Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Expert validation of AI-generated clinical assessment items for physician assistant certification: a multi-dimensional rubric and dual- reliability method

I Made Agus Setiawan
made.agus@pitt.edu

University of Pittsburgh
Maria Yuliana
University of Pittsburgh
Bayu Aryoyudanta
University of Pittsburgh
Firdaus A Indradhirmaya
University of Pittsburgh
Haomin Hu
University of Pittsburgh
Dipu Patel
University of Pittsburgh
David C. Beck
University of Pittsburgh
Andi Saptono
University of Pittsburgh
Bambang Parmanto
University of Pittsburgh

Article

Keywords:

Posted Date: August 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10475937/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. The generation engine evaluated in this study was developed by the authors' research group, and the group has filed an invention disclosure with the University of Pittsburgh related to the technology. This work was supported in part by a Pitt.INC award, a program that funds projects with a commercial development endpoint. Some authors also served as expert raters (two of four in the Round 1 pilot and

one of six in Round 2); to limit this overlap, the study analyst was not a rater, and instrument development, rating, and analysis were kept as separate roles. No author has a personal financial interest, such as equity, patent royalties, or consulting income, related to the technology at this time

Abstract

Creating high-quality clinical questions for certification exam preparation takes substantial expert time and does not scale to meet learner demand. Generative AI could produce such content at scale, but in high-stakes certification contexts AI-generated content must be checked against the same quality standard as expert-written material, and no established method exists for validating it systematically.

We present a validation instrument and method for AI-generated certification exam content, demonstrated on a generation engine we developed for Physician Assistant National Certifying Examination (PANCE) preparation. The approach mirrors the human item-development workflow: items are generated against published item-writing guidelines, then judged by an expert panel against those same guidelines. An initial pilot rounds exposed recurring content problems and guided refinement of the instrument. A second round then validated the refined five-dimension rubric, which use a two-track quality gate separating question construction from explanation quality. In the second round, six PA faculty reviewed 55 items across 14 of the 15 PANCE organ systens. Content validity was strong (overall S-CVI/Ave of 0.9285), and dual inter-rater reliability reporting revealed a prevalence paradox: Fleiss' κ was low at 0.1201 while Gwet's AC1 was high at 0.8653, a gap reflecting skewed rating distribution rather than genuine disagreement. The same paradox appeared in both rounds under different rating designs. In a preliminary blinded student perception study (n = 10), AI and human-written items gave near-identical answering accuracy (70.0% versus 69.3%, p = 0.89) and equivalence clarity and realism, with source detection followed a response-bias pattern rather than real discrimination. The instrument is designed to transfer to any certification exam with a content blueprint, a curated knowledge base, and item-writing guidelines.

Introduction

Programs preparing students for high-stakes certification exams require large banks of high-quality practice questions, more than expert faculty can produce by hand. Each item demands clinical knowledge, item-writing training, and iterative review against established construction standards. The supply is further constrained by the constant need to update items as clinical guidance evolves. This demand is especially strong for physician assistant (PA) programs, where performance on the Physician Assistant National Certifying Examination (PANCE) carries programmatic weight. A sustained decline in first-time pass rates can draw accreditation scrutiny. The pressure became concrete during the COVID-19 pandemic[1], when national analysis of first-time examinees found that PANCE pass rates were affected in the pandemic's third year, with 1.24 times higher odds of failing in 2022 than before the pandemic, which has made PA faculty look more closely at curriculum and student preparedness.

This constraint is not specific to one certification or setting. Any program preparing students for a blueprint-driven certification exam, whether in Physician Assistant education, nursing, medical specialties, or allied health, faces the same shortage. Generative AI offers a path to scalable content production. Large language models (LLMs) can produce clinical vignettes, answer options, and explanations at a rate and cost that faculty-driven development cannot match, and retrieval-augmented generation (RAG) can ground that output in curated source material[1]. These items are meant to serve two purposes: practice testing and remediation, with the aim of supporting candidates preparing for certification, primarily first-time test-takers. For remediation, the explanation accompanying each item addresses not only why the correct option is right, but also why the other options are wrong. Both uses depend on quality: a practice item must be well constructed, and a remediation explanation must be sound. For high-stakes certification preparation, though, AI-generated content must be validated against the same quality standards as expert-written material before it can be responsibly given to students.

Human-written items already have a validation model: items are constructed against a published item-writing standard, and an expert panel then evaluates them against that same standard before use. The question this work addresses is how to adapt that model to AI-generated items. No established methodology exists for systematically validating AI-generated assessment content for clinical education. Published work has focused on the deployment and use of AI-generated educational content, or on how well LLMs perform as exam-takers, rather than on the quality of AI-generated items as assessment instruments. That gap is what this work sets out to fill.

Three lines of prior work converge on the problem this study addresses. The first is the use of LLMs to answer and generate medical assessment items. QUEST-AI (Bedi et al., 2025) generated and verified USMLE-style items, and several studies have characterized LLM performance on USMLE, PANCE, and NCLEX-style examinations. More recent work has compared AI-generated and human-written items directly. Linde et al.[2] found that GPT-4o and human MCQs in imaging specialties had comparable difficulty and discrimination and could not be told apart above chance, and Qian et al.[3] reported a similar result for DeepSeek-generated radiology items, with residents unable to identify the source, though AI items were weaker on higher-order questions. Çelik et al[4], studying how well LLMs answer cardiology in-service items, showed that model answering accuracy tracked item difficulty and that strong answering accuracy did not guarantee a good explanatory rationale. Prior work therefore establishes that AI can engage with assessment content, but it does not provide a way to certify their content quality of AI-generated items against published item-writing guidelines[5] before use.

The second line is human validation of AI-generated educational content. Nam et al[6] validated a multi-LLM generation pipeline with radiologist raters and concluded that the error rate did not support fully automated deployment, indicating that human-in-the-loop validation is a requirement rather than an option. Thesen and Park[7] deployed an AI-generated teaching assistant and studied how students adopted and trusted it, but focused on its delivery and use rather than on the validity of the content it produced. Together, these show that validation of AI output is recognized as necessary, yet no shared instrument exists to carry it out.

The third line is content-validation methodology. The Content Validity Index[8–10] is a well-established basis for expert content validation in nursing and health-professions research, but it has not been applied systematically to AI-generated assessment content. The tools exist. The application is missing.

These three lines leave a specific gap. There is no reusable, standard-grounded instrument for validating AI-generated assessment content, and no established practice for reporting inter-rater agreement when expert acceptance is high and chance-corrected indices become unreliable.

This study addresses both. We developed a validation instrument and method for AI-generated certification exam content, grounded in published NCCPA[5] item-writing guidelines so that AI-generated items are judged against the same standards expected of human item writers. The instrument score five dimensions through a two-track quality gate that separates question construction from explanation quality, and it is designed to transfer to any blueprint-driven certification exam given a content blueprint, a curated knowledge base, and the relevant item-writing guidelines. Applying it across two rounds of expert validation, we find that validation of high-quality AI content produces a prevalence paradox in inter-rater agreement, high observed agreement alongside a low chance-corrected Fleiss' κ driven by skewed ratings, and we report dual reliability indices (Fleiss' κ with Gwet's AC) so that genuine disagreement can be distinguished from this skew. We demonstrate the instrument on content from a retrieval-augmented, agentic generation engine we developed, and extend the evidence with a preliminary blinded student perception study comparing AI-generated and human-written items.

Results

Content validity

Content validity was strong by CVI criteria. The overall S-CVI/Ave across all five dimensions was 0.9285, meeting the 0.90 threshold for strong content validity[9]. The QA composite (Dimensions 1 to 4) was 0.9356, and the Explanation dimension (Dimension 5) was 0.9000, so both tracks of the two-track gate cleared the threshold.

