MCQ | | Yes/No | | Open | |
|---|---|---|---|---|---|---|---|---|
| | Value | Δ | Value | Δ | Value | Δ | Value | Δ |
| Qwen3.5-397B-A17B (*T*) | 71.45 | — | 65.94 | — | 83.89 | — | 64.51 | — |
| ↪ *without question* | 35.84 | -35.61 | 55.99 | -9.95 | 51.54 | -32.35 | 0.00 | -64.51 |
| ↪ *without image* | 31.80 | -39.65 | 29.08 | -36.86 | 54.20 | -29.69 | 12.11 | -52.40 |
| Gemma-4-31B-it (*T*) | 66.78 | — | 64.07 | — | 80.53 | — | 55.72 | — |
| ↪ *without question* | 34.58 | -32.20 | 51.79 | -12.28 | 51.96 | -28.57 | 0.00 | -55.72 |
| ↪ *without image* | 28.54 | -38.24 | 24.11 | -39.96 | 54.06 | -26.47 | 7.46 | -48.26 |
| GPT-5.4 | 65.93 | — | 62.52 | — | 78.71 | — | 56.55 | — |
| ↪ *without question* | 32.91 | -33.02 | 47.74 | -14.78 | 50.98 | -27.73 | 0.00 | -56.55 |
| ↪ *without image* | 27.94 | -37.99 | 23.95 | -38.57 | 52.24 | -26.47 | 7.63 | -48.92 |
| Claude Sonnet 5 | 65.57 | — | 61.28 | — | 80.53 | — | 54.89 | — |
| ↪ *without question* | 32.80 | -32.77 | 53.03 | -8.25 | 45.38 | -35.15 | 0.00 | -54.89 |
| ↪ *without image* | 17.89 | -47.68 | 4.35 | -56.93 | 48.32 | -32.21 | 1.00 | -53.89 |
| Gemma-4-31B-it (*I*) | 58.60 | — | 56.45 | — | 76.89 | — | 42.45 | — |
| ↪ *without question* | 30.86 | -27.74 | 39.50 | -16.95 | 53.08 | -23.81 | 0.00 | -42.45 |
| ↪ *without image* | 21.45 | -37.15 | 11.82 | -44.63 | 50.70 | -26.19 | 1.82 | -40.63 |
| Qwen3.5-397B-A17B (*I*) | 58.25 | — | 45.26 | — | 77.59 | — | 51.91 | — |
| ↪ *without question* | 30.55 | -27.70 | 39.97 | -5.29 | 51.68 | -25.91 | 0.00 | -51.91 |
| ↪ *without image* | 26.25 | -32.00 | 18.97 | -26.29 | 50.00 | -27.59 | 9.78 | -42.13 |
| Random | 16.79 | — | 0.38 | — | 50.00 | — | 0.00 | — |

Table 3: Model accuracy by question type on the test split, under the full-input setting and two input ablations, with all values reported as percentages. Unindented rows report results with the complete input. Indented rows report results after removing either the image (*without image*) or the question (*without question*). For model variants, (*T*) denotes *Thinking* mode and (*I*) denotes *Instruct* mode.

**Input ablations.** Table 3 reports performance under the full-input setting and after removing either the image or the question for top performing models. Removing the image reduces overall macro accuracy by 32.00–47.68 pp across models. Yes/no accuracy falls to 48.32–54.20%, open-ended accuracy to 1.00–12.11%, and multiple-choice accuracy to 4.35–29.08%. The substantial performance drop after removing the image indicates that models rely heavily on visual information to answer questions in PoVisLE.

Removing the question, while retaining the image and mcq options, reduces open-ended accuracy to 0% in reported models and leaves yes/no accuracy close to chance. Multiple-choice accuracy remains relatively high at 39.50–55.99%, decreasing by only 5.29–16.95 points compared with the full-input setting. This shows that the image and answer options are often sufficient to identify the expected answer without access to the question. The result may reflect image–answer compatibility and differences in distractor plausibility.

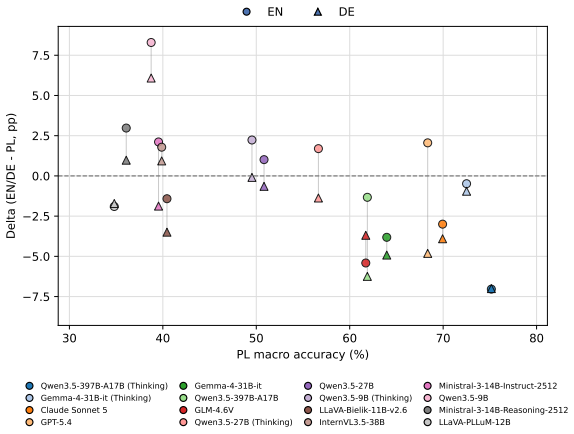


Figure 3: Translation effect by model. Macro accuracy in Polish (PL) is compared with the accuracy change after switching to English (EN) and German (DE).

**Impact of prompt language.** Figure 3 (detailed results in Table 14 of Appendix I) shows a relationship between performance on the original Polish questions and the effect of translation. Models with higher Polish accuracy generally perform worse when the questions are translated into English or German. Qwen3.5-397B-A17B *Thinking*, the strongest model on the Polish subset, loses approximately 7 pp in both languages, while Claude Sonnet 5, Gemma-4-31B-it, and GLM-4.6V also show lower translated performance. In contrast, several lower-performing models improve with English prompts. The largest gain is observed for Qwen3.5-9B, which improves by 8.29 pp in English and 6.08 pp in German. This trend suggests that weaker general-purpose models may benefit from input in a language more strongly represented in their training data, whereas stronger models make better use of the original Polish formulation. GPT-5.4 is an exception, improving in English despite its high Polish performance, while the Polish-oriented LLaVA-Bielik and LLaVA-PLLuM models perform best in Polish.

## 6 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) intro-

duced 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 the evaluation towards external and common sense knowledge, while recent culturally grounded VQA benchmarks investigate whether models can interpret culturally specific symbols, practices, social norms, and geographically localized knowledge

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 Polish 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 include a comparison of PoVisLE with 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.

## 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.

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. At the same time, they expose a fundamental challenge for evaluation. Exact-match scoring provides a transparent, reliable, and reproducible assessment framework, but it necessitates relatively constrained answer formats. We currently lack robust methodologies for reliably evaluating semantically equivalent yet linguistically diverse responses, particularly in culturally grounded settings. Developing such methods remains an important direction for future research.

## Acknowledgements

This work was supported by the Polish Ministry of Digital Affairs (subsidy no. 4/WII/DBI/2026). The computational resources were provided by the Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) under computational grant no. PLG/2026/019138.

## 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.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

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. Evaluat-

ing 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*.

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.

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.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pages 19426–19454.

