

Programmatic Tiered Prompting for LLM Generation of ELA & Math Items¹

David Whitcomb

Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas

Marjorie Wine

Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas

Alexander Hoffman

AleDev Research & Consulting

Abstract

This study uses the *Low-Bar Item Draft Acceptability Tool* (LBIDAT) to evaluate automatically generated 8th grade assessment items from three Large Language Models and six prompt engineering tiers. Using bias-reduced ordinal logistic regression, we tested main effects of prompt tier and model selection on defect severity across 8th grade English Language Arts (ELA) and Mathematics items; neither significantly affected defect severity in either domain. Results instead reveal stark cross-domain disparities: ELA items consistently exhibited poor quality and high defect rates across all models and prompt tiers, whereas mathematics items achieved high rubric acceptability by defaulting to familiar tasks that exist in broader training data. However, even math items generally fell short of desired cognitive targeting.

1 Introduction & Framework

Automated item generation (AIG) using LLMs holds immense potential for scaling educational assessment, yet automatically generated items often fail professional review for quality, alignment, or fairness, even when they initially appear properly constructed (Kurdi et al., 2020; Tan et al., 2025). Ensuring proper alignment to specific content standards at appropriate grade levels remains the core challenge of item development—whether using automated or conventional workflows. With LLM-based workflows, items frequently fail to meet these requirements because they lack accurate cognitive architectures that recognize the goal of each alignment reference and how those learning goals connect learning sequences across levels of

mastery and/or years with modern learning standards. Therefore, the items produced do not elicit evidence of targeted cognition for typical test takers (Wine & Hoffman, 2024; Lane & Marion, 2025).

The *Low-Bar Item Draft Acceptability Tool* (LBIDAT; Hoffman & Wine, 2025) evaluation framework provides a non-compensatory, rubric-based protocol aimed at identifying item drafts with potentially fatal design or execution issues. It was designed to evaluate item drafts/starters from item writers or automated workflows. It categorizes generated item quality by five primary dimensions, each scored on a 3-point scale. Excessive critical issues in any one dimension are disqualifying, regardless of evaluation in other dimensions.

To test LLM adaptability, prompts were programmatically structured into progressive tiers ranging from just merely standards' codes and text in Tier 2 up to (in Tier 7) comprehensive scaffolds incorporating both medium length affirmative and exclusionary instructions, and even the full LBIDAT rubric guidance itself. Outputs were generated via Python-based API pipelines across Claude, Gemini, and GPT.

2 Research Questions

This study addresses three primary research questions:

- RQ1: How, if at all, do increasingly detailed tiers of zero-shot prompt engineering significantly alter the odds of item defect severity across individual LBIDAT dimensions?
- RQ2: How, if at all, does item defect severity on individual LBIDAT dimensions significantly differ across Large Language Models?

¹ Paper presented at the National Council on Measurement in Education's second Artificial Intelligence in Measurement and Education Conference.

- RQ3: How, if at all, do item defect severity patterns and cognitive complexity limits appear outside the dimensions of the LBIDAT?

3 Methodology & Statistical Model

3.1 Item Generation Pipeline & Prompt Design

To systematically test the limits of LLM-based item generation (AIG), item production was structured using a Python-based pipeline to access three LLMs' (Anthropic claude-opus-4-1-20250805, Google gemini-3-pro-preview, OpenAI gpt-5.1) APIs directly. Prompts were executed using a zero-shot prompting framework across two fixed hyperparameter configurations (model temperature locked at 0.7 and 0.8) to balance structural adherence with some creative variability. The pipeline programmatically populated structured JSON payloads enforcing strict output schemas. This compelled models to return explicit JSON keys for the item stem, answer options, key rationale, distractor rationales, and (for mathematics) LaTeX formatting instructions and figure descriptions. Math outputs were then rendered to produce a single graphic file with stem, answer options, figures and distractors that could be used for item evaluation.

