# CAN LLMS SERVE AS REFERENCE ANCHORS FOR THE GERMAN KANGAROO COMPETITION?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Benchmarks for mathematical reasoning rarely preserve the incentives, visual content, and long-running human reference data of real exams, especially outside English. We introduce Kangaroo, a longitudinal benchmark built from the German Mathematical Kangaroo archive from 1998–2025: 140 exams and 3,886 multiple-choice items across five grade groups, 1,746 of them multimodal, with official human results for 115 exams across 23 years. We evaluate four frontier vision–language models in a single-turn exam mode with explicit abstention under the contest's marking scheme, which implies negative points for false answers. The best model scores 88.34% accuracy, but every model loses between 23 and 43 percentage points on multimodal items. We then demonstrate that model scores cannot be used as reference anchors to decide objectively whether human performance improves or degrades over time. Across the 115 exams, human and model scores are negatively correlated with the correlation shrinking toward zero once we account for grade and the share of multimodal items. Calibrations fit on earlier years predict later years no better than the historical grade mean. We therefore reject the reference anchor hypothesis and conclude that humans and models do not align on perceived difficulty for these problems.

## 1 INTRODUCTION

Benchmarks shape how progress in large language models is measured and may even influence their development. But many mathematical evaluations remain text-only and English-centered (Cobbe et al., 2021; Shi et al., 2023). Such datasets leave out behavior that matters in real exams: reading diagrams, comparing graphical answer choices, and choosing between abstention for zero points, or selecting an answer when points are deducted for wrong answers.

The German Mathematical Kangaroo contest is a setting in which these behaviors can be studied together. It is a yearly, standardized exam series with a public archive since 1998, five grade groups, and many diagram-based problems. Its scoring rule encourages selectively giving answers under uncertainty: a wrong answer loses a quarter of the item's points while no answer receives zero points. Many questions require diagram reading, spatial inference, and reasoning over image-based answer choices, which remain hard for current multimodal models (Lu et al., 2023; Zhang et al., 2024a).

The contest also comes with a rare kind of human reference data. Official German summaries report the mean score of the pupils who took each exam for 23 years, with cohorts of up to 881,000 participants. Such a long record raises a question that matters for mathematics education: do pupils get better or worse over the years? The mean scores alone cannot answer it, because each year brings a new exam, and a lower mean can reflect a harder exam as well as a weaker cohort. Separating the two requires a *reference anchor*: an examinee who takes every exam and whose score shows how hard each exam is for pupils. Frontier models are promising candidates, because they can take every exam in the archive under the same conditions. To serve as an anchor, a model's score must move with the pupils' mean score from exam to exam and help predict it on later exams. What matters is therefore not how well a model solves an exam, but whether its score tracks how the pupils fare. If model scores meet these conditions, they offer a way to compare pupils across contest years. If they do not, the comparison still shows which features of an exam affect models and pupils differently, and that a model's score cannot stand in directly for the pupils' score.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We build the German Kangaroo benchmark for 1998–2025 and evaluate four frontier vision–language models under contest-faithful scoring. The benchmark contains 140 exams and 3,886 multiple-choice items, 44.93% of them multimodal. For the 115 exams with official human results, we test these conditions in three steps: whether model and human scores move together within a grade group, whether the model score adds anything beyond the share of multimodal items in the exam, and whether a calibration fit on earlier years predicts later years better than the grade group's historical mean.

Our contributions are threefold. First, we release a longitudinal multimodal benchmark of German school mathematics that keeps the contest's original negative-marking incentives. Second, we evaluate frontier models in single-turn exam mode with explicit abstention, and we test the two main alternative explanations for our results, the German evaluation language and the difficulty of multimodal items, with paired comparisons. Third, we show that frontier models score highly but remain much weaker on multimodal items, answer or abstain in model-specific ways under penalties, and do not serve as reference anchors: the raw negative correlation with human scores shrinks toward zero once the share of multimodal items is accounted for, and calibrations fit on earlier years predict later years no better than historical grade means.

## 2 Related Work and Positioning

Kangaroo intersects four lines of work. Text-based school-math benchmarks such as GSM8K established widely used tests of mathematical reasoning, and MGSM extended a translated subset to ten languages, including German (Cobbe et al., 2021; Shi et al., 2023). Benchmarks based on official exams, including AGIEval and GAOKAO-Bench, add authentic human assessment settings, while MATH and FrontierMath emphasize harder competition-style or expert-level problems (Zhong et al., 2024; Zhang et al., 2023; Hendrycks et al., 2021; Glazer et al., 2024).

Multimodal mathematics benchmarks have exposed persistent weaknesses in diagram understanding and visual reasoning (Lu et al., 2023; Zhang et al., 2024a). Multiple-choice evaluation is efficient, but models can be sensitive to option identifiers, order, count, and wording (Zhang et al., 2024b; Zheng et al., 2024; Song et al., 2025). Public archives also raise validity concerns: rotating or newly released tasks reduce contamination risk, and controlled numerical variants test whether success survives small changes to a problem (White et al., 2024; Balunović et al., 2025; Mirzadeh et al., 2024). Finally, evaluation rules affect whether models guess or abstain, and explicit error penalties provide a direct test of this behavior (Kalai et al., 2026). These findings motivate the benchmark's contest-faithful scoring and its robustness checks.

The reference-anchor question connects to work that uses models as stand-ins for human measurement: as judges of response quality (Zheng et al., 2023), as simulated participants (Argyle et al., 2023; Dillion et al., 2023), and as predictors of item difficulty (Benedetto et al., 2023; Yaneva et al., 2024). Closest to our question, LLaSA prompts language models to answer as students of various levels and estimates item difficulty from the simulated responses (Park et al., 2024), and Säuberli et al. (2025) find that such model response distributions resemble human ones only weakly without calibration. Comparing cohorts across exams is the problem of test equating, which requires common items or a reference population (Kolen & Brennan, 1995; Lord, 1980). A model that scored exams as pupils do would supply a fixed reference examinee, in the sense in which Hardt (2026) treats benchmarks as measurement instruments. We evaluate ordinary exam-mode scores as predictors of the human mean score of whole exams, with calibration tested on later years. We do not evaluate methods that simulate pupil behavior, and our analyses do not establish measurement invariance (Millsap, 2011) or equate exams.

Kangaroo is distinct in combining a long-running public exam archive with multimodal items, contest-faithful negative marking, and official human results. Its yearly administration lets us test whether model scores remain useful as anchors over time.

# 3 THE KANGAROO BENCHMARK

## 3.1 CORPUS

The benchmark covers the German Mathematical Kangaroo archive from 1998–2025. We define an exam as the paper for one grade group in one year. The archive spans five grade groups (3–4, 5–6, 7–8, 9–10, 11–13), 140 exams, and 3,886 multiple-choice items (Appendix A). The official solution sheet[1] identifies two tasks that cannot be scored, one in each of two exams. We credit both so that each exam keeps its published maximum (Appendix F). The dataset is a Parquet file with question text, answer options, gold labels, item value, grade group, year, modality flags, and embedded images for questions and options.

We built the dataset from the original PDFs. Each question region is located, passed through OCR, and paired with its visual content. The rendered question region, any separate diagram, and any image-based answer options are stored as image bytes, so the benchmark preserves the visual form of the original questions (Appendix H). Every stored problem number and gold letter was checked against the official solution sheet.

We call an item *multimodal* if its question has a separate diagram or any answer option is an image, and *text-only* otherwise. By this definition, 1,746 items (44.93%) are multimodal and 2,140 (55.07%) are text-only. Of the multimodal items, 1,627 have a question diagram and 393 (10.11%) have image-based answer choices. The *visual share* of an exam is the proportion of its items that are multimodal. It rises from 31% in 1998 to 70% in 2025 (Figure 1a).

## 3.2 CONTEST-FAITHFUL SCORING

The benchmark keeps the official scoring rule. Items are worth 3, 4, or 5 points, in equal thirds on complete exams, and every participant starts with one point per item. A correct answer earns the item's points, a wrong answer loses a quarter of them, and a blank answer earns zero. For an item worth $p$ points, blind guessing is break-even in expectation, $\frac{1}{5}p - \frac{4}{5} \cdot \frac{p}{4} = 0$, so deciding whether to answer is part of the task (Kalai et al., 2026).