## A Dataset examples and statistics visualization

Figure 4 presents representative examples from the dataset, illustrating the diversity of question types and linguistic formulations, while Figure 5 visualizes this distribution in a hierarchical form, offering an intuitive overview of the balance between high-level categories and their internal structure.

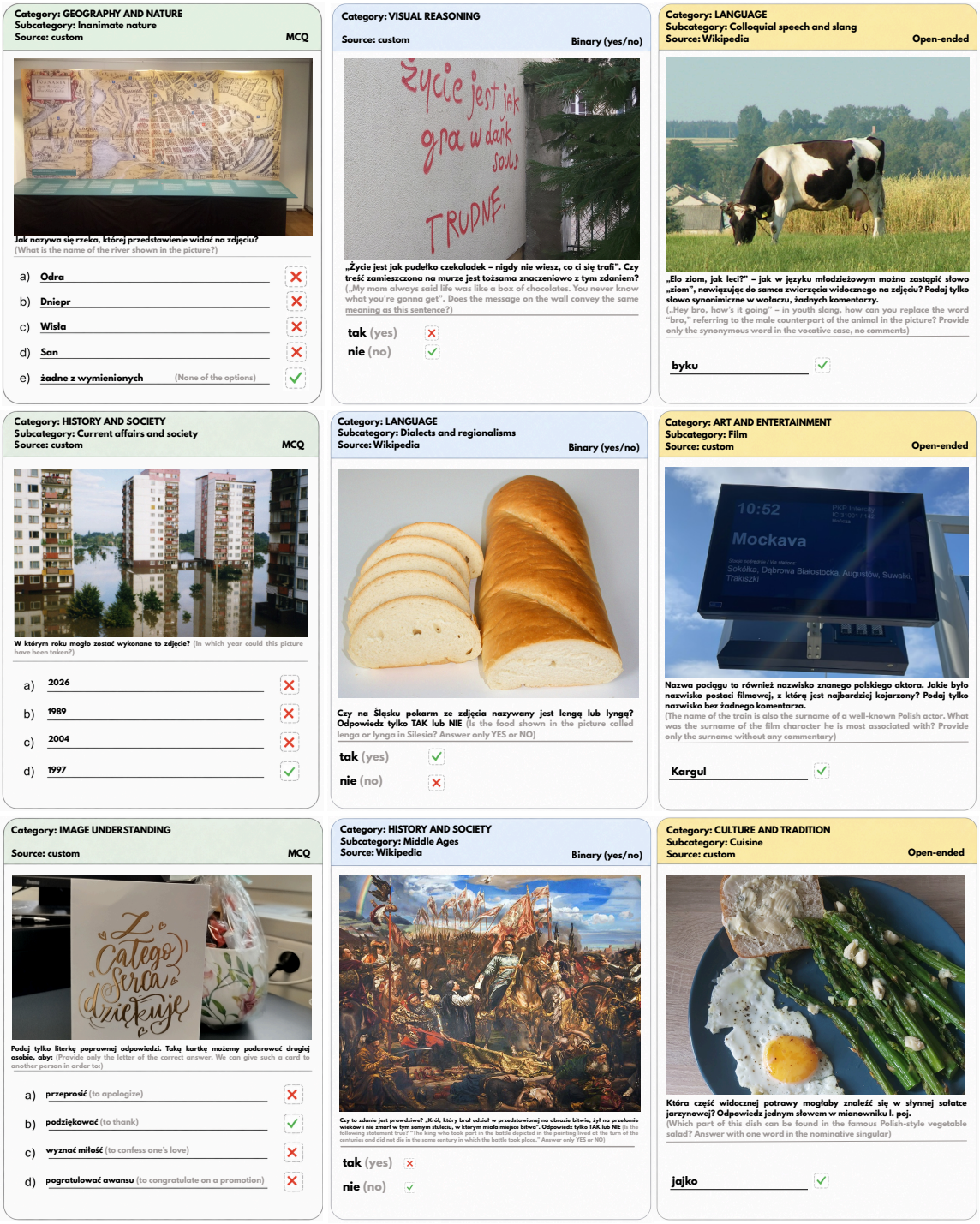


Figure 4: Examples from the dataset illustrating multiple-choice, binary (yes/no), and open-ended questions across all seven main categories. Image sources include Wikimedia Commons and annotators' personal collections contributed to the project. English translations are omitted for Polish-specific named entities and for open-ended answers requiring Polish vocabulary.


Figure 5: Distribution of 2,198 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.

## 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), limiting their usefulness for stylometric analysis. For MCQ tasks, options provided are excluded and not considered part of the question.

The average question length is 18.7 tokens (Figure 6). The distribution is right-skewed, with most questions containing between approximately 12 and 25 tokens and a long tail extending toward longer questions. This indicates that while the majority of questions are of moderate length, a smaller number of substantially longer questions are also present in the dataset.

### 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 4.

Interestingly, we observe a minimal difference between the surface-form type–token ratio (L_TTR_IA) and its lemmatized counterpart (L_TTR_LA) (0.850 vs. 0.841), 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 named entity ratio (0.031) 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 nearly identical content word incidence and content word types (0.5996 vs. 0.5967), 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 (0.099 vs. 0.096), 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 relatively high proportion of tokens associated with declarative constructions (0.626 vs. 0.349). This can be partly attributed to strict 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 significant proportion of declarative structures. A smaller, yet non-negligible proportion of tokens corresponds to negative constructions (0.035), indicating the presence of adversarial or contrastive question formulations.

Finally, high values for words within modifiers (0.296) and words in nominal phrases (0.613) 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 capital-

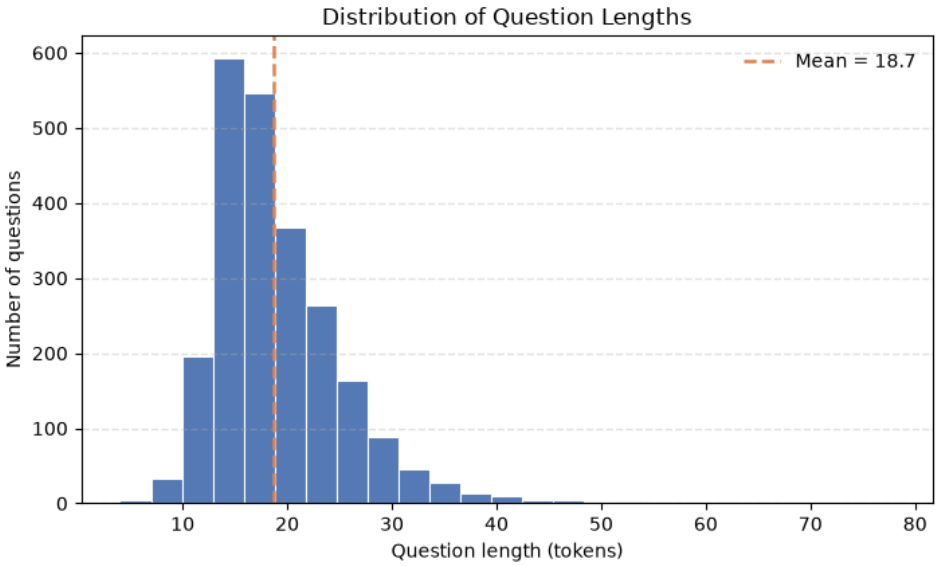


Figure 6: Distribution of question lengths (in tokens) across the dataset. The dashed vertical line indicates the mean question length (18.7 tokens).

ized 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 = 0.334, SCP = 0.081, and SP = 0.384. This also reveals an overlap between the two indicators, corresponding to questions that contain both a relative clause and a subordinate conjunction (3.09% 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.