Dimension-level S-CVI/Ave values ranged from 0.8545 for Answer Option Quality (Dimension 3) to 0.9758 for Language and Clarity (Dimension 4). Clinical-Educational Alignment (Dimension 1) was 0.9727, Scenario Completeness (Dimension 2) was 0.9394, and Explanation Quality (Dimension 5) met the threshold at 0.9000. Answer Option Quality was the only dimension below 0.90.

S-CVI/UA is the proportion of items with universal agreement, meaning all six raters endorse the dimension. Two dimensions were at or above the 0.80 mark: Language and Clarity (0.8909, 49 of 55 items) and Clinical-Educational Alignment (0.8545, 47 of 55). Three fell below it: Scenario Completeness (0.7636, 42 of 55), Explanation Quality (0.6182, 34 of 55), and Answer Option Quality (0.4182, 23 of 55). The wide gap between S-CVI/Ave and S-CVI/UA on Answer Option Quality and Explanation Quality indicates that most raters endorsed these items, but full six-rater agreement was harder to achieve. This pattern connects directly to the inter-rater reliability results below.

At the item level, 41 of 55 items (74.5%) passed all five dimensions at I-CVI at least 0.78. Another 6 items (10.9%) passed four of five, 5 items (9.1%) passed three of five, and 3 items (5.5%) passed two of five. No item passed only one dimension or none. Under the Track 1 question-construction criterion (Dimensions 1 to 4), item pass counts by dimension were 54 of 55 for Dimension 1, 51 of 55 for Dimension2, 45 of 55 for Dimension 3, and 53 of 55 for Dimension 4. Dimension 5 passed 47 of 55. Table 1 reports the complete Round 2 CVI metrics.

Table 1. Round 2 content validity across 55 items evaluated by six PA faculty raters. S-CVI/Ave is the mean I-CVI across items (threshold ≥ 0.90[10]). S-CVI/UA is the proportion of items with universal agreement (all six raters endorsing that dimension), reported per dimension only; it has no standard cross-dimension composite form, unlike the mean-based S-CVI/Ave. Items pass is the count of items meeting I-CVI ≥ 0.78 for that dimension.

Dimension	Track	S-CVI/Ave	S-CVI/UA	Items pass (I-CVI ≥ 0.78)	Pass %
Dimension 1: Clinical-Educational Alignment	Q+A	0.9727	0.8545	54	98.2
Dimension 2: Scenario Completeness	Q+A	0.9394	0.7636	51	92.7
Dimension 3: Answer Option Quality	Q+A	0.8545	0.4182	45	81.8
Dimension 4: Language and Clarity	Q+A	0.9758	0.8909	53	96.4
QA composite (Dimensions 1–4)	**Q+A**	**0.9356**			
Dimension 5: Explanation Quality	Explanation	0.9000	0.6182	47	85.5
Overall (all 5 dimensions)		**0.9285**		41	74.55

Inter-rater reliability and the prevalence paradox

Aggregate observed agreement was high (P_o = 0.8832). Fleiss' κ was low (aggregate κ = 0.1201, 95% CI 0.0611 to 0.1792). The interval excludes zero, but every dimension-level κ falls in the poorest range on the Landis and Koch scale[11]. Gwet's AC1 was high (aggregate 0.8653, 95% CI 0.8352 to 0.8953) and tracked closely with observed agreement.

This split between high observed agreement, near-zero Fleiss' κ, and high Gwet AC1 is the signature of the prevalence paradox[12–14]. When ratings cluster in the Acceptable and Excellent categories, the chance-correction in Fleiss' κ inflates expected agreement and pushes the κ estimate down, even though raw agreement is high. Gwet's AC1 uses a different chance-agreement model that holds up under skewed distributions, so it stays close to the observed agreement in high-prevalence settings[12,15].

We first saw this pattern in the Round 1 pilot. Using a binary acceptable-versus-not judgment on 128 complete cases, raters agreed on 89.5% of items (P_o = 0.8945), yet Fleiss' κ was slightly negative (κ = -0.0557), while Gwet's AC1 on the same data was high (0.8828). The negative κ is instructive. Expected agreement under κ's chance model (0.9001) slightly exceeded the observed 0.8945, so κ fell below zero even though raters agreed on nearly nine of ten items. Round 2 reproduced the divergence under a different design, a five-dimensional ordinal rubric and a six-rater panel, with binary P_o = 0.8832, κ = 0.1201, and AC1 = 0.8653. Both rounds report binary Fleiss' κ and Gwet's AC1 computed the same way, so the two coefficients are directly comparable across rounds, even though the rating task differed (a single holistic disposition per item in Round 1 and dimension-level scoring in Round 2). The same split appeared in both rounds: high observed agreement, κ near or below zero, and AC1 close to 0.88 (0.8828 in Round 1, 0.8653 in Round 2). This points to the high-acceptance condition as the cause, not any single rubric. When a pipeline produces items that experts mostly accept, the ratings are skewed by design, and κ understates agreement.

The dimension-level results show the paradox is not uniform, which matters for reading it correctly. Dimension 1 (Clinical-Educational Alignment) showed P_o = 0.9479, κ = 0.0177, AC1 = 0.9450, a textbook paradox signature. Dimension 4 (Language and Clarity) was similar and strongest overall (P_o = 0.9564, κ = 0.0776, AC1 = 0.9542). Dimension 3 (Answer Option Quality) broke the pattern: P_o = 0.7673, κ = 0.0638, AC1 = 0.6903. The lower AC1 here is the important part. Where AC1 stays high, the low κ is a skew artifact. Where AC1 itself drops, as on Dimension 3, there is real rater disagreement, not just a prevalence effect. Therefore, the AC1 level distinguishes genuine disagreement from the artifact, and it flags Answer Option Quality as the dimension on which experts actually disagreed.

Gwet's AC2 with linear and quadratic weighting captured agreement on the ordinal 3-point gradient. Aggregate AC2 was 0.6363 (linear) and 0.7232 (quadratic). The gap between AC1 (binary) and AC2 (ordinal) reflects variation in how valid raters judged items: steady endorsement at the binary level, with movement between Acceptable and Excellent at the ordinal level. Dimension 4 had the highest AC2 quadratic (0.8407) and Dimension 3 the lowest (0.4849), matching the AC1 ranking and again marking Dimension 3 as the main site of substantive disagreement. Table 2 reports the full IRR comparison.

As a robustness check against the inclusion of an author among the raters, we recomputed Round 2 with the single author-rater excluded (five external raters). Content validity was essentially unchanged (overall S-CVI/Ave 0.9222 versus 0.9285; QA composite 0.9318 versus 0.9356; 43 versus 41 of 55 items passing all five dimensions). The

prevalence paradox persisted, with Gwet's AC1 remaining high (0.8565 versus 0.8653) and Fleiss' κ remaining low (0.1436 versus 0.1201). The conclusions do not depend on the author-rater.

Table 2. Inter-rater reliability for Round 2 expert validation across 55 items and six PA faculty raters. Observed agreement (P_o) and Fleiss' κ (with 95% CI) are reported alongside Gwet's AC1 (binary Valid / Not Valid coding) and AC2 (ordinal 3-point coding, linear and quadratic weighting). Under the high acceptance prevalence in Round 2, Fleiss' κ is deflated by distribution skew, while AC1 tracks observed agreement more faithfully (the prevalence paradox). The gap between AC1 and AC2 reflects variability in the ordinal Acceptable-to-Excellent gradient beyond the binary validity threshold. Following Vach and Gerke[16], the Landis and Koch[11] benchmark labels are applied only to Fleiss' κ and not to the AC coefficients.