Prompts were engineered with persona framing into six progressive Prompt Tiers (Tiers 2-7) to evaluate whether increasing prompt specificity reduces item quality concerns. Tier 1—just a standard's code—was removed as effectively duplicative in an earlier phase of this project (Hoffman, et al. 2025).

- Tier 2: Included only the official Common Core State Standards (CCSS; NGA/CCSSO, 2010) code and verbatim standard text
- Tiers 3 & 4: Added short (50-100 words) or medium (100-200 words) standard-specific affirmative instructions written by Subject Matter Expert (SME) Content Development Professionals (CDPs) outlining expected item behavior.
- Tiers 5 & 6 (Exclusionary Constraints): added short (50-100 words) or medium (100-200 words) negative/exclusionary directions (what the item *must* not do)
- Tier 7: added common standard-specific student misconceptions to guide stem and distractor construction.

(Each tier included versions that did not (e.g., Tier 4a) and did (e.g., Tier 4b) include the LBIDAT rubric and full explanation. This had no impact on item quality, so these versions were combined at each tier.)

3.2 Sample & Target Standards Selection

This project purposefully targeted 8th grade standards across Mathematics and English Language Arts (ELA). This level was selected to represent moderately advanced content that balanced complexity/non-trivial content demand and potential clarity to non-SMEs. All selected standards were taken directly from the Common Core State Standards.

The selected mathematics standards cover statistics, algebra and geometry.

- 8.F.B.4: Constructing simple linear models and understanding rate of change and initial value.
- 8.G.C.9: Knowing and using the formulas for the volumes of cones, cylinders, and spheres.
- 8.SP.A.4: Analyzing simple bivariate categorical data, two-way tables and relative frequencies.

The selected ELA standards include both informational and literary reading.

- RI 8.2: Determining central ideas and analyzing their development over the course of an informational text.
- RL 8.1: Citing strong textual evidence to support explicit statements and inferences in literary text.
- RL 8.4: Determining word/phrase meanings, figurative/connotative tone, and literary analogies

ELA prompts included designated reading stimuli from publicly released 8th grade assessment passages from six states (MA, MN, NJ, TN, TX, UT).

In total, 284 math and 456 ELA items were generated, from which ~27% overall were randomly selected for evaluation (n=110 ELA, n=90 math).

3.3 Analysis

Randomly selected items were reviewed by 1-2 SME CDPs with the generating LLM, temperature, and prompt tier obscured. Items were evaluated using the LBIDAT, a non-compensatory dimensional

rubric measuring defect severity across five categories: Stem & Task, Answer Key, Distractors, Stimulus, and Fairness. In each content area, both raters independently scored a common set of items. Consistent with consensus-based qualitative rating traditions, discrepant ratings were treated as material for discussion rather than error to be quantified, and calibration concluded when raters reached consensus interpretations of each dimension. Remaining items were then divided between raters and scored independently.

Each dimension was scored on a 3-point ordinal scale:

- ✓ No critical issues identified; item draft likely usable or easily refinable.
- ~ Significant execution issues present; item draft potentially refinable.
- !! Severe alignment or validity flaws; item draft likely *unsalvageable*.

Dimension ratings (✓ > ~ > !!) were modeled with ordinal logistic regression—an adjacent-categories logit with a parallel-odds constraint—as a function of LLM engine, prompt tier, and standard.

$$\ln\left(\frac{P(Y = y_{m+1})}{P(Y = y_m)}\right) = \beta_{0m} + \beta_1(\text{LLM}_j) + \beta_2(\text{Tier}_k) + \beta_3(\text{Standard}_l)$$

Where:

- β_{0m} represents the intercept for each adjacent-category boundary (✓ vs ~; ~ vs !!).
- β_1 isolates the main effect of model selection (LLM_j: Claude, Gemini, GPT).

- β_2 isolates the main effect of zero-shot prompt engineering depth (Tier_k: Tiers 2 through 7).
- β_3 accounts for baseline standard difficulty (Standard_l).

