Evaluating LLM Workflows for Generating Clinical Communication Assessment Items: A Comparative Study with Subject-Matter Experts

Christopher Runyon, Peter Baldwin, Ian Micir, Kevin Frome
Stephanie Mann, Saed Rezayi, Keelan Evanini, Victoria Yaneva

NBME

(crunyon, pbaldwin, imicir, kfrome,
semann, srezayidemne, kevanini, vyaneva)@nbme.org

Abstract

Generative AI is increasingly used to accelerate assessment content development, yet its effectiveness for generating content used in complex assessment tasks for knowledge-rich domains such as medical education is unclear. This study evaluates automated LLM-supported workflows for generating patient-centered communication assessment items that allow students to practice their communication skills. We compared two content generation approaches—constrained linear and exploratory branching—each implemented with and without anchoring in vetted multiple-choice questions (MCQs). Ten subject-matter experts (SMEs) evaluated 80 communication items across six quality dimensions using structured rubrics. The constrained linear approach yielded better ratings than exploratory branching approaches, particularly for medical accuracy and alignment with learning objectives and patient-centered behaviors. MCQ anchoring did not improve medical accuracy. Only a minority of items met all criteria without requiring revision, and no items were unanimously approved by all SMEs. These findings underscore the importance of workflow design in LLM-supported assessment content generation, the continued need for human oversight, and the current limitations of automated content generation in medical education.

1 Introduction

Recent advances in generative AI have made AI-assisted content development one of the fastest-growing educational applications of this technology, with numerous examples reported in both research and operational contexts (see Bouguettaya et al. (2025) for an overview). Owing to their ability to generate coherent and grammatically well-formed narratives, generative models have been used primarily in language-focused assessment domains such as language learning (Attali et al., 2022; LaFlair et al., 2023) and reading comprehension (Lin and Chen, 2024; Säuberli and Clematide, 2024; Uto et al., 2023). In these settings, generative AI typically serves to accelerate item development even while extensive fairness and bias reviews and item performance analyses remain necessary to ensure that the resulting items function appropriately (Belzak et al., 2023).

In contrast, the application of generative AI to content development in knowledge-rich domains such as medicine, law, and engineering, among others, remains comparatively underexplored. These domains impose additional requirements beyond grammatical accuracy and narrative coherence. Generated content must accurately represent complex domain knowledge, including factual correctness (e.g., medical accuracy), currency (reflecting rapidly evolving knowledge bases), and domain plausibility (e.g., clinically realistic scenarios). Meeting these requirements poses a substantial generation challenge. Moreover, *evaluating* AI-generated content in knowledge-rich domains requires specialized expert judgment, making the design and execution of human evaluation studies complex and resource-intensive. As a result, robust methodological approaches for generating and validating assessment content in such domains require focused attention from the educational research community.

In this study, we investigate the use of generative AI (GPT5; OpenAI, 2025) to support content development in the domain of medical education. Specifically, we focus on developing and evaluating methods for generating clinical communication assessment items for the Communication Learning Assessment (CLA) tool. CLA is a mobile application designed to help learners practice clinical communication skills through interactions with virtual patients. The system presents learners with brief clinical scenarios and patient prompts, and learners record spoken responses that demon-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 153–173

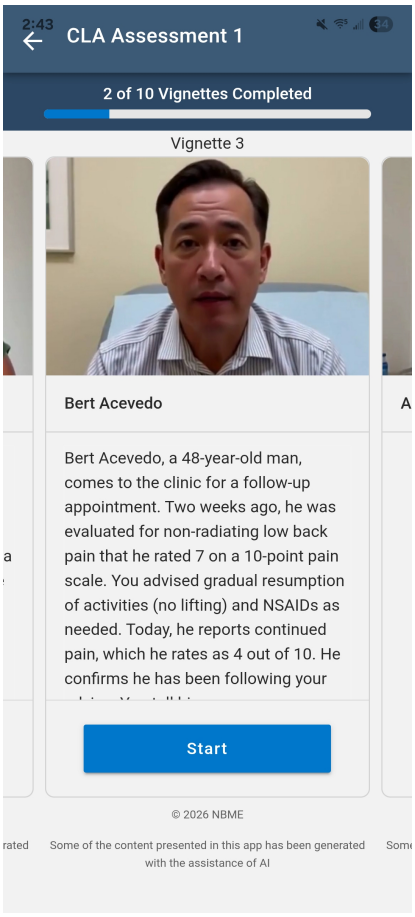

Figure 1: Screenshot of the Communication Learning Assessment tool.

strate their ability to use appropriate communication strategies (see Figure 1 for an illustration). The application then provides feedback to support skill development. For example, a learner may receive background information describing a clinical scenario such as a patient expressing concern over their weight-loss efforts plateauing (vignette background), followed by a patient statement or question (learner prompt). The learner is then asked to record a response that demonstrates their ability to use appropriate communication skills for that situation.

