

LLM Difficulty Prediction Across Three L2 Task Types and Two Examination Forms

Meng Lyu
Boston College

Moses Oluoke Omojekunola
HSE University

Feng Wang
University of Massachusetts Amherst

Abstract

Large language models (LLMs) have been proposed as a substitute for field trials in item-difficulty estimation, but evidence about where they succeed is usually drawn from a single test form. We compare LLM difficulty predictions across three task types that make different language demands, grammar cloze, semantic cloze and reading comprehension, on two operational Grade 9 English examinations sat by the same cohort in China (1,199 and 1,166 examinees). Three LLMs estimated difficulty for 60 MCQ items under the same standardized protocol. On Form 1, grammar cloze had the highest Pearson correlation for all three models; the Fisher-averaged correlations were .81 for grammar cloze and .34 for semantic cloze. On Form 2, reading had the highest correlation for all three models (Fisher-averaged .78), and the grammar–semantic difference was smaller and inconsistent in direction. With ten items per task type, a task-type advantage seen on one form needs testing on other forms before it is read as a property of the model.

1 Introduction

Knowing how difficult an item is before administration supports test assembly, item-bank triage and adaptive testing, yet empirical difficulty is normally recovered only after costly field trials (Benedetto et al., 2023). Traditional readability formulas and NLP indices predict difficulty weakly, especially for second language (L2) items, where difficulty depends on construct-specific demands that surface features miss (Benedetto et al., 2023; Hwang et al., 2025).

Large language models offer an alternative, and several studies report moderate correlations between LLM difficulty estimates and empirical difficulty across domains and prompting strategies (Caines et al., 2023; Raina and Gales, 2024; Razavi and Powers, 2025; Kolesnikova et al.,

2026). Beyond overall prediction accuracy, a natural next question concerns how model performance varies across different types of items.

We compare LLM difficulty predictions across grammar cloze, semantic cloze and reading comprehension, and examine whether task-type differences persist across two examination forms. The three task types are chosen because they make different language demands. A grammar cloze blank is resolved by applying a morphosyntactic rule; a semantic cloze blank requires recovering meaning from the surrounding context; a reading item requires integrating information across sentences and, often, inference beyond the literal text. The two cloze sections carry an additional advantage for comparison: they share a four-option format and appear in the same paper, so a difference between them is not a difference of response format.

Evidence on this question is usually produced from one test form, a few dozen items, one prompt and one run per item, with sub-group claims read off subsets of ten or twenty items. We therefore ask the same question on two forms of the same testing program, with three models and three runs per item, under one documented protocol.

This paper contributes: (i) a two-form, three-model, three-run comparison of LLM difficulty validity across three L2 task types on an operational Chinese Grade 9 English examination; (ii) a psychometric audit of a four-administration item bank documenting when real difficulty can serve as a criterion; and (iii) a documented protocol, prompt set and per-item prediction record.

2 Aims

This study examines whether LLM predictions of item difficulty vary across grammar cloze, semantic cloze, and reading comprehension, and whether these patterns are consistent across two examination forms from the same testing program. We

consider both relative validity, defined as how well predictions track differences in empirical difficulty, and absolute accuracy, defined as how closely predicted proportions correct match observed values. The study addresses two research questions:

RQ1. How well do three LLMs predict empirical difficulty in each of the three task types?

RQ2. Do the differences between task types observed on one test form also appear on a second form from the same testing program?

3 Data

3.1 Two examination forms

The primary data are two operational English examinations sat by a single Grade 9 cohort (approximately CEFR B1) at a public school in a western city in China. Form 1 is the mid-term; Form 2 is a subsequent mock examination. From each form we analyze 30 four-option items: grammar cloze, semantic cloze and reading comprehension. Form 1 was taken by 1,199 students, of whom 1,168 answered every analyzed item; item difficulties are computed on all valid responses to each item ($n = 1,192$ to $1,199$). For Form 2, after removing rows with no responses to the analyzed items and one class-average summary row, 1,166 examinees answered every analyzed item, and every item difficulty is based on those 1,166 responses.