Dimension	P_o	Fleiss' κ	95% CI	AC1	95% CI	AC2 Linear	95% CI	AC2 Quadratic	95% CI
Dimension 1: Clinical-Edu. Alignment	0.9479	0.0177	[−0.056, 0.091]	0.9450	[0.906, 0.984]	0.7684	[0.710, 0.827]	0.8429	[0.791, 0.895]
Dimension 2: Scenario Completeness	0.9055	0.1697	[−0.003, 0.342]	0.8933	[0.832, 0.955]	0.6647	[0.582, 0.747]	0.7612	[0.684, 0.839]
Dimension 3: Answer Option Quality	0.7673	0.0638	[−0.040, 0.168]	0.6903	[0.589, 0.792]	0.4215	[0.305, 0.538]	0.4849	[0.352, 0.618]
Dimension 4: Language and Clarity	0.9564	0.0776	[−0.007, 0.163]	0.9542	[0.916, 0.993]	0.7521	[0.693, 0.811]	0.8407	[0.790, 0.892]
Dimension 5: Explanation Quality	0.8388	0.1044	[0.021, 0.187]	0.8034	[0.716, 0.891]	0.5206	[0.410, 0.631]	0.6053	[0.484, 0.727]
Aggregate	**0.8832**	**0.1201**	**[0.061, 0.179]**	**0.8653**	**[0.835, 0.895]**	**0.6363**	**[0.596, 0.676]**	**0.7232**	**[0.683, 0.763]**

Two-track quality gate

The two-track gate separates question-construction validity (Dimensions 1 to 4) from explanation validity (Dimension 5), and both tracks cleared the threshold. Track 1 (the QA composite) was 0.9356, and Track 2 (Explanation) was 0.9000. The 0.0356 gap between the tracks indicates that explanation quality, while above threshold, is the weaker component. This is consistent with Answer Option Quality (Dimension 3) being the lowest individual dimension and with the thematic analysis, where explanation-related concerns appeared in 7% of coded comments.

The item-level cross-tabulation (Table 3) shows the joint outcome. Items passing both tracks are deployable for both practice-question and remediation use. Items passing Track 1 only are usable as practice questions while their explanations need revision. Items passing Track 2 only are an unusual pattern, pointing to explanation-specific issues that the question construction did not reveal. Items failing both need full revision. Of the 55 items, 41 passed both tracks, 1 passed Track 1 only, 6 passed Track 2 only, and 7 failed both.

Table 3. Two-track quality gate cross-tabulation across 55 Round 2 items. Track 1 (question and answer construction) requires Dimensions 1 to 4 all at least Acceptable. Track 2 (explanation quality) requires Dimension 5 at least Acceptable. Item dispositions follow from the cell: items passing both tracks are ready for full deployment; items passing Track 1 but not Track 2 are usable as practice questions while their explanations are revised; items passing Track 2 but not Track 1 show an explanation-specific pattern where the question construction needs work; items failing both require full revision.

	Track 2 Pass	Track 2 Fail	Row Total
Track 1 Pass	41	1	42 (76.4%)
Track 1 Fail	6	7	13 (23.6%)
Column Total	47 (85.5%)	8 (14.5%)	55

Round 1 pilot findings

Round 1 served as the pilot. Majority acceptance among the four raters was near universal, so at the level of overall usability the engine's output was already largely acceptable. The value of the pilot was not in this acceptance rate but in the specific concerns raised on the items that were declined, which were few and concentrated. These concerns are what shaped the instrument and the engine changes for Round 2.

Three types of concern recurred in the decline rationales. The first was content relevance, for example items judged too specific or outside common entry-level PA knowledge. The second was answer option validity, such as options that conflicted or more than one defensible correct answer, illustrated by a rater noting that two options were both correct. The third was question clarity, where a vignette left out detail needed to choose an answer, for example missing findings needed to rule out an orbital fracture.

Two of these patterns are worth carrying forward. Some concerns reflected a textbook-versus-practice gap, for example a rater arguing that antibiotics for a sinus infection were not wrong in practice even if the textbook key differed, or that paronychia needs no diagnostic test. And the answer-option concerns map directly onto Answer Option Quality, the dimension that later showed the most genuine rater disagreement in Round 2.

Diagnostic tags and comment themes

During validation, raters applied structured diagnostic tags and could add a free-text comment. The tags most often flagged multiple correct answers possible (35 applications), other (32), contains test-wise clues (19), missing critical clinical detail (14), and incomplete or inaccurate explanation (14). The free-text comments, analyzed with a rule-based lexical content analysis, described the same concerns in the raters' own words. Answer Option Integrity was the most frequent substantive theme (20% of comments; distractor-quality concerns in 17 comments and multiply-defensible keys in 6), alongside a cross-cutting Actionable Revision Suggestions category (22%). The remaining themes were Clinical-Educational Alignment (13%), Positive Appraisal (11%), Scenario Completeness (9%), Language and Clarity (8%), and Explanation Quality (7%). Positive appraisals were present but less frequent than corrective feedback, as expected in expert validation.

Together, the tags and comments locate the raters' concerns on answer-option construction and, secondarily, on clinical detail in the vignette, consistent with the lower content-validity and reliability scores on Answer Option

Quality (Dimension 3), and with Scenario Completeness (Dimension 2) as the second target for engine improvement. The full theme and diagnostic-tag frequency tables are in the Supplement S3.

Student perception (preliminary)

The results below come from the current sample of 10 participants who completed all perception tasks. Data collection toward the target of 30 is ongoing, and the results below are therefore preliminary. Preliminary status is noted throughout rather than in a single closing statement.

Participants. 49 students were screened and 10 eligible participants completed all three tasks to form the analytic sample (33% of the eligible participants). Six who initiated but did not complete were excluded under the 80% task-completion rule. Recruitment is ongoing. The analytic sample is 10 clinical-year PA students, eight from the University of Pittsburgh PA program and two from other programs, most with extensive prior standardized-exam exposure, prior use of AI tools. Figure 1 shows the full CONSORT flow, and full demographics are in Supplementary S4.

Knowledge equivalence. This is the most directly informative finding at the current sample. On 300 answering trials, overall accuracy was 69.67% (95% CI 64.24 to 74.59), and essentially identical by source: AI items 70.0%, human items 69.33%, a 0.7-point difference (Table 4a). A logistic regression with student and item fixed-effect adjustments showed no source effect on correctness ($\beta = 0.036$, SE = 0.275, $p = 0.89$; OR = 1.04, 95% CI 0.60 to 1.78). With 10 participants and 300 trials, the estimate is near zero at both the point and interval level, so it is unlikely to shift much at the target sample. In this sample, AI items functioned as assessment items at the same difficulty as human-written items.

Perceptual equivalence. Clarity and realism ratings on 5-point scales were near-identical across sources (Table 4a). Clarity was 4.21 (SD 0.68) for AI items and 4.14 (SD 0.72) for human items, a source effect of 0.07 points (90% CI -0.04 to 0.19). Realism was 3.76 (SD 0.97) and 3.75 (SD 0.90), a source effect of 0.01 points (90% CI -0.15 to 0.16). Two one-sided tests against a pre-specified margin of 0.5 pooled SD rejected at $\alpha = 0.05$ for both, supporting perceptual equivalence in this sample. Equivalence claims are sample-size sensitive, so intervals will tighten and point estimates may move a little as the sample grows, but the preliminary magnitudes are small enough that a reversal is unlikely.