Developing suitable content for assessing clinical communication includes the generation and evaluation of multiple item components. To guide the reader's understanding of these, an example CLA item is shown in Appendix A.1. First, each item must include a medically-accurate and plausible clinical scenario appropriate for a target learner's level of training (e.g., first-year medical students). A given clinical scenario is called a *vignette*, as shown in Appendix A.1. The vignette must align with specific learning objectives (LOs) for the assessment (e.g., "respond to emotion"), and be designed to elicit appropriate contextualized communication behaviors. These communication behaviors, also shown in the example, are referred to as *patient-centered behaviors* (PCBs; e.g., "provide empathic reassurance that her fear is understandable and manageable"). In addition, the learning experience concludes with a reflection on two exemplary "gold standard" responses (*exemplars*) that demonstrate how to effectively and appropriately communicate within the given scenario, each of which integrate all the targeted PCBs. Exemplars are a crucial aspect of content development that require alignment with the scenario and the targeted LOs and PCBs while maintaining a high degree of realism. In addition to medical accuracy, plausibility, and appropriate difficulty targeting, all components of the item—the vignette, LOs, PCBs, and exemplars—must be internally consistent and must not reinforce or perpetuate unsupported demographic associations or stereotypes (i.e., links not grounded in established clinical evidence). CLA also requires high-quality content covering a range of clinical communication scenarios. Together, these requirements make the development of high-quality CLA content challenging.

This study makes the following original contributions in "jump-starting" the development of clinical communication vignettes via AI:

- Propose and compare two fully automated content-generation methods that reflect different conceptual approaches to producing interdependent content: constrained linear versus exploratory branching.
- Evaluate whether grounding the generation process in a vetted multiple-choice question (MCQ) improves the medical accuracy of the output.
- Develop detailed evaluation rubrics and guidelines tailored for assessing AI-generated clinical communication items.
- Recruit and train ten subject-matter experts to evaluate 80 AI-generated items, each consisting of a vignette, LOs, PCBs, and exemplars.
- Analyze the evaluation results to assess the extent to which AI can reduce human effort in item development and identify areas for improving the generation process.
- Provide a detailed discussion of the contributions, limitations, and ethical considerations of using generative AI to assist in clinical con-

tent development.

Before we introduce details about the current study, Section 2 presents related work in AI-assisted content development in medicine for item formats beyond MCQs.

2 Background

2.1 Related Research

A robust medical education program teaches learners to do more than select a correct answer to an MCQ. Students must also learn to reason through complex diagnostic scenarios, such as interactions with standardized patients in objective structured clinical examinations (OSCEs). While LLMs have been used productively in MCQ item writing by supporting item development workflows (Kıyak and Emekli, 2024; Artsi et al., 2024), complex performance tasks pose a greater challenge. Yet, it is *because* these tasks are typically more resource-intensive to develop than their MCQ analogs that makes them natural—although more difficult—candidates for LLM-assisted content generation (Cusimano et al., 1994; Kıyak and Emekli, 2024; Miri et al., 2024).

Recent literature has begun to examine the potential for LLMs to assist with the generation of performance task materials, especially those materials related to OSCE and standardized-patient (SP) tasks. Misra and Suresh describe how ChatGPT can support OSCE preparation—including drafting cases, SP training materials, and grading rubrics—while emphasizing that medical accuracy, security, and compliance require systematic human validation (Misra and Suresh, 2024). Zafar et al. provide practical guidance for using AI to generate high-quality OSCE stations, recommending outcome alignment, structured prompts, iterative refinement, and faculty oversight to manage hallucinations and construct-validity threats (Zafar et al., 2026). Complementing peer-reviewed studies, the AAMC has released a workflow for generating curriculum-aligned, role-locked virtual patient profiles for simulation, explicitly positioning AI generation as scalable but educator-supervised (Quiroga Velasquez, 2026). Moving from static materials to interactive performance tasks, Brugge et al. show that LLM-simulated history-taking encounters paired with AI-generated structured feedback can improve learners' clinical decision-making indicators, suggesting that LLMs may help generate and sustain practice-ready performance-assessment experiences when embedded in governed workflows (Brügge et al., 2024). Prior work has explored using LLMs to support communication tasks, but these models have primarily served as the delivery mechanism for the task and its feedback (Stamer et al., 2023). In contrast, the use of LLMs to generate the content for communication tasks themselves remains underexplored.

