SEGDATA: A Standardised German Visual Question Answering Benchmark from Educational Assessments

Imge Yüzüncüoglu[1], Fabio Barth[1], René Wolf[1], Philippe Thomas[1]
[1]German Research Center for Artificial Intelligence
firstname.lastname@dfki.de

Abstract

The performance of Vision-Language Models is primarily assessed using benchmark datasets, particularly in the form of Visual Question Answering. Recent studies have shown that, although Vision-Language Models perform well on standard benchmarks, they struggle with non-English inputs and specialised tasks. However, as existing benchmarks are largely derived from English-centric, web-scale data, there is a lack of coverage for multilingual settings and more specialised reasoning tasks. To address these gaps, we introduce SEGDATA, a manually curated, mainly German multiple-choice Visual Question Answering dataset that covers diverse domains and a range of competency levels. The dataset is based on standardised educational assessment tasks developed by the German Institute for Quality Assurance in Education, ensuring high quality and well-defined evaluation criteria. We evaluate a range of recent state-of-the-art Vision-Language Models on it and show that even strong models struggle with the tasks included, particularly when accounting for varying question types and competence requirements. SEGDATA is publicly available.

1 Introduction

Visual-Language Models (VLMs), also referred to as Large Multimodal Models (Yue et al., 2024), represent a paradigm shift in artificial intelligence by integrating visual perception with linguistic reasoning. Although these models have achieved remarkable milestones, their capabilities are largely the product of extensive training in English-centric web-scale Visual Question Answering (VQA) datasets (Antol et al., 2015; Bordes et al., 2024; Fu et al., 2024). Consequently, the development of robust benchmarks has become essential to evaluate their performance, identify remaining gaps in multimodal understanding, and guide the further direction of training (Fu et al., 2024; Li et al., 2025).

Recent surveys (Fu et al., 2024; Li et al., 2025) categorise existing benchmarks as targeting fundamental capabilities, model introspection, or downstream applications. Despite this diversity, previous work consistently highlights two central limitations of current benchmarks: *i)* The majority of benchmarks are primarily constructed from English-language data derived from the web, which leads to a systematic bias in the evaluation process, as VLMs are largely trained using English-centric corpora (Antol et al., 2015). Their performance on non-English input is often weaker, and this area is still insufficiently studied (Parashar et al., 2024; Sugiura et al., 2026). This lack of high-quality benchmarks in other languages makes it challenging to conduct a comprehensive assessment and quantification of this gap. *ii)* Since many benchmarks rely on large-scale web scraping or automated data augmentation, issues such as noise and imbalance, as well as limited control over data distribution arise (Parashar et al., 2024; Fu et al., 2024). This can result in inflated performance on frequently represented concepts, whereas deficiencies on less common or more complex tasks are masked. Furthermore, a potential overlap between the benchmark and training data raises concerns about data leakage and overly optimistic evaluation results.

To address these gaps, we introduce **SEGDATA** (**S**tandardised **E**ducational **G**erman **Dataset**), a mainly German multiple-choice VQA dataset for evaluating VLMs. The tasks are derived from standardised educational assessment tasks for school grades of Year 3 and Year 8, developed by the Institute for Quality Assurance in Education (IQB) and deployed throughout Germany. These tasks are carefully curated and aligned with defined competency levels (Stanat et al., 2023), providing a controlled and interpretable benchmark for model evaluation. We transform the original IQB ma-

terials into a format suitable for vision-language research. SEGDATA spans multiple school subjects, allowing the evaluation of various model capabilities while reducing the risk of task-specific bias (Parashar et al., 2024). Additional optional cognitive complexity labels allow for a systematic analysis of model performance at various levels of reasoning complexity.

Our contributions can be summarised as follows:

- We release SEGDATA, a high-quality German dataset for the evaluation of VLMs, derived from standardised educational assessment tasks.
- We provide expert-verified competency annotations for each task, allowing for a granular analysis of model performance across developmental stages (School Years 3 and 8) and reasoning complexities.
- We establish baseline results for multiple open-weight VLMs on the dataset, demonstrating that standardised educational tasks remain a challenge for current multimodal architectures. Particularly in spatial and geometric reasoning tasks.

2 Related Work

Recent progress in VLMs has led to a broad ecosystem of benchmarks for multimodal evaluation. Early large-scale datasets such as VQAv2 (Jia et al., 2024), GQA (Hudson and Manning, 2019), and TextVQA (Singh et al., 2019) established standard settings for object recognition, compositional reasoning, and text understanding in images. More recent comprehensive benchmarks, including MMBench (Xu et al., 2023) and MMMU (Yue et al., 2024), extended evaluation toward expert knowledge, cross-domain reasoning, instruction following, and broader real-world competence.

Several recent benchmarks focus specifically on mathematical and scientific multimodal reasoning. MathVista (Lu et al., 2024) evaluates chart interpretation, geometry, symbolic reasoning, and quantitative problem solving. MathVerse (Zhang et al., 2024) and HC-M3D (Liu et al., 2025) further increase image dependence and reasoning difficulty. These resources substantially improve domain specificity and cognitive depth, but they remain predominantly English-centric and therefore provide limited evidence on multilingual robustness.

Multilingual VLM evaluation has so far developed along three main directions: translated VQA datasets, multilingual OCR benchmarks, and cross-lingual reasoning benchmarks. A recent example is PISA-Bench (Haller et al., 2025), which derives multimodal tasks from PISA assessments and provides parallel data in six languages. It reports substantial performance degradation on non-English splits and exposes weaknesses in spatial and geometric reasoning. However, its primary objective is cross-lingual comparability across translated parallel sets rather than calibrated competence assessment within each language. In addition, automatically translated samples may introduce linguistic artifacts and reduce benchmark fidelity.

Other multilingual benchmarks, such as M3Exam (Zhang et al., 2023) and MultiMath (Peng et al., 2024), source items from authentic school or national examinations across languages. Their focus, however, is largely restricted to mathematical reasoning and exam-style problem solving. Moreover, those benchmarks do not cover German. EXAMS-V (Das et al., 2024) similarly draws on school examinations from multiple countries and includes German, but covers only a limited set of domains, primarily Physics and Business/Economics.