Source detection. Detection is best read as a qualifier to the equivalence findings. In 300 detection trials, students identified the source correctly on 158 (52.67%, 95% CI 47.02 to 58.25). A one-sided binomial test against chance was not significant ($p = 0.19$). Taken alone, that looks like chance performance. But the overall rate hides a clear asymmetry. On human items, students identified the source correctly 59.3% of the time (89 of 150). On AI items, they were correct 46.0% of the time (69 of 150), which is below chance. A logistic regression with student and item fixed-effect adjustments confirmed the asymmetry (source $\beta = -0.60$, 95% CI -1.07 to -0.12; OR = 0.55, 95% CI 0.34 to 0.89; $p = 0.014$). Genuine discrimination would raise accuracy on both sources; instead, above-chance accuracy on human items with below-chance accuracy on AI items, and chance-level accuracy overall, is the signature of a response bias toward the “human” label rather than real detection (Table 4b). Students leaned toward the "human" label, which made overall accuracy look like chance while they systematically mislabeled AI items as human. The asymmetry points to engine-refinement targets rather than broad detectability.

Synthesis. Three directional conclusions hold at the current sample. Knowledge equivalence is the most robust: near-identical accuracy (70.0% versus 69.3%, OR 1.04, $p = 0.89$), unlikely to move at the target sample. Perceptual

equivalence is supported by two one-sided tests (TOST) on both clarity and realism. Detection does not support a simple "at chance" reading: the 52.67% overall masks a 13.3-point asymmetry where students cannot identify AI items as AI (46.0%) but can somewhat identify human items as human (59.3%), a response-bias pattern that fits the equivalence findings and points to specific refinement targets. These results are preliminary and underpowered against the target of 30; intervals will narrow, and the detection asymmetry needs confirmation at full sample with per-item sensitivity analysis to check whether it arises from a subset of items with distinctive human cues.

Table 4a. [PRELIMINARY] Student perception ratings and answering accuracy by source. Ratings were on 5-point Likert scales. The source effect was estimated from linear regression with student and item fixed-effect adjustments (90% CI). The TOST equivalence margin was ±0.5 pooled SD. The answering-accuracy source effect is from logistic regression with student and item fixed-effect adjustments, reported as an odds ratio (OR) with 90% CI. Sample: n = 10 of 30 target, 150 ratings per source.

Panel	Metric	AI Mean (SD)	Human Mean (SD)	Source Effect (90% CI)	TOST / Test Result
Clarity & Realism	Clarity	4.21 (0.68)	4.14 (0.72)	0.07 [−0.04, 0.19]	Margin ±0.35, equivalence supported ($p < 0.001$)
	Realism	3.76 (0.97)	3.75 (0.90)	0.01 [−0.15, 0.16]	Margin ±0.47, equivalence supported ($p < 0.001$)
Answering Accuracy	Correctness	0.70 (105/150)	0.69 (104/150)	OR = 1.04 [0.61, 1.78]	$p = 0.895$

Table 4b. [PRELIMINARY] Student perception, source detection across 300 trials (10 participants, 30 items each**).** Top panel: overall detection accuracy against 0.50 chance. Middle panel: per-source accuracy and logistic regression source asymmetry with student and item fixed-effect adjustments (AI relative to human as reference).

Panel	Metric	n trials	Correct	Accuracy	95% CI
Overall Detection	All trials	300	158	52.67%	[47.02%, 58.25%]
	Binomial *p* versus. chance				$p = 0.193$
Source Asymmetry	Human items	150	89	59.33%	[51.33%, 66.87%]
	AI items	150	69	46.00%	[38.22%, 53.98%]
	Asymmetry (Human − AI)			+13.33 pp	
	Source coeff. β			−0.60	[−1.07, −0.12]
	OR			0.55	[0.34, 0.89]
	p				0.014

Discussion

This study makes two separate contributions. The first is a validation instrument and method: a five-dimension rubric with a two-track quality gate, diagnostic tags, and dual inter-rater reliability reporting, grounded in published NCCPA item-writing guidelines. The second is a methodological finding on how to measure agreement when the content being judged is mostly good.

On the instrument, Round 2 gave strong evidence. Across 55 items and six PA faculty raters spanning 14 of the 15 PANCE organ systems, the overall S-CVI/Ave was 0.9285, above the 0.90 threshold for strong content validity. Both tracks of the gate cleared that threshold on their own, the question track at 0.9356 and the explanation track at 0.9000, and 41 of 55 items passed all five dimensions. The instrument did not merely pass content. It also showed where content was weakest, with Answer Option Quality consistently the lowest-scoring dimension, which made that dimension visible as a specific engine-refinement target rather than as undifferentiated noise.

The second finding is the prevalence paradox. In both validation rounds, observed agreement among raters was high while Fleiss' κ was near zero, and the prevalence-robust Gwet's AC1 stayed high (0.8828 in the pilot, 0.8653 in Round 2). We did not expect this. We first observed it in the pilot, recognized it as the known paradox, and reported dual indices in Round 2, where the same pattern returned under a different panel and a different rating design. The rating task differed between rounds, a single holistic judgment per item in the pilot, and dimension-level scoring in Round 2, so the recurrence does not depend on one particular rubric. It reflects the condition the paradox needs: when a generation pipeline is working, experts accept most items, ratings skew toward acceptance, and κ is deflated by construction. Both rounds report binary Fleiss' κ and Gwet's AC1 computed the same way, so the two are directly comparable.

Preliminary student data offer a second line of evidence. In a blinded sample of 10 students, AI and human items produced near-identical answering accuracy (70.0% versus 69.3%, $p = 0.89$) and were rated equally clear and realistic, with equivalence supported by TOST on both. Students could not reliably tell which items were AI. When they guessed wrong, they leaned toward labeling AI items as human, so the AI items were judged human-written rather than machine-made. These findings are preliminary and will be revisited at the target sample.

Recent work has largely treated the quality of AI-generated content as something to prove in a single study, rather than validated using a reusable instrument. QUEST-AI validated LLM-generated USMLE items, and several studies have compared AI-generated and human-written assessment items on psychometric properties. Linde and colleagues[2] compared GPT-4o and human-written multiple-choice items in imaging specialties and found comparable difficulty and discrimination with source detection at chance. Çelik and colleagues[4] reported similar psychometric comparisons in a cardiology in-service examination. This body of work establishes that AI items can perform like human items, but it does not provide a standard, reusable way to certify the quality of their content before use. Our contribution fills that gap by creating a validation instrument and method that another program can adopt, with dimensions anchored to a published item-writing standard, a two-track gate that separates question quality from explanation quality, and dual IRR reporting. The prevalence paradox and Gwet's AC are established in the methodology literature. Our contribution is to demonstrate that this paradox is a predictable feature of AI-content validation, and to report dual indices so that genuine disagreement can be told apart from a false signal caused by skewed ratings.

Telling genuine disagreement apart from a false signal caused by skewed ratings has practical weight. Linde and colleagues reported low chance-corrected agreement among their expert raters, with ICC values from 0.07 to 0.18, and interpreted it as substantive disagreement about item quality. In the high-acceptance setting that a working pipeline produces, such low values are the expected signature of the paradox rather than evidence of true

disagreement and distinguishing the two requires paradox-robust indices. Work on AI in medical education has also focused on delivery rather than validation. Thesen and Park[7] deployed a RAG-constrained teaching assistant in a medical course and studied how students adopted and trusted it. Our work addresses that complementary and still-open problem. The two lines are related: one asks whether students will use AI support, the other asks whether the AI content meets the standard required before students rely on it.

Two implications follow. The first is that content validation should come before deployment. AI can generate assessment content at scale, but scale is not the constraint that matters in a certification setting; quality is. Deploying unvalidated AI content in exam preparation carries pedagogical and ethical risks, and the instrument described here provides programs with a way to check content against an expert standard before students use it. The second is that agreement reporting in this area needs to change. As AI-generated content improves, expert acceptance rises, and rating distributions skew toward acceptance, and under that skew, Fleiss' κ and similar chance-corrected indices are deflated by construction, so a study can show strong expert agreement and still report a κ near zero. This is the expected state for any validation of good AI content, not a rare edge case, and reporting a prevalence-robust index such as Gwet's AC alongside κ should therefore be standard practice in content-validation studies of AI-generated educational material.