3 Method

3.1 Generation Methods

The four generation methods are outlined below: 1a, 1b, 2a, and 2b. The core distinction between the "a" and "b" versions of each method is what anchors the generation at the start. In the "a" versions, the workflow begins with a diagnosis and communication task extracted from a retired clinical MCQ from a medical licensure examination. In the "b" versions, the workflow also includes the full text of the associated MCQ. Because these seed MCQs were previously developed in conjunction with physicians, the intent is that clinically accurate scenario details and logic can be inherited by the CLA item within the workflow. Each generation method developed a set of 20 items for a total of 80 items to be reviewed by subject-matter experts (SMEs). SMEs evaluated all four versions of each item, which were presented in randomized order.

3.1.1 Methods 1a/1b

Method 1 treats item writing as a progressive, linear build, similar to drafting a single coherent case from start to finish[1]. It begins by establishing key guardrails for what the item should measure (the construct), either explicitly through encounter features defined up front (1a) or implicitly through the narrative constraints embedded in a retired MCQ stem (1b). The item is then generated in sequence, with each component checked for consistency with the selected learning objectives and intended tone/affect. Quality gates are applied throughout (e.g., accuracy and internal-consistency checks followed by revision as needed), and a final pass focuses on naturalness and readability. The build proceeds in the following steps, using the information from the preceding steps to generate subsequent content: identifying appropriate LOs based on the communication task and diagnosis → generating the communication vignette (including site of care and patient age) → generating the

[1]Method 1 used gpt5-2024-12-01-preview

PCBs → generating the exemplars → stereotype check, selection of remaining demographics, and final refinement. The prompts used for Method 1a are available in the Appendix A.2.

3.1.2 Methods 2a/2b

Method 2 was designed as a deliberate alternative to Method 1 and emphasizes an exploratory branching search and refinement rather than early commitment to a single narrative[2]. Instead of progressing along one path, it begins by generating multiple candidate cases using multiple agents (using the same LLM) and then converges through structured selection and improvement, such as ranking, justification, critique from different perspectives, and targeted revision. Early in the workflow, learning objective–mapped PCBs are used as a temporary scaffold while the vignette remains flexible, and the process relies on detailed component descriptions and repeated examples to guide both generation and critique. The workflow explores alternatives by branching across aspects of the vignette background before selecting and polishing the strongest candidate, and it delays introducing details such as site of care and patient demographics to reduce the chance that surface features steer construct specification too early. Overall, Method 2 treats alignment as something achieved through convergent optimization across the vignette, PCBs, and exemplars, rather than as a one-pass enforcement of early constraints.

3.2 Subject-Matter Expert Evaluation

Ten SMEs were recruited via NBME's Assessment Alliance network[3] to provide feedback on the items. The SMEs were selected based on their involvement in the teaching of communication skills to medical students, a demonstrated research history in this area, experience in developing assessments in medical education, or a combination of these factors; nine were licensed medical doctors and the other was a PhD RDN with expertise in communication and sociocultural aspects of health care.

The SMEs evaluated each item based on the following item quality dimensions: *medical accuracy, LO alignment, PCB alignment, Exemplar alignment, presence of harmful stereotypes,* and *difficulty appropriateness* (see Appendix A.3 for a full definition of each dimension). These dimensions were chosen because they reflected the most common areas of editorial work identified by the test development team responsible for communication item development.

For each dimension, SMEs were provided with a forced-choice Likert scale response (see Appendix A.4 for exact phrasing of the item-evaluation questions). For the first four item dimensions, the response options were about the degree of changes necessary for the accuracy / alignment, with options of "No Changes", "Minimal", "Moderate", "Major". Additional guiding language provided contextual information on response options. For stereotype inclusion, SMEs simply indicated if the vignette perpetuated any harmful stereotypes (Yes/No), and for difficulty appropriateness SMEs identified if the item was too easy, too difficult, or at the appropriate level for the student given the level of medical training targeted by the item.

It is important to note that in cases where an SME would judge a vignette to be medically inaccurate to the point of being not salvageable, they were permitted to opt out of providing ratings for subsequent item dimensions. This was a deliberate feature of the evaluation design rather than an incidental source of nonresponse. The intent was to avoid collecting downstream judgments for items that would not realistically proceed in an editorial workflow. For this reason, later-dimension analyses should be interpreted conditionally, that is, as characterizing items that satisfied the medical accuracy criterion.