Item counts and starting points vary by grade and year, so raw scores are not comparable across exams. We therefore divide each exam total by the maximum achievable score of that exam, $\%max = 100 \times$ total points/maximum possible points, and use this scale for models and pupils alike.

## 3.3 HUMAN REFERENCE DATA

Official German contest summaries exist for 2001–2019 and 2022–2025 and report, for each exam, the mean score and a histogram of scores. The 2020 and 2021 summaries were withheld because pupils worked at home, and the 2022 summary covers pupils who took the contest in person. Participation rose from about 104,000 pupils in 2001 to about 881,000 in 2025 (Mathematikwettbewerb Känguru e.V., 2025). We call the mean %max of the pupils who took an exam its *human mean score*. Where a summary lists grades separately, we weight them by participant count. The 2001–2012 summaries report only histograms. For those 60 exams we use the bin-midpoint mean, which differs from the reported mean by 0.13 points on average and at most 0.94 where both exist, and changes the prediction errors of Section 6.5 by at most 0.03 points (Appendix B). This gives 115 exams with human results across 23 years, covering 15.6 million participations (median 86,221 per exam).

The human mean score depends on the exam and on the pupils who took it, so it is not a pure measure of exam difficulty. It is, however, the quantity a reference anchor would have to predict. Because the summaries are aggregates, models and pupils can be compared on exam totals, but not item by item.

[1] https://www.mathe-kaenguru.de/wettbewerb/loesung/kaenguru_loesungen_alle.pdf

# 4 EVALUATION PROTOCOL

We evaluate models in a zero-shot exam mode that mirrors the contest. Each item is shown once with the OCR question text, a rendered image of the question region, the five answer choices A–E, and any separate diagram or option images. The model must end with a single answer letter or the explicit abstention token "Declined" (Appendix G). There is no tool use, no multi-turn interaction, and no second look at an item.

Scoring follows Section 3.2: a correct answer earns the item's points, a wrong or malformed answer loses a quarter of them, and an explicit abstention earns zero. For accuracy, abstentions count as incorrect.

We evaluate four frontier vision–language models: GPT-5, Qwen3-VL 235B Thinking, Grok 4 Fast, and Claude Sonnet 4.5. All runs went through OpenRouter-hosted APIs at temperature zero and top-$p = 1$, with prompts built directly from the benchmark fields. We archive one final response per item. Items that a provider error left without a final answer were re-queried under the same settings. The ablations of Section 6.3 ran through the same provider with the same settings. Model versions were the provider's current ones at the time of each run. We report the four models separately and, as a secondary series, the equal-weight ensemble, the mean of their four %max values.

# 5 TESTING SCORES AS COHORT-PERFORMANCE PREDICTORS

A reference anchor must move with the pupils' mean score from exam to exam and help predict it on later exams. We test both conditions on the 115 exams with official human results. Let $H_{gt}$ be the human mean score of grade group $g$ in year $t$ and $M_{gt}$ a model's %max on the same exam. The unit of analysis is the exam.

**C1: Do the scores move together?** We fit $H_{gt} = \alpha_g + \beta M_{gt} + \varepsilon_{gt}$ by ordinary least squares on the 115 exams, for each model and for the ensemble. A positive slope means that exams that are easy for pupils are also easy for models. A negative slope could still be useful for prediction, which C3 tests.

**C2: Does the model score add to the visual share?** We add the exam's visual share to C1 and, in a further specification, replace it with the separate shares of items with a question diagram and with image-based options. For each specification we report $\beta$ and the increase in $R^2$ from adding the model term.

**C3: Does a calibration predict later years?** Within each grade group we fit the calibration on the first eight reference years and predict the ninth, then extend the training window one year at a time. This gives 15 predictions per grade group and 75 in total (2009–2019 and 2022–2025). We compare four predictors: the grade group's mean over earlier years (the baseline), the exam's visual share alone, the model score alone, and the model score plus the visual share. The primary metric is the mean absolute error (MAE) and its difference from the baseline on the same exams. A forward $R^2 = 1 - \mathrm{SSE}_{\text{model}}/\mathrm{SSE}_{\text{baseline}}$ is secondary. The test is retrospective: it uses the 2025 model scores on every exam.

**Uncertainty.** Exams from the same year share a contest administration, so intervals for C1–C3 come from a year-cluster bootstrap that resamples years with replacement (2,000 draws, resampling 23 years for C1 and C2 and 15 held-out years for C3 with the predictions held fixed). The paired item-level comparisons of Section 6.3 resample items instead. All intervals are 95% percentile intervals.

# 6 EMPIRICAL FINDINGS

## 6.1 OVERALL PERFORMANCE

All four models score highly. GPT-5 is strongest at 88.34% accuracy, followed by Qwen3-VL at 86.31%. The order is the same under contest scoring: of the 19,440 points available across the 140

Table 1: Accuracy on text-only and multimodal items (2,140 and 1,746 items). Abstentions and malformed responses count as incorrect. The gap is in percentage points. The decline rate is the share of explicit abstentions over all 3,886 items.

| Model | Overall | Text-only | Multimodal | Gap | Decline rate |
|---|---|---|---|---|---|
| GPT-5 | 88.34% | 98.88% | 75.43% | 23.45 | 3.01% |
| Qwen3-VL | 86.31% | 98.41% | 71.48% | 26.93 | 0.00% |
| Grok 4 Fast | 78.33% | 97.66% | 54.64% | 43.02 | 11.84% |
| Claude Sonnet 4.5 | 77.35% | 92.62% | 58.65% | 33.97 | 0.75% |

exams, GPT-5 earns 17,299.5 and wins 85 exams, Qwen3-VL wins 39, Grok 4 Fast 9, and Claude Sonnet 4.5 7 (Appendix Table 5). GPT-5, Qwen3-VL, and Claude Sonnet 4.5 exceed the human mean score on all 115 exams with human results, and Grok 4 Fast on 113, by 35.2 to 47.2 points of %max on average (Appendix B). On text-only items, the models are close to the ceiling (Table 1).

### 6.2 Multimodal Items and Abstention

All four models do much better on the 2,140 text-only items than on the 1,746 multimodal items. GPT-5 falls from 98.88% to 75.43% and Grok 4 Fast from 97.66% to 54.64%. Table 1 gives all four models.

Where the visual content sits matters. GPT-5 reaches 80.78% on items with a question diagram only, 62.18% with image options only, and 54.74% with both. The other models show the same order, though the image-option subsets are smaller (Appendix Figure 4). The gap persists within every point tier and every grade group, so neither point tier nor grade explains it.

Abstention adds a second, model-specific behavior. Qwen3-VL never declines but produces four unresolved non-answer outputs. GPT-5 and Grok 4 Fast decline more often on multimodal items: 6.13% and 24.80% against 3.01% and 11.84% overall. Under contest scoring, the decision to answer matters as much as accuracy.

Within a grade group, exams with more multimodal items are harder for models but easier for pupils. The within-grade standard deviations across exams are 4.3 points for the human mean score, 5.9 for the ensemble, and 8.6 for Claude Sonnet 4.5. A 10-point higher visual share goes with a 2.78-point lower ensemble %max and a 1.35-point higher human mean score. The individual models range from $-3.88$ to $-1.70$ (Figure 1b,c). Adding the visual share to grade effects raises $R^2$ by 0.226 $[0.149, 0.298]$ for the ensemble and by 0.128 $[0.009, 0.317]$ for the human mean score (Appendix Table 11). The visual share rises over the years, so we also add year effects: the human coefficient falls to $+0.38$ $[-0.65, 1.40]$ while the ensemble's stays at $-2.23$ $[-3.47, -0.78]$ (Appendix Table 12). The model-side association is therefore robust to calendar time. The human-side one is not.

### 6.3 Language and Visual-Grounding Ablations

Two ablations, run under the same protocol and decoding settings, test the two main alternative explanations for the results above: that evaluating in German lowers model scores, and that the multimodal gap reflects harder items rather than reliance on the image. They cover GPT-5, Qwen3-VL, and Claude Sonnet 4.5. Grok 4 Fast had been retired by the provider (Appendix D).