## C Comparison to Other Datasets

Table 5 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

| Metric | Description | Score |
|---|---|---|
| **Grammatical Forms** | | |
| G_N | Nouns | 0.278 |
| G_V | Verbs | 0.128 |
| G_ADJ | Adjectives | 0.096 |
| G_ADV | Adverbs | 0.027 |
| G_PRO | Pronouns | 0.099 |
| G_PRO_NEG | Negative pronouns | 0.003 |
| G_PRO_REL | Relative pronouns | 0.009 |
| G_PRO_INT | Interrogative pronouns | 0.034 |
| G_CONJ | Conjunctions | 0.029 |
| G_CCONJ | Coordinating conjunctions | 0.026 |
| G_SCONJ | Subordinating conjunctions | 0.003 |
| **Punctuation** | | |
| PUNCT_TOTAL | Total punctuation | 0.133 |
| **Syntactic** | | |
| SY_MOD | Words within modifiers | 0.296 |
| SY_NPHR | Words in nominal phrases | 0.613 |
| SY_S_DE | Words in declarative sentences | 0.349 |
| SY_S_IN | Words in interrogative sentences | 0.626 |
| SY_S_NEG | Words in negative sentences | 0.035 |
| SY_QUOT | Words in quotation marks | 0.001 |
| **Lexical** | | |
| L_TTR_IA | Type-token ratio for non-lemmatized tokens | 0.850 |
| L_TTR_LA | Type-token ratio for lemmatized tokens | 0.841 |
| L_CONT_A | Incidence of content words | 0.600 |
| L_CONT_T | Content word types | 0.600 |
| L_FUNC_A | Incidence of function words | 0.242 |
| L_FUNC_T | Function words types | 0.233 |
| L_NAME_ENT | Named entities | 0.031 |
| L_NAME | Proper names | 0.015 |
| L_PERSN | Person names | 0.005 |
| L_PLACEN_GEOG | Place and geographical names | 0.009 |

Table 4: Stylometric feature distribution of the dataset across grammatical, syntactic, and lexical categories.

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 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.[3] While it is presented as a benchmark for Polish cultural understanding in vision-language models, as of August 2026 no documentation regarding its construction, annotation methodology, or evaluation protocol is publicly available, which prevents a meaningful and reproducible comparison.

[3] 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** – artistic, cultural, and media-related content.
  - **Architecture** – famous buildings, monuments, and structural design.
  - **Film** – movies, scenes, actors, and film-related references.
  - **Literature** – books, authors, and literary history.
  - **Media** – press, television, and digital media.
  - **Music** – musical works, performers, and related cultural references.
  - **Paintings** – paintings, artists, and related artistic phenomena.
  - **Sculpture** – sculptures, sculptors, and related artistic works.
  - **Sport** – sports activities, events, and figures.
- **Geography & Nature** – physical environment and spatial context.
  - **Animate nature** – animals and living organisms.
  - **Inanimate nature** – landscapes and natural formations.
  - **Man-made** – human-made geographical, urban, industrial, and infrastructural elements, including landmarks.
  - **Socio-political** – regions, borders, and administrative entities.
- **Culture & Tradition** – customs, practices, and shared cultural symbols.
  - **Cuisine** – food, beverages, and culinary traditions.
  - **Pop culture** – contemporary cultural trends and references.
  - **Regional and ethnic cultures** – local traditions and cultural variation.

| 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 | ✓ | ✗ | ✓ | ✗ | ✗ |
| Afri-MCQA | 12 (African countries) | 3,000 (?) | 7,642 | EN + 15 African languages | ✓ | ✓ | ✓ | ✓ | ✗ |
| PLCC | 1 (PL) | — | 600 | PL | ✓ | ✓ | ✓ | — | — |
| **PoVisLE** | **1 (PL)** | **1,117** | **2,366** | **PL** | ✓ | ✓ | ✓ | ✓ | ✗ |

Table 5: 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.

  - **Religion and tradition** – rituals, beliefs, and heritage practices.

- **History & Society** – historical events and social context.
  - **Middle Ages** – medieval historical period.
  - **Early modern and modern history** – post-medieval developments to 20th century.
  - **World War II** – events and figures related to WWII.
  - **Post-war history** – developments after 1945.
  - **Current affairs and society** – contemporary social and political issues.

- **Language** – linguistic structure, meaning, and variation.
  - **Colloquial speech and slang** – informal language use.
  - **Dialects and regionalisms** – region-specific language variation.
  - **Grammar** – syntactic structure and correctness.
  - **Language basics and phonetics** – pronunciation, sounds, and basic elements of language.
  - **Orthography** – spelling and writing conventions.
  - **Phraseology** – idioms and fixed expressions.
  - **Rhetorical figures** – stylistic and figurative language.
  - **Semantics** – word meaning and interpretation.

- **Image Understanding** – direct recognition and interpretation of visual content.

- **Visual Reasoning** – reasoning about relationships and context within an image.

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 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, combining multiple pieces of information from the image or integrating visual cues with culturally grounded or general knowledge. Such questions should go beyond direct recognition and involve interpretation, inference, or linking distinct visual elements.

  *Example:* "Which sport is this man's wife known for?"

  *Interpretation:* This question constitutes a multi-hop reasoning task when the sport is not directly observable in the image and the model must: (i) recognize the man (e.g., a public figure), (ii) infer the identity of his spouse, and (iii) recall the sport discipline in which she is active.
- **Use of associations:** Treat the image as a visual anchor that can connect to broader cultural or knowledge contexts. Questions may refer to entities or concepts associated with the image content rather than only to what is directly visible. However, the image must provide the essential context needed to identify the relevant entity, object, place, or event and answer the question.
- **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, each annotated instance is reviewed by another annotator through careful manual inspection, following a structured set of validation criteria.