Model parameters were estimated with bias-reducing adjusted score equations (Kosmidis & Firth, 2021), which yield finite estimates under the sparse-cell conditions that produce separation under maximum likelihood, permitting a single estimator across all dimensions and both content domains. Likelihood Ratio Chi-Square (χ^2) tests of each main effect (prompt tier, df=5; LLM, df=2) were conducted at the $\alpha=0.05$ significance threshold, adjusted for the study's 20 comparisons using Bonferroni correction. The parallel-odds constraint was assessed for each dimension against boundary-specific fits; it was rejected only for ELA Stem & Task ($p < .001$), and relaxing the constraint there left both main-effect tests non-significant.

4 Results

4.1 Research Question 1 (Prompt Tiers): Impact of Prompt Engineering Depth

Surprisingly, controlling for LLM engine and target standard, zero-shot prompt engineering depth (Prompt Tier) did not produce a statistically significant main effect on item defect severity across the vast majority of LBIDAT dimensions in either content domain (Table 1).

English Language Arts (ELA): Prompt tier selection yielded no statistically significant reduction in defect odds across any individual dimension.

Figure 1. Fatal Defect Rate (!!) by LBIDAT Dimension and Content Domain

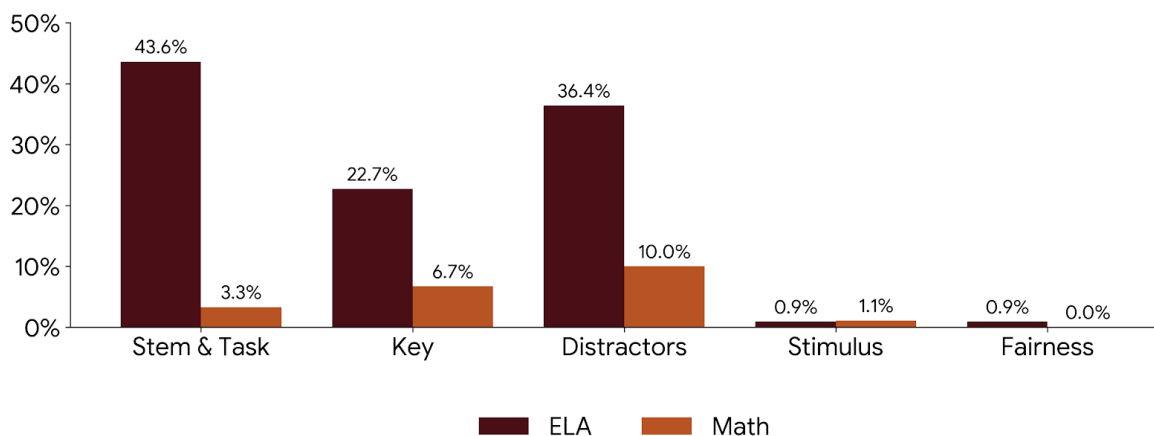


Table 1. Likelihood-Ratio Tests of Prompt Tier and LLM Main Effects on Defect Severity, by LBIDAT Dimension and Content Domain

Dimension	ELA		Mathematics	
	Tier $\chi^2(5)$	LLM $\chi^2(2)$	Tier $\chi^2(5)$	LLM $\chi^2(2)$
Stem & Task	4.21 (.52)	1.16 (.56)	1.28 (.94)	2.06 (.36)
Answer Key	3.02 (.70)	0.74 (.69)	3.53 (.62)	2.00 (.37)
Distractors	2.45 (.78)	1.72 (.42)	1.38 (.93)	11.15 (.0038)
Stimulus	0.00 (>.99)	0.69 (.71)	13.55 (.019)	0.00 (>.99)
Fairness	0.00 (>.99) ^a	0.01 (.99) ^a	— ^b	— ^b

Note. Values are likelihood-ratio χ^2 statistics with p-values in parentheses. No test was significant at the Bonferroni-adjusted threshold ($\alpha = .05/20 = .0025$).

^a No ELA item received a ~ Fairness rating; the Fairness model reduces to its ✓ vs. !! boundary.