The instrument’s design also carries lessons beyond this study. The central one is that explicit dimensions turn rater variability into something interpretable. A holistic accept-or-decline judgment produces variation that cannot be traced to any particular cause, while scoring the same item on defined dimensions does not eliminate disagreement but locates it. A compact instrument helps more than a comprehensive one. The final design uses five dimensions on a three-point scale, which requires fifteen scale-point judgments per item. This is low enough to keep multi-rater validation practical, and a three-point scale of Poor, Acceptable, and Excellent aligns with the judgments reviewers can consistently make, rather than asking for finer distinctions they cannot apply reliably. The instinct to add a dimension for every quality concern should be resisted in favor of clear boundaries between fewer dimensions. Two further principles apply. Explanation quality matters enough to score on its own, because the content is meant both to test students and to teach them. That is why it is a separate dimension and its own track in the gate, not something combined with the other quality judgments. The diagnostic tags work alongside the dimension ratings. The ratings show the overall quality of an item, while the tags flag specific known problems, such as test-wise cues or more than one defensible answer, that a rating alone would miss.

The instrument is demonstrated on PANCE, but its design is not specific to PA certification. Its authority comes from being anchored to a published item-writing standard, and every major clinical certification body publishes comparable construction guidance and a content blueprint. Adapting the instrument to another exam therefore needs three inputs and no change to its structure: the relevant content blueprint, a curated knowledge base for that domain, and the item-writing guidelines of the certifying body. The five dimensions, the two-track gate, the diagnostic tags, and the dual-IRR practice transfer without modification, and dimension wording might gain a domain-specific element, for example a patient-safety dimension in nursing. This makes the applicability claim a matter of design rather than breadth of testing. We validated on one exam. The reason the approach should transfer is not that we tried many domains, but that the instrument's dimensions derive from a class of standard that exists across medical and health-professions certification. Natural next targets include NCLEX-RN and NCLEX-PN in nursing, the USMLE and COMLEX-USA sequences in medicine, and certification exams in physical therapy, occupational therapy, and speech-language pathology.

Several limitations bear on how these findings should be read. The expert panel was recruited through professional networks rather than at random, which limits how far the validation generalizes. Blueprint coverage was 14 of the 15

PANCE organ systems, with Professional Practice excluded because no seed material was available at the time, so the instrument was demonstrated on 14 systems rather than the full blueprint. Six raters are adequate for content-validity computation but modest for reliability estimation, and the reliability figures should be read with that in mind. The generation engine is described at a functional level with full specifications proprietary, so a reader can apply the validation instrument to any generation source but reproducing this engine's content specifically requires engine access. The student perception results are preliminary at 10 of a target 30 participants and were analyzed with student and item fixed-effect adjustments rather than mixed-effects models, which suits the current sample size but limits inference to this sample. Cross-domain applicability is argued from the instrument's design but has not been tested outside PA certification. Finally, no psychometric item analysis of difficulty, discrimination, or differential item functioning was performed at this stage, which requires larger student samples and is deferred to later validation.

The most immediate step is completing the perception study, at which point the analysis can move from fixed-effect adjustments to mixed-effects models that become better identified at the larger sample. Beyond that, applying the instrument to a second domain, for example NCLEX-RN with a nursing blueprint, knowledge base, and item-writing guidelines, would be the first direct test of the cross-domain claim. An automated scorer calibrated against expert ratings, producing a per-dimension confidence score, could route only uncertain items to expert review and make validation of large content banks more tractable. Developing and validating Professional Practice content would close the one remaining gap in PANCE blueprint coverage.

Taken together, this study contributes a validation instrument and method for AI-generated clinical assessment items, together with a way of reporting inter-rater reliability that stays accurate when experts accept most items. Demonstrated on content from the generation engine we developed for PA certification across 55 items, 14 organ systems, and six expert raters, the instrument reached an overall S-CVI/Ave of 0.9285 with both tracks of the gate clearing the 0.90 threshold for strong content validity, while the prevalence paradox appeared on two separate rounds and marks high expert agreement coexisting with a low chance-corrected Fleiss' κ as a general feature of validating good AI output. Its design is not tied to PA certification and adapting it to another exam requires only a content blueprint, a curated knowledge base, and the relevant item-writing guidelines, which makes it a practical basis for validating AI-generated assessment content wherever such content is used.

Methods

Study context and workflow

Certification-exam content is normally produced in two steps. Item writers construct questions against a published item-writing standard, and an expert panel then reviews those questions against the same standard before they are used. The approach in this study mirrors that workflow. An AI engine generates questions grounded in the item-writing standard, and an expert panel validates them against it using a defined instrument. The contribution is the instrument and the validation method, not the generator.

The instrument's design is not specific to one exam. It applies to any certification exam that uses structured clinical questions with a defined content blueprint, including PANCE, NCLEX-RN, NCLEX-PN, USMLE, and COMLEX. Adapting it to a new domain requires three inputs: a content blueprint defining the distribution of clinical topics and cognitive tasks, a curated knowledge base of authoritative reference material, and the item-writing guidelines for the target exam.

This study uses Physician Assistant certification (PANCE) as the demonstration context. The NCCPA PANCE Content Blueprint[17] defines 15 organ systems, each weighted by proportional representation, crossed with 8 clinical task categories (history-taking, physical examination, diagnostic studies, diagnosis, clinical intervention, pharmacotherapy, patient education, and professional practice). The study used NCCPA item-writing guidelines as the source of construction rules, alongside established test-development standards[18] and a faculty-curated knowledge base drawn from standard PA program references. Figure 2 summarizes the two rounds of validation, the between-round refinement, and the preliminary student perception study.

AI item generation

The generation engine is a retrieval-augmented, agentic engine we developed for this study. It produces Question-Answer-Explanation (QAE) triplets, each a clinical question stem with answer options, an answer key, and a rationale explanation, through retrieval-augmented generation (RAG)[1] with automated content quality assurance (Figure 3). A representative validated QAE is shown in Supplementary S5. The knowledge sources are a faculty-curated textbook knowledge base, the NCCPA item-writing guidelines[5], and the PANCE Content Blueprint of 15 organ systems by 8 task categories. Faculty-curated and public-source QAs serve as seeds, classified against the blueprint at the organ system by task category level. The automated content quality assurance layer was added for Round 2 and was not present in Round 1. The engine is described at a functional level. Its internal architecture is proprietary. The subject of this study is the validation instrument applied to the engine's output, not the engine itself. The functional description here and the high-level architecture in Figure 3 are what the validation results depend on, and they are sufficient to interpret those results.

Validation instrument

The validation instrument scores each QAE on five dimensions using a 3-point ordinal scale (Poor, Acceptable, Excellent). Four dimensions cover question construction: Clinical-Educational Alignment, Scenario Completeness, Answer Option Quality, and Language and Clarity. The fifth dimension, Explanation Quality, covers the rationale explanation. The four construction dimensions operationalize the core principles of the NCCPA Item-Writing Workshop[5]. Appropriate scope and cognitive level (Clinical-Educational Alignment), complete and relevant clinical vignettes (Scenario Completeness), defensible keys with plausible distractors and no test-wise cues (Answer Option Quality), and focused, unambiguous lead-in phrasing (Language and Clarity). Grounding the dimensions in a published standard means the instrument checks whether an AI-generated item follows the same rules a human item writer is expected to follow.

The instrument uses a two-track quality gate. Track 1 covers question construction and is met when all four construction dimensions (Dimensions 1 to 4) are rated at least Acceptable. Track 2 covers the explanation and is met when Dimension 5 is rated at least Acceptable. The two tracks are judged separately because a question can be sound while its explanation needs work, or the reverse. Keeping them apart lets each component be accepted or sent back on its own. Acceptance is derived analytically from the dimension ratings rather than collected as a separate overall judgment. This removes the inconsistency that arises when a holistic decision conflicts with the underlying dimension ratings and lets the content validity analysis run directly on the rubric data.