Each form supplies one passage for the grammar cloze section, one for the semantic cloze section, and three for the reading section. Task type and passage are therefore confounded within a form. Using a second form gives each task type a second set of passages, which reduces the chance that a pattern depends on one particular text, but does not separate task type from passage (§7).

The two forms were written to different difficulty targets. Form 1 is a routine mid-term, set at the difficulty of everyday classroom testing; Forms 2 to 4 are mock papers set at the difficulty of the high-school entrance examination itself. This follows the regional norm in the western city in China, where teaching materials and routine assessment are deliberately pitched above the entrance examination, so that candidates meet an easier paper on the day.

3.2 Difficulty range across the four administrations

The same cohort sat four curriculum-sequential forms. Table 1 summarises all four and Figure 1 shows the two that enter the analyses, because the difficulty range of a form determines whether it can serve as a criterion at all. After excluding rows with no responses to the analysed items, and one class-average summary row present in each of the Form 2–4 exports, the numbers of examinees answering every analysed item are 1,168, 1,166, 1,162 and 1,151.

Form 1 spans the full difficulty range. Forms 2–4 are compressed against the ceiling, with 60–76% of items above $p = .90$ and 34–47% of examinees answering every analysed item correctly. Mean $SE(p)$ in those forms is .008, so item difficulty is estimated precisely over a narrow range. KR-20 for the analysed items remains high in all four forms (.87–.90), because the full-sample total-score variance is large, while the marginal reliability of a unidimensional 2PL model fitted to the same items by marginal maximum likelihood falls from .85 in Form 1 to .65–.68 in Forms 2–4. The two indices answer different questions and we report both.

We use Forms 1 and 2 for the LLM analyses and treat Forms 3 and 4 as documentation of the difficulty range. Forms 3 and 4 replace the two ten-item cloze sections with a single fifteen-item cloze, so they do not support a grammar–semantic comparison within the same paper.

4 Method

4.1 Criterion and scale

The criterion is the classical proportion correct p for each item in its own form. The primary scale is proportion correct; correlations, bias and absolute error are reported on it. As a secondary check we repeat the correlations on the logit scale, transforming predicted and empirical values in the same direction; predictions are bounded to $[-.01, .99]$ by the prompt and the largest empirical p is .989, so no boundary handling is required.

Because the two forms were written to different difficulty targets and share no items, we do not place their difficulty parameters on a common scale and we do not pool raw difficulty. All primary analyses are within form. A secondary pooled analysis standardises empirical and predicted proportion correct within each form-by-

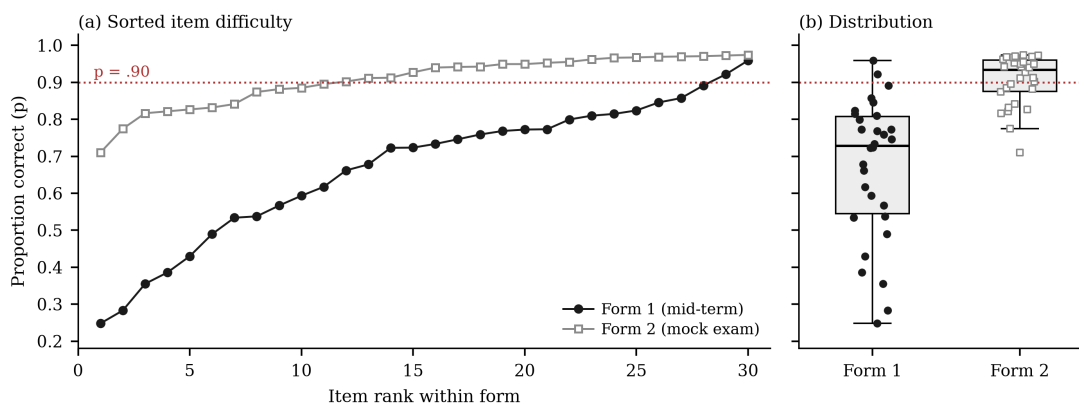


Figure 1: Empirical item difficulty across the two examination forms. Panel (a) plots proportion correct sorted within each form; panel (b) shows the corresponding distributions with every item overlaid. Only Forms 1 and 2 enter the LLM analyses.