- **Linguistic and logical correctness:** Verify that the question is grammatically correct, natural in phrasing, and logically well-formed, i.e., it is internally consistent, does not contain contradictory or ill-defined references, and clearly specifies the entity or concept being asked about.
- **Answer correctness and unambiguity:** Ensure that the answer is factually accurate and unequivocal, with no plausible alternative interpretations. Ensure that the answer can be inferred from the image and any relevant contextual or textual clues provided in the question. Consider potential biases arising from regional or cultural assumptions, and verify whether the interpretation is broadly understandable; if it relies on specific local knowledge, either reformulate the question or make the required context explicit (e.g., when referring to a regional term, indicate that it may be non-standard or region-specific).
- **Visual grounding:** Ensure that the answer can be derived solely from the visual content and that the question cannot be answered without access to the image.
- **Compliance with VQA criteria:** Confirm that the question adheres to the defined task requirements, including its Polish cultural relevance and appropriate category assignment.
- **Metadata verification:** Ensure that all required metadata is complete and that licensing terms have been correctly applied. In particular, verify that no copyrighted material is included in violation of the dataset's licensing constraints.

If any issues or inaccuracies are identified, apply appropriate revisions:

- Reformulate the question and the answer for logical clarity or linguistic correctness

- Adjust the question type if necessary
- Replace the question if it does not meet the guidelines
- Delete the image or question if it is found to be non-compliant with VQA criteria

**Escalation:** In cases of uncertainty, consult the super-annotator to ensure consistency with the overall annotation standards. Any full replacement or deletion must be explicitly approved by the super-annotator.

**Super-annotator interaction:** Be aware that the super-annotator may review selected samples throughout the annotation process. Annotators are expected to incorporate their feedback, including suggested corrections, reformulations, and improvements. In particular, annotators may be asked to revise questions that exhibit recurring issues, lack clarity, or reflect systematic inconsistencies, in order to maintain overall dataset quality and consistency.

### D.4 Iterative Quality Control

As part of the iterative quality control process, annotators will receive dedicated review sheets containing the results of validation analyses conducted on a selected subset of 11 LLMs. These sheets serve as supplementary material and identify VQA instances that may require revision. Any necessary changes should be implemented directly in the annotation tool.

- **Multiple-choice question difficulty assessment:** Review VQA instances flagged on the basis of image-and-options-only evaluations, in which LLMs were provided with the image and answer options but not the corresponding question. Questions for which the models achieved a high success rate (approximately above 60%) should be considered potentially too easy. When revising such questions, possible actions include:
  - replacing distractors with more plausible/challenging alternatives,
  - increasing the number of answer options,
  - adding alternatives such as "none of the options" when appropriate,
  - converting the question into an open-ended format when a single clear and unambiguous answer can reasonably be expected.
- **Visual grounding assessment:** Review questions flagged on the basis of text-only evaluations, in which LLMs were given the question without access to the image. Questions that can be answered correctly without visual information should be considered insufficiently grounded in the image content. Such questions should be reformulated, replaced, or removed.
- **Open-ended answer validation:** Analyze review sheets containing open-ended questions, LLM-generated responses, and the corresponding automated evaluation results. Each sample should be inspected manually to verify that correct answers are consistently scored as correct and that incorrect answers are not accepted by the evaluation protocol. When reviewing a sample, take into account the following:
  - identify plausible alternative correct answers that are not currently accepted and add them to the list of accepted responses when appropriate,
  - verify compliance with grammatical and orthographic requirements,
  - ensure that answers containing orthographic errors are not accepted, even if they refer to the correct entity or event,
  - reformulate the question if the current evaluation protocol does not allow for unambiguous assessment of responses,
  - check for ambiguities or inconsistencies in the answer-matching rules that go beyond the adjustments annotators can make and may lead to incorrect scoring; such issues should be reported to the evaluation team so that the matching rules can be revised accordingly.

### D.5 Annotators' demographics

The annotation team consisted of 16 participants representing four age groups. The distribution was balanced across age categories, with four annotators (25%) in each group: 20–25, 25–30, 30–35, and 35–40 years. All annotators have resided in Poland.

Annotators were assigned one of three roles. Two *Primary* annotators were responsible for the core annotation process, one *Super-annotator* oversaw annotation quality and guideline compliance,

and the remaining thirteen *Auxiliary* annotators contributed by supplying their own images as candidate visual question answering (VQA) instances, thereby increasing the regional diversity of the dataset.

The annotators originated from 11 of Poland's 16 voivodships, providing broad geographic coverage. In addition to their region of origin, annotators reported a *current/familiarized region*, defined as a voivodship in which they had lived, studied, or worked for a substantial period of time. This distinction allowed us to account not only for birthplace but also for regional familiarity acquired through migration and long-term residence. The demographic characteristics of the annotation team are presented in Table 6.

# E Implementation Details

Below we provide details on specific model versions, prompts, decoding parameters and computational costs.

## E.1 Model Versions

Table 7 lists the exact model versions and identifiers used in our experiments. Open-weight models were served locally with vLLM, whereas proprietary models were accessed through OpenRouter. We report the corresponding repository or detail model version in API.

## E.2 Random baseline

We compute the random baseline using the expected accuracy of a uniformly random answer. For yes/no questions, the expected accuracy is $0.5$. For open-ended questions, it is $0$, since random generation is not expected to match the reference answer exactly. For a multiple-choice question with $k$ options (where $k \in 2, \ldots, 8$), the expected accuracy is $1/k$. With circular evaluation, the answer must be correct under all $k$ cyclic rotations, so the expected accuracy is $(1/k)^k$. We then aggregate these expected per-example accuracies using the same macro-averaging procedure as for model results and report theoretical random accuracy.

## E.3 Prompts and Decoding Parameters

For open-ended and yes/no examples, for the prompt we provide the raw question:

```
{question}
```

For multiple-choice examples, the answer options are appended below the question with letter labels:

```
{question}
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
...
```

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:

```
<image>
{prompt}
```

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. For our experiments, we use greedy decoding. Model specific decoding parameters are provided in Table 8.

## E.4 Computational Details

Open-weight models were evaluated using 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. For proprietary models, we estimate our cost around $150.

