

Predicting Item-to-Range Performance Level Descriptor Matches with Structured Language Model Reasoning

Kang Xue, M. Christina Schneider, Sangdon Lim, Garron Gianopulos, and Alexandra Perez
Cambium Assessment, Inc.

Abstract

Range performance level descriptors (RPLDs) connect assessment items to claims about what students at different achievement levels know and can do. Retrospectively assigning RPLDs to a large item bank is valuable but labor intensive. This work-in-progress study evaluates whether small and large language models can assist expert item-to-RPLD matching for 524 Grade 3–5 English language arts items. We compared direct classification with structured prompts that guide a model to analyze the knowledge, skills, evidence, and cognitive demand required by an item. We also examined self-consistency voting and a rater-informed prompting. Preliminary results indicate that larger hosted models achieved the closest overall performance to humans, although some locally hosted models produced comparable results. Voting improved prediction reliability but did not consistently increase agreement with human scores, whereas training models with human scores generally improved classification accuracy. Model performance also tended to decline as grade level increased.

1 Background

Principled assessment design (Huff et al., 2025) connects claims about student proficiency to the evidence elicited by assessment tasks. Within this approach, Range performance level descriptors (RPLDs) (Egan et al., 2012) support item development and score interpretations by defining evidence characteristics that distinguish increasingly sophisticated levels of student understanding and cognitive processing. Difficulty drivers and other construct-relevant item features can therefore be used intentionally to engineer items for targeted performance definitions found in the RPLDs.

Content specialists must invest considerable time to document alignments between items and RPLDs, which can create human resource challenges. When a state adopts new standards and rewrites RPLDs,

existing items may need to be retroactively re-aligned to new RPLDs and standards. For computer adaptive testing programs, the resulting workload can involve thousands of items. Content specialists must analyze the item, identify the evidence a student demonstrates when responding correctly, compare that evidence with adjacent descriptors, and resolve ambiguous boundary cases to conduct the alignment process. Language models may reduce this burden, but an apparently plausible classification is insufficient: the classification must be accurate, reliable, and relatively consistent when compared to human judgment.

The present work evaluates language models as decision-support tools for item-to-RPLD matching. It asks four questions:

1. How closely do language-model performance-level classifications agree with resolved scores relative to two human raters in terms of exact and adjacent agreement?
2. Does expert-informed prompting improve agreement with resolved expert performance-level classifications relative to direct classification?
3. Does self-consistency voting improve classification accuracy or agreement with resolved expert performance level classifications?
4. Where do model-expert discrepancies occur, particularly among items with lower human consensus or disputed first-round ratings?

We used an evaluation framework that treats RPLD prediction as an ordinal, expert-referenced task rather than ordinary multiclass classification. In addition to exact agreement, we evaluated adjacent-level agreement, quadratic weighted kappa, directional error, coverage, disagreement by human-consensus status, and computational cost.

The current results are preliminary because the experiment matrix remains unbalanced across model families and prompting conditions.

2 Data

The dataset consisted of English Language Arts items from Grades 3–5. Each item record included the stimulus, when applicable, item stem, response options, answer key, relevant standard information, the full set of RPLD level definitions, two initial human RPLD ratings, and a final resolved rating. The final analytic sample comprised 524 items with sufficient item-content text and relevant standard information for LLM-based classification: 195 Grade 3 items, 125 Grade 4 items, and 204 Grade 5 items.

RPLD levels were ordered as Below, Approaching, At, and Above. Each item was initially rated by two content experts. Where agreement was not exact, ratings went through resolution for a final alignment.

3 Method

3.1 Models and Experimental Conditions

We evaluated instruction-tuned models from the Qwen2.5 (Yang et al., 2024), Gemma 2 (Gemma Team et al., 2024), Llama 3.1 and 3.2 (Grattafiori et al., 2024; Meta AI, 2024), and DeepSeek-R1-Distill (DeepSeek-AI et al., 2025) families, with parameter sizes ranging from approximately 1B to 32B. We also evaluated the hosted Claude Sonnet 5 model (Anthropic, 2026), accessed through Amazon Bedrock. All open-weight models were deployed on an NVIDIA L40S 48GB GPU. To accommodate the memory requirements of models of different scales, smaller models were executed using FP16 precision, whereas models with 14B parameters or more were executed using FP8 quantization. For each run, the experimental pipeline recorded the model identifier, prompt condition, sampling configuration, quantization setting, random seed, inference time, and token usage when available.