Form	Analysed	Cloze	Reading	<i>n</i> complete	Mean <i>p</i> (SD)	Range of <i>p</i>	KR-20
Form 1 (mid-term)	30	20	10	1,168	.67 (.19)	.25–.96	.87
Form 2 (mock)	30	20	10	1,166	.91 (.07)	.71–.97	.90
Form 3	25	15	10	1,162	.89 (.08)	.65–.98	.89
Form 4	25	15	10	1,151	.92 (.07)	.72–.99	.87

Table 1: Item bank composition and empirical difficulty by administration. *p* = classical proportion correct. *n* complete = examinees answering every analysed item; item *p* is computed on all valid responses to each item, and KR-20 on complete responses. Forms 1 and 2 supply the LLM analyses; Forms 3 and 4 document the difficulty range only.

task-type cell before pooling (§5.2).

4.2 Prompting protocol

The protocol was written and fixed before any model was called. Each item is one independent call; no conversation state is carried between items. The prompt contains the full passage, the sentence containing the target blank or the reading stem, and the four options, together with a fixed context block naming the population (Grade 9 Chinese EFL learners, approximately CEFR B1), the task (estimate the expected proportion correct), and the required JSON response. All cloze blanks in the passage are shown open, as the students saw them. Chinese glosses printed in the original papers are removed from the model input, so that the two forms are presented on the same terms.

The same template is used for every item and every model, so the only thing that varies between calls is the item. The model is not told which option is correct, so the estimate is not conditioned on the answer. The prompt carries no expert meta-data such as the grammar point, its position in the national curriculum, or the position of the blank in the passage: such features exist for some task

types and not others, and supplying them would make the task types differ in the information given rather than in the item itself. The prompt also does not name the task type, since that label names one of the comparisons under test.

Three models were used: Claude Opus 4.8, GPT-5.5 and Gemini 3.1 Pro. Each item was run three times per model at default sampling settings; the reported prediction is the mean of the three runs and we report the within-item standard deviation across runs. Raw responses, resolved version strings and parameters are archived for all 540 calls.

Sixty items were run. Form 2 item 50 asks for the order of an experiment whose four options are sequences of three pictures labelled a, b and c; the pictures were not supplied to the models, which received only the letter sequences, so the item is excluded from all analyses. The main analyses therefore use 59 items.

5 Results

5.1 Validity by task type and form

Table 2 and Figure 2 give the main result. On Form 1, all three models produced their highest

Pearson correlation for grammar cloze ($r = .72, .83, .85$), and for every model grammar cloze exceeded semantic cloze. The models did not agree beyond that: two ranked reading above semantic cloze and one ranked semantic cloze above reading. Averaged across models by Fisher z , the Form 1 values are grammar .81, reading .46 and semantic .34.

On Form 2, all three models produced their highest Pearson correlation for reading ($r = .80, .83, .71$). The grammar–semantic difference was smaller than on Form 1 and not consistent in direction: two models placed grammar above semantic cloze and one placed semantic cloze above grammar. The Fisher means are reading .78, grammar .37 and semantic .23.

Part of the change in the cloze sections is consistent with restriction of range: the standard deviation of empirical p falls between the two forms from .257 to .064 for grammar cloze and from .170 to .044 for semantic cloze. For reading the standard deviation also falls (from .142 to .089) while the coefficients rise, so range alone does not account for the pattern. Every confidence interval in Table 2 is wide, and most of them overlap.

5.2 The grammar–semantic contrast across forms

Table 3 expresses the grammar-versus-semantic comparison as a Fisher z difference within each form and model. On Form 1 the difference is positive for all three models ($\Delta z = 0.89, 0.53, 0.86$). On Form 2 it is smaller and, for one model, negative ($0.35, -0.08, 0.19$). Every interval in the table includes zero.