Overall, existing benchmarks typically optimise either broad English evaluation, multilingual transfer, or narrow domain reasoning. They rarely combine authentic multilingual data, diverse subject coverage, and a controlled range of competency levels within a unified benchmark for German. Our benchmark addresses this gap by evaluating VLMs across multiple domains and difficulty levels using high-quality native-language data rather than translated surrogates.

3 SEGDATA Description

SEGDATA is based on the **VERA** (**VER**gleichs**A**rbeiten) assessment material provided by the IQB, a German research institute responsible for defining and evaluating nationwide educational standards across all 16 federal states of Germany (Burblies et al., 2026). They comprise standardised assessments that measure students' competency levels in relation to predefined educational objectives, in order to observe educational trends in Germany. SEGDATA consists of these assessments designed for primary school pupils in Year 3 and lower secondary school pupils in Year

8, covering the subjects German and mathematics. A smaller subset of tasks in English and French is included for lower secondary education. As a result, this dataset primarily provides high-quality German-language material for VLM evaluation as well as a standardised competency categorisation, enabling limited multilingual analysis and direct comparison of model performance across various proficiency levels.

We preprocess the original IQB material by filtering out non-visual tasks and converting the remaining content into a format suitable for VLM evaluation (see Section 4). The resulting dataset comprises 170 task descriptions, 395 questions, and 175 images, organised across 141 task sheets. German accounts for the majority of the dataset, covering 80.14% of all questions, while English and French constitute the remaining portion (see Table 1).

Within SEGDATA multiple questions are often associated with the same visual input, which requires models to reason about a shared context across different queries. Furthermore, VQA tasks span a wide range of domains, including geometry, text comprehension, and diagram interpretation. By preserving these characteristics, the benchmark reflects realistic educational assessment scenarios and captures a broad range of reasoning skills, facilitating fine-grained analyses of model behaviour and cross-task generalisation (Radford et al., 2021). Rather than serving as a large-scale pretraining dataset, SEGDATA is designed as a high-quality, standardised benchmark for the controlled and interpretable evaluation of VLMs. Its diverse tasks and varying levels of reasoning complexity support analyses that go beyond a single aggregated accuracy score.

Metric	German	English	French
Documents	113	15	13
Tasks	142	15	13
Questions	289	55	51

Table 1: Detailed breakdown of the language distribution within SEGDATA. Even though English and French tasks are included within the dataset, they only represent a minor subset of SEGDATA.

4 Dataset Construction

The IQB provided a dataset comprising PDF files. For each assessment, there is a handout describing the tasks for the students, another that contains the solutions, and a third with educational explanations, that delineates the competencies demanded by a given task and the potential challenges that pupils may face. The construction process of SEGDATA is visualised in Figure 1 and is described below.

4.1 Task Selection and Filtering

We manually identified which provided tasks could be completed using visual input (pictures, tables, maps, diagrams, etc.) only. Tasks that use an image solely for illustration purposes or could be solved without an image, were omitted. This particularly applies to tasks that focus exclusively on reading comprehension. Tasks that are meant to be auditive tasks and those where the solution had to be presented in the form of a sketch (e.g. drawing an axis of reflection) were also left out, as this would test capabilities other than visual processing.

4.2 Task Formatting

The remaining tasks were converted into JSON format, in which meta-, task- and question-based information is listed. To this end, the task document and the pedagogical commentary were entered into the standard ChatGPT web interface alongside a prompt (see Appendix A) to generate an initial draft of the JSON structure. As OpenAI automatically updates the model deployed via the web interface over time, different model iterations (ranging from GPT-4o to GPT-5.1) have been used. This reduced the amount of manual work required, particularly with regard to redundant processes. The JSON format generated by ChatGPT was then manually corrected. This manual process involved *i)* correctly partitioning the tasks and questions, *ii)* accurately incorporating the meta information and task descriptions, *iii)* formatting the questions into multiple-choice questions, and *iv)* generating image files. A work example for steps *i)* to *iii)* is attached in the Appendix B.

For task formatting, each task comprises all assignments that depend on the same image. As soon as a different or additional image is required to solve a task or answer a question, a new task is created. Each created task contains a general introduction to the task, as well as a list of questions to be answered based on the task description and the provided image. After the tasks and questions have been separated correctly, the next step follows: formatting the questions into multiple-choice tasks.

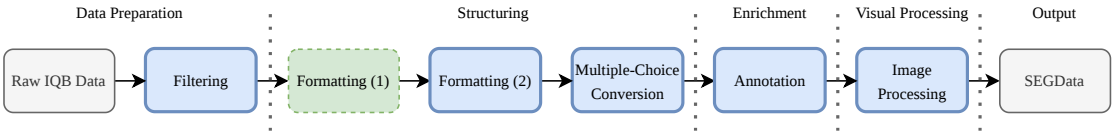


Figure 1: Overview of the dataset construction pipeline described in Section 4. It illustrates the process from the raw IQB data to the final SEGDATA. Light gray boxes denote input and output data, blue boxes represent manually curated processing steps, and the single green box indicates the semi-automated step using ChatGPT.

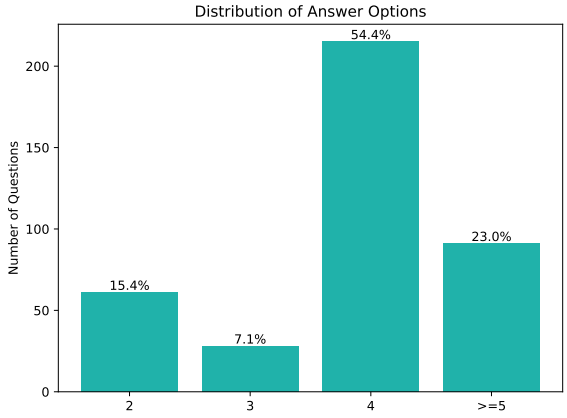


Figure 2: Distribution of answer options across the SEGDATA dataset.