# F Open-ended Task Error Analysis

Open-ended questions constitute the most challenging task type in PoVisLE, as they require models to generate the target answer directly, in the required grammatical form. To keep the primary benchmark scoring deterministic and reproducible, we constrain the expected output format in the

| Annotator ID | Role | Age Group | Voivodship of Origin | Current/Familiarized Region |
|---|---|---|---|---|
| P1 | Primary | 20–25 | opolskie | mazowieckie |
| P2 | Primary | 35–40 | małopolskie | opolskie |
| S1 | Super-annotator | 35–40 | dolnośląskie | wielkopolskie |
| A1 | Auxiliary | 20–25 | dolnośląskie | dolnośląskie |
| A2 | Auxiliary | 25–30 | podlaskie | mazowieckie |
| A3 | Auxiliary | 20–25 | wielkopolskie | wielkopolskie |
| A4 | Auxiliary | 25–30 | mazowieckie | mazowieckie |
| A5 | Auxiliary | 35–40 | pomorskie | pomorskie |
| A6 | Auxiliary | 30–35 | świętokrzyskie | mazowieckie |
| A7 | Auxiliary | 30–35 | lubelskie | mazowieckie |
| A8 | Auxiliary | 30–35 | śląskie | mazowieckie |
| A9 | Auxiliary | 30–35 | mazowieckie | mazowieckie |
| A10 | Auxiliary | 25–30 | dolnośląskie | mazowieckie |
| A11 | Auxiliary | 35–40 | podlaskie | mazowieckie |
| A12 | Auxiliary | 20–25 | warmińsko-mazurskie | mazowieckie |
| A13 | Auxiliary | 25–30 | dolnośląskie | mazowieckie |

Table 6: Demographic characteristics of the annotation team.

| Model | Size | Type | Backend | Thinking | ID |
|---|---|---|---|---|---|
| GPT-5.4 | – | proprietary | OpenRouter | | `openai/gpt-5.4-20260305` |
| Claude Sonnet 5 | – | proprietary | OpenRouter | | `anthropic/claude-sonnet-5-20260630` |
| Qwen3.5-9B | 9B | open weight | vLLM | ✓ | `Qwen/Qwen3.5-9B` |
| Qwen3.5-27B | 27B | open weight | vLLM | ✓ | `Qwen/Qwen3.5-27B` |
| Qwen3.5-397B-A17B | 397B | open weight | vLLM | ✓ | `Qwen/Qwen3.5-397B-A17B` |
| Gemma-4-31B-it | 31B | open weight | vLLM | ✓ | `google/gemma-4-31B-it` |
| Mistral-Medium-3.5-128B | 128B | open weight | vLLM | | `mistralai/Mistral-Medium-3.5-128B` |
| Ministral-3-14B | 14B | open weight | vLLM | ✓ | `mistralai/Ministral-3-14B-Reasoning-2512` |
| GLM-4.6V | 106B | open weight | vLLM | | `zai-org/GLM-4.6V` |
| LLaVA-PLLuM-12B | 12B | open weight | vLLM | | `NASK-PIB/LLaVA-PLLuM-12b-nc-instruct-250715` |
| LLaVA-Bielik-11B-v2.6 | 11B | open weight | vLLM | | `NASK-PIB/LLaVA-Bielik-11b-v2.6-instruct` |
| InternVL3.5-38B | 38B | open weight | vLLM | | `OpenGVLab/InternVL3_5-38B` |

Table 7: Specific model versions used in our experiments. The Thinking column indicates models evaluated also in a thinking configuration.

prompt, including the required answer length and, where relevant, grammatical form. We then use deterministic exact-match evaluation against the reference answer, requiring correct Polish diacritics in all cases and correct capitalization where it is part of the expected answer.

We additionally inspect open-ended answers to check whether the scoring mechanism behaves as intended and to understand what types of mistakes models make. We therefore conduct a post-hoc analysis of incorrect or partially invalid open-ended answers. This analysis is not used for benchmark scoring; it is intended only to characterize model behavior and inspect the evaluation protocol.

First, we manually inspected samples of model responses and derived a compact set of recurring error classes:

- HALLUCINATION: the answer introduces an incorrect entity, event, place, object, or cultural association.
- NON-ADHERENCE TO INSTRUCTION: the output does not follow the required format, for example by providing explanations, multiple candidates, overly long answers, or a different grammatical form than requested.
- LANGUAGE SWITCH: the answer is produced partly or fully outside Polish.
- MISSPELLING: the answer contains orthographic errors, including missing or incorrect diacritics.
- REJECTION: the model refuses to answer, states that it cannot answer, or claims that the answer cannot be determined despite the task requiring a direct answer.
- NO ERROR: the model answers correctly and in the required form.

Since a single answer may contain more than one problem, the error classes are treated as multi-

| Model | Max tok. | Temp. | Top-$p$ |
|---|---|---|---|
| GPT-5.4 | 64 | – | – |
| Claude Sonnet 5 | 64 | – | – |
| Qwen3.5-9B | 64 | 0.0 | 1.0 |
| Qwen3.5-9B (*Thinking*) | 16384 | 0.0 | 1.0 |
| Qwen3.5-27B | 64 | 0.0 | 1.0 |
| Qwen3.5-27B (*Thinking*) | 16384 | 0.0 | 1.0 |
| Qwen3.5-397B-A17B | 64 | 0.0 | 1.0 |
| Qwen3.5-397B-A17B (*Thinking*) | 16384 | 0.0 | 1.0 |
| Gemma-4-31B-it | 64 | 0.0 | 1.0 |
| Gemma-4-31B-it (*Thinking*) | 16384 | 0.0 | 1.0 |
| Mistral-Medium-3.5-128B | 64 | 0.0 | 1.0 |
| Ministral-3-14B | 64 | 0.0 | 1.0 |
| Ministral-3-14B (*Thinking*) | 16384 | 0.0 | 1.0 |
| GLM-4.6V | 16384 | 0.0 | 1.0 |
| LLaVA-PLLuM-12B | 64 | 0.0 | 1.0 |
| LLaVA-Bielik-11B-v2.6 | 64 | 0.0 | 1.0 |
| InternVL3.5-38B | 64 | 0.0 | 1.0 |

Table 8: Decoding parameters used in our experiments. A temperature of 0.0 corresponds to greedy decoding.

label classifications. We then apply an LLM-as-a-Judge method using `DeepSeek-V4-Pro` [4], asking the judge model to assign one or more of the above classes to each open-ended prediction. The prompt used for the judge model is shown in Table 9.

The resulting label frequencies are shown in Table 10. The results indicate that hallucination is the most frequent error label for open-ended answers. Across all evaluated models, 49.45% of open-ended predictions are labeled as hallucinated, and the rate exceeds 50% for several weaker models, including Qwen3.5-9B, Ministral-3-14B-2512, LLaVA-Bielik-11B-v2.6, InternVL3.5-38B, Ministral-3-14B-2512 (*Thinking*), and LLaVA-PLLuM-12B. This suggests that open-ended errors often involve incorrect entities, events, places, objects, or cultural associations.