The second form therefore gives limited support to the Form 1 pattern. We do not treat the three models as independent samples: they predict the same items, so their errors are shared. Two sets of passages per task type cannot support a passage-level inference.

Pooling the two forms, after standardising empirical and predicted p proportion correct within each form-by-task-type cell, gives 20 items per cloze task type and 19 for reading. The pooled correlations are .56 to .62 for grammar cloze, .47 to .74 for reading and .05 to .47 for semantic cloze across the three models: the separation between grammar cloze and reading that is visible within each form does not survive pooling. We report this as a secondary summary, since two sets of passages per task type do not constitute passage-level

replication.

5.3 Calibration

Relative validity and absolute calibration are separate questions, and both matter for triage. Table 4 reports calibration by form and model, and Figure 3 plots every analysed item.

On Form 1, where empirical difficulty varies widely, all three models compress the range: predicted difficulty carries 40 to 53 per cent of the dispersion of the empirical values, and two of the three place items too easy on average. On Form 2, where empirical difficulty is itself narrow, dispersion and bias vary by model.

The last column of Table 4 gives the error of a constant prediction at the form mean. Only one model is more accurate than that baseline on each form. Because the baseline uses labels from the same sample it is a descriptive reference rather than a predictor anyone could deploy, but it shows that a correlation of .7 or .8 with empirical difficulty need not yield usable absolute accuracy.

6 Discussion

Across the two forms, the task type for which LLM difficulty estimates correlated most strongly with empirical difficulty was not the same. On the first form all three models produced their highest correlation for grammar cloze, the section whose difficulty is governed by a morphosyntactic rule visible in the item text, and each of them produced a lower correlation for the meaning-dependent cloze section. That pattern is compatible with an account in terms of learner interlanguage: meaning-based difficulty may depend on which lexical and inferential associations a particular population has consolidated, which the item text does not carry. On the second form all three models produced their highest correlation for reading, and the grammar–semantic difference was smaller and inconsistent in direction.

We are not in a position to say which picture is the more representative. Each rests on ten items per task type, and the intervals in Table 2 are wide and mostly overlapping. The reading we would offer is narrow: a task-type advantage observed on a single form is a description of that form, and needs testing on further forms before it is used to decide where automated difficulty triage can be trusted.

Reading deserves a separate note. On Form 2 all three models produced their highest correlation

Form	Model	Grammar cloze	Semantic cloze	Reading	All items	Mean run <i>SD</i>
Form 1	Claude Opus 4.8	.72* [.16, .93]	.02 [−.62, .64]	.32 [−.39, .79]	.39	.017
	GPT-5.5	.83** [.42, .96]	.58 [−.07, .89]	.15 [−.53, .71]	.44	.022
	Gemini 3.1 Pro	.85** [.46, .96]	.37 [−.34, .81]	.77** [.26, .94]	.64	.022
	Mean (Fisher <i>z</i>)	.81	.34	.46	.50	
Form 2	Claude Opus 4.8	.41 [−.30, .82]	.08 [−.58, .68]	.80* [.28, .96]	.44	.014
	GPT-5.5	.29 [−.41, .78]	.36 [−.35, .81]	.83** [.36, .96]	.52	.017
	Gemini 3.1 Pro	.40 [−.31, .82]	.23 [−.46, .75]	.71* [.08, .93]	.54	.020
	Mean (Fisher <i>z</i>)	.37	.23	.78	.50	

Table 2: LLM difficulty validity by task type and form. Pearson correlations between predicted and empirical proportion correct, 95% CIs in brackets. $k = 10$ per cell, except Form 2 reading ($k = 9$, picture item excluded). * $p < .05$, ** $p < .01$. Run *SD* is the mean within-item standard deviation across three runs.