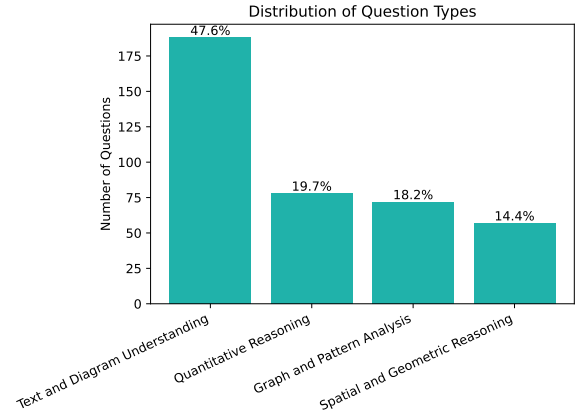


Figure 3: Distribution of the different question types within the dataset.

Multiple-Choice Some of the tasks in the dataset are cloze tests or reasoning tasks rather than multiple-choice questions. Where possible, we converted these into a standardised four-option format (one correct answer and three distractors). This design choice aligns with observed educational assessment practices from the IQB and ensures comparability with established VQA and reasoning benchmarks (e.g. Zhang et al. (2023) and Lu et al. (2024)), while providing a clear and consistent evaluation. Although four-option items constitute the majority of the dataset (see Figure 2), the collection also includes binary decisions, as well as tasks with three, five, or more alternatives. If available, the incorrect answer options (distractors) are taken directly from those provided in the solution document. If not, they are manually constructed to appear plausible and closely related to the correct answer. Examples include adjacent entries in tables or diagrams and spatially close points in coordinate systems. The correct answer labels were randomly shuffled to avoid systematic positional biases. This prevents models from exploiting answer-position patterns and encourages predictions based on the actual visual and textual content of the task.

Question Classification As the dataset contains tasks spanning different forms of reasoning, including mathematical reasoning, textual comprehension, and the interpretation of diagrams and other visual representations, each question was assigned to one of four categories based on Haller et al. (2025): *Spatial and Geometric Reasoning*, *Quantitative Reasoning*, *Graph and Pattern Analysis*, and *Text and Diagram Understanding*. The categorisation was initially determined by ChatGPT based on the didactic commentary and was subsequently manually checked. The distribution of task types is shown in Figure 3.

Class Level As the data provided by the IQB consists of compulsory achievement tests designed to assess subject-specific competencies of students and allow cross-state comparisons, we have additionally supplemented the tasks with the respective competence level assigned to each. This extension makes it possible to analyse the performance of VLMs in a differentiated manner along established cognitive difficulty levels and to interpret their results in relation to human competence profiles. Competency levels were taken from the official IQB website. A total of 47 task documents

have been assigned to grade 3 and 83 to grade 8, whilst 11 documents could not be clearly assigned to any grade because they were not listed (see Table 2). An additional overview of the distribution of the question types covered for each class grade is provided in Table 3.

Metric	Grade 3	Grade 8	Unknown
Documents	47	83	11
Tasks	56	103	11
Questions	120	237	38

Table 2: Detailed breakdown of SEGDATA granularity. The distribution shows the volume of documents, unique visual tasks, and individual questions segmented by target class grade.

4.3 Image Extraction

The images were generated in three steps: First, a screenshot was taken of the image within the PDF file. Second, a white background was added. Third, the image was centered and, if necessary, arranged side by side or below another required image. We intentionally retain variations in image resolution.

5 Experiments

We generate first baseline results on SEGDATA using a selection of recent open-weight VLMs, including models from Gemma 3 (Team et al., 2025), Gemma 4 (Gemma Team, 2026), and Qwen 3 (Bai et al., 2025) families (see Table 4). These models were selected to represent a variety of competitive architectures that are commonly used in current research (Choi et al., 2026; Nowicki et al., 2026), allowing us to assess model performance on the proposed dataset meaningfully and enable deterministic long-term reproducibility without relying on closed-source APIs. For readability, we refer to models using shortened names indicating family and parameter size (e.g., Gemma 4-31B). Full model identifiers are provided in Table 7 in the Appendix.

5.1 Experimental Setup

All experiments were conducted using the `vLLM` inference framework (Kwon et al., 2023) to ensure reproducible and efficient decoding. We evaluate the models in a **zero-shot** setting, meaning no task-specific examples were provided in the prompt. We use a structured German prompt that provides the model with a general instruction, the task description, the question, and the available multiple-choice options. Models were explicitly instructed to output only the letter corresponding to the correct answer. The full prompt template and an example of the input format are provided in Appendix C.

To eliminate parsing errors (e.g., the model providing conversational filler instead of a single letter), we utilise a constrained decoding approach. We extract the logits for the first generated token and restrict the output space to the set of valid answer keys $\{A, B, C, \dots\}$ defined for each question. This ensures that the model's performance is measured based on its internal probability distribution over the allowed answer space rather than its ability to follow formatting constraints. For all models, we use a temperature of 0 to ensure deterministic outputs.

5.2 Evaluation Metrics

To provide a nuanced assessment of model performance, we employ three complementary metrics that capture both classification success and probabilistic calibration.

Accuracy We calculate the standard accuracy as the proportion of correctly answered questions:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i = y_i) \tag{1}$$

where $\hat{y}_i$ is the predicted label and y_i the ground truth.

Chance-Corrected Accuracy As the number of answer options in SEGDATA varies, we report a chance-corrected accuracy to account for differing guessing probabilities. Adapting standard chance-correction principles (e.g., Cohen (1960)), we normalise the score relative to the random guessing probability $p_i = 1/n_i$, where n_i is the number of options for question i:

$$\text{Chance-Corrected Acc.} = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{Acc}_i - p_i}{1 - p_i} \tag{2}$$

Brier Score To evaluate the quality of the predicted probability distributions, we report the Brier score (Glenn, 1950). A lower Brier score indicates better calibration and higher confidence in correct predictions:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(f_{ij} - o_{ij})^2 \tag{3}$$

Question Type	Grade 3	Grade 8	UNK
Spatial and Geometric Reasoning	22	50	0
Quantitative Reasoning	35	43	0
Graph and Pattern Analysis	25	32	0
Text and Diagram Understanding	38	112	38

Table 3: Distribution of question types across class grades. The labels are based on the available information of IQB's website.

where f_{ij} is the predicted probability for option j, and $o_{ij} \in \{0, 1\}$ indicates the ground truth. Probabilities are computed via a softmax over the restricted logit space of valid answer tokens.