The voting condition was included to improve prediction consistency and reduce reliance on any single language model response. By aggregating ten independently generated performance level classifications, majority voting can mitigate run-to-run variability and produce a more stable final prediction. This condition was applied to non-DeepSeek model configurations for which repeated inferences were computationally feasible.

DeepSeek-R1-Distill models were evaluated only with single-run predictions because their reasoning-oriented generation produced substantially longer inference times than comparably sized instruction-tuned models, making ten-vote self-consistency too computationally time-consuming for the present experiment. The majority voting approach is particularly relevant for models whose temperature (i.e., settings that control the degree of randomness an LLM uses when generating a response) cannot be directly controlled; the repeated inference and aggregation provide an alternative mechanism for assessing and improving output stability (Xue and Appleton, 2026).

We evaluated two prompting conditions. In the direct condition, the model assigned one of four performance levels based on the item and the corresponding standard alignment metadata. In the prior-knowledge condition, the model received content specialist-developed guidance for interpreting item demands and differentiating adjacent RPLD levels. This guidance directed the model to prioritize evidence from the current item and its descriptors, evaluate cognitive demand, information scope, and option plausibility, and examine the boundaries surrounding the best-fitting level. Worked examples were provided solely for calibration rather than as answer keys. In both conditions, the model returned a single normalized RPLD label; missing or unparseable responses were treated as coverage failures.

For local models, deterministic single-run predictions were generated with a temperature of 0. In the voting condition, ten sampled predictions were aggregated by majority vote, with temperature set to 0.3 for models smaller than 8B parameters and 0.6 for models with 8B parameters or more. These settings were selected to introduce sufficient response diversity for self-consistency voting while limiting excessive variation, particularly among smaller models. Because temperature was not configurable for the Amazon Bedrock implementation used in this study, Claude Sonnet 5 was evaluated under the provider-defined low- and high-effort settings. These settings control the amount of computational effort allocated to reasoning: the low-effort setting favors faster, less resource-intensive inferences, whereas the high-effort setting permits more extensive reasoning at greater computational cost.

Model family	Model variants	Parameter size	Deployment	Precision or quantization
Qwen 2.5 Instruct	1.5B, 3B, 7B, 14B, 32B	1.5–32B	L40S 48GB	FP16 or INT8
Gemma 2 Instruct	2B, 9B, 27B	2–27B	L40S 48GB	FP16 or BF16
Llama Instruct	Llama 3.2 1B and 3B; Llama 3.1 8B	1–8B	L40S 48GB	FP16 or BF16
DeepSeek R1 Distill	Qwen 1.5B, 7B, 14B, and 32B; Llama 8B	1.5–32B	L40S 48GB	FP16 or INT8
Claude Sonnet 5	Low- and high-effort settings	Not disclosed	Amazon Bedrock	

Table 1: Models included in the experiment.

3.2 Evaluation

Model performance was evaluated against the Round 2 resolved human classification using four primary metrics: exact agreement, exact plus adjacent agreement, Pearson correlation, and quadratic weighted kappa. Exact agreement represents the proportion of model predictions that matched the resolved performance level, whereas exact plus adjacent agreement represents the proportion that either matched the resolved level or differed by no more than one level. Pearson correlations were used to assess the strength of the linear association between model-assigned and resolved performance levels. The same metrics were calculated for first-round human ratings to provide a descriptive benchmark for contextualizing model performance.

4 Results

4.1 Overall Performance Relative to Human Benchmarks

We used the Round 2 resolved ratings as the human benchmark. Across Grades 3–5, individual human raters remained substantially closer to the resolved labels than any model configuration. However, the Round 2 resolved level was established through human review and discussion of the initial Round 1 ratings. Because the resolved level was informed by these initial human judgments, it is not independent of them, which partly explains the high accuracy and correlations observed for the human raters. Table 2 reports all available human-rater comparisons and the best-performing configuration from each model family, organized within each grade. Human exact agreement, adjacent agreement, Pearson correlations, and quadratic weighted kappa values remained higher than all model-family values.