3.3 Research Questions

- RQ1: How well does each generation method produce content that is aligned with each of the six item quality dimensions?
- RQ2: Does anchoring an item generation workflow with an MCQ improve the medical accuracy of the generated items?
- RQ3: How many items require no changes across all item dimensions, and thus could be directly used without human refinement?

3.4 Statistical Analysis

SME ratings of AI-generated content were analyzed using cumulative link mixed models (CLMMs; Christensen, 2019), an extension of ordinal logistic regression that accommodates clustered measures designs. CLMMs are appropriate when outcomes are ordinal categories with a natural ordering but spacing between levels is not assumed

[2]Method 2 used gpt5-2025-01-01-preview.

[3]https://www.nbme.org/research/research-collaborations/assessment-alliance

to be consistent; that is, the difference between "None" and "Minimal" is not assumed to be the same as the difference between "Moderate" and "Major".

CLMMs model the cumulative probability of observing a response at or below each category. The model estimates threshold parameters that define boundaries between adjacent categories on a latent continuous scale, allowing the data to determine category spacing rather than imposing assumptions. CLMMs extend standard ordinal regression by incorporating random effects, which account for systematic variation due to nested observations. In our design, all 10 raters evaluated all 80 case-method combinations (a fully-crossed design), introducing dependencies that would violate independence assumptions of standard regression. The CLMM addresses this by estimating random intercepts for raters and cases, capturing individual rater tendencies (e.g., overall harshness or leniency) and case-level variation in quality.

CLMMs also partition variance into components attributable to random effects (raters and cases) and residual variance, providing insight into sources of rating variability. Variance decomposition follows standard practice for logistic mixed models (Hedeker and Gibbons, 2006); residual variance is fixed at $\pi^2/3$ and random effect variances are estimated via maximum likelihood. The proportion of variance attributable to each random effect is computed as its estimated variance divided by the total (rater variance + case variance + $\pi^2/3$). The proportion of variance attributable to raters reflects consistency in how raters applied evaluation criteria: lower rater variance indicates that raters evaluated content similarly, while higher rater variance suggests systematic differences in stringency or leniency across raters. The proportion attributable to cases reflects genuine variability in content quality across the 20 diagnosis-communication task pairs common to all generation methods. Together, these variance components offer a complementary perspective on rating consistency alongside the fixed effects of primary interest.

Model coefficients are expressed as log-odds, with exponentiated values interpretable as odds ratios. An odds ratio greater than 1 indicates higher odds of receiving a more severe rating relative to the reference method; values less than 1 indicate lower odds. The proportional odds assumption, required for CLMMs, states that predictor effects (generation methods) are consistent across category thresholds; this assumption was evaluated using likelihood ratio tests. Models were fit using the `ordinal` package in R with Laplace approximation for random effects integration (Christensen, 2019).

The cumulative link mixed model (Christensen, 2018) is specified as:

$$\text{logit}[P(Y_{ijk} \leq c)] = \theta_c - (\beta_k + u_i + v_j) \quad (1)$$

where Y_{ijk} is the ordinal rating from rater i for case j under generation method k; θ_c are threshold parameters separating adjacent response categories, constrained such that $\theta_1 < \theta_2 < \cdots < \theta_{C-1}$; β_k is the fixed effect for method k (with one method serving as reference); and $u_i \sim N(0, \sigma^2_{\text{rater}})$ and $v_j \sim N(0, \sigma^2_{\text{case}})$ are crossed random intercepts for raters and cases, respectively.

4 Results

4.1 Medical Accuracy

Table 1: Medical Accuracy

Method	OR (95% CI)	p
Method 1b	1.28 (0.80, 2.04)	0.29
Method 2a	2.18 (1.39, 3.42)	<0.001
Method 2b	2.98 (1.90, 4.67)	<0.001

We fit a cumulative link mixed model estimating the generation method as a fixed effect and random intercepts for raters and cases. Method 1a served as the reference for all analyses. Table 1 shows that items generated by Methods 2a and 2b were significantly more likely to receive ratings indicating they required additional changes for medical accuracy. Method 1b did not significantly differ from Method 1a. The proportional odds assumption was supported ($\chi^2(6) = 5.99, p = .42$). Descriptive statistics for expert ratings for each generation method across all item dimensions are reported in Appendix A.5.

About one-third of the variance in ratings (33.1%) was due to systematic differences between raters, indicating notable variability in rater stringency. Case-level differences accounted for only 8.3% of variance, suggesting that content quality was relatively consistent across cases. The majority of variance (58.6%) was residual, reflecting rating-level variability not captured by rater or case effects.