The language ablation evaluates 200 German–English item pairs with author-reviewed translations. The items were drawn from the dataset with proportional quotas over grade group, point value, visual-content type, and year, so the subset keeps the composition of the dataset. Both arms show the text in the respective language with the diagrams and image-based options. Each model answers one fewer item correctly in English (a decrease of 0.50 percentage points), with paired intervals of $[-3.5, 2.5]$ for GPT-5, $[-5.0, 4.0]$ for Qwen3-VL, and $[-4.5, 3.5]$ for Claude Sonnet 4.5, and the discordant pairs are nearly symmetric (Appendix Table 16). German evaluation does not lower scores for these models.

The visual-grounding ablation re-evaluates the 1,353 items with a question diagram and textual answer options without any image, keeping the OCR question text and the five options, and compares

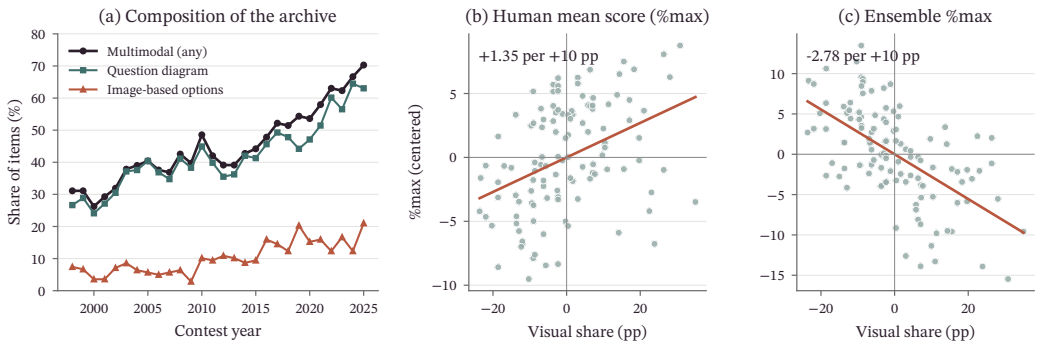


Figure 1: Exam composition and its opposite-signed associations. (a) Yearly share of the 3,886 items that are multimodal, that have a question diagram, and that have image-based options. The categories overlap, since an item may have both. (b, c) Human mean score and equal-weight ensemble %max against the exam's visual share on the 115 exams, all centered within grade group. Lines are least-squares fits annotated with the change per 10-point increase in the share.

Table 2: Visual comparison on the 1,353 question-diagram items with textual options. The difference is without minus with images in percentage points with a percentile bootstrap interval over items (2,000 draws). Abstention rates are explicit abstentions over all 1,353 items.

| Model | With images | Without images | Diff. (pp) [95% CI] | Lost / gained | Abstentions with / without |
|---|---|---|---|---|---|
| GPT-5 | 80.78% | 39.25% | $-41.54$ $[-44.3, -38.8]$ | 575 / 13 | 4.43% / 54.62% |
| Qwen3-VL | 77.90% | 56.25% | $-21.66$ $[-24.3, -18.9]$ | 357 / 64 | 0.00% / 7.10% |
| Claude Sonnet 4.5 | 64.30% | 44.79% | $-19.51$ $[-22.5, -16.6]$ | 365 / 101 | 0.96% / 25.28% |

the result with the main runs on the same items. Accuracy falls by 41.54 percentage points for GPT-5, 21.66 for Qwen3-VL, and 19.51 for Claude Sonnet 4.5, with paired item-bootstrap intervals of $[-44.3, -38.8]$, $[-24.3, -18.9]$, and $[-22.5, -16.6]$ points (Table 2). The flips are strongly asymmetric: GPT-5 loses 575 items and gains 13. The multimodal gap in Table 1 therefore reflects dependence on the image, not item difficulty alone.

Removing the image also changes whether models answer. The explicit-decline rate on these items rises from 4.43% to 54.62% for GPT-5, from 0.96% to 25.28% for Claude Sonnet 4.5, and from 0.00% to 7.10% for Qwen3-VL. Image access affects both accuracy and the decision to answer.

### 6.4 Human–Model Alignment and Exam Composition (C1, C2)

C1 asks whether exams that are easy for pupils are also easy for models. They are not. Across the 115 exams, the ensemble score and the human mean score are negatively correlated: Pearson $r = -0.490$ $[-0.639, -0.291]$ and Spearman $\rho = -0.457$ $[-0.617, -0.259]$, with year-cluster bootstrap intervals. Each model separately has $r$ between $-0.340$ and $-0.454$, and each grade group separately has $r$ between $-0.315$ and $-0.576$ (Appendix Table 10). Within grade groups, a 10-point higher ensemble %max goes with a 2.67-point lower human mean score, $[-4.49, -0.80]$, and an $R^2$ increase of 0.109 over grade effects alone. The estimate is stable when any single year is left out (range $-3.35$ to $-2.19$) or the two exams with a credited task are excluded ($-2.59$ $[-4.45, -0.59]$).

C2 asks whether the model score adds anything beyond the visual share. It does not. Adding the exam's visual share moves the ensemble slope to $-1.48$ $[-3.36, 0.26]$ with an $R^2$ increase of 0.022. The separate diagram and image-option shares move it to $-0.40$ $[-2.40, 1.44]$ with an increase of 0.001. Every individual model shows the same pattern, with intervals spanning zero and $R^2$ increases of at most 0.021 (Figure 2 and Appendix Table 8). Year fixed effects, which use only the differences between grade groups within a year, give $-0.32$ $[-1.99, 1.33]$ for the ensemble (Appendix Table 9). The negative correlation is largely a composition effect: exams with more multimodal items are harder for models and easier for pupils. There is no stable mapping from model score to human score.

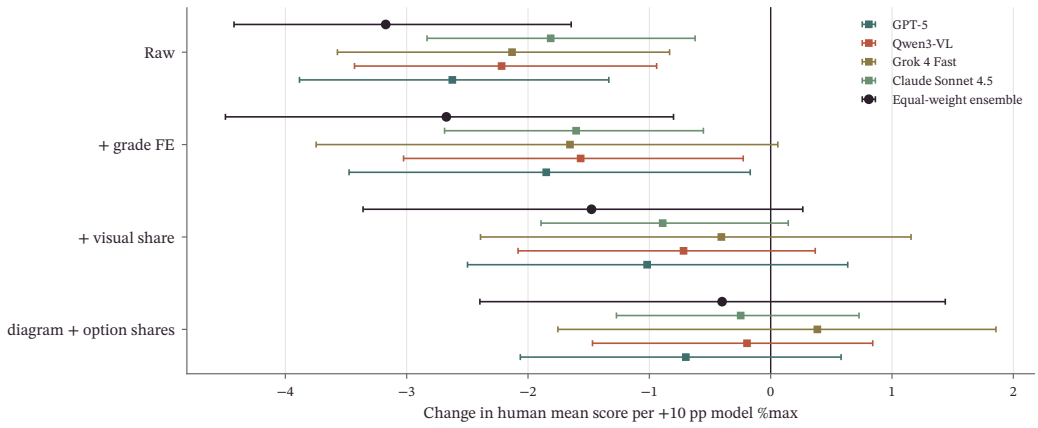


Figure 2: Change in the human mean score per 10-point increase in model %max on the 115 exams, for each model and the equal-weight ensemble. The four specifications are raw, grade fixed effects, grade effects plus the exam's visual share, and grade effects plus the separate shares of items with a question diagram and with image-based options, which replace the total share. Bars are year-cluster bootstrap 95% intervals.