Instruction non-adherence (19.70%) is the second most frequent class, appearing more often in lower-performing models, which produce explanations, hedged answers, or responses that do not match the requested concise format. This is visible, for example, for LLaVA-PLLuM-12B (42.79%), Qwen3.5-9B (*Thinking*) (34.83%), and Qwen3.5-27B (*Thinking*) (28.36%). Other errors, such as language switching (3.01%), misspellings (3.19%), and rejection (4.30%), are also observed, but much less often. This suggests that open-ended questions are difficult mainly because models give wrong answers or do not follow the requested format, rather

[4] https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro

```
You're a helpful assistant specialized in error analysis.
You will be given:
- a question,
- accepted gold answers,
- an LLM answer.
Your task is to compare the LLM answer with the accepted gold
answers and identify all applicable error types.
The task is multi-label classification: an answer may have more
than one error type. Assign all labels that apply.

Use the following labels:

1) Hallucination
The answer provided by the LLM contradicts the gold answer, contains
factually incorrect information, or gives an entity/value that is not
among the accepted answers.
Example:
Question: Jak nazywa się król przedstawiony na banknocie widocznym
na obrazku?
Gold answer: Zygmunt I Stary
LLM answer: Bolesław Krzywousty
Label: Hallucination

2) Non-adherence to instruction
The answer provided by the LLM can be considered factually correct,
but it does not follow specific instructions included in the prompt
(e.g. incorrect format, too many words, wrong grammatical form, not
answering exactly as requested, answering using whole sentences
instead of one phrase or word, etc.).
Example:
Question: Do jakiej dynastii należał król przedstawiony na banknocie?
Odpowiedz jednym słowem w mianowniku.
Gold answer: Jagiellonowie
LLM answer: Król Zygmunt I Stary należał do dynastii Jagiellonów.
Label: Non-adherence to instruction

3) Language-switching
The answer provided by the LLM can be considered factually correct,
but it is not in Polish.
Example:
Question: Do jakiej dynastii należał król przedstawiony na banknocie?
Odpowiedz jednym słowem w mianowniku.
Gold answer: Jagiellonowie
LLM answer: Jagiellonian dynasty
Label: Language-switching

4) Misspellings
The LLM answer is factually correct, but it contains spelling or
orthographic errors, such as incorrect capitalization or missing
diacritics, and is therefore considered incorrect.
Example:
Question: Jak nazywa się król przedstawiony na banknocie widocznym
na obrazku?
Gold answer: Zygmunt I Stary
LLM answer: Zygmunt I stary
Label: Misspellings

5) Rejection
The LLM answer refuses to answer, states that it cannot identify/
recognize the person, object, or place, says there is not enough
information, or otherwise avoids providing the requested answer.
Example:
Question: Jak nazywa się osoba przedstawiona na zdjęciu? Podaj tylko
imię i nazwisko tej osoby, nic poza tym.
Gold answer: Jan Kowalski
LLM answer: Nie jestem w stanie zidentyfikować osób na podstawie zdjęć.
Label: Rejection

Classification rules:
- Assign Hallucination whenever the answer is factually incorrect or
  contradicts the gold answer.
- Assign Non-adherence to instruction whenever the answer violates
  formatting or instruction requirements specified in the question.
- Assign Language-switching whenever the answer, or a substantial part
  of it, is not in Polish.
- Assign Misspellings whenever the answer is malformed or misspelled.
- Assign Rejection whenever the answer refuses to answer or says it
  cannot identify someone/something.
- Multiple labels may be assigned to the same answer.
- Return valid JSON only, using this schema:
  {"labels": ["..."], "rationale": "one short sentence"}

Question: {question}
Accepted gold answers: {accepted_answers}
LLM answer: {prediction}
```

Table 9: Prompt used for classifying open-ended answer errors with the judge model.

| Model | No error | Halluc. | Non-adh. instr. | Lang. switch | Misspell. | Reject. |
|---|---|---|---|---|---|---|
| Qwen3.5-397B-A17B (*Thinking*) | 64.51 | 29.19 | 8.46 | 1.16 | 2.82 | 0.00 |
| Gemma-4-31B-it (*Thinking*) | 55.72 | 31.67 | 14.43 | 1.00 | 1.49 | 7.79 |
| GPT-5.4 | 56.55 | 33.00 | 8.96 | 0.33 | 3.65 | 2.82 |
| Claude Sonnet 5 | 54.89 | 26.70 | 18.24 | 0.50 | 2.82 | 7.30 |
| Gemma-4-31B-it | 42.45 | 50.25 | 13.43 | 1.82 | 1.16 | 1.00 |
| Qwen3.5-397B-A17B | 51.91 | 41.63 | 11.61 | 0.33 | 2.32 | 0.17 |
| GLM-4.6V | 48.92 | 42.29 | 14.76 | 3.15 | 4.64 | 2.99 |
| Qwen3.5-27B (*Thinking*) | 42.95 | 39.47 | 28.36 | 9.45 | 2.82 | 11.94 |
| Qwen3.5-27B | 33.17 | 58.37 | 15.42 | 1.16 | 2.65 | 0.00 |
| Qwen3.5-9B (*Thinking*) | 31.84 | 51.08 | 34.83 | 13.10 | 3.32 | 11.94 |
| Qwen3.5-9B | 25.54 | 66.83 | 16.75 | 1.66 | 2.49 | 0.17 |
| Ministral-3-14B-2512 | 23.55 | 67.33 | 19.57 | 2.16 | 4.81 | 1.33 |
| LLaVA-Bielik-11B-v2.6 | 22.39 | 58.04 | 25.54 | 2.16 | 3.15 | 6.47 |
| InternVL3.5-38B | 14.59 | 69.65 | 22.39 | 3.32 | 6.97 | 4.48 |
| Ministral-3-14B-2512 (*Thinking*) | 20.07 | 70.65 | 19.73 | 5.64 | 3.81 | 0.50 |
| LLaVA-PLLuM-12B | 15.42 | 55.06 | 42.79 | 1.16 | 2.16 | 9.95 |
| **Total** | 37.78 | 49.45 | 19.70 | 3.01 | 3.19 | 4.30 |

Table 10: Classified error label frequencies for open-ended predictions on the test split. Values indicate the percentage of predictions assigned each label.

than because of minor spelling or language-form issues.

## G Results on the Validation Split

This section reports supplementary results for the validation split. As shown in Section 3, splitting the validation set by category results in small groups. To obtain more stable estimates, we therefore report confidence intervals only for overall macro accuracy, shown in Figure 7. Table 11 gives accuracy by dataset category, and Table 12 gives accuracy by task type.

## H Results on the Test Split

This section reports supplementary results for the test split. Figure 8 shows category-level macro accuracy with confidence intervals. Table 13 gives accuracy by task type.

## I Detailed Translation Results

Table 14 reports overall accuracy for the Polish sample and its English and German translations. Deltas are computed relative to the Polish version.