The instrument also includes twelve optional diagnostic tags, available for any item regardless of its ratings. The tags mark known item-construction problems, such as multiple correct answers, test-wise clues, or missing clinical

detail, as well as explanation-specific issues. The tags do not affect the acceptance decision. They record why an item fell short, which supports targeted revision of the engine and of individual items.

Instrument provenance

The instrument and the refined engine both emerged from a pilot validation (Round 1) conducted before the NCCPA item-writing guidelines were available to the team. In Round 1, four PA faculty raters reviewed 130 items across three organ systems (Cardiovascular, EENT, and Dermatologic), produced by an early version of the engine with only basic quality filtering. The 130 items were the validated set drawn from 160 generated candidates, after an LLM-ensemble check of answer keys. Raters gave a holistic Accept, Revise, or Decline judgment through an internally developed validation platform. Ratings were independent and blinded to other raters. The raters were PA faculty, and some were members of the study team. The engine developers did not rate.

Round 1 exposed systematic content problems (reported in Results) and, because it ran before the guidelines were in hand, it made clear what a grounded instrument would need. Once the collaborators provided the NCCPA guidelines and additional reference materials, both the engine and the validation instrument were aligned with that standard for Round 2. The full instruments used in both rounds are provided in Supplementary S1.

Round 2 validation

Round 2 validated the refined engine's output using the finished instrument described above, with a larger, mostly external expert panel and broader organ-system coverage.

Participants. Six PA faculty served as raters: one from Round 1 and five recruited from other PA programs. Median clinical experience was 15.5 years, and median teaching experience was 7.8 years, and three of the six had prior certification-exam item-review experience. The five-of-six external composition supports generalizability beyond the originating program. Some raters were also study authors: two of the four raters in the Round 1 pilot, and one of the six raters in Round 2 (the rater retained from the pilot for continuity across rounds). The author-raters contributed expert ratings only. The study analyst was not a rater in either round, and no author-rater took part in computing or analyzing the reliability or content-validity results.

Materials. Fifty-five QAEs spanning 14 of the 15 PANCE organ systems and seven task categories were generated by the refined engine. Professional Practice was excluded because no seed material was available at the time of engine development. Generation provenance: 1,382 seed QAs, expanded from Round 1's 58 through additional faculty curation and public-source integration across the 14 systems, produced a candidate pool of 809 QAEs after automated quality assurance, from which 55 items were drawn by stratified systematic sampling.

Sampling used four sequential gates within each system: technical quality (LLM-ensemble verification with at least all models agreeing on the key answer, complete QAE format, correct blueprint classification), task-type diversity, a spread across difficulty levels where the per-system count allowed, and systematic selection from the eligible pool rather than curator-preferred items. Difficulty levels for this stratification were assigned by an LLM-assisted annotation using a predefined three-level rubric and were used only to spread the selection across difficulty, not as a validation measure. Per-system quotas balanced blueprint coverage against rater workload. The full selection procedure and candidate-pool provenance are given in the Supplement S2. Table 5 reports the per-system allocation.

Table 5. Expert validation item allocation across 14 PANCE organ systems in Round 2. Professional Practice was excluded because no seed material was available at engine-development time. Item counts followed a floor of three items per system, the minimum for a meaningful item-level content validity index with six raters[8], with additional items assigned toward higher-blueprint-weight systems and to the three systems also covered in the pilot (Cardiovascular, EENT, Dermatologic), which were sampled at five items for continuity across rounds. This balanced blueprint coverage against the review burden on the six-rater panel. The 55 items span 14 of the 15 organ systems, representing approximately 94% of the PANCE blueprint weight.

System	PANCE Weight	Items
Cardiovascular †	11%	5
EENT †	6%	5
Dermatologic †	4%	5
Pulmonary	9%	5
Gastrointestinal / Nutrition	8%	4
Musculoskeletal	8%	4
Neurologic	7%	4
Psychiatric / Behavioral	7%	4
Infectious Diseases	7%	4
Reproductive	7%	3
Endocrine	6%	3
Hematologic	5%	3
Renal	5%	3
Genitourinary	4%	3
Total	**94%**	**55**

† Systems also included in the Round 1 pilot.

Instrument. Raters used the five-dimension rubric described above, with the 3-point scale and the diagnostic tags, and rated difficulty on the separate instrument. The two-track gate was applied as defined: Track 1 (all of Dimensions 1 to 4 at least Acceptable) and Track 2 (Dimension 5 at least Acceptable), operating independently.

Analysis. Content validity was assessed with the content validity index (CVI): the item-level CVI (I-CVI, target at least 0.78), the scale-level CVI averaging method (S-CVI/Ave, target at least 0.90), and the scale-level CVI universal-agreement method (S-CVI/UA, target at least 0.80), computed per dimension and overall, with modified kappa per item to adjust I-CVI for chance agreement[10]. Inter-rater reliability was reported in two forms: Fleiss' κ with 95% confidence intervals as the conventional chance-corrected measure[19], and Gwet's AC1 (on dichotomized Valid/Not Valid ratings) with AC2 (on the 3-point ordinal scale, linear and quadratic weights) as the prevalence-robust check[12,13]. Analyses were performed in Python; the agreement coefficients (Fleiss' κ, Gwet's AC1, and AC2) were

computed with the irrCAC and statsmodels library, and the content-validity indices were computed using standard CVI formulas. As a sensitivity analysis, content validity and inter-rater reliability were recomputed with the single author-rater excluded (five external raters). Rater free-text comments were analyzed with a rule-based lexical content analysis, in which a predefined codebook of ten codes grouped into seven themes assigned each comment to one or more themes by case-insensitive regular-expression matching; the structured diagnostic tags raters applied were tabulated alongside.

Student perception study

Design. A blinded within-subject comparison of AI-generated and human-written questions. Each participant rated 30 questions (15 AI-generated, 15 human-written, matched by organ system, task category and difficulty) in randomized order. The source was withheld throughout. Participants were told that the items came from various sources and were not primed to distinguish AI-generated from human-generated items.

Procedure and recruitment. This study was conducted entirely online. Participants were recruited through flyers and email invitations sent to PA program coordinators, outreach through faculty professional networks, and in-person study session with clinical-year students at which the study was introduced. Interested students accessed the study through a web link and completed an eligibility screener. Eligible participants were then directed to provide demographic information and complete the perception tasks. All study activities were completed online, with no in-person data collection. Participation was voluntary, and completion was tracked through the internally developed study platform.

Participants. The target is 30 completed responses from PA students recruited from the University of Pittsburgh PA program and other PA programs, with an enrollment buffer for attrition. Participant flow and the analytic sample are reported in the Results.

Materials. Thirty questions: 15 AI items selected from the Round 2 validated bank (from items passing both tracks at unanimous consensus) and 15 human-written items contributed by the six Round 2 expert validators, matched one-to-one with the AI items by organ system, task type, and difficulty level.

Task. Participants completed three tasks in sequence. In Task 1, they saw each item as question-and-answer options only, without the explanation, and for each item rated clarity and realism on 5-point scales, gave a binary source guess (human or AI), rated confidence in that guess on a 5-point scale, and optionally gave a free-text reason. In Task 2, they answered all 30 items and saw the correct answer for each. Task 3 was a short open-response questionnaire on unusual item features, willingness to use AI-generated questions, and any further comments. Task 1 excludes the explanation so that clarity, realism, and detection reflect the question itself rather than the question-plus-explanation.