### 6.5 Predicting Later Years (C3)

C3 asks whether a calibration fit on earlier years predicts the human mean score of later years better than the grade group's own history. It does not. Over the 75 held-out exams of Table 3, the grade-mean baseline has an MAE of 4.16 points. That is about the size of the year-to-year spread of the human mean score within a grade group, whose standard deviation is 3.7 to 4.6 points (Appendix Table 6), so the baseline leaves the year-to-year variation to be explained. The ensemble calibration reaches 4.12, a difference of $-0.04$ $[-0.68, 0.54]$, and the individual models range from 4.10 to 4.25, all with intervals spanning zero and forward $R^2$ between $-0.057$ and $0.009$. The visual share alone predicts better than any model, with an MAE of 3.83, a forward $R^2$ of 0.094, and a lower error than the grade mean on 47 of the 75 exams. Its difference from the baseline is $-0.33$, with a 95% confidence interval of $[-1.16, 0.50]$ that also includes zero. Adding a model score to the visual share does not help either: the change is $+0.05$ $[-0.22, 0.31]$ for the ensemble and $-0.15$ $[-0.43, 0.11]$ for Qwen3-VL, the only model with a negative estimate. Splitting each grade series at 2012 and fitting in both directions agrees: nineteen of twenty out-of-sample $R^2$ values are negative (Appendix Table 13). Figure 3 shows the observed human mean scores against both forecasts for each grade group.

The per-grade results (Appendix Table 14) rest on 15 predictions each. The ensemble beats the grade mean in four grade groups and is markedly worse in grades 5–6, where in 2024 the pupils scored

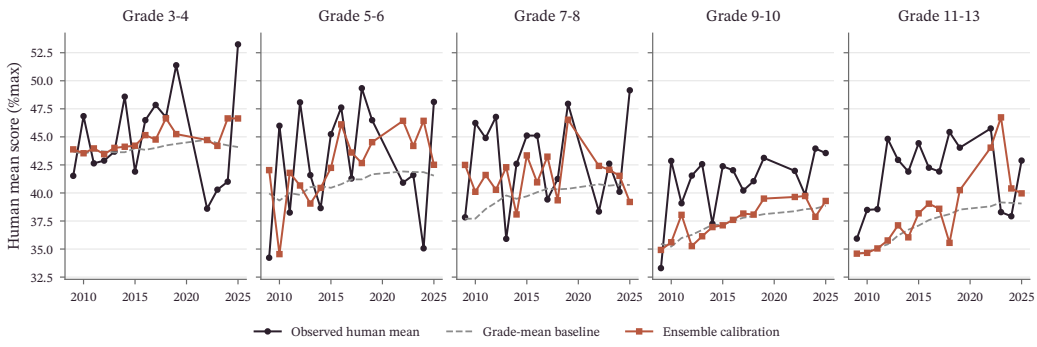


Figure 3: Observed human mean scores, grade-mean forecasts, and ensemble-score forecasts for the 15 held-out years of each grade group.

Table 3: Prediction of the human mean score on 75 held-out exams (expanding window within each grade group, starting from the first eight reference years). MAE differences are relative to the grade-mean baseline in the upper block and to the visual-share predictor in the lower block, with year-cluster bootstrap intervals. Forward $R^2$ and wins use the grade-mean baseline in both blocks. The visual-share predictions are the same for every model series.

| Predictor | MAE | ΔMAE [95% interval] | Forward $R^2$ | Wins (of 75) |
|---|---|---|---|---|
| Grade mean (baseline) | 4.159 | – | 0.000 | – |
| Visual share of the exam | 3.827 | $-0.331$ $[-1.16, 0.50]$ | 0.094 | 47 |
| GPT-5 score | 4.162 | $+0.004$ $[-0.45, 0.45]$ | $-0.015$ | 32 |
| Qwen3-VL score | 4.104 | $-0.054$ $[-0.51, 0.36]$ | $-0.007$ | 40 |
| Grok 4 Fast score | 4.161 | $+0.002$ $[-0.64, 0.66]$ | $-0.057$ | 41 |
| Claude Sonnet 4.5 score | 4.252 | $+0.093$ $[-0.27, 0.43]$ | 0.009 | 37 |
| Ensemble score | 4.115 | $-0.043$ $[-0.68, 0.54]$ | $-0.006$ | 43 |
| *Model score + visual share, difference relative to visual share alone* | | | | |
| GPT-5 score + visual share | 3.995 | $+0.168$ $[-0.12, 0.46]$ | 0.018 | 37 |
| Qwen3-VL score + visual share | 3.679 | $-0.149$ $[-0.43, 0.11]$ | 0.069 | 45 |
| Grok 4 Fast score + visual share | 3.892 | $+0.065$ $[-0.17, 0.32]$ | 0.009 | 45 |
| Claude Sonnet 4.5 score + visual share | 4.027 | $+0.200$ $[-0.10, 0.51]$ | $-0.001$ | 43 |
| Ensemble score + visual share | 3.875 | $+0.048$ $[-0.22, 0.31]$ | 0.047 | 43 |

35.07 %max while the grade mean predicted 41.85, the ensemble calibration 46.43, and the visual share 47.18. Two alternative specifications leave the picture unchanged (Appendix Table 15): pooling the grade groups with one shared slope lowers the ensemble's MAE to 3.91 ($-0.25$ $[-0.75, 0.29]$ against the grade mean), and a baseline that uses only the previous three reference years reaches 3.55, below every model calibration. A reference anchor would have to predict later exams better than the grade group's own history and add information beyond the visual share. The tested calibrations do neither. Current model scores cannot serve as reference anchors for human exam performance.

## 7 Discussion

Kangaroo is a benchmark for multimodal mathematical reasoning under real scoring incentives and a probe of how models differ from pupils. High raw accuracy shows that frontier models solve most school-style mathematics items. The multimodal gap shows that diagrams and image-based choices remain a major source of error, and the no-image rerun shows that the models depend on the supplied image. The missing link between model and human scores shows that what makes an exam hard for a model is not what makes it hard for pupils.

**Why model scores fail as anchors.** Exam composition is the main reason. The share and location of visual content change from year to year, model accuracy falls sharply on multimodal items, and the human mean score moves the other way. Once the visual share is accounted for, the association between model and human scores shrinks toward zero. Model-specific abstention under negative marking adds further variation that pupils do not share. Our conclusion rests on the prediction test, not on the sign of the raw correlation. Two cheap predictors already exist, the grade group's historical mean and the exam's visual share, and a model score is useful only if it beats them on later exams. None does.

**What would make an anchor work.** Three things are missing. Item-level human responses would show which items separate pupils and whether the text-only slice, where models are near the ceiling, tracks human difficulty at all. Methods that simulate pupils rather than solve exams, such as LLaSA (Park et al., 2024), target the right quantity and could be tested on the same 115 exams. A predictor that uses item content, for example the visual share together with item features, could be tested with the same three checks. Until one of these passes C3, historical grade means and exam composition remain the better forecasts.

**What Kangaroo measures well.** The benchmark is most informative on whether a model can solve realistic school-math items that mix text, diagrams, and the decision to answer. Unlike text-only

benchmarks it captures the interaction between perception and reasoning, and its organization by exam supports analysis of whole papers rather than isolated items. The three checks apply to any exam archive with published human results and can be rerun as models change.

## 8 Limitations and Future Directions

The two ablations have limited reach. The German–English comparison covers 200 items and three models, so effects of a few percentage points, effects specific to the full corpus, and languages other than English remain untested. The no-image rerun covers items with a question diagram and textual options. The 393 items with image-based answer options cannot be tested the same way. The archived runs do not test option-order robustness (Appendix E reports the gold-position marginals), and the public availability of old exams means that training exposure cannot be ruled out. Problem numbers and gold letters were checked against the official solution sheet, but residual OCR errors in statements remain possible.

The official summaries give exam-level means and histograms, not item-level responses. This rules out matched human slices, item-response modeling, and a direct test of whether the text-only slice tracks human item difficulty. The participants are contest entrants, not a random sample of German pupils, and their number, preparation, and grade mix change across years. Dividing by the exam maximum equates scales, not difficulty. We evaluated four models, once each, through one provider, with one prompt. The results describe those runs. They concern linear calibrations of ordinary exam-mode scores and do not assess pupil-simulation methods or predictors trained on item content.