Model performance was meaningfully lower than the human benchmarks, although several configurations produced useful levels of ordinal agreement. The strongest model result in Table 2 was obtained by Claude Sonnet 5. Under the prior-knowledge voting condition, Claude Sonnet 5 reached exact agreement of .708 in Grade 3 and .574 in Grade 5; in Grade 4, its strongest listed

Grade	Group	Configuration	Valid n	Exact	Adjacent	Pearson r	QWK
G3	Human R1	Human rating	145	.876	.979	.914	.910
G3	Human R2	Human rating	100	.850	.960	.877	.869
G3	Human R3	Human rating	145	.917	.972	.900	.896
G3	Claude Sonnet 5	Prior knowledge, voting	192	.708	.917	.662	.661
G3	DeepSeek R1 Distill	Qwen-32B, prior knowledge	193	.637	.870	.547	.528
G3	Qwen 2.5 Instruct	14B, prior knowledge, voting	193	.622	.845	.439	.387
G3	Qwen 2.5 Instruct	27B, direct	195	.590	.856	.407	.309
G3	Llama Instruct	3.1-8B, voting	190	.568	.805	.341	.214
<i>Grade 4</i>							
G4	Human R1	Human rating	83	.940	.976	.906	.905
G4	Human R2	Human rating	116	.957	.983	.942	.940
G4	Human R3	Human rating	51	.804	.902	.704	.702
G4	Claude Sonnet 5	Prior knowledge, single	125	.736	.880	.557	.556
G4	DeepSeek R1 Distill	Qwen-32B, prior knowledge	125	.504	.856	.411	.406
G4	Qwen 2.5 Instruct	32B, prior knowledge	125	.560	.888	.467	.449
G4	Gemma 2 Instruct	27B, prior knowledge, voting	112	.366	.777	.276	.198
G4	Llama Instruct	3.2-3B, direct, voting	122	.377	.869	-.046	-.037
<i>Grade 5</i>							
G5	Human R1	Human rating	134	.896	.940	.884	.882
G5	Human R2	Human rating	138	.833	.935	.855	.843
G5	Human R3	Human rating	136	.934	.985	.919	.919
G5	Claude Sonnet 5	Prior knowledge, voting	202	.574	.906	.568	.561
G5	DeepSeek R1 Distill	Qwen-32B, prior knowledge	204	.505	.882	.542	.528
G5	Qwen 2.5 Instruct	32B, prior knowledge, voting	203	.448	.911	.464	.424
G5	Gemma 2 Instruct	9B, prior knowledge, voting	203	.399	.833	.113	.098
G5	Llama Instruct	3.2-3B, direct, voting	196	.403	.929	.407	.345

Table 2: Human ratings and best-performing model-family results by grade. QWK denotes quadratic weighted kappa.

result was the prior-knowledge single-run condition, with exact agreement of .736. Adjacent agreement for the Claude configurations was .917, .880, and .906 across Grades 3–5, respectively, indicating that many model errors were within one RPLD level. Among local models evaluated across all three grades, the strongest prior-knowledge results were concentrated among larger models. DeepSeek-R1-Distill-Qwen-32B achieved exact accuracy of .637, .504, and .505 in Grades 3, 4, and 5, respectively, whereas Qwen2.5-32B-Instruct achieved .560 in Grade 4 under prior knowledge and .448 in Grade 5 under prior knowledge with voting. In Grade 3, the best listed Qwen2.5 result was the 14B prior-knowledge voting configuration, with exact accuracy of .622.

Within most model families, the strongest listed configuration tended to involve a larger parameter variant, including DeepSeek-R1-Distill-Qwen-32B and Qwen2.5-32B-Instruct in Grades 4 and 5. The Llama family showed a different pattern in the table: the best listed Llama result was Llama 3.1-8B with voting in Grade 3, whereas Llama 3.2-3B with direct voting (zero-shot voting) was the best listed Llama configuration in Grades 4 and 5. This pattern should be interpreted cautiously because the comparison confounds parameter size with model

generation, prompting condition, and post-training differences; the Grade 4 and Grade 5 advantage of Llama 3.2-3B may therefore reflect differences in calibration to this four-level ordinal classification task rather than a general benefit of smaller model size.

Performance also varied by grade. Grade 3 generally showed stronger model results than Grade 5 for the leading hosted and local configurations: Claude Sonnet 5 declined from .708 exact agreement in Grade 3 to .574 in Grade 5 under prior-knowledge voting, and DeepSeek-R1-Distill-Qwen-32B declined from .637 to .505 under prior knowledge. Qwen2.5 results showed the same general direction when comparing the strongest listed Grade 3 Qwen2.5 configuration (.622 for 14B prior-knowledge voting) with the Grade 5 Qwen2.5-32B prior-knowledge voting result (.448). This pattern suggests that agreement with resolved RPLD performance levels became more difficult in the upper-grade items set represented in the current table.