Analysis. Clarity and realism equivalence was tested with linear regression using student and item fixed-effect adjustments, followed by two one-sided tests (TOST) against a margin of 0.5 pooled SD[20]. Detection was tested with an aggregate binomial test against 50% chance, with a logistic regression using student and item fixed-effect adjustments as the sensitivity analysis for source asymmetry. Knowledge equivalence was tested with logistic regression using student and item fixed-effect adjustments on item correctness. Given the preliminary sample size, inferential models used student and item fixed-effect adjustments rather than random intercepts, which avoids poorly identified variance components at this N=10. Mixed-effects estimation is planned for the full sample. At the

current sample, all planned analyses were run, with results reported under preliminary framing and inference kept to this sample.

Declarations

Ethics

Both sub-studies received an exempt determination from the University of Pittsburgh Institutional Review Board (Protocol #STUDY26020168), under 45 CFR 46.104(d)(1) for research in educational settings. In Sub-Study 1, expert validators act as content evaluators rather than research subjects, consistent with standard practice in content-validation studies. In Sub-Study 2, participation is voluntary, data are de-identified, and taking part has no effect on academic standing. Before beginning any study activity, participants reviewed an information statement describing the study purpose, procedures, expected time commitment, and research team contact information, and proceeded to the tasks by their own action. Signed documentation of consent was not collected. Eligibility screening uses a sequential two-item gate in Microsoft Forms. Respondents who do not meet criteria are anonymous by design, with no contact information collected for ineligible respondents, consistent with 45 CFR 46.104(d). The exemption approach follows published precedent[7].

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy or ethical restrictions.

Code Availability

The underlying code used in this study is not publicly available but may be provided to qualified researchers upon reasonable request to the corresponding author.

Acknowledgements

The authors thank the PA faculty experts who served as reviewers for this study. Their clinical and educational experience was essential to the validation process. The authors also thank the physician assistant program faculty at the University of Pittsburgh and other participating programs for their support, and the physician assistant students who volunteered for the perception study. This work was supported by University of Pittsburgh through a 2025 Innovation in Education Awards from the Provost's Advisory Council on Innovation in Education (ACIE) and the 2025 Pitt Idea Navigation to Commercialization (Pitt.INC) program.

Author contributions

Conceptualization: I.M.A.S., D.P., D.C.B., A.Sa., B.P. Data curation: I.M.A.S., M.Y., B.A. Formal analysis: I.M.A.S., M.Y. Funding acquisition: I.M.A.S., D.P., D.C.B., A.Sa., B.P. Investigation: I.M.A.S., M.Y., B.A., F.A.I. Methodology: I.M.A.S., M.Y. Project administration: A.Sa., B.P. Resources: D.P., D.C.B., A.Sa., B.P. Software: I.M.A.S., M.Y., B.A., F.A.I., H.H.

Supervision: D.P., D.C.B., A.Sa., B.P. Validation: I.M.A.S., M.Y. Visualization: I.M.A.S., F.A.I. Writing – original draft: I.M.A.S. Writing – review and editing: all authors. All authors reviewed and approved the final manuscript.

Competing interests

The generation engine evaluated in this study was developed by the authors' research group, and the group has filed an invention disclosure with the University of Pittsburgh related to the technology. This work was supported in part by a Pitt.INC award, a program that funds projects with a commercial development endpoint. Some authors also served as expert raters (two of four in the Round 1 pilot and one of six in Round 2); to limit this overlap, the study analyst was not a rater, and instrument development, rating, and analysis were kept as separate roles. No author has a personal financial interest, such as equity, patent royalties, or consulting income, related to the technology at this time

References

1. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. in *Advances in Neural Information Processing Systems* vol. 33 9459–9474 (Curran Associates, Inc., 2020).
2. Linde, P. *et al.* Psychometric properties and detectability of GPT-4o–generated multiple-choice questions compared with human-authored items across imaging specialties. *npj Digit. Med.* **9**, 132 (2026).
3. Qian, W. *et al.* Performance of DeepSeek in the generation of in-training examination questions in radiology resident education. *npj Digit. Med.* **9**, 384 (2026).
4. Çelik, A., Kırış, T., Kocabaş, U., Özdemir, E. & Karaca, M. Psychometric characterization of human and artificial intelligence performance on cardiology residency in-service examination items. *npj Digit. Med.* https://doi.org/10.1038/s41746-026-03002-9 (2026) doi:10.1038/s41746-026-03002-9.
5. NCCPA. Item-writing workshop. (2020).
6. Nam, Y. *et al.* Human-in-the-loop validation of a sequential multi-LLM medical education pipeline. *npj Digit. Med.* https://doi.org/10.1038/s41746-026-02978-8 (2026) doi:10.1038/s41746-026-02978-8.
7. Thesen, T. & Park, S. H. A generative AI teaching assistant for personalized learning in medical education. *npj Digit. Med.* **8**, 627 (2025).
8. Lynn, M. Determination and Quantification of Content Validity Index. *Nursing Research* **35**, 382–386 (1986).
9. Polit, D. F. & Beck, C. T. The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health* **29**, 489–497 (2006).
10. Polit, D. F., Beck, C. T. & Owen, S. V. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res Nurs Health* **30**, 459–467 (2007).
11. Landis, J. R. & Koch, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **33**, 159–174 (1977).
12. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology* **61**, 29–48 (2008).
13. Wongpakaran, N., Wongpakaran, T., Wedding, D. & Gwet, K. L. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol* **13**, 61 (2013).

14. Feinstein, A. R. & Cicchetti, D. V. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology* **43**, 543–549 (1990).
15. Quarfoot, D. & Levine, R. A. How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution? *The American Statistician* **70**, 373–384 (2016).
16. Vach, W. & Gerke, O. Gwet's AC1 is not a substitute for Cohen's kappa – A comparison of basic properties. *MethodsX* **10**, 102212 (2023).
17. NCCPA. PANCE Blueprint. *NCCPA* https://www.nccpa.net/become-certified/pance-blueprint/.
18. Haladyna, T. M. & Rodriguez, M. C. *Developing and Validating Test Items*. (Routledge, New York, 2013). doi:10.4324/9780203850381.
19. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382 (1971).
20. Lakens, D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Soc Psychol Personal Sci* **8**, 355–362 (2017).