4.1.1 Effect of Using Full MCQs on Medical Accuracy

Table 2: Effect of MCQ on Medical Accuracy

IV	OR (95% CI)	*p*
Method 2	2.18 (1.39, 3.42)	<0.001
"b" Methods	1.28 (0.81, 2.04)	0.29
Interaction	1.07 (0.57, 1.98)	0.84

To isolate the effect of incorporating MCQs into the generation workflow (RQ2), a separate CLMM was estimated with generation approach (Method 1 vs. Method 2), starting point (diagnosis and communication task only—the "a" methods—vs. full MCQ item content—the "b" methods), and their interaction included as fixed effects. Method 1 and the "a" generation methods were the reference groups. For the main effect of method, Method 2 produced content requiring significantly more revision than Method 1 (Table 2). The main effect for the inclusion of the full MCQ item content did not significantly affect medical accuracy ratings, and no interaction was observed, indicating that the advantage of Method 1 over Method 2 for medical accuracy was consistent regardless of generation workflow starting point. Variance decomposition indicated that rater tendencies (33.1%) exceeded case differences (8.3%), with 58.6% residual variance. The proportional odds assumption was supported for both generation method ($\chi^2(1) = 0.06, p = .80$) and starting point ($\chi^2(1) = 0.07, p = .80$).

4.2 Learning Objective Alignment

Table 3: Learning Objective Alignment

Method	OR (95% CI)	*p*
Method 1b	1.14 (0.45, 2.84)	0.78
Method 2a	14.25 (6.51, 31.18)	<0.001
Method 2b	4.73 (2.12, 10.57)	<0.001

For learning objective alignment (Table 3), generation Methods 2a and 2b showed significantly poorer alignment than Method 1a. Method 1b did not differ significantly from Method 1a. Variance decomposition indicated that rater tendencies (22.9%) and case differences (23.5%) contributed approximately equally to random variance, with 53.6% residual variance. The proportional odds assumption was supported ($\chi^2(6) = 8.01, p = .23$).

4.3 Patient-Centered Behavior Alignment

Table 4: Patient-Centered Behavior Alignment

Method	OR (95% CI)	*p*
Method 1b	0.56 (0.30, 1.04)	0.07
Method 2a	5.57 (3.29, 9.41)	<0.001
Method 2b	3.56 (2.09, 6.06)	<0.001

Here, the proportional odds assumption showed evidence of violation ($\chi^2(3) = 9.43, p = .02$). Given the modest departure and to maintain consistency with analyses of other outcomes, we retained the proportional odds model; however, results should be interpreted with appropriate caution.

For PCB alignment, Table 4 shows that content from Method 2a and Method 2b received significantly worse ratings than Method 1a content. Method 1b showed a trend toward better performance than Method 1a, though this was not statistically significant. Variance decomposition indicated that case differences (16.5%) exceeded rater tendencies (8.5%), with 75% residual variance, suggesting that raters applied criteria consistently while cases varied meaningfully in quality. As a reminder, the proportional odds assumption was violated and readers are cautioned against over interpreting these findings.

4.4 Exemplar Alignment

Table 5: Exemplar Alignment

Method	OR (95% CI)	*p*
Method 1b	1.36 (0.90, 2.07)	0.14
Method 2a	0.81 (0.51, 1.25)	0.34
Method 2b	0.74 (0.47, 1.18)	0.21

For exemplar alignment, no significant differences were observed between development methods (all $p \geq .14$). Variance decomposition revealed that case differences accounted for only 0.9% of variance, suggesting that content was uniformly aligned with exemplars regardless of case or method. Rater tendencies contributed 16.3% of variance, with residual variance comprising 82.8%. The proportional odds assumption was again supported ($\chi^2(6) = 9.38, p = .15$).

4.5 No Stereotypes

Stereotype inclusion was rare, flagged in only 13 of 723 ratings (1.80%) across all methods. Due to this low base rate, formal inferential comparisons were not conducted. Descriptively, rates were similarly low across methods (Method 1a: 2.7%;

Method 1b: 2.7%; Method 2a: 1.1%; Method 2b: 0.6%). Although stereotype flags were too rare to permit formal comparisons across methods, the potential consequences of stereotype inclusion are serious, underscoring the importance of maintaining vigilance throughout both item generation and subsequent human review.

Notably, the study design may have contributed to some stereotype flags. Because experts evaluated four versions of each case with potentially similar demographic characteristic combinations (race, gender, diagnosis), this may have prompted some raters to perceive stereotyping that reflected accumulated pattern recognition rather than content within any single version.