## 9 Conclusion

Kangaroo covers 3,886 items from 140 German Mathematical Kangaroo exams from 1998–2025 under the contest's negative marking, with official human results for 115 exams. We asked whether frontier models can serve as reference anchors, fixed examinees whose scores show how hard each exam is for pupils. Frontier models score highly but remain much weaker on multimodal items. Their scores are negatively correlated with the human mean score, and the correlation shrinks toward zero once the share of multimodal items is accounted for. Calibrations fit on earlier years predict later years no better than the historical grade mean on 75 held-out exams. Current model scores therefore cannot serve as reference anchors and should not replace human reference data for comparing pupils across years. The main reason is exam composition: exams with more multimodal items are harder for models but not for pupils. The benchmark and its three checks can be rerun as models change and applied to any exam archive with published human results.

### AI Use Statement

In this work, we used generative AI tools as assistants: to help refine the study hypotheses and conceptual framework, to give feedback on the research methodology and statistical analysis, to assist in implementing the analysis and evaluation code, to assist with cleaning and reformatting the dataset, to assist with the German–English translation of the 200-item comparison subset, to help interpret results, to assist in creating and modifying figures, to help draft and edit parts of the paper, to help identify and summarize related literature, and to format references. The code assistance included a test that constructs synthetic panels with known relationships, used only to verify the analysis code and never as empirical observations. We have not otherwise used generative AI tools to generate synthetic data, and we have not used them to develop theoretical models or to formulate or prove mathematical claims. Qualitative or thematic analysis, surveys, and transcription are not applicable to this work. We have reviewed all AI-assisted work: the analysis code recovers planted effects on synthetic data, every statistical table and figure is produced by the analysis code, the translations were reviewed item by item by the authors, and citations were checked against the cited sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### Ethics Statement

The source material is the published archive of a school contest. The organizers have approved redistribution of the dataset under a CC BY-NC 4.0 license, and the code is released under MIT terms. The official results contain aggregate statistics, not participant-level records. The human mean score of each exam, derived from them, is reported in Appendix B. The raw organizer summaries are not redistributed, in line with the organizers' terms, and are available to reviewers on request. The benchmark is intended for evaluation and measurement research, not for grading pupils. High model accuracy alone should not be interpreted as evidence of pupil-like reasoning or accurate cohort forecasts.

### Reproducibility Statement

The dataset (3,886 items) is released under CC BY-NC 4.0. An anonymized code package accompanies the submission with the evaluation runner and prompt (Appendix G), the final predicted letter of every item in the four October 2025 runs and the nine comparison runs together with each run's configuration, and the analysis code. The English text of the 200 translated items is released with the dataset. The commands `uv run reproduce-diagnostics` and `uv run reproduce-surrogate` score the archived outputs against the dataset keys and regenerate every statistical table and figure in this paper. The 115 human mean scores are tabulated in Appendix B and included at full precision in the package. The raw official summary files, which the commands read, are not redistributed and are available to reviewers on request. Full regression, prediction, sensitivity, and decomposition tables are in Appendix C.

## References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37, 2023.

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

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Moritz Hardt. *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653:1047–1051, 2026. doi: 10.1038/s41586-026-10549-w.

Michael J. Kolen and Robert L. Brennan. *Test Equating: Methods and Practices*. Springer, New York, 1995.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Mathematikwettbewerb Känguru e.V. Känguru der Mathematik in Deutschland: Chronik. `https://www.mathe-kaenguru.de/chronik/index.html`, 2025. Accessed 2026-05-06.

Roger E. Millsap. *Statistical Approaches to Measurement Invariance*. Routledge, New York, 2011.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. Large language models are students at various levels: Zero-shot question difficulty estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8157–8177, 2024.

Andreas Säuberli, Diego Frassinelli, and Barbara Plank. Do LLMs give psychometrically plausible responses in educational assessments? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pp. 266–278, 2025.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*, 2023.

Seok Hwan Song, Mohna Chakraborty, Qi Li, and Wallapak Tavanapong. Is large language model performance on reasoning tasks impacted by different ways questions are asked? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22066–22081. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1138.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. LiveBench: A challenging, contamination-free LLM benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pp. 470–482, 2024.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024a.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on GAOKAO benchmark. *arXiv preprint arXiv:2305.12474*, 2023.

Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. Multiple-choice questions are efficient and robust LLM evaluators. *arXiv preprint arXiv:2405.11966*, 2024b.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2299–2314. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.149.

# A Corpus Statistics and Item-Level Results

Table 4: Corpus statistics of the released dataset.

| | |
|---|---|
| Years | 1998–2025 |
| Grade groups | 3–4, 5–6, 7–8, 9–10, 11–13 |
| Exams | 140 |
| Evaluated items | 3,886 (1,296 / 1,296 / 1,294 three-, four-, five-point) |
| Multimodal items | 1,746 (44.93%) |
| with a question diagram | 1,627 (1,353 with textual options) |
| with image-based options | 393 (10.11%) |
| Text-only items | 2,140 (55.07%) |
| Human reference coverage | 2001–2019 and 2022–2025 (23 years) |
| Exams with official human results | 115 |

Table 5: Overall performance on the 3,886 items and all 140 exams. Contest scores include the starting points and penalties for wrong answers. An exam is won by the model with the highest %max, with exact ties resolved by the fixed model order. The last two columns are explicit abstention rates over all items and over the 1,746 multimodal items.

| Model | Accuracy | Contest score | Exams won | Abstentions (all) | Abstentions (multimodal) |
|---|---|---|---|---|---|
| GPT-5 | 88.34% | 17,299.5 | 85 | 3.01% | 6.13% |
| Qwen3-VL | 86.31% | 16,803.75 | 39 | 0.00% | 0.00% |
| Grok 4 Fast | 78.33% | 15,762.5 | 9 | 11.84% | 24.80% |
| Claude Sonnet 4.5 | 77.35% | 14,958.25 | 7 | 0.75% | 0.97% |

Qwen3-VL never abstains explicitly but produced four non-answer outputs, which are scored as wrong answers. Figure 4 splits accuracy by where the visual content sits. The gap in Table 1 persists for every model within each point tier and within each grade group.

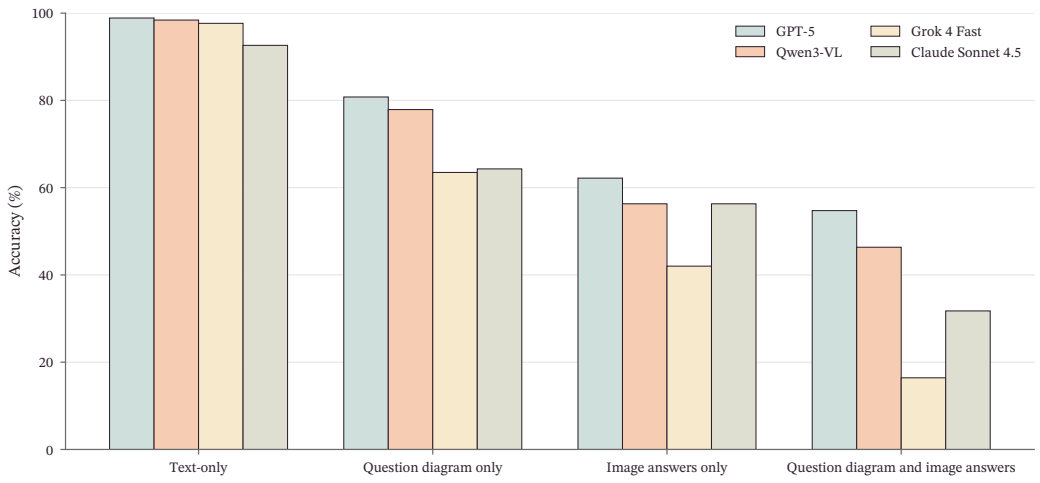


Figure 4: Accuracy by where the visual content sits: 2,140 text-only items, 1,353 with a question diagram only, 119 with image-based options only, and 274 with both. Image-based answer choices come with the largest drop, especially when the question also contains a diagram.

## B Human Reference Series

Table 6 summarizes the official human results by grade group, and Table 7 lists the human mean score for each of the 115 exams. Every model scores above the human mean on almost every exam: GPT-5, Qwen3-VL, and Claude Sonnet 4.5 on all 115, Grok 4 Fast on 113. The average margin ranges from 35.2 points of %max for Claude Sonnet 4.5 to 47.2 for GPT-5.