6 Results

In this section, we present the evaluation results across the evaluated models based on the introduced metrics. While disentangling specific error sources (e.g., separating German linguistic comprehension from visual grounding or domain-specific reasoning) would provide valuable insights, we deliberately focus this work on establishing a standardised performance baseline, leaving fine-grained error diagnostic frameworks for future work.

6.1 Overall Evaluation Results

Table 4 reports the performance of all evaluated models on SEGDATA. Across model families, performance generally improves with scale: larger models achieve higher accuracy and chance-corrected accuracy, while yielding lower (i.e., better) Brier scores. Gemma 4-31B achieves the best overall performance, reaching 74.9% accuracy and 64.6% chance-corrected accuracy. Qwen 3-235B-A22B performs similarly strongly across metrics and attains the lowest Brier score (0.42). Notably, Gemma 4-31B outperforms Qwen 3-235B-A22B despite having fewer parameters. Comparing model families, both Gemma and Qwen models demonstrate robust performance, with Qwen models showing particularly strong results at smaller parameter sizes. In contrast, smaller models across all families exhibit lower chance-corrected accuracy, highlighting the increased difficulty of the dataset when accounting for varying numbers of answer options. The gap between standard accuracy and chance-corrected accuracy is substantial across all models, indicating that performance is strongly influenced by the number of answer choices and that chance-corrected evaluation provides a more informative measure of model capability. Finally, the Brier scores show that larger models not only make more correct predictions but also produce better calibrated probability estimates, suggesting improved confidence estimation alongside accuracy gains. This shows that SEGDATA poses non-trivial challenges even for strong open-weight VLMs, particularly when accounting for varying answer options.

6.2 Results on Question Types

The accuracy scores for the different question types presented in Table 5 reveal several notable patterns in model performance. Overall, larger models tend to achieve higher scores across most categories. However, no single model consistently achieves the highest performance across all question types. Across all evaluated models, the highest accuracy is achieved on *Text and Diagram Understanding*, whereas *Spatial and Geometric Reasoning* consistently yields the lowest scores. Performance on *Quantitative Reasoning* and *Graph and Pattern Analysis* exhibits considerably greater variation between models, indicating that these categories differentiate model capabilities more strongly. When comparing model families, Qwen models generally perform best on *Quantitative Reasoning* (79.5% with Qwen 3-235B-A22B) whereas Gemma models achieve consistently strong performance across categories and attain the highest scores on *Text and Diagram Understanding* (79.9% with Gemma 4-31B). Furthermore, several mid-sized models outperform larger counterparts in individual categories. Interestingly, within the Qwen family, performance on *Spatial and Geometric Reasoning* decreases with increasing model size.

These findings suggest that improvements in overall model size do not translate uniformly across all reasoning skills. Instead, model architecture and training data appear to influence performance on specific reasoning tasks. Consequently, evaluating VLMs across diverse question categories provides

Model	Acc. ↑	Chance-Corrected Acc. ↑	Brier ↓
Gemma 3 Series			
Gemma 3-4B	43.5	22.8	1.10
Gemma 3-12B	53.7	35.4	0.84
Gemma 3-27B	61.8	47.6	0.73
Gemma 4 Series			
Gemma 4-E2B	41.0	18.9	1.07
Gemma 4-E4B	53.4	35.9	0.75
Gemma 4-26B-A4B	68.6	56.3	0.56
Gemma 4-31B	**74.9**	**64.6**	0.47
Qwen 3 Series			
Qwen 3-8B	59.2	44.6	0.71
Qwen 3-32B	70.1	59.1	0.55
MoE Variants			
Qwen 3-30B-A3B	61.3	46.9	0.59
Qwen 3-235B-A22B	72.4	61.5	**0.42**

Table 4: Comparison of Gemma 3 and Qwen 3 variants on SEGDATA. Metrics are split into hard classification accuracy and probabilistic scoring. MoE models are grouped by active parameter scaling.

Model	Graph and Pattern Analysis	Quantitative Reasoning	Spatial and Geometric Reasoning	Text and Diagram Understanding
Gemma 3-4B	45.6	38.5	45.8	44.4
Gemma 3-12	52.6	39.7	45.8	63.1
Gemma 3-27B	64.9	59.0	51.4	65.8
Gemma 4-E2B	35.1	37.2	36.1	46.5
Gemma 4-E4B	40.4	48.7	54.2	58.8
Gemma 4-26B-A4B	63.2	66.7	59.7	74.3
Gemma 4-31B	**75.4**	78.2	58.3	**79.7**
Qwen 3-8B	50.9	53.8	**63.9**	62.6
Qwen 3-32B	68.4	64.1	61.1	76.5
Qwen 3-30B-A3B	56.1	57.7	58.3	65.8
Qwen 3-235B-A22B	**75.4**	**79.5**	51.4	76.5

Table 5: Systematic evaluation of Gemma and Qwen model families across the four SEGDATA question categories. Reported are the accuracy results (%).

a more informative assessment than relying solely on an aggregated accuracy score.

6.3 Performance Across Grade Levels

Table 6 reports model performance across different question competence levels. We observe a general drop in performance from Grade 3 to Grade 8 across most evaluated models, indicating that tasks in higher grades are typically more challenging for VLMs. Notably, Gemma 3-12B (+2.2%) and Gemma 4-E2B (+4.4%) present exceptions to this overall trend, achieving higher accuracy on Grade 8 assessments compared to Grade 3. This outlier behavior suggests that performance across grade levels does not strictly correlate with model size or general capacity, but may also depend on specific pre-training distributions or task alignment in individual models. Furthermore, while larger models achieve higher overall accuracy, they exhibit greater performance drops across grade levels. For instance, Gemma 3-27B (-11.5%), Gemma 4-E4B (-12.5%), and Qwen 3-235B (-9.7%) suffer substantial declines, demonstrating that Grade 8 tasks pose a non-trivial challenge even to high-capacity architectures. Overall, these results highlight that SEGDATA enables differentiated evaluation across task types, revealing performance variations that are not captured by aggregate accuracy alone.