^b No mathematics item received a non-✓ Fairness rating; no model was fit.

Increasing prompt detail from basic standard text (Tier 2) to full affirmative and negative guidance, plus common student errors (Tier 7) failed to lower the odds of severe item defects. (Stimulus defects were rare—3 of 110 items—as ELA stimuli were provided from released passages rather than generated content.)

Mathematics: Prompt tier showed no significant main effect on any testable dimension.

The nominal Stimulus result reflects eight defective items concentrated in specific tiers rather than any gradient with tier and does not approach the Bonferroni-adjusted threshold.

4.2 Research Question 2: Model Performance Differences

English Language Arts (ELA): Model choice (Claude, Gemini, GPT) exerted no statistically significant effect on defect severity for any LBIDAT dimension. Comparing engines against the baseline (Claude), neither Gemini nor GPT significantly altered the odds of receiving higher defect ratings.

Mathematics: Model selection produced one nominally significant result, which does not survive the Bonferroni-adjusted threshold ($\alpha = .0025$).

The nominal Distractors result reflected greater defect severity among Gemini-generated distractors relative to Claude; given the study's 20 comparisons, one nominally small p-value is consistent with chance, and we do not interpret it as a model effect.

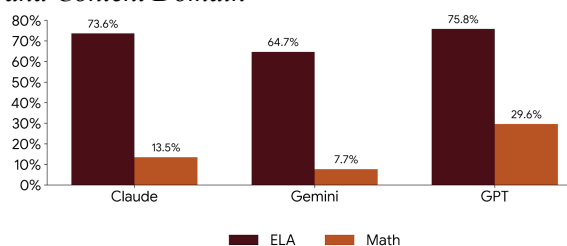
4.3 Research Question 3: Cross-Domain Divergence

Over 73% of ELA items received fatal defect (!!) ratings in at least one dimension, often driven by *anchor standard traps*—where models generated generic, off-grade reading comprehension questions aligned (at best) to broad anchor concepts such as broad, lower-grade reading comprehension, literal text recall, or basic vocabulary identification rather than grade-appropriate 8th-grade KSAs. Rather than operationalizing the nuanced cognitive demands of grade-specific standards, expanded prompts merely caused the models to append superficial, rubric-aligned language onto fundamentally off-grade items. Thus, these failed to meet even the low bar of the *Low-Bar Item Draft Acceptability Tool*.

For example, RI 8.2 items often required test takers to recognize the most accurate plot summary, rather than analyze the relationship of a central idea to supporting ideas. RL 8.4 items targeted words whose in-context meanings were not definitively determinable, therefore offering multiple plausible readings with a merely *arguable* key—with other items labeling alternative readings as the key!

A *high-score illusion* in mathematics is quite striking. Over 75% mathematics items across all tiers and models achieved acceptable ratings with the LBIDAT's *low bar*. Qualitative CDP review re-

Figure 2. Fatal Defect Rate (!!) by LLM Engine and Content Domain



vealed that this high acceptability rate masked severe cognitive complexity and proficiency deficiencies. Models consistently defaulted to basic algorithmic procedures and simple formula applications, completely avoiding the conceptual modeling and multi-step problem solving required by *many* 8th-grade standards. While mathematics items passed *surface* rubric checks, they would fail to elicit evidence of *targeted grade-level* cognition.

For example, bivariate data was always binary, despite prompts suggesting that more than two possible values were also appropriate. Items were so conceptually simple that distractors were often the product of elementary grade computation mistakes rather than grade-level misunderstanding or misapplications.

This evaluative blindness extended to the models' self-assessments. When prompted tiers included the LBIDAT rubric itself (the b-variants), models occasionally LBIDAT scores for their own drafts—and awarded top marks across every dimension, regardless of the item's actual ratings, including items human review deemed unsalvageable. The models applied the rubric's language fluently while detecting none of the defects it defines.