**Estimating the mean from histograms.** The 2001–2012 summaries report binned score histograms without a mean, so for the 60 exams of those years we use the frequency-weighted midpoint of the bins. On the 55 exams that report a mean, the midpoint estimate differs from it by 0.13 points of %max on average and by at most 0.94. Using the midpoint estimate for every exam changes the pooled MAE of Table 3 from 4.159 to 4.187 for the grade mean, from 3.827 to 3.837 for the visual share, and from 4.115 to 4.138 for the ensemble calibration.

Table 6: Official human mean score (%max) by grade group over the 23 reference years, and participants per exam.

| Grade group | Forms | Mean %max | SD | Range | Median participants | Range |
|---|---|---|---|---|---|---|
| 3–4 | 23 | 44.5 | 4.5 | 35.9–53.2 | 226,847 | 17,690–351,356 |
| 5–6 | 23 | 41.8 | 4.5 | 34.2–49.3 | 323,191 | 37,059–387,880 |
| 7–8 | 23 | 41.1 | 4.6 | 31.6–49.2 | 159,429 | 27,218–191,562 |
| 9–10 | 23 | 39.1 | 3.7 | 32.5–44.0 | 71,960 | 15,370–92,159 |
| 11–13 | 23 | 39.2 | 4.6 | 30.8–45.7 | 16,735 | 6,457–29,236 |

Table 7: Official human mean score (%max) for each of the 115 exams.

| Year | 3–4 | 5–6 | 7–8 | 9–10 | 11–13 |
|---|---|---|---|---|---|
| 2001 | 35.9 | 37.2 | 35.7 | 32.5 | 32.2 |
| 2002 | 47.8 | 41.2 | 36.5 | 35.6 | 34.2 |
| 2003 | 44.5 | 38.7 | 40.5 | 33.1 | 30.9 |
| 2004 | 36.6 | 35.0 | 31.6 | 33.7 | 30.8 |
| 2005 | 42.7 | 41.4 | 43.9 | 36.8 | 39.3 |
| 2006 | 46.2 | 40.8 | 36.9 | 36.4 | 37.6 |
| 2007 | 46.1 | 39.9 | 36.1 | 37.7 | 35.5 |
| 2008 | 49.7 | 45.5 | 40.4 | 37.9 | 36.1 |
| 2009 | 41.5 | 34.2 | 37.8 | 33.3 | 35.9 |
| 2010 | 46.8 | 46.0 | 46.2 | 42.9 | 38.5 |
| 2011 | 42.6 | 38.3 | 44.9 | 39.1 | 38.6 |
| 2012 | 42.9 | 48.1 | 46.8 | 41.5 | 44.8 |
| 2013 | 43.7 | 41.6 | 35.9 | 42.6 | 42.9 |
| 2014 | 48.6 | 38.7 | 42.6 | 37.3 | 41.9 |
| 2015 | 41.9 | 45.2 | 45.1 | 42.4 | 44.4 |
| 2016 | 46.5 | 47.6 | 45.1 | 42.0 | 42.3 |
| 2017 | 47.8 | 41.3 | 39.4 | 40.2 | 41.9 |
| 2018 | 46.7 | 49.3 | 41.3 | 41.1 | 45.4 |
| 2019 | 51.4 | 46.5 | 47.9 | 43.1 | 44.0 |
| 2022 | 38.6 | 40.9 | 38.3 | 42.0 | 45.7 |
| 2023 | 40.3 | 41.6 | 42.6 | 39.8 | 38.3 |
| 2024 | 41.0 | 35.1 | 40.1 | 44.0 | 37.9 |
| 2025 | 53.2 | 48.1 | 49.2 | 43.6 | 42.9 |

# C Full Regression, Sensitivity, Decomposition, and Prediction Tables

Table 8: Change in the human mean score per 10-point increase in model %max on the 115 exams, with year-cluster bootstrap intervals and the increase in $R^2$ from adding the model term. The last specification replaces the total visual share with the separate diagram and image-option shares.

| Series | Specification | Effect per 10 pp | 95% interval | $\Delta R^2$ |
|---|---|---|---|---|
| GPT-5 | Raw | $-2.62$ | $[-3.88, -1.33]$ | 0.184 |
| | Grade FE | $-1.85$ | $[-3.47, -0.17]$ | 0.058 |
| | + visual share | $-1.02$ | $[-2.50, 0.64]$ | 0.015 |
| | Diagram + option shares | $-0.70$ | $[-2.06, 0.58]$ | 0.007 |
| Qwen3-VL | Raw | $-2.22$ | $[-3.43, -0.94]$ | 0.161 |
| | Grade FE | $-1.57$ | $[-3.03, -0.23]$ | 0.057 |
| | + visual share | $-0.72$ | $[-2.08, 0.37]$ | 0.010 |
| | Diagram + option shares | $-0.20$ | $[-1.47, 0.84]$ | 0.001 |
| Grok 4 Fast | Raw | $-2.13$ | $[-3.57, -0.83]$ | 0.206 |
| | Grade FE | $-1.65$ | $[-3.75, 0.06]$ | 0.071 |
| | + visual share | $-0.41$ | $[-2.39, 1.16]$ | 0.003 |
| | Diagram + option shares | $0.38$ | $[-1.75, 1.86]$ | 0.002 |
| Claude Sonnet 4.5 | Raw | $-1.81$ | $[-2.83, -0.62]$ | 0.115 |
| | Grade FE | $-1.60$ | $[-2.69, -0.55]$ | 0.085 |
| | + visual share | $-0.89$ | $[-1.89, 0.14]$ | 0.021 |
| | Diagram + option shares | $-0.25$ | $[-1.27, 0.73]$ | 0.001 |
| Equal-weight ensemble | Raw | $-3.17$ | $[-4.42, -1.64]$ | 0.240 |
| | Grade FE | $-2.67$ | $[-4.49, -0.80]$ | 0.109 |
| | + visual share | $-1.48$ | $[-3.36, 0.26]$ | 0.022 |
| | Diagram + option shares | $-0.40$ | $[-2.40, 1.44]$ | 0.001 |

Table 9: Robustness of the grade-fixed-effects slope. The first check drops the two exams with a credited task. The second adds year fixed effects, so only differences between grade groups within a year identify the slope. The third refits the model leaving out each of the 23 years in turn and reports the mean and range of the estimates in place of an interval.

| Series | Check | Exams | Effect per 10 pp | 95% interval or range |
|---|---|---|---|---|
| GPT-5 | Excluding the two credited exams | 113 | $-1.78$ | $[-3.56, -0.08]$ |
| | Grade + year FE | 115 | $-0.25$ | $[-1.53, 1.13]$ |
| | Leave one year out (mean, range) | 110 | $-1.85$ | $[-2.20, -1.23]$ |
| Qwen3-VL | Excluding the two credited exams | 113 | $-1.58$ | $[-3.14, -0.18]$ |
| | Grade + year FE | 115 | 0.09 | $[-1.21, 1.20]$ |
| | Leave one year out (mean, range) | 110 | $-1.57$ | $[-2.15, -1.24]$ |
| Grok 4 Fast | Excluding the two credited exams | 113 | $-1.60$ | $[-3.62, 0.16]$ |
| | Grade + year FE | 115 | $-0.41$ | $[-2.06, 0.87]$ |
| | Leave one year out (mean, range) | 110 | $-1.66$ | $[-2.78, -1.23]$ |
| Claude Sonnet 4.5 | Excluding the two credited exams | 113 | $-1.54$ | $[-2.63, -0.53]$ |
| | Grade + year FE | 115 | $-0.20$ | $[-1.28, 0.84]$ |
| | Leave one year out (mean, range) | 110 | $-1.60$ | $[-1.88, -1.32]$ |
| Equal-weight ensemble | Excluding the two credited exams | 113 | $-2.59$ | $[-4.45, -0.59]$ |
| | Grade + year FE | 115 | $-0.32$ | $[-1.99, 1.33]$ |
| | Leave one year out (mean, range) | 110 | $-2.67$ | $[-3.35, -2.19]$ |