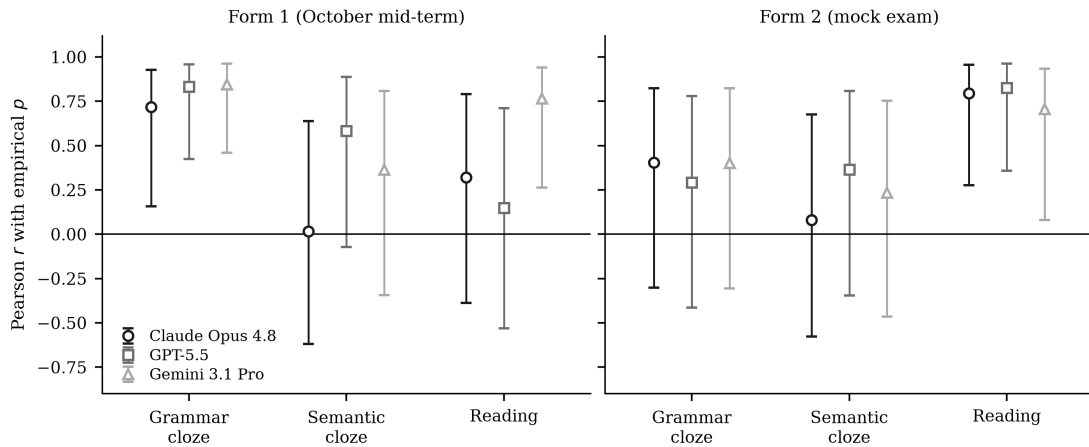


Figure 2: Validity by task type and form with 95% confidence intervals. On Form 1 every model is highest on grammar cloze; on Form 2 every model is highest on reading.

Form	Model	Δz	95% CI
Form 1	Claude Opus 4.8	0.89	[−0.16, 1.93]
	GPT-5.5	0.53	[−0.52, 1.57]
	Gemini 3.1 Pro	0.86	[−0.19, 1.90]
Form 2	Claude Opus 4.8	0.35	[−0.70, 1.40]
	GPT-5.5	−0.08	[−1.13, 0.97]
	Gemini 3.1 Pro	0.19	[−0.86, 1.24]

Table 3: Grammar versus semantic cloze. Δz = Fisher-transformed correlation for grammar minus semantic, within form and model. Descriptive: the three models share items and are not independent samples.

for reading, on a section whose empirical difficulties are themselves tightly clustered, so an easy item set can still carry between-item variation that the models track. The Form 1 reading coefficients, however, ranged from .15 to .77 across the three models, so these data do not support a general claim that reading difficulty is easier to predict than cloze difficulty.

Three of the four forms in this bank are compressed against the ceiling. They still yield precise

item difficulties, and the reading result shows that a compressed form can carry usable signal, but they offer limited discrimination among higher-ability examinees, and a model whose predictions cluster in the moderately-easy range can show small absolute errors on them for reasons that have little to do with the quality of its judgements.

7 Limitations

Ten items per cell yield intervals wide enough that most contrasts in this paper are consistent with a wide range of true values. The three models predict the same items and are not independent replications. The data come from one cohort at a single school, so the results describe that testing program rather than L2 examinations in general. Form 2 item 50 was excluded because its options are pictures that were not supplied to the models.

8 Conclusion

On two forms of an operational Grade 9 English examination, three models produced difficulty es-

Form	Model	Bias	MAE	RMSE	SD ratio	Baseline
Form 1	Claude Opus 4.8	−.01	.145	.175	.53	.187
	GPT-5.5	+.15	.183	.226	.46	
	Gemini 3.1 Pro	+.17	.179	.229	.40	
Form 2	Claude Opus 4.8	−.16	.162	.198	1.91	.066
	GPT-5.5	−.03	.055	.073	1.01	
	Gemini 3.1 Pro	−.02	.047	.064	.89	

Table 4: Absolute calibration by form and model. Bias = mean predicted minus empirical *proportion correct*. SD ratio = standard deviation of predictions divided by standard deviation of empirical difficulty. Baseline = RMSE of predicting the form mean for every item, a within-sample descriptive reference. Bold marks the only model that beats the baseline on each form.