5 Discussion

The findings of this study demonstrate that zero-shot prompt engineering depth—spanning basic standard text (Tier 2) to hundreds of words of standard-specific expert item writing guidance (Tier 7)—failed to reduce the frequency of excessive critical issues in *any* of the LBIDAT's five dimensions in either English Language Arts (ELA) or Mathematics. Prompt expansion did not resolve core structural defects in either domain.

A primary driver of both the ELA *anchor standard traps* and overall prompt insensitivity is likely the fundamental limitation of the underlying training corpus. LLM prompt engineering necessarily can only leverage embedded training rather than incredibly specialized assessment design expertise. Because existing web corpora are heavily populated with general reading comprehension questions, basic text summaries, and elementary-level exercises, LLMs can not help but defer to the long history of surface-level reading items. But this was not limited to ELA items. Math items also tended to be considerably below true grade-level expectations for proficiency, just targeting the day 2 level of a 5-day unit. In Math, surface-level acceptability

often masked cognitive simplification. ELA's domain model sees such items as below grade level, and Mathematics unfortunately accepts them as just mediocre—for essentially the same issue.

Including a handful of examples with the LLM prompts is the obvious—and *ultimately quite unwise*—next step. This will only produce the kind of templated predictable items that are incredibly subject to inappropriate test prep (Koretz, 2008, 2017; Linn, 2000). Over-constraining LLMs through rigid, highly templated prompt structures—the direct descendant of item shells and item models (Haladyna & Shindoll, 1989; Gierl & Lai, 2012)—risks inducing repetitive item formats, predictable distractor patterns, and diminished item diversity. Templating can constrain a model's generative scope, producing mechanically sound items that lack creative context or authentic problem scenarios. This approach entirely undermines assumptions of sampling that are absolutely necessary to generalize from performance on individual items to statements about levels of proficiency regarding skills, abilities, understanding and judgments (Kane, 2013).

These corpus limitations point toward fine-tuning rather than deeper prompting—a path we take up in the recommendations below. (Consider the top marks the LLMs gave their own low-quality items in this study: even vetting a fine-tuning corpus cannot yet be automated.)

6 Limitations

Several limitations bound these findings. Though 740 items were generated, evaluation was conducted on random subsamples of 110 ELA and 90 mathematics items—roughly 15–18 items per prompt tier within each domain. Individual tier contrasts therefore carry wide confidence intervals, and no single test can definitively rule out a moderate effect on any one dimension; nor could the design detect tier effects confined to a single model, as tier-by-model cells were too small to test interactions. That said, the null reading does not rest on any single test. Across ten parallel omnibus tier tests spanning each domain—each estimated on the full domain sample with all three models pooled—nine of ten likelihood-ratio statistics fell below their expectation under the null, and the one exception (mathematics Stimulus) reflected eight defective items concentrated in specific tiers rather than any gradient, a pattern indicating essentially no dose-response relationship in the data rather

than an effect straining against limited power. Descriptive defect rates remained flat across tiers, where any practically meaningful improvement would be visible regardless of significance.

The design's scope was deliberately narrow: one grade level, one standards framework, six standards, and three commercial models at a single point in time. Successor models may behave differently, though we have argued the underlying corpus limitation is structural rather than a matter of model maturity. The study also tested only zero-shot prompting; few-shot approaches were excluded on the templating grounds discussed above. Finally, the two temperature settings (0.7 and 0.8) were pooled in analysis; the contrast is narrow and temperature was not modeled as a factor.

7 Conclusion and Recommendations

While Large Language Models can rapidly generate candidate assessment content that appears adequate on the surface, even quick (i.e., <5 minute) examination by content experts can make clear that the candidates are inappropriate. This study suggests some obvious recommendations.

7.1 Do Not Be Fooled or Distracted

First, LLMs naturally excel at producing credible-seeming...anything. The surface gloss and shallow credibility should not be mistaken for substantive quality.

Second, LLM-produced items are less problematic for mathematics than for ELA. No one should extrapolate their current abilities with simple procedural math items to other content areas. AIG researchers should evaluate their outputs across multiple content areas.