7 Conclusion

We introduced **SEGDATA**, a diverse, mostly German dataset for evaluating VLMs, based on high-quality, standardised assessment tasks developed by the IQB. The dataset enables evaluation across a range of task types and reasoning levels, providing a controlled alternative to existing web-derived benchmarks. We further presented baseline results for recent state-of-the-art VLMs, showing that even strong models struggle with the challenges posed by SEGDATA. Particularly when accounting for varying answer spaces and reasoning complexity. Overall, our findings highlight the need for more diverse, structured, and multilingual datasets to better assess and improve the capabilities of VLMs.

Data and Code Availability

To support the research community and ensure the reproducibility of our results, we make all resources publicly available. The copyright of the original data is held by the IQB. SEGDATA is hosted on **Hugging Face** (https://huggingface.co/datasets/DFKI-SLT/SEGData). To facilitate standardised benchmarking, the dataset has been integrated into the **LM Evaluation Harness** framework (Gao et al., 2023), allowing for automated evaluation of compatible models. The complete evaluation pipeline, including the restricted logit processing logic, the custom vLLM-based inference scripts, and the preprocessing steps are available on **GitHub** (https://github.com/imgeyuez/SEGData).

Limitations

The original IQB materials were transformed into a VLM-compatible dataset format (see Section 4) by an internal domain expert and then reviewed by a second expert. While most of the dataset is based on objective task content, some annotations, such as the assignment of task types, may involve a degree of subjectivity despite being guided by didactic descriptions. Consequently, minor annotation bias cannot be entirely excluded. However, we expect its impact to be limited due to the review process and the structured nature of the underlying source material. This potential bias could be reduced further in future work by incorporating multiple independent annotators. Furthermore, due to the limited number of annotators, SEGDATA's overall size is smaller than that of large-scale web-derived datasets. Therefore, it is primarily intended for evaluation and controlled analysis, rather than large-scale pre-training.

Acknowledgements

We would like to thank the IQB for providing us with the original task documents. These include the tasks descriptions, solutions and pedagogical comments.
We would also like to thank Björn Lünsdorf for drawing our attention to this gap regarding VLMs. This research was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Model	3	8	Δ
Gemma 3-4B	46.7	41.5	-5.2
Gemma 3-12B	50.8	53.0	+2.2
Gemma 3-27B	68.3	56.8	-11.5
Gemma 4-E2B	37.5	41.9	+4.4
Gemma 4-E4B	61.7	49.2	-12.5
Gemma 4-26B-A4B	70.8	67.8	-3.0
Gemma 4-31B	77.5	**73.3**	-4.2
Qwen 3-8B	61.7	57.6	-4.1
Qwen 3-32B	75.0	66.5	-8.5
Qwen 3-30B	67.5	58.5	-9.0
Qwen 3-235B	**78.3**	68.6	-9.7

Table 6: Accuracy results (%) across question-classes 3 and 8. The Δ column represents the performance shift $(8 - 3)$, with green indicating improvement and red indicating regression. Bold values denote best-in-class performance.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-VL Technical Report. *Preprint*, arXiv:2511.21631.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, and 22 others. 2024. An Introduction to Vision-Language Modeling. *Preprint*, arXiv:2405.17247.

Jule J Burblies, Edna Grewers, Benjamin Becker, Florian Enke, Nicklas J Hafiz, Rebecca Schneider, Karoline A Sachse, Sebastian Weirich, and Stefan Schipolowski. 2026. IQB-Bildungstrend 2021.

Byungjin Choi, Seongsu Bae, Sunjun Kweon, and Edward Choi. 2026. KorMedMCQA-V: A Multimodal Benchmark for Evaluating Vision-Language Models on the Korean Medical Licensing Examination. *Preprint*, arXiv:2602.13650.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. 2024. MME-Survey: A Comprehensive Survey on Evaluation of Multimodal LLMs. *arXiv preprint arXiv:2411.15296*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. A Framework for Few-Shot Language Model Evaluation.

Gemma Team. 2026. Gemma 4 Model Card. `https://ai.google.dev/gemma/docs/core/model_card_4`. Accessed: 2026-04-25.

W Brier Glenn. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly weather review*, 78(1):1–3.

Patrick Haller, Fabio Barth, Jonas Golde, Georg Rehm, and Alan Akbik. 2025. PISA-Bench: The PISA Index as a Multilingual and Multimodal Metric for the Evaluation of Vision-Language Models. *Preprint*, arXiv:2510.24792.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. *Preprint*, arXiv:1902.09506.

Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. 2024. VQA^2: Visual Question Answering for Video Quality Assessment. *Preprint*, arXiv:2411.03795.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.

Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1587–1606.

Yufang Liu, Yao Du, Tao Ji, Jianing Wang, Yang Liu, Yuanbin Wu, Aimin Zhou, Mengdi Zhang, and Xunliang Cai. 2025. The Role of Visual Modality in Multimodal Mathematical Reasoning: Challenges and Insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22596–22611, Vienna, Austria. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations (ICLR)*.

Filip Nowicki, Hubert Marciniak, Jakub Łączkowski, Krzysztof Jassem, Tomasz Górecki, Vimala Balakrishnan, Desmond C. Ong, and Maciej Behnke. 2026. Visual Affect Analysis: Predicting Emotions of Image Viewers with Vision-Language Models. *Preprint*, arXiv:2602.00123.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. 2024. The Neglected Tails in Vision-Language Models. *Preprint*, arXiv:2401.12425.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318.

Petra Stanat, Stefan Schipolowski, Rebecca Schneider, Sebastian Weirich, Sofie Henschel, and Karoline A Sachse. 2023. IQB-Bildungstrend 2022. *IQB-Bildungstrend 2022. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich*, 9.