Adjacent agreement often presented a more favorable picture than exact agreement. Several configurations with modest exact agreement still achieved adjacent agreement at or above .85. However, Pearson correlations and QWK values remained below the human benchmarks and did not meet the criteria typically required for operational deployment in high-stakes automated scoring. For example, the Claude Sonnet 5 configurations produced QWK values of .661, .556, and .561 across Grades 3–5. Nevertheless, these models were developed using a relatively small set of expert-annotated training examples, and the grade-level samples functioned as validation-style evaluation sets. Given the low-stakes context, the observed exact agreement provides preliminary evidence that Claude Sonnet 5 could serve as one initial rater, provided that a human completes the other rating and disagreements are referred for resolution.

4.2 Effect of Self-Consistency Voting

Self-consistency voting did not consistently improve performance across grades, models, or prompting conditions as shown in Figure 1. In most paired comparisons, the differences between single-run and voting predictions were small, and improvements in exact accuracy were often accompanied by minimal gains, no change, or declines in adjacent accuracy, Pearson correlation, or quadratic weighted kappa. For example, under the prior-knowledge condition, Claude Sonnet 5 showed al-

most no change in Grade 3 exact agreement (.708 to .708) or QWK (.659 to .661), improved exact agreement slightly in Grade 5 (.562 to .574), but decreased in Grade 4 (.734 to .701). Similar mixed patterns appeared for local models: Qwen2.5-32B-Instruct showed small gains in Grade 3 and Grade 5 under prior knowledge, but a slight decline in Grade 4.

These results suggest that voting should not be interpreted as a general accuracy-improving strategy for this task. Instead, its main value is diagnostic: repeated predictions make it possible to evaluate whether a model converges on the same RPLD label across independent generations or whether its classifications are unstable near adjacent descriptor boundaries. When the underlying model already produces stable and relatively unbiased predictions, majority voting may preserve or slightly smooth performance; when the model has a systematic bias, however, voting may simply reinforce the same incorrect label. Therefore, in the present task, self-consistency voting appears more useful for assessing prediction stability and identifying uncertain cases for review than for substantially increasing agreement with resolved expert labels.

4.3 Effect of Prior-Knowledge Prompting

In contrast to self-consistency voting, the prior-knowledge prompt produced more consistent evidence of performance improvement. Across paired direct and prior-knowledge comparisons, adding expert-derived rating guidance generally increased exact agreements, adjacent agreements, Pearson correlations, and quadratic weighted kappas, particularly in Grades 4 and 5 and for larger local models. The largest gains occurred when the direct prompt appeared poorly calibrated to the RPLD scale: for example, Qwen2.5-7B-Instruct in Grade 4 improved from .281 to .469 in exact accuracy under the single-run condition, and Qwen2.5-14B-Instruct in Grade 5 improved from .216 to .343. Similar gains were observed for Claude Sonnet 5, which improved from .664 to .734 in Grade 4 and from .490 to .562 in Grade 5 under single-run comparisons.

Prior-knowledge benefits were not limited to exact agreement. In several cases, the prior-knowledge prompt produced larger gains in Pearson correlation than in exact agreement, suggesting that content specialist-derived guidance helped models better preserve the ordering of the four performance levels described by the RPLDs. For in-

stance, Qwen2.5-32B-Instruct in Grade 5 under voting increased QWK by .231, and Qwen2.5-32B-Instruct in Grade 4 under voting increased QWK by .241. These improvements indicate that the prompt was especially useful for aligning responses with adjacent descriptor boundaries, even when it did not eliminate all exact-agreement errors.

However, the prior-knowledge prompt was not uniformly beneficial. Some Grade 3 configurations declined after adding the expert-derived guidance, most notably Qwen2.5-7B-Instruct, which decreased from .559 to .292 in exact accuracy under the single-run condition, and Llama-3.1-8B-Instruct, which showed a substantial decline in Pearson correlation and QWK. One plausible explanation is that the calibration examples used in the prior-knowledge prompt were drawn from Grade 6 items. Although Grade 3 items were generally simpler, they may occupy a developmental boundary between foundational reading skills and higher-level comprehension or analysis, making Grade 6-based calibration less directly transferable. As a result, the same expert-derived guidance may have aligned better with Grade 4 and Grade 5 items while shifting some Grade 3 models away from the label distribution implied by the item and descriptors. Thus, prior-knowledge prompting appears more effective as a structured task representation than as a guaranteed improvement mechanism.