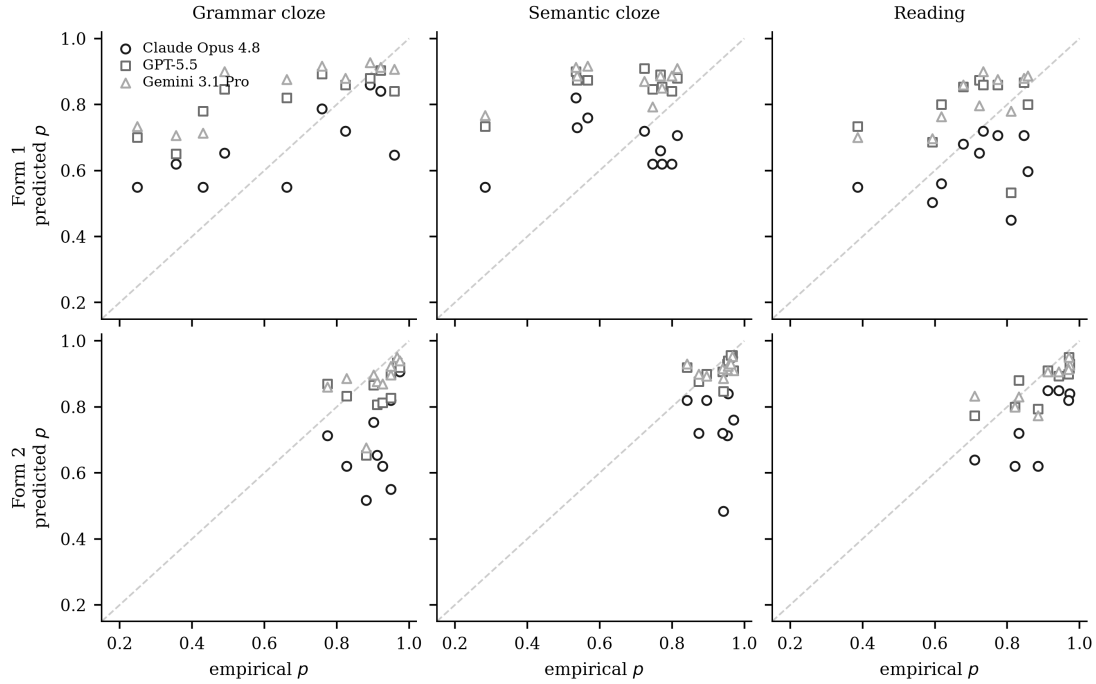


Figure 3: Predicted against empirical *proportion correct* for every analysed item, by form and task type. The dashed line is perfect calibration. On Form 1 the predictions span a narrower range than the empirical values; on Form 2 the empirical values are themselves narrow and the relation between the two dispersions varies by model.

estimates whose correlation with empirical difficulty varied by task type and by form: grammar cloze had the highest correlation for every model on the first form, reading for every model on the second. These data suggest that a task-type advantage observed on a small item set from one form needs checking on other forms before it is treated as a property of the model. The protocol, prompts and per-item prediction records are retained so that the estimates reported here can be checked and extended.

References

- Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. [A survey on recent approaches to question difficulty estimation from text](#). *ACM Computing Surveys*, 55(9):1–37.
- Andrew Caines, Luca Benedetto, Shiva Taslimipour, Christopher Davis, Yuan Gao, Øistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. [On the application of large language models for language teaching and assessment technology](#). *Preprint*, arXiv:2307.08393.
- Seonjeong Hwang, Hyounghun Kim, and Gary Geunbae Lee. 2025. [Can LLMs estimate cognitive complexity of reading comprehension items?](#) *Preprint*, arXiv:2510.25064.
- Daria Kolesnikova, Kirill Fedyanin, Abe D. Hofman, Matthieu J. S. Brinkhuis, and Maria Bolsinova. 2026.

Estimating item difficulty with large language models as experts. *Preprint*, arXiv:2605.18562.

Vatsal Raina and Mark Gales. 2024. Question difficulty ranking for multiple-choice reading comprehension. *Preprint*, arXiv:2404.10704.

Pooya Razavi and Sonya Powers. 2025. Estimating item difficulty using large language models and tree-based machine learning algorithms. *Preprint*, arXiv:2504.08804.