Table 10: Raw Pearson correlations between the ensemble score and the human mean score on the 23 exams of each grade group, with year-cluster bootstrap intervals.

| Grade group | Pearson $r$ | 95% interval |
|---|---|---|
| 3–4 | $-0.399$ | $[-0.726, 0.070]$ |
| 5–6 | $-0.315$ | $[-0.672, 0.117]$ |
| 7–8 | $-0.576$ | $[-0.834, -0.068]$ |
| 9–10 | $-0.316$ | $[-0.632, 0.064]$ |
| 11–13 | $-0.321$ | $[-0.672, 0.058]$ |

Table 11: Variance decomposition of exam-level %max on the 115 exams. $\Delta R^2$ is the increase from adding the visual share to grade fixed effects, with a year-cluster bootstrap interval. SD is the within-grade standard deviation of %max across exams. The last column is the fitted change per 10-point increase in the share. The $R^2$ columns use grade fixed effects alone, grade effects plus the total visual share, and grade effects plus the diagram and image-option shares, which replace the total share.

| Outcome | SD | $R^2$: grade | + visual | diagram + option | Visual $\Delta R^2$ [95% CI] | Per 10 pp |
|---|---|---|---|---|---|---|
| Human mean | 4.30 | 0.177 | 0.305 | 0.432 | 0.128 $[0.009, 0.317]$ | $+1.35$ |
| GPT-5 | 6.17 | 0.366 | 0.442 | 0.448 | 0.076 $[0.023, 0.167]$ | $-1.70$ |
| Qwen3-VL | 7.21 | 0.294 | 0.414 | 0.449 | 0.120 $[0.056, 0.208]$ | $-2.37$ |
| Grok 4 Fast | 7.65 | 0.427 | 0.659 | 0.684 | 0.232 $[0.121, 0.318]$ | $-3.88$ |
| Claude Sonnet 4.5 | 8.65 | 0.053 | 0.253 | 0.322 | 0.199 $[0.104, 0.294]$ | $-3.16$ |
| Equal-weight ensemble | 5.87 | 0.359 | 0.585 | 0.629 | 0.226 $[0.149, 0.298]$ | $-2.78$ |

Table 12: Change in %max per 10-point increase in the exam's visual share under grade fixed effects, with year fixed effects added, and with a linear calendar trend added instead, with year-cluster bootstrap intervals. The share rises over the years, so the calendar controls absorb any common trend.

| Outcome | Grade FE | + year FE | + linear trend |
|---|---|---|---|
| Human mean | $+1.35\ [0.36, 2.53]$ | $+0.38\ [-0.65, 1.40]$ | $-0.19\ [-1.32, 0.92]$ |
| GPT-5 | $-1.70\ [-2.87, -0.95]$ | $-1.39\ [-2.77, 0.12]$ | $-1.42\ [-2.93, 0.03]$ |
| Qwen3-VL | $-2.37\ [-3.38, -1.64]$ | $-1.08\ [-2.70, 0.60]$ | $-1.08\ [-2.47, 0.17]$ |
| Grok 4 Fast | $-3.88\ [-4.91, -2.73]$ | $-3.97\ [-6.08, -1.85]$ | $-4.08\ [-5.77, -2.33]$ |
| Claude Sonnet 4.5 | $-3.16\ [-4.43, -2.26]$ | $-2.46\ [-5.09, 0.45]$ | $-1.57\ [-3.69, 0.66]$ |
| Equal-weight ensemble | $-2.78\ [-3.44, -2.32]$ | $-2.23\ [-3.47, -0.78]$ | $-2.04\ [-3.02, -0.89]$ |

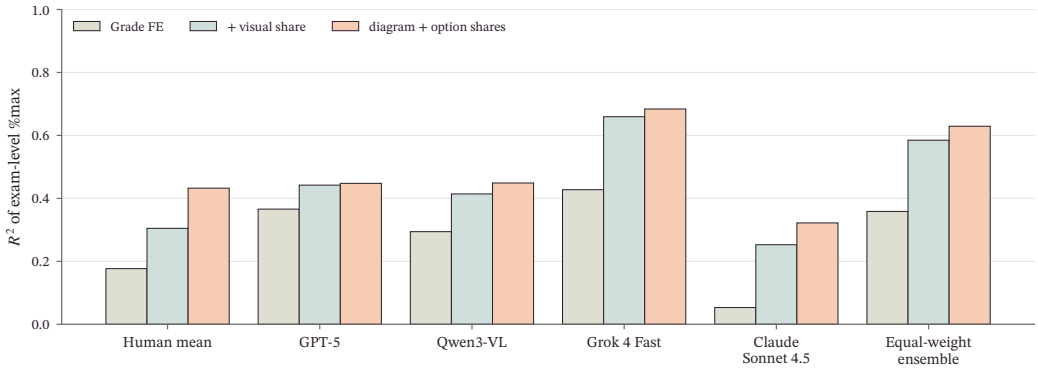


Figure 5: Share of the variance in %max across the 115 exams explained by grade fixed effects alone, grade effects plus the total visual share, and grade effects plus the diagram and image-option shares, which replace the total share, for the human mean score and each model series.

Table 13: Early/late check: out-of-sample $R^2$ of the calibration fit on one half of each grade series, split at 2012, and applied to the other half, for GPT-5 and the ensemble. Unlike the forward $R^2$ of Section 6.5, the denominator here is the squared deviation from the held-out half's own mean. In the early-to-late direction 9 of 10 values are negative.

| Grade group | Series | Early $\rightarrow$ late | Late $\rightarrow$ early |
|---|---|---|---|
| 3–4 | GPT-5 | $-0.067$ | $-0.073$ |
| | Ensemble | $0.003$ | $-0.316$ |
| 5–6 | GPT-5 | $-0.765$ | $-1.229$ |
| | Ensemble | $-0.381$ | $-1.471$ |
| 7–8 | GPT-5 | $-1.271$ | $-0.770$ |
| | Ensemble | $-0.449$ | $-0.572$ |
| 9–10 | GPT-5 | $-9.219$ | $-3.071$ |
| | Ensemble | $-8.593$ | $-3.164$ |
| 11–13 | GPT-5 | $-6.758$ | $-6.631$ |
| | Ensemble | $-8.551$ | $-7.842$ |

Table 14: Prediction of later years by grade group for the equal-weight ensemble, 15 held-out years each: MAE of the grade-mean baseline, of the visual-share predictor, and of the model-score calibration. Forward $R^2$ and wins compare the model-score calibration with the grade mean.

| Grade group | Baseline MAE | Visual-share MAE | Ensemble MAE | Forward $R^2$ | Wins (of 15) |
|---|---|---|---|---|---|
| 3–4 | 3.55 | 3.60 | 3.17 | 0.191 | 10 |
| 5–6 | 4.31 | 4.22 | 5.00 | −0.321 | 7 |
| 7–8 | 4.20 | 4.30 | 4.04 | 0.151 | 7 |
| 9–10 | 3.96 | 2.86 | 3.60 | 0.060 | 10 |
| 11–13 | 4.77 | 4.15 | 4.76 | −0.025 | 9 |

Table 15: Two alternative specifications on the 75 held-out exams of Table 3: a recent-history baseline that predicts each exam by the mean of the previous three reference years of its grade group, and calibrations that pool the five grade groups with grade intercepts and one shared model-score slope. MAE differences are relative to the grade-mean baseline and to the visual-share predictor, with year-cluster bootstrap intervals. Forward $R^2$ and wins use the grade-mean baseline.

| Predictor | MAE | ΔMAE vs. grade mean | ΔMAE vs. visual share | Forward $R^2$ | Wins (of 75) |
|---|---|---|---|---|---|
| Recent mean (previous three years) | 3.550 | $-0.608$ $[-1.36, 0.12]$ | $-0.277$ $[-0.83, 0.34]$ | 0.138 | 46 |
| GPT-5 score, shared slope | 4.065 | $-0.093$ $[-0.54, 0.29]$ | $+0.238$ $[-0.47, 0.87]$ | 0.022 | 33 |
| Qwen3-VL score, shared slope | 3.988 | $-0.170$ $[-0.47, 0.18]$ | $+0.161$ $[-0.48, 0.77]$ | 0.061 | 46 |
| Grok 4 Fast score, shared slope | 4.026 | $-0.133$ $[-0.74, 0.58]$ | $+0.199$ $[-0.26, 0.65]$ | 0.020 | 45 |
| Claude Sonnet 4.5 score, shared slope | 4.021 | $-0.137$ $[-0.40, 0.13]$ | $+0.194$ $[-0.52, 0.91]$ | 0.109 | 44 |
| Ensemble score, shared slope | 3.910 | $-0.248$ $[-0.75, 0.29]$ | $+0.083$ $[-0.44, 0.58]$ | 0.127 | 45 |

## D Ablation Protocols and Paired Results

Both comparisons use the evaluation protocol and decoding settings of Section 4 (temperature zero, top-$p = 1$, no token cap) and ran through the same provider on August 7–8, 2026. Grok 4 Fast had been retired by the provider and could not be re-evaluated.