Overall, these findings support the value of incorporating domain-specific rating knowledge into the prompt, particularly for Grade 4 and Grade 5 items and for models with sufficient capacity to use the additional information. At the same time, the mixed Grade 3 results indicate that prior knowledge should be empirically validated for each model and grade rather than assumed to transfer uniformly across developmental levels. Future analyses should examine whether the negative Grade 3 cases reflect prompt sensitivity, label-distribution shifts, or a mismatch between Grade 6 calibration examples and the foundational comprehension boundary represented in lower-grade RPLD distinctions.

4.4 Consensus Disagreement and Efficiency

Both model and human performance differed substantially between consensus and disputed items. For Claude Sonnet 5 under the prior-knowledge single-run condition, exact agreement declined by 11.3–40.0 percentage points across grades, adjacent agreement declined by 3.6–15.5 percentage points,

and Pearson correlations decreased by .386–.715. Human raters also performed less accurately on disputed items, with mean exact agreement ranging from .448 to .463, adjacent agreement from .753 to .869, and Pearson correlations from .348 to .458. Although exact agreement decreased for both humans and the model, relatively high adjacent accuracy indicated that many disagreements remained within one RPLD level of the resolved classification.

These findings suggest that disputed items represent meaningful RPLD boundary cases rather than random disagreement. Future versions of the prior-knowledge prompt should therefore incorporate carefully selected examples of human disagreement and their subsequent resolution to better illustrate ambiguous boundaries and expert decision strategies. To prevent information leakage, these calibration examples should remain separate from the evaluation set.

5 Conclusion and Discussion

The preliminary findings provide qualified evidence that structured, domain-informed prompting can improve item-to-RPLD classification. Improvements were most evident when prompts incorporated expert-developed guidance for interpreting item demands and distinguishing adjacent RPLD levels. This finding suggests that model performance depends not only on model capacity but also on how closely the prompt represents the reasoning required by the assessment framework.

Nevertheless, all model configurations remained below human performance across grades. Claude Sonnet 5, accessed through Amazon Bedrock, achieved the strongest overall model performance, but its agreement with the resolved labels did not reach that of the individual human raters. Although adjacent accuracy was relatively high for several configurations, exact agreement and QWK remained lower than the human benchmarks. These results indicate that identifying a nearby RPLD level is substantially easier for the models than reproducing the specific classification selected through expert review.

From an operational perspective, the observed error patterns support using large language models—particularly Claude Sonnet 5—as initial-rater decision-support tools rather than autonomous RPLD coders. Claude Sonnet 5 achieved approximately 71% exact agreement with the final resolved

ratings, suggesting that models could generate preliminary tags, identify plausible classifications, or flag unstable cases for expert review, thereby concentrating human effort on items that are difficult to classify. In such a workflow, model confidence should not be inferred solely from the fluency of the generated rationale. Agreement across repeated predictions, consistency across prompting conditions, and empirically estimated error rates may provide more defensible indicators for prioritizing human review.

In this study, voting did not consistently improve classification accuracy but provided a useful mechanism for assessing prediction stability and identifying uncertain classifications. For hosted models accessed through online services, where sampling parameters may not be directly configurable, aggregating repeated predictions can reduce reliance on a single response and provide a practical measure of output stability. For locally deployed models, voting involves a trade-off between response diversity and consistency. Selecting an appropriate temperature is therefore important for generating sufficient variation to make voting informative without introducing excessive randomness.

Finally, locally deployed open-weight models may provide a practical compromise between predictive performance and operational security. Although the hosted model achieved the strongest overall results, some larger local models approached its performance, particularly when provided with task-specific prior knowledge. These models captured a meaningful degree of ordinal information while allowing secure assessment content to remain within locally controlled infrastructure. Local deployment may therefore be suitable for pre-tagging, triage, and reviewer prioritization when data security, cost, latency, or infrastructure control are important considerations. However, the observed performance gap indicates that both local and hosted models should remain subject to expert review before their classifications are used operationally.