4.6 Appropriate Difficulty

Table 6: Appropriate Difficulty

Gen Method	OR (95% CI)	*p*
Method 1b	0.86 (0.48, 1.54)	0.62
Method 2a	0.66 (0.37, 1.16)	0.14
Method 2b	0.52 (0.29, 0.92)	0.02

Response options for difficulty were "too easy," "appropriate," and "too difficult," providing granular insight into the generation process. However, our primary interest was whether each method produced items at an appropriate difficulty level. We therefore collapsed responses into a binary outcome: appropriate versus off-level (combining "too easy" and "too difficult"). Across all generation methods, the majority of items were judged to be at an appropriate difficulty level (80.8%). When items were off-level, they were more commonly rated as too difficult (17.5%) than too easy (1.7%).

A generalized linear mixed model with binomial family was fit with random intercepts for raters and cases (Table 6). Method 2b produced content significantly less likely to be rated at the appropriate difficulty level compared to Method 1a; Method 1b and Method 2a did not differ significantly from Method 1a. Variance decomposition indicated that rater tendencies (16.1%) exceeded case differences (8.3%), with 75.6% residual variance. Model assumptions were evaluated and met: no overdispersion (ratio = 0.83), random effects showed no significant departures from normality, and residual distributions were acceptable.

4.7 Full Item Evaluation

Table 7: All Alignment

Gen Method	OR (95% CI)	*p*
Method 1b	0.63 (0.40, 1.01)	0.06
Method 2a	0.40 (0.25, 0.66)	<0.001
Method 2b	0.38 (0.23, 0.52)	<0.001

To evaluate the total item quality (RQ3), a composite variable was created for each item / rater combination that indicated if a particular rater felt that item needed no corrections across all 6 dimensions (medically accurate; LO, PCB, and exemplar all aligned; no stereotype; appropriate difficulty). Of the 800 item/generation method/rater combinations, only 222 (27.8%) met this criterion. A generalized linear mixed model with binomial family (outcome was binary: all aligned vs. not) was fit with random intercepts for raters and cases. Table 7 indicates that Method 2a and Method 2b had significantly lower odds of meeting all criteria compared with Method 1a. Method 1b also demonstrated a lower success rate; however, this difference was not statistically significant. Variance decomposition indicated that rater tendencies (27.9%) exceeded case differences (4.7%), with 67.4% residual variance. Model assumptions were again evaluated and met: no overdispersion (ratio = 0.94), random effects showed no significant departures from normality, and residual distributions were acceptable.

This analysis does not capture the extent of rater agreement that all item dimensions were aligned within a given item–generation method pair. To assess this directly, we summed the number of the 10 SMEs who agreed that a given item required no changes for each of the 80 item–method combinations. Table 8 shows the distribution of rater agreement across the 80 item-method combinations. For example, 11 item-method combinations received no full alignment ratings from any rater, while only 1 combination was rated as fully aligned by 7 raters—the maximum observed; no more than 7 raters agreed that an item-method combination required no edits. An average of 2.78 raters agreed all criteria were met for any item-method combination. Figure 2 in the Appendix shows the count of raters indicating full alignment by case and generation method.

Table 8: Distribution of Items by Number of Raters Indicating Full Alignment

Raters in Full Alignment	Count	Percent (%)
0	11	13.75
1	15	18.75
2	13	16.25
3	10	12.5
4	12	15
5	12	15
6	6	7.5
7	1	1.25
8	0	0
9	0	0
10	0	0

5 Discussion

This study examined the viability of fully automated generative AI workflows for producing clinical communication items suitable for formative assessment in medical education. Across three research questions, the findings provide converging evidence that generation method matters substantially for item quality, that anchoring generation in vetted multiple-choice questions does not meaningfully improve medical accuracy, and that only a modest proportion of generated items are currently suitable for direct use without human refinement, suggesting human review will remain an important aspect of AI-assisted content generation in the medical communication domain.

The results for RQ1 show consistent differences between generation methods in their ability to produce content aligned with the six item quality dimensions. Constrained linear methods (Method 1a and 1b) outperformed exploratory branching methods (Method 2a and 2b) across several core dimensions, most notably medical accuracy, learning objective alignment, and patient-centered behavior alignment. These dimensions are foundational to the validity of clinical communication assessments, as they ensure that scenarios are both medically plausible and capable of eliciting the intended communication behaviors.

In contrast, exemplar alignment showed no significant differences across methods. This finding suggests that once a successful vignette has been generated, LLMs are generally capable of producing exemplar responses that require very little revision, if any.