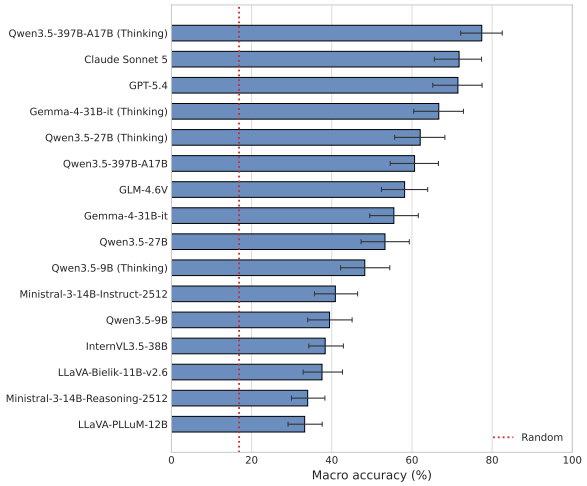


Figure 7: Overall macro accuracy on the validation split with 95% image-cluster bootstrap confidence intervals. The red dotted line denotes the random baseline.

| Model | Overall | Art & Entert. | Culture & Trad. | Geogr. & Nature | History & Society | Language | Image Und. | Visual Reas. |
|---|---|---|---|---|---|---|---|---|
| **PROPRIETARY MODELS** | | | | | | | | |
| Claude Sonnet 5 | 71.76 | 59.19 | 51.94 | 91.22 | 74.50 | 73.39 | **100.00** | 86.67 |
| GPT-5.4 | 71.47 | **62.73** | 83.33 | **94.76** | 69.01 | 69.12 | 86.67 | **93.33** |
| **OPEN-WEIGHTS MODELS** | | | | | | | | |
| Qwen3.5-397B-A17B (*Thinking*) | **77.42** | 60.22 | **86.94** | 91.98 | **88.10** | **77.79** | **100.00** | 80.00 |
| Gemma-4-31B-it (*Thinking*) | 66.70 | 54.72 | 69.44 | 54.35 | 71.24 | 71.73 | 93.33 | 63.33 |
| Qwen3.5-27B (*Thinking*) | 62.06 | 42.07 | 66.11 | 81.54 | 64.94 | 66.50 | **100.00** | 63.33 |
| Qwen3.5-397B-A17B | 60.66 | 58.21 | 44.72 | 54.66 | 65.89 | 52.06 | 86.67 | 80.00 |
| GLM-4.6V | 58.18 | 54.06 | 66.11 | 52.96 | 66.94 | 49.34 | 80.00 | 63.33 |
| Gemma-4-31B-it | 55.49 | 40.99 | 62.22 | 53.27 | 53.56 | 55.07 | 93.33 | 76.67 |
| Qwen3.5-27B | 53.28 | 38.99 | 41.11 | 82.30 | 54.78 | 51.72 | 93.33 | 73.33 |
| Qwen3.5-9B (*Thinking*) | 48.24 | 35.96 | 40.56 | 43.59 | 60.56 | 44.59 | 86.67 | 56.67 |
| Ministral-3-14B-2512 | 40.89 | 35.40 | 35.00 | 77.55 | 33.62 | 37.16 | 86.67 | 26.67 |
| Qwen3.5-9B | 39.44 | 29.66 | 29.44 | 40.05 | 37.18 | 36.29 | 86.67 | 66.67 |
| InternVL3.5-38B | 38.36 | 36.48 | 43.61 | 39.11 | 24.55 | 36.65 | 71.67 | 43.33 |
| LLaVA-Bielik-11B-v2.6 | 37.57 | 33.05 | 38.89 | 45.12 | 28.85 | 31.04 | 71.67 | 26.67 |
| Ministral-3-14B-2512 (*Thinking*) | 33.99 | 28.17 | 28.89 | 37.14 | 34.77 | 36.73 | 43.33 | 33.33 |
| LLaVA-PLLuM-12B | 33.24 | 32.13 | 32.22 | 35.57 | 29.14 | 30.45 | 51.67 | 13.33 |
| Random | 16.88 | 16.92 | 16.83 | 16.79 | 16.79 | 16.97 | 16.80 | 16.95 |

Table 11: Model accuracy by dataset category on the validation split, 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.

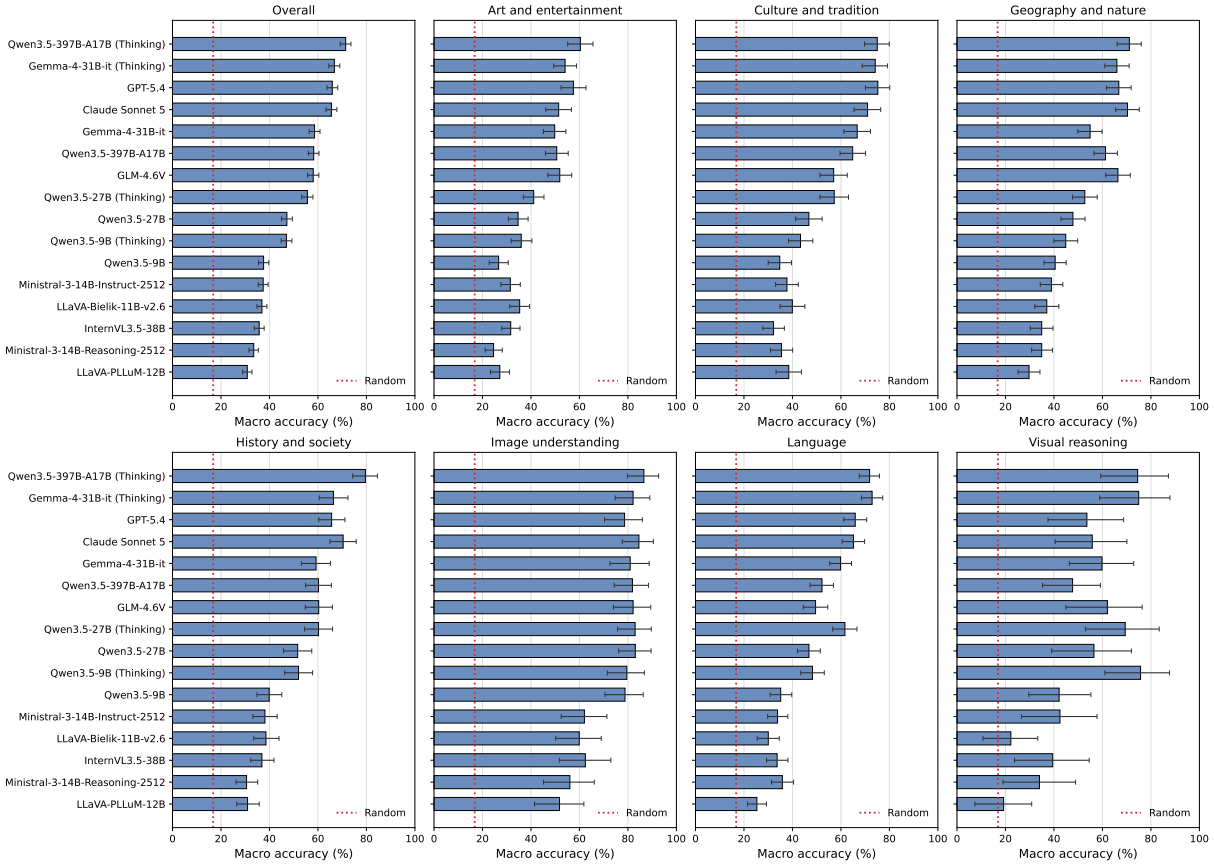


Figure 8: Macro accuracy by category with 95% image-cluster bootstrap confidence intervals on the test split. The red dotted line denotes the random baseline.