Issa Sugiura, Keito Sasagawa, Keisuke Nakao, Koki Maeda, Ziqi Yin, Zhishen Yang, Shuhei Kurita, Yusuke Oda, Ryoko Tokuhisa, Daisuke Kawahara, and Naoaki Okazaki. 2026. Jagle: Building a Large-Scale Japanese Multimodal Post-Training Dataset for Vision-Language Models. *Preprint*, arXiv:2604.02048.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Cheng Xu, Xiaofeng Hou, Jiacheng Liu, Chao Li, Tianhao Huang, Xiaozhi Zhu, Mo Niu, Lingyu Sun, Peng Tang, Tongqiao Xu, Kwang-Ting Cheng, and Minyi Guo. 2023. MMBench: Benchmarking End-to-End Multi-modal DNNs and Understanding Their Hardware-Software Implications. In *2023 IEEE International Symposium on Workload Characterization (IISWC)*, pages 154–166.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, henzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? *arXiv preprint arXiv:2403.14624*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Initial Generation Prompt

The first step in converting the IQB data into a format suitable for VLMs involves a semi-automated stage that uses ChatGPT (versions ranging from GPT-4o to GPT-5.1.). This step aims to reduce redundant effort when creating the overall JSON structure for each document, as well as avoiding the need for manual transcription of all task content into this format. The prompt below, together with the respective documents, was used to generate an initial draft of the JSON structure for this purpose. This instructs the model to extract task- and question-level information from the source material and organise it into a predefined schema. These outputs were not used directly, but served as a preliminary scaffold. Each generated JSON file was subsequently manually reviewed and corrected by the annotator, including accurately separating tasks, refining task descriptions and verifying the extracted content, as well as all the further steps described in Section 4.

Generation Prompt (German)

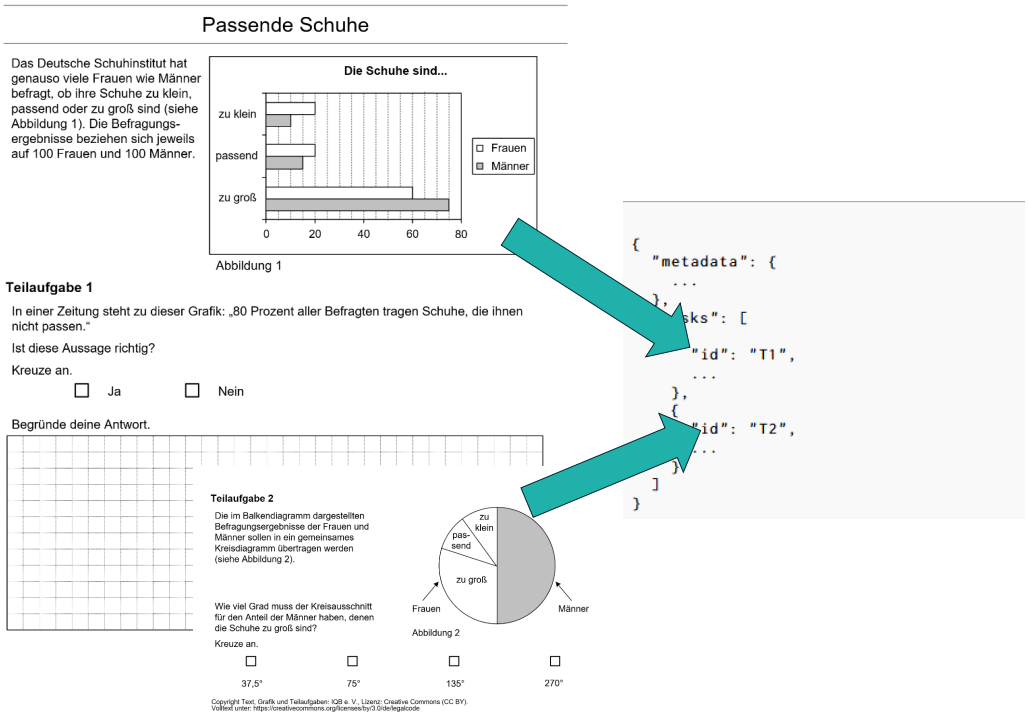
```
Analysiere das folgende Dokument.
- Jedes Dokument beschreibt eine Aufgabe, die wiederum mehrere Fragen enthalten kann.
- Extrahiere die Aufgabe(n) in folgendes JSON-Format:


{
  "metadata": {
    "filename": "<Dateiname>",
    "title": "<Titel falls vorhanden>",
    "source": "<Quelle falls vorhanden>"
  },
  "tasks": [
    {
      "id": "T1",
      "description": "<Beschreibung der Aufgabe>",
      "image": "",
      "questions": [
        {
          "id": "Q1",
          "added_multiplechoice": <True oder False>,
          "question_type": "<Fragetyp>",
          "text": "<Frage oder Antwortmöglichkeit>",
          "answers": ["<Antwortoption 1>", "..."],
          "gold_answer": ""
        }
      ]
    }
  ]
}
- Füge das Feld "image" als Platzhalter ein mit "image": "".
- Lasse "gold_answer" leer.
- Nutze eine eindeutige ID für jede Aufgabe (T1, T2, ...) und jede Frage (Q1, Q2, ...).
- Kategorisiere die Aufgabe in eine der folgenden Typen:
spatial and geometric reasoning, quantitative reasoning,
graph and pattern analysis, text and diagram understanding.
Verwende dafür ebenfalls den Ausschnitt aus der didaktischen Kommentierung.
- Falls die Frage in keine der angegebenen Typen passt, lass das Feld leer.
Gib nur die json aus.
```