Stereotype flags were uncommon across all workflows. Item difficulty targeting was generally insensitive to generation method; however, Method 2b items were more likely to be judged as falling outside the intended difficulty range, most often being too difficult.

Collectively, these results suggest that while certain classes of failure were less likely to arise, medical accuracy and alignment among items' component parts remain highly sensitive to the structure of the generation workflow.

RQ2 addressed whether anchoring generative workflows in vetted multiple-choice questions improves the medical accuracy of the generated clinical communication items. Contrary to expectations, no evidence was found that incorporating full MCQs as starting points improved medical accuracy. This null effect held across both linear and branching generation methods, and no interaction between anchoring and method was observed.

These findings suggest that the presumed benefits of grounding generation in validated assessment content may not readily transfer from selected-response contexts to more open-ended performance tasks. Although MCQs encode medically correct information, their structure and purpose differ substantially from those of clinical communication items. Expanding an MCQ into a fully developed communication item requires substantial inference and contextual detail that the MCQ format does not provide or constrain. As a result, anchoring on MCQs may offer less practical value than anticipated for supporting accuracy in complex generative tasks.

The third research question (RQ3) focused on the practical yield of the generation workflows: how many items met *all* quality criteria without requiring human revision. Across methods, only a small minority of items satisfied this stringent standard, and items produced by branching generation workflows were significantly less likely to do so than those produced by linear workflows. Importantly, the number of raters agreeing that an item-method combination required no changes varied widely and there were no items for which all raters (or even 8/10 raters) agreed that all elements were aligned. These results point to the persistent limitations of fully automated generation for high-stakes or high-fidelity assessment content in medicine, while also revealing meaningful efficiency differences across workflows.

Importantly, the composite outcome integrates judgments across all six quality dimensions, reflecting the holistic standard that assessment developers must meet in practice. That some items—particularly from constrained linear work-

flows—were judged acceptable without revision indicates that AI-assisted generation can reduce human development effort, even if it does not eliminate the need for expert review.

Overall, these findings suggest that generative AI can play a productive but circumscribed role in the development of clinical communication assessment content. Workflow design emerges as a central determinant of quality, with simpler, more structured generation approaches outperforming more complex branching strategies. The results further caution against assuming that validated content from one assessment format can be straightforwardly repurposed to support generation in another.

From a broader perspective, this study illustrates the importance of evaluating AI-generated assessment content not only for surface-level plausibility but also for deep alignment among medical context, learning objectives, and elicited behaviors, when developing complex performance tasks. As generative models continue to improve, careful attention to workflow structure and evaluation criteria will remain essential for ensuring that efficiency gains do not come at the expense of validity.

6 Limitations and Ethical Considerations

Several limitations deserve comment. First, the evaluation relied on SME judgment across six dimensions of item quality. Although SMEs completed an orientation session and practice review before formal rating began, a single training session may not have been sufficient to ensure shared interpretations of all rubric categories. The nontrivial rater-level variability observed in several analyses may, in some cases, reflect construct-relevant differences in professional judgment among SMEs; it may also reflect differences in rating severity or rubric interpretation. Interestingly, the medical accuracy dimension showed the most variance across raters, although this dimension is arguably the most objective of all dimensions. Future work would benefit from additional training and rater calibration before the formal review task begins, both to refine rubric wording and to ensure a stronger shared understanding of the evaluation criteria.

Relatedly, we report variance decomposition rather than relying exclusively on traditional inter-rater agreement statistics because our data structure made single-summary agreement coefficients poorly aligned with our inferential goals. The combination of multi-point ordinal scales, a large rater panel, and ratings concentrated in modal categories created conditions under which chance-corrected agreement statistics can be difficult to interpret. Kappa-family statistics, including Fleiss's kappa, are sensitive to prevalence and marginal distributions and can yield low estimates even when observed agreement is high (Feinstein and Cicchetti, 1990). Krippendorff's alpha is more flexible with respect to number of raters and scale type, but it also summarizes reliability as observed disagreement relative to expected disagreement derived from the overall distribution of ratings, which can be difficult to interpret when rating distributions are highly restricted or skewed (Hayes and Krippendorff, 2007). Gwet's AC2 is less sensitive to prevalence effects than kappa-based statistics (Gwet, 2002), but none of these coefficients separate systematic rater severity differences from residual disagreement about item quality. Variance decomposition better matched our goals because it partitioned variation into method, item, rater, and residual components, allowing us to evaluate method differences while explicitly modeling rater variability. This approach allowed us to examine whether raters exhibited consistent relative patterns across methods despite differences in absolute rating standards. Future work could develop or evaluate agreement indices better suited to ordinal, multi-rater designs with skewed response distributions and systematic rater severity effects.