Third, LLMs' ability to produce cliché procedural items should not be mistaken for an ability to produce items that generate evidence of conceptual understanding, let alone the critical KSAs of the CCSS's *Standards for Mathematical Practice*. Few assessments focus on factual recall or trivial examples of procedural application. AIG researchers should be sure to evaluate their outputs across the breadth of cognitive demand from a domain model.

Fourth, there are any number of ways to evaluate the products of AIG processes. No one should be satisfied with high success rates on *shallow* criteria, even as explicit as the LBIDAT's five low bars.

Fifth, efficiency is not a form of validity evidence. The attractions of time or dollar savings

should not distract from validity-based evaluation. That is, validity does not care how much it costs.

7.2 Build Strong Teams

Educational measurement and assessment development remains a highly multi-disciplinary endeavor (Huff et al., 2025; Wine & Hoffman, 2024). While automated workflows may seek to shift many contributions to faster AI technologies with lower marginal costs, the development of such workflows clearly requires the wider breadth of contributions. That is, research and development teams must include AI/LLM specialists, psychometricians, test designers *and* content development professionals—at the very least. Even if test users are not included on these teams, they must include professionals with the expertise with both the content and the cognition that items aim to address—and how to elicit evidence of that intersection.

7.3 Consider the Context

The available training corpora are almost certainly dominated by low quality items and assessments. Even professionally developed examples are viewed poorly by educators and the public (Sireci, 2021). Relying on the most readily available corpus is simply not appropriate for highly specialized work, like content development for large scale assessment.

This makes it difficult to use proprietary frontier models for AIG. The low quality examples in their training corpora are simply too much to overcome through prompt engineering. Rather, some level of customization through fine tuning seems to be necessary, which requires open-weight models (or expensive supervised fine-tuning) and vetting for the highest quality items that truly elicit evidence of grade level proficiencies—something that cannot yet be done in an automated fashion.

To date, no one has collected truly high quality items that meet the highest standards to form the fine tuning corpus, for a wide variety of reasons. Test development organizations might comb their item banks for such items, but it would require more stringent examination than the LBIDAT approach. Falling short of this can only produce more of the same items, and we already know how products based upon them are received by the public.

Furthermore, the widespread use of inappropriate test preparation is a product of predictable items. The fact that such test prep works contributes to the poor reputation of today's assessments.

This is the context of most large scale assessment, and those facts should strongly influence the type of tests we use automated workflows to generate.

References

- Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing*, 12(3), 273–298.
- Haladyna, T. M., & Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation & the Health Professions*, 12(1), 97–104.
- Hoffman, A., & Wine, M. (2025). *The Low Bar Item Draft Acceptability Tool & Creativity-Based Item Draft Acceptability Tool* [White Paper]. The Rigorous Test Development Project.
- Hoffman, A., Whitcomb, D., & Wine, M. (2025). Item Quality Consideration in Automatic Item Generation. Paper presented at the Artificial Intelligence in Measurement and Education Conference, Pittsburgh, PA.
- Huff, K., Nichols, P., & Schneider, M. C. (2025). Designing and developing educational assessments. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 441–511). Oxford University Press.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.
- Koretz, D. (2017). *The testing charade: Pretending to make schools better*. University of Chicago Press.
- Kosmidis, I., & Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108(1), 71–82.
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1), 121–204.
- Lane, S., & Marion, S. F. (2025). Validity and validation. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 191–275). Oxford University Press.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29(2), 4–16.
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics and for English Language Arts*. Washington, DC.
- Sireci, S. G. (2021). NCME Presidential Address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice*, 40(1), 7–16. <https://doi.org/10.1111/emip.12415>.
- Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2025). A review of AIG techniques leveraging LLMs. *International Journal of Assessment Tools in Education*, 12(2), 317–340.
- Wine, M., & Hoffman, A. (2024). Commentary: What is the Breadth of “Educational Measurement?” *Educational Measurement: Issues and Practice*, 43(3), 23–26. <https://doi.org/10.1111/emip.12627>