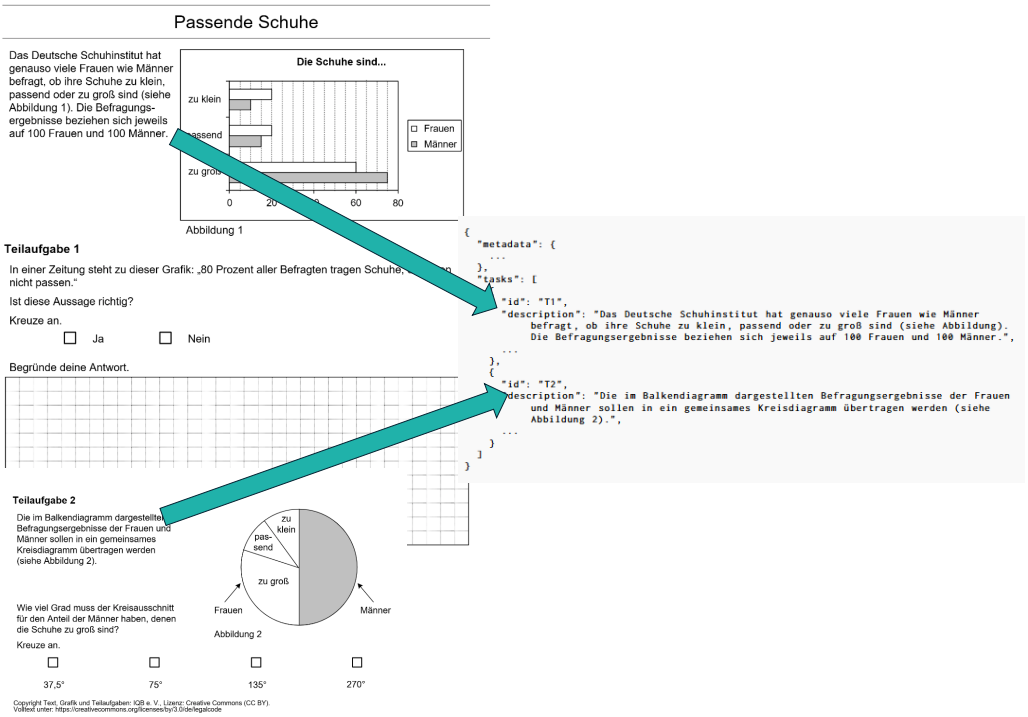
B Dataset Construction - Work Example

The output generated in Appendix A undergoes several manual refinement steps before being included in the final benchmark. As outlined in Section 4.2, this process comprises multiple stages. The following worked example demonstrates the transformation from the original PDF document to the final JSON format, visualising the processing steps *Formatting (2)* and *Multiple-Choice Conversion*, illustrated in Figure 1. The shown example is the IQB assessment *Passende Schuhe* conceived for grade 8, shown below.

Task Formatting (2.1) - Task Structure Once the raw ChatGPT output has been obtained, the extracted content must be reorganised into the final task structure. Each image-dependent problem is represented as an individual task. As the source document contains two independent images, each of which requires a separate solution, the corresponding JSON file is divided into two tasks.

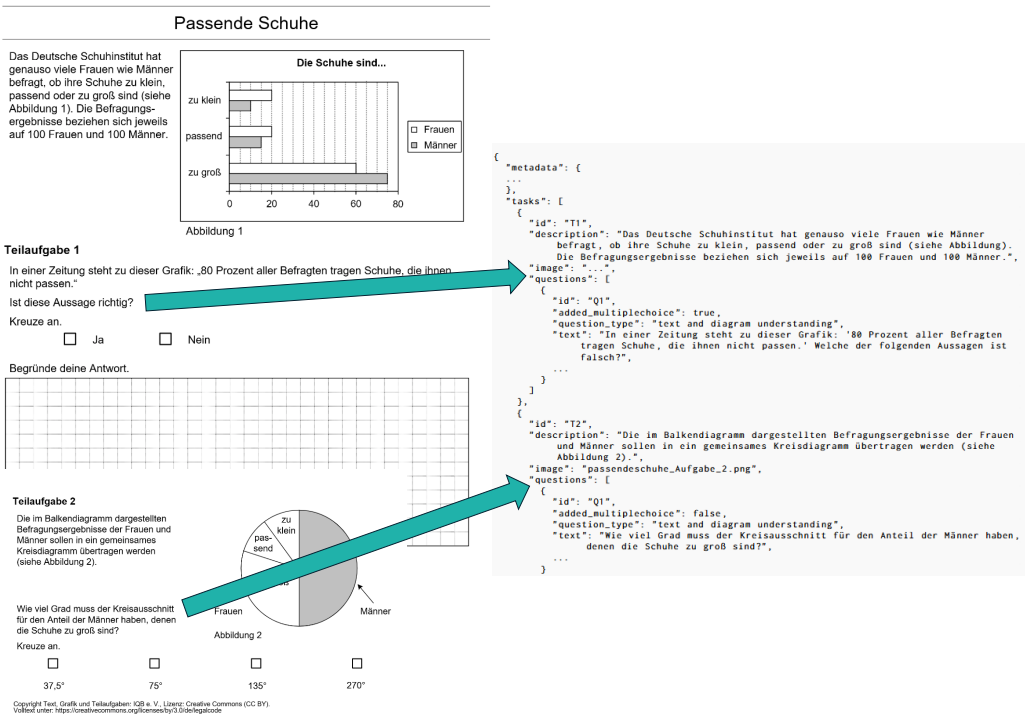


Task Formatting (2.2) - General Task Description Once the tasks have been categorised, a general description is assigned to each one. In many cases, as in this example, the description can be taken directly from the original document. The output of the initial prompt however, paraphrases or modifies the original wording. In these instances, the original task description is manually restored to ensure fidelity to the source material.