6 Limitations and Ongoing Work

This is a work-in-progress analysis with several limitations. First, the experiment matrix is unbalanced: some structured-skill runs currently cover only one grade, and not every model appears in every prompt and voting condition. Second, the resolved label is not independent of the original

human ratings, so human-to-resolution agreement is a contextual benchmark rather than a direct ceiling. Third, coverage differs across reviewers and some model runs. Fourth, the current results are descriptive and do not yet include paired significance tests or bootstrap confidence intervals. Fifth, findings are limited to Grade 3-5 English language arts items and the RPLDs represented in this item bank. Finally, prompts include operational assessment content and expert-derived guidance; replication will require careful protection of secure material and documentation of prompt versions.

The current benchmark also does not directly quantify the gap between LLM predictions and the two initial human ratings from Round 1, because model performance was evaluated primarily against the Round 2 resolved ratings. Subsequent analyses will therefore compare each model directly with the initial human ratings, using a common item set, to provide a clearer assessment of how model agreement differs from first-round human agreement.

Ongoing work will complete the crossed model-by-prompt-by-voting design on a common item set; estimate paired differences with item-level resampling; analyze confusion matrices and directional bias by grade and RPLD level; characterize disputed items qualitatively; and evaluate selective prediction policies that defer uncertain cases to human reviewers. We will also examine whether lightweight adaptation offers gains beyond prompt engineering, provided that a leakage-resistant train-validation-test split can be constructed.

7 Conclusion

Language models show promise as supplemental tools for retrospective RPLD tagging, particularly when their reasoning is structured around the knowledge, skills, evidence, and boundary distinctions used by expert raters. In the current experiments, a rater-informed hosted model improved over direct prompting but remained well below human agreement. These results favor a human-in-the-loop workflow in which models support prioritization and second review, while experts retain responsibility for final classifications and ambiguous cases.

References

Anthropic. 2026. [Introducing claude sonnet 5](#).

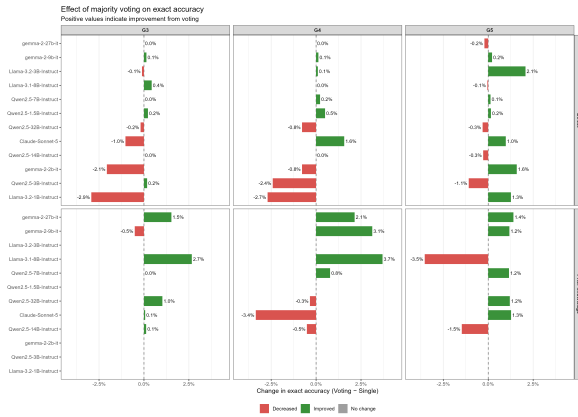


Figure 1: Changes in model performance from single-run prediction to self-consistency voting across grades and evaluation metrics.

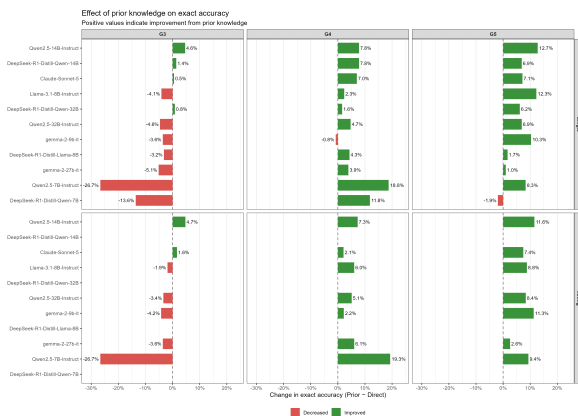


Figure 2: Changes in model performance from direct prompting to prior-knowledge prompting across grades and evaluation metrics.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.

Kang Xue and James J. Appleton. 2026. Evaluating general-purpose multimodal AI for Q-matrix generation from math items: A cognitive diagnostic modeling exploration. *Journal of Educational Measurement*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. 2025. [DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.

Katherine L. Egan, M. Christina Schneider, and Steve Ferrara. 2012. Performance level descriptors. In Gregory J. Cizek, editor, *Setting Performance Standards: Foundations, Methods, and Innovations*, 2 edition, pages 79–106. Routledge.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, et al. 2024. [Gemma 2: Improving open language models at a practical size.](#) *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.

Kristen Huff, Paul Nichols, and M. Christina Schneider. 2025. [Designing and developing educational assessments](#). In Linda L. Cook and Mary J. Pitoniak, editors, *Educational Measurement*, 5 edition, pages 441–511. Oxford University Press.