**Language comparison.** The 200-item subset was selected from the dataset with proportional quotas for each year and each combination of grade group, point value, and visual-content type, so it keeps the composition of the dataset. Nine items whose visual content could not be translated were excluded beforehand. Translations were machine-generated and then reviewed item by item by the authors. One option and one statement were revised, and one item was replaced by another from the same sampling stratum and translated manually. Both arms show the statement and the options as text in the respective language, with the extracted diagrams and image-based options. Option order and gold letters are unchanged.

Table 16: Language comparison on 200 matched German–English pairs. The difference is English minus German in percentage points with a percentile bootstrap interval over items (2,000 draws). Discordant counts are items answered correctly in exactly one arm.

| Model | German | English | Diff. (pp) [95% CI] | DE-only / EN-only correct |
|---|---|---|---|---|
| GPT-5 | 90.00% | 89.50% | $-0.50\ [-3.5, 2.5]$ | 5 / 4 |
| Qwen3-VL | 85.00% | 84.50% | $-0.50\ [-5.0, 4.0]$ | 11 / 10 |
| Claude Sonnet 4.5 | 78.00% | 77.50% | $-0.50\ [-4.5, 3.5]$ | 9 / 8 |

**Visual comparison.** The no-image arm covers all 1,353 items with a question diagram and textual answer options. Prompts drop every image and keep the OCR question text and the five options. The with-image reference is the October 2025 evaluation of the same items.

## E Option-Position and Recency Diagnostics

**Option position.** Gold letters are close to uniform: A 17.91%, B 20.87%, C 21.15%, D 21.90%, E 18.17%. Among valid A–E selections, the share of C is 21.92% (GPT-5), 22.83% (Grok 4 Fast), 23.24% (Qwen3-VL), and 24.16% (Claude Sonnet 4.5). Accuracy conditioned on the gold position spans 86.9–89.7% for GPT-5, 82.6–89.1% for Qwen3-VL, 75.6–82.0% for Grok 4 Fast, and 74.6–82.1% for Claude Sonnet 4.5, in each case lowest when the gold option is A or E. These are descriptive marginals from the archived runs.

**Recent years.** On the 87 text-only items from the 2024 and 2025 contests, accuracy is 100% for GPT-5 and Grok 4 Fast, 98.85% for Qwen3-VL, and 93.10% for Claude Sonnet 4.5, against 98.83%, 97.56%, 98.39%, and 92.60% on the 2,053 such items from 1998–2023. This comparison cannot establish the absence of training exposure.

## F Credited Tasks and Grade Aggregation

**Credited tasks.** The 2001 grades 9–10 exam had 29 tasks, and the solution sheet credits every participant with five points for task 30. Task 21 of the 2003 grades 3–4 exam is marked flawed with no valid answer, and the 2003 results summary reports the full 105-point maximum. For both exams we add one starting point and five task points for the missing task, so the exam keeps its published maximum. The 2001 credit is explicit in the source. For 2003 we follow the same rule, and Table 9 shows that excluding both exams leaves the result unchanged.

**Grade aggregation.** Official summaries for 2007–2016 list grades 11, 12, and 13 separately. We weight the three by participant count.

## G Evaluation Prompt

The evaluation prompt was German for all main runs. The English arm of the language comparison used English instructions. The system message, with line breaks added for display, was:

```
Du bist ein hilfreicher Assistent für Multiple-Choice-Aufgaben.
Bewertung: Eine korrekte Antwort erhält die vollen Punkte der Aufgabe;
eine falsche oder nicht eindeutig auswertbare Endantwort zieht ein
Viertel der Aufgabenpunkte ab; 'Ich entscheide mich, nicht zu
antworten.' ergibt 0 Punkte ohne Abzug.
Wähle die beste Option (A-E). Wenn die richtige Option nicht mit
ausreichender Sicherheit bestimmbar ist oder die Aufgabe mehrdeutig ist,
kannst du ausdrücklich mit 'Ich entscheide mich, nicht zu antworten.'
antworten. Begründungen sind optional.
Füge am Ende eine einzelne Zeile an: Final answer: A|B|C|D|E|Declined.
```

Each item was then rendered from the benchmark fields:

```
Frage: <problem_statement>

Fragebild: <image, when present>
Zusatzbild 1: <image, when present>

Antwortmöglichkeiten:
A) <option text or attached option image>
B) <option text or attached option image>
C) <option text or attached option image>
D) <option text or attached option image>
E) <option text or attached option image>

Erinnere dich: Beende deine Antwort mit einer Zeile in diesem Format:
Final answer: A|B|C|D|E|Declined.
```

We parse the final line first. Narrow fallbacks apply only to truncated or malformed outputs. A missing or ambiguous final selection is scored as a wrong answer.

## H Example Items

Figures 6–8 show representative extracted items. The multimodal examples include the separately stored image-based answer options below the question region.

**9.** Sonja hat eine weiße, eine rote und eine grüne Dose. In einer hat sie ihre Kaugummis, in einer ihre Schokolade, die dritte Dose ist leer. Als ihr großer Bruder sie um einen Kaugummi bittet, sagt sie keck: „Du darfst dir einen nehmen. Sie sind entweder in der weißen oder in der roten Dose, und die Schokolade ist weder in der weißen noch in der grünen Dose.“ Welche Farbe hat die Kaugummidose?

(**A**) weiß (**B**) rot (**C**) grün (**D**) rot oder grün (**E**) unbestimmt

Figure 6: Extracted item from the 2009 exam, grades 5–6, question 9.

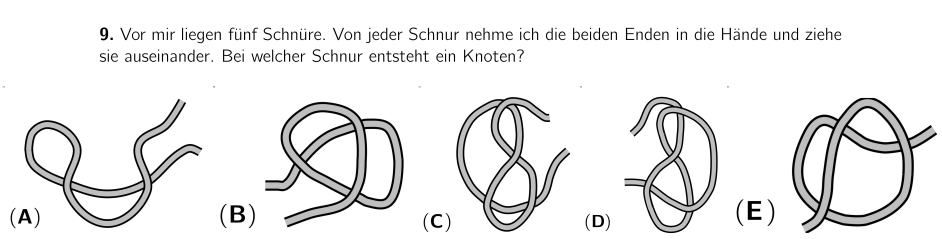


Figure 7: Extracted multimodal item for the 2010 exam, grades 5–6, question 9, including image-based answer options.

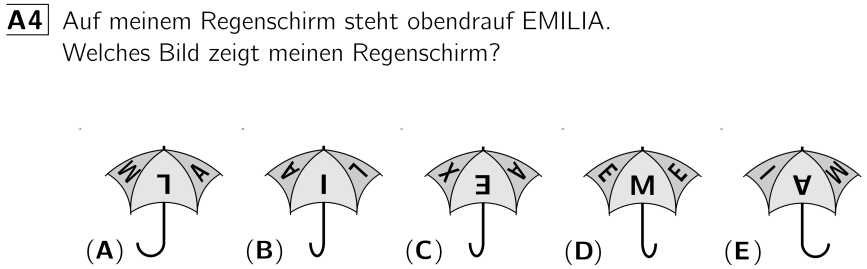


Figure 8: Extracted multimodal item for the 2015 exam, grades 3–4, question A4, including image-based answer options.