Figures

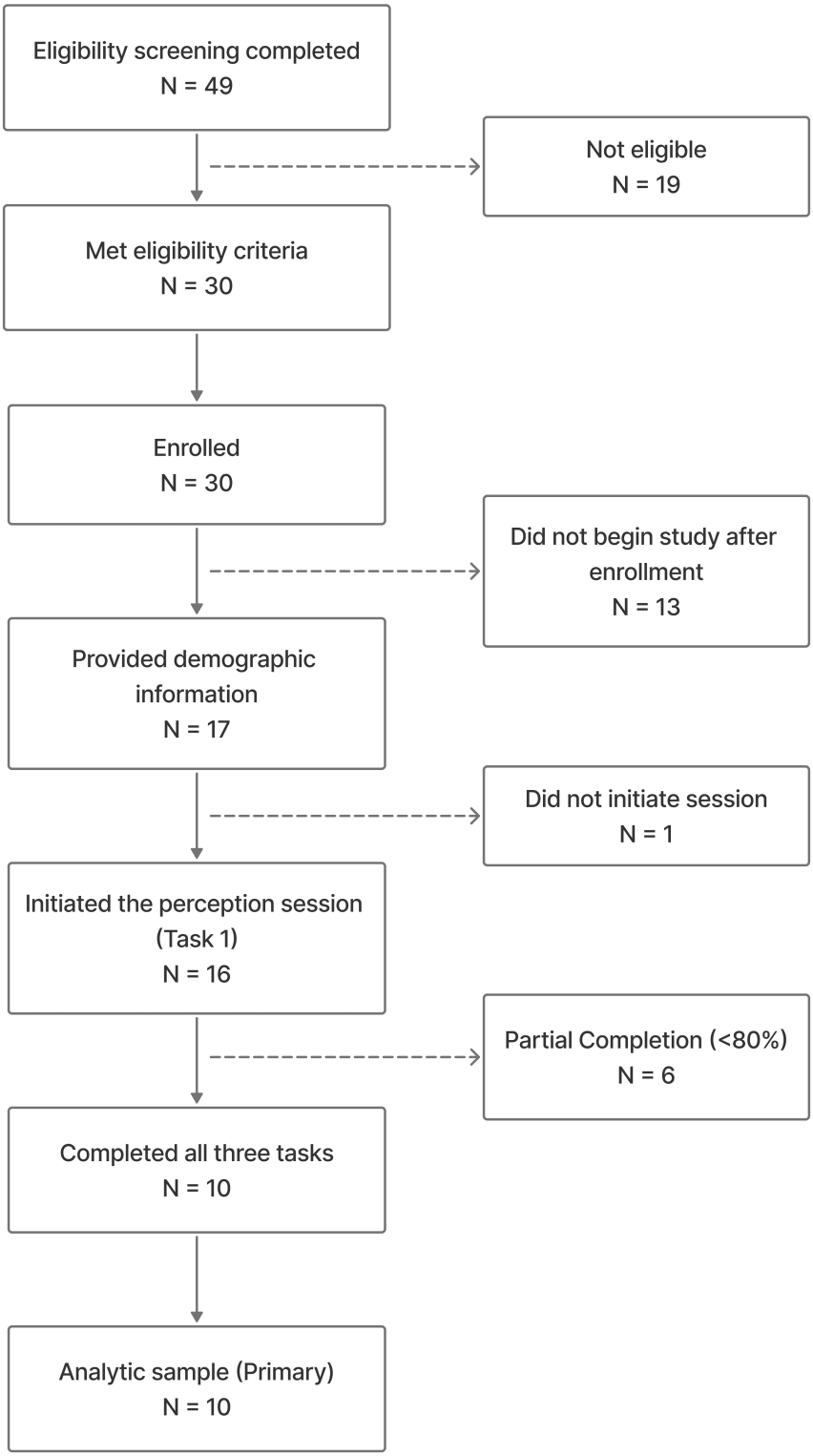


Figure 1

CONSORT flow diagram for Sub-Study 2. Of 49 participants who completed the eligibility screener, 30 met the eligibility criteria. Seventeen provided demographic information, 16 initiated the perception session, and 10 completed all three tasks. The analytic sample (n = 10) excludes six participants with partial task completion (below 80%). Exclusions at each stage were instrument-enforced per study protocol.

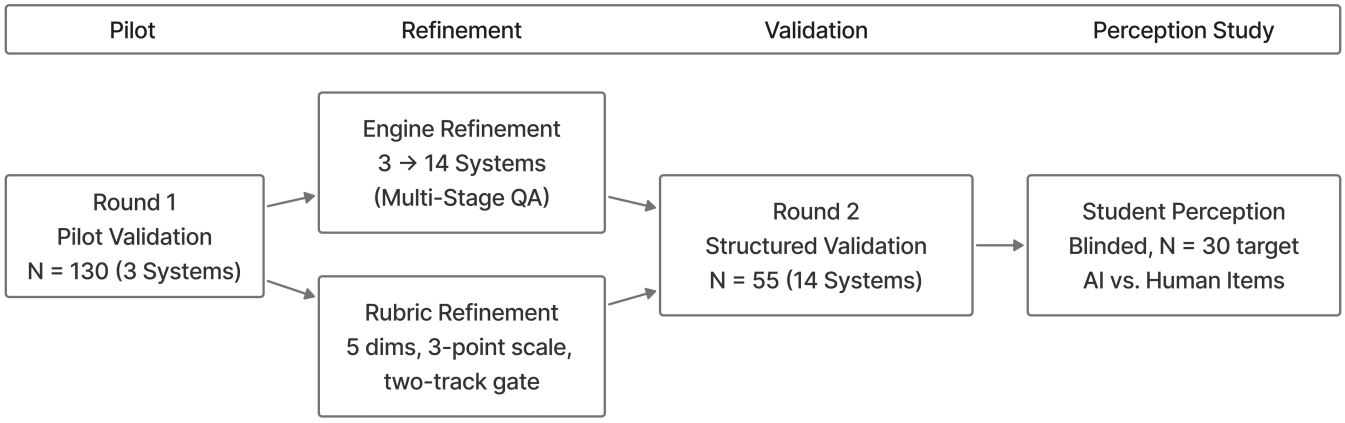


Figure 2

Study workflow. The validation instrument and the generation engine were developed and applied across two expert-validation rounds with a refinement stage in between. In Round 1 (pilot validation), four PA faculty reviewed 130 engine-generated items across three organ systems, and the concerns raised guided refinement of both the generation engine and the validation instrument. Round 2 (structured validation) then applied the refined five-dimension rubric and two-track quality gate: six PA faculty reviewed 55 items across 14 of the 15 PANCE organ systems. A preliminary blinded student perception study (n = 10 of 30 target) compared AI-generated and human-written items on answering accuracy, clarity, realism, and source detection. Arrows indicate the forward flow of the study; the pilot informed the refinement, which produced the content validated in Round 2.

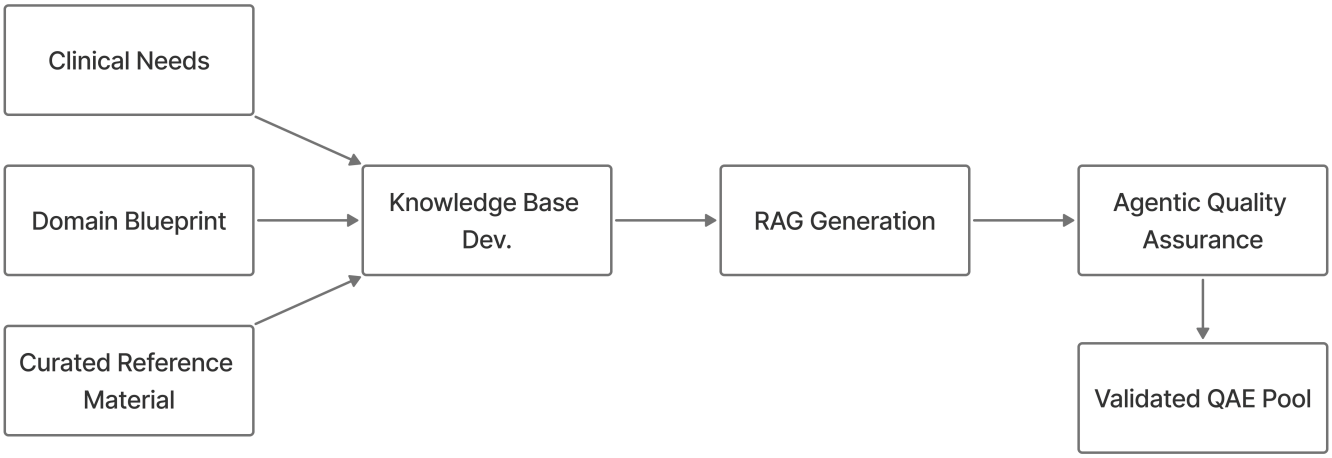


Figure 3

Overview of the generation engine for AI-generated clinical assessment item production. Clinical seeds, a domain blueprint, and curated reference materials are ingested through knowledge base development, retrieval-augmented generation, and agentic quality assurance, producing a validated QAE pool.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExpertvalidationofAIgeneratedclinicalassessmentitemsSetiawanSupplementary.docx