Second, the review process followed a planned conditional evaluation design. SMEs were permitted to stop rating later dimensions when an item was judged medically inaccurate to the point of being unsalvageable, on the rationale that downstream judgments are of limited practical value for content unlikely to advance in an editorial workflow. This approach increases the practical relevance of the evaluation by aligning it with how content would be screened in operational development, but it also changes the scope of inference for later outcomes. Certain constructs—such as vignette alignment—are not meaningfully defined for medically unsalvageable cases, as alignment is a property of viable cases only. Accordingly, later evaluations are best interpreted not as judgments conditional on medical accuracy, but rather as judgments about medically accurate items.

Third, despite clear workflow differences across several outcomes, each method produced only a modest number of items that required no revisions. Therefore, the absence of a measurable benefit from

anchoring generation in retired MCQs should be interpreted as a lack of evidence for improvement under the present design rather than as definitive evidence that anchoring cannot be useful. The usefulness of anchoring likely depends in part on the source items selected: some MCQs may provide a richer substrate for communication-focused vignette generation than others, depending on the amount of clinically relevant context available and the extent to which that content helps the LLM perform the generation task. Future research could consider broader and more systematic sampling of MCQ items. Likewise, statistically significant differences across workflows should not be taken to imply that the absolute quality of generated items was high or that the observed advantages would necessarily translate into large practical gains in operational item development.

Fourth, the present comparison is bounded by the specific model and implementation used in the study. All items were generated using a single LLM (GPT5) and a fixed set of workflow designs. The findings should therefore not be generalized to all models, prompting strategies, or medical education contexts. The observed advantage of the constrained linear workflow over the exploratory branching workflow may reflect a broader principle—that more tightly constrained generation is beneficial when item components must remain strongly aligned—but it may also depend on particulars of the prompts, revision logic, and model behavior used here. More generally, differences across methods may reflect implementation details as much as any intrinsic advantage of one workflow family over another. Replication across models and alternative implementations of the same workflow concepts will be necessary to determine which conclusions are robust and which are due to specific characteristics of a given implementation. For this, it may be beneficial to include factors beyond content quality that nevertheless have practical importance such as the relative costs and run times across methods. In addition, the study relied on private LLM deployments that did not log data or use prompts for model training. Such secure deployment options may not be available in all settings, which could limit the feasibility of similar workflows when safeguarding assessment content is essential.

These limitations also point to several ethical considerations. Most importantly, the fluency of generated clinical content should not be mistaken for evidence that the content is adequate for a particular assessment purpose. Validity evidence remains necessary to support the intended interpretations and uses of any assessment. Fairness considerations are likewise central. Although stereotype flags were too rare to support formal comparisons across methods, even infrequent instances can be consequential, particularly in educational materials involving patient demographics, diagnoses, and communication challenges. Ethical oversight should therefore extend beyond item-level review to consideration of aggregate patterns in representation across generated content pools. As a whole, these considerations suggest that the most defensible near-term role for generative AI in this context is to support early drafting while preserving expert review, fairness screening, and institutional control over source materials and final content decisions.

References

Yaara Artsi, Vera Sorin, Eli Konen, Benjamin S. Glicksberg, Girish Nadkarni, and Eyal Klang. 2024. Large language models for generating medical examinations: systematic review. *BMC Medical Education*, 24:354.

Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077.

William CM Belzak, Ben Naismith, and Jill Burstein. 2023. Ensuring fairness of human-and ai-generated test items. In *International conference on artificial intelligence in education*, pages 701–707. Springer.

Sirine Bouguettaya, Francesco Pupo, Min Chen, and Giancarlo Fortino. 2025. A meta-survey of generative ai in education: Trends, challenges, and research directions. *Big Data and Cognitive Computing*, 9(9).

Emilia Brügge, Sarah Ricchizzi, Malin Arenbeck, Marius Niklas Keller, Lina Schur, Walter Stummer, Markus Holling, Max Hao Lu, and Dogus Darici. 2024. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC medical education*, 24(1):1391.

Rune Haubo B Christensen. 2018. Cumulative link models for ordinal regression with the r package ordinal. *Submitted in J. Stat. Software*, 35:1–46.

Rune Haubo Bojesen Christensen. 2019. *ordinal: Regression Models for Ordinal Data*. R package version 2019.12-10.

Michael D. Cusimano, R. Cohen, W. Tucker, J. Murnaghan, R. Kodama, and R. Reznick. 1994. A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). *Academic Medicine*, 69(7):571–576.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Kilem L. Gwet. 2002