| Model | Overall | MCQ | Yes/No | Open |
|---|---|---|---|---|
| **PROPRIETARY MODELS** | | | | |
| Claude Sonnet 5 | 71.76 | 67.92 | 79.87 | 67.50 |
| GPT-5.4 | 71.47 | **70.75** | 81.17 | 62.50 |
| **OPEN-WEIGHTS MODELS** | | | | |
| Qwen3.5-397B-A17B (*Thinking*) | **77.42** | 68.40 | **86.36** | **77.50** |
| Gemma-4-31B-it (*Thinking*) | 66.70 | 66.98 | 83.12 | 50.00 |
| Qwen3.5-27B (*Thinking*) | 62.06 | 46.23 | 82.47 | 57.50 |
| Qwen3.5-397B-A17B | 60.66 | 49.06 | 77.92 | 55.00 |
| GLM-4.6V | 58.18 | 56.60 | 77.92 | 40.00 |
| Gemma-4-31B-it | 55.49 | 57.55 | 71.43 | 37.50 |
| Qwen3.5-27B | 53.28 | 49.06 | 70.78 | 40.00 |
| Qwen3.5-9B (*Thinking*) | 48.24 | 36.79 | 77.92 | 30.00 |
| Ministral-3-14B-2512 | 40.89 | 32.55 | 70.13 | 20.00 |
| Qwen3.5-9B | 39.44 | 27.83 | 62.99 | 27.50 |
| InternVL3.5-38B | 38.36 | 41.98 | 65.58 | 7.50 |
| LLaVA-Bielik-11B-v2.6 | 37.57 | 40.57 | 57.14 | 15.00 |
| Ministral-3-14B-2512 (*Thinking*) | 33.99 | 25.00 | 69.48 | 7.50 |
| LLaVA-PLLuM-12B | 33.24 | 34.43 | 57.79 | 7.50 |
| Random | 16.88 | 0.65 | 50.00 | 0.00 |

Table 12: Model accuracy by task type on the validation split, 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.

| Model | Overall | MCQ | Yes/No | Open |
|---|---|---|---|---|
| **PROPRIETARY MODELS** | | | | |
| GPT-5.4 | 65.93 | 62.52 | 78.71 | 56.55 |
| Claude Sonnet 5 | 65.57 | 61.28 | 80.53 | 54.89 |
| **OPEN-WEIGHTS MODELS** | | | | |
| Qwen3.5-397B-A17B (*Thinking*) | **71.45** | **65.94** | **83.89** | **64.51** |
| Gemma-4-31B-it (*Thinking*) | 66.78 | 64.07 | 80.53 | 55.72 |
| Gemma-4-31B-it | 58.60 | 56.45 | 76.89 | 42.45 |
| Qwen3.5-397B-A17B | 58.25 | 45.26 | 77.59 | 51.91 |
| GLM-4.6V | 58.06 | 49.77 | 75.49 | 48.92 |
| Qwen3.5-27B (*Thinking*) | 55.65 | 45.41 | 78.57 | 42.95 |
| Qwen3.5-27B | 47.19 | 43.55 | 64.85 | 33.17 |
| Qwen3.5-9B (*Thinking*) | 47.00 | 35.77 | 73.39 | 31.84 |
| Qwen3.5-9B | 37.59 | 26.44 | 60.78 | 25.54 |
| Ministral-3-14B-2512 | 37.41 | 24.11 | 64.57 | 23.55 |
| LLaVA-Bielik-11B-v2.6 | 36.89 | 30.02 | 58.26 | 22.39 |
| InternVL3.5-38B | 35.76 | 30.79 | 61.90 | 14.59 |
| Ministral-3-14B-2512 (*Thinking*) | 33.51 | 19.13 | 61.34 | 20.07 |
| LLaVA-PLLuM-12B | 30.85 | 23.64 | 53.50 | 15.42 |
| Random | 16.79 | 0.38 | 50.00 | 0.00 |

Table 13: Model accuracy by task type on the test split, 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.

| Model | PL | EN | | DE | |
|---|---|---|---|---|---|
| | | Value | Δ | Value | Δ |
| Qwen3.5-397B-A17B (*Thinking*) | 75.16 | 68.12 | -7.04 | 68.15 | -7.01 |
| Gemma-4-31B-it (*Thinking*) | 72.51 | 72.02 | -0.48 | 71.55 | -0.95 |
| Claude Sonnet 5 | 69.94 | 66.94 | -3.00 | 66.03 | -3.91 |
| GPT-5.4 | 68.35 | 70.40 | +2.06 | 63.53 | -4.82 |
| Gemma-4-31B-it | 63.96 | 60.15 | -3.82 | 59.04 | -4.92 |
| Qwen3.5-397B-A17B | 61.89 | 60.56 | -1.33 | 55.64 | -6.25 |
| GLM-4.6V | 61.72 | 56.31 | -5.41 | 58.03 | -3.69 |
| Qwen3.5-27B (*Thinking*) | 56.65 | 58.35 | +1.70 | 55.28 | -1.37 |
| Qwen3.5-27B | 50.83 | 51.84 | +1.01 | 50.18 | -0.65 |
| Qwen3.5-9B (*Thinking*) | 49.54 | 51.77 | +2.23 | 49.44 | -0.09 |
| LLaVA-Bielik-11B-v2.6 | 40.44 | 39.02 | -1.42 | 36.94 | -3.50 |
| InternVL3.5-38B | 39.88 | 41.66 | +1.79 | 40.81 | +0.93 |
| Ministral-3-14B-2512 | 39.54 | 41.65 | +2.11 | 37.66 | -1.88 |
| Qwen3.5-9B | 38.75 | 47.04 | +8.29 | 44.83 | +6.08 |
| Ministral-3-14B-2512 (*Thinking*) | 36.08 | 39.06 | +2.97 | 37.06 | +0.98 |
| LLaVA-PLLuM-12B | 34.80 | 32.90 | -1.90 | 33.08 | -1.72 |

Table 14: Overall model accuracy on the Polish sample and translated English and German samples, with all values reported as percentages. Deltas are computed relative to PL.