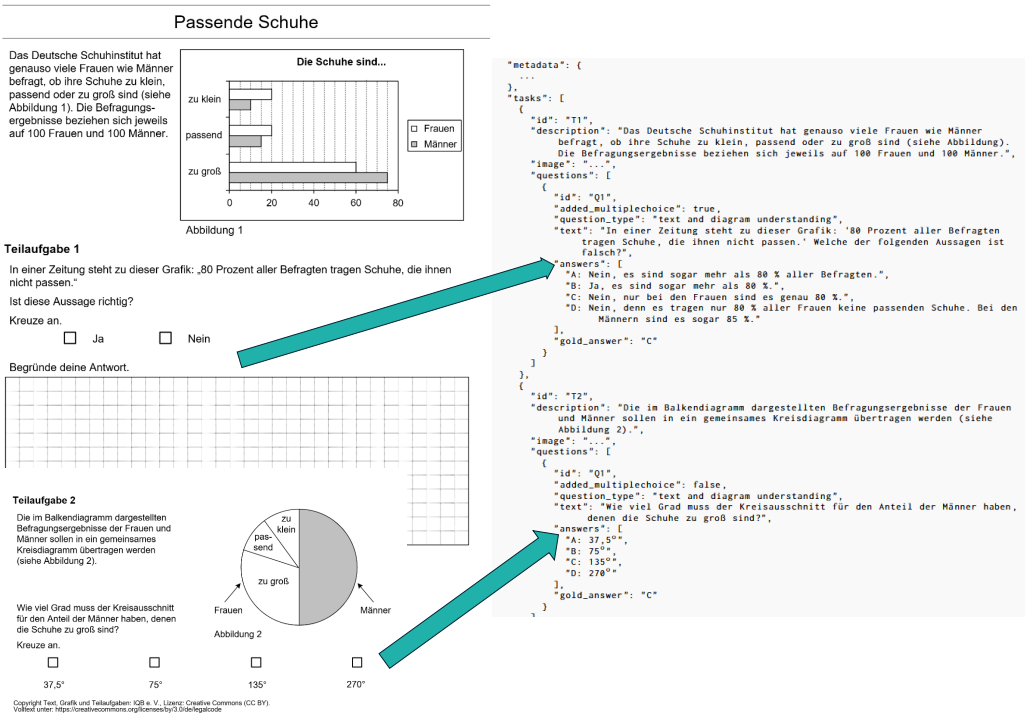


Task Formatting (2.3) - Question Formulation In this example, each task contains a single question. Similarly to the previous step, the question can largely be adopted from the original document, as there is a clear distinction between the general task description and the specific question. The second subtask of *Teilaufgabe 1* is an open-ended reasoning task. Since our initial task schema is designed exclusively for multiple-choice questions, it needs to be adapted if possible or omitted otherwise. To preserve the additional reasoning component of the original exercise, the justification is merged with the preceding binary decision, resulting in a single combined question. With respect to the multiple-choice conversion described in the next step, the question is also reformulated to identify the incorrect answer.

For *Teilaufgabe 2*, only a minor modification is required. The instruction "Tick the correct answer" is omitted, as VLMs do not interact with the worksheet directly. Instead, the model is expected to return the corresponding answer option (e.g., the letter of the correct choice).



Multiple-Choice Conversion To convert *Teilaufgabe 1* into a multiple-choice task, we refer to the official IQB solution guide. For this example, the solution provides a list of several acceptable justifications, as well as examples of borderline reasoning and incorrect reasoning. As there are multiple correct justifications but only one incorrect justification, asking directly for the correct answer would either result in a multiple-select task, or we would need to create distractors. In order to preserve the multiple-choice format with a single correct option, and to remain as faithful as possible to the IQB documents, the question is reformulated to ask which of the following statements is incorrect.




```
"metadata": {
  ...
},
"tasks": [
  {
    "id": "T1",
    "description": "Das Deutsche Schuhinstitut hat genauso viele Frauen wie Männer
        befragt, ob ihre Schuhe zu klein, passend oder zu groß sind (siehe Abbildung).
        Die Befragungsergebnisse beziehen sich jeweils auf 100 Frauen und 100 Männer.",
    "image": "...",
    "questions": [
      {
        "id": "Q1",
        "added_multiplechoice": true,
        "question_type": "text and diagram understanding",
        "text": "In einer Zeitung steht zu dieser Grafik: '80 Prozent aller Befragten
            tragen Schuhe, die ihnen nicht passen.' Welche der folgenden Aussagen ist
            falsch?",
        "answers": [
          "A: Nein, es sind sogar mehr als 80 % aller Befragten.",
          "B: Ja, es sind sogar mehr als 80 %.",
          "C: Nein, nur bei den Frauen sind es genau 80 %.",
          "D: Nein, denn es tragen nur 80 % aller Frauen keine passenden Schuhe. Bei den
              Männern sind es sogar 85 %."
        ],
        "gold_answer": "C"
      }
    ]
  },
  {
    "id": "T2",
    "description": "Die im Balkendiagramm dargestellten Befragungsergebnisse der Frauen
        und Männer sollen in ein gemeinsames Kreisdiagramm übertragen werden (siehe
        Abbildung 2).",
    "image": "...",
    "questions": [
      {
        "id": "Q1",
        "added_multiplechoice": false,
        "question_type": "text and diagram understanding",
        "text": "Wie viel Grad muss der Kreisausschnitt für den Anteil der Männer haben,
            denen die Schuhe zu groß sind?",
        "answers": [
          "A: 37,5°",
          "B: 75°",
          "C: 135°",
          "D: 270°"
        ],
        "gold_answer": "C"
      }
```


C Evaluation Prompt

The models were evaluated using a zero-shot prompting strategy. Each instance was presented with a system instruction in German, followed by the task description, the question text, the available multiple-choice options, and the image. The following template was used for all experiments:

Evaluation Template (German)

```
Ich zeige dir gleich eine Aufgabe, welche aus allgemeiner Beschreibung, Frage und
einem Bild besteht. Antworte auf die Frage NUR mit dem korrekten Buchstaben (A, B, C,
D, ...). Gib nur den Buchstaben aus, nichts anderes.

Beschreibung: {task_description}
Frage: {question_text}
Antwortmöglichkeiten: {answer_options}
Antwort:
```

The placeholders were dynamically filled with the corresponding dataset fields. To ensure consistency, the `answer_options` were provided as a comma-separated list of letters and their respective labels (e.g., *"A: Option 1, B: Option 2"*).

D Model Details

Short Name	Full Identifier
Gemma 3-4B	GEMMA3_google_gemma-3-4b-it
Gemma 3-12B	GEMMA3_google_gemma-3-12b-it
Gemma 3-27B	GEMMA3_google_gemma-3-27b-it
Gemma 4-E2B	GEMMA4_google_gemma-4-E2B-it
Gemma 4-E4B	GEMMA4_google_gemma-4-E4B-it
Gemma 4-26B-A4B	GEMMA4_google_gemma-4-26B-A4B-it
Gemma 4-31B	GEMMA4_google_gemma-4-31B-it
Qwen 3-8B	Qwen3VL_Qwen_Qwen3-VL-8B-Instruct
Qwen 3-30B-A3B	Qwen3VL_Qwen_Qwen3-VL-30B-A3B-Instruct-FP8
Qwen 3-32B	Qwen3VL_Qwen_Qwen3-VL-32B-Instruct-FP8
Qwen 3-235B-A22B	Qwen3VL_Qwen_Qwen3-VL-235B-A22B-Instruct-FP8

Table 7: List of the full model identifiers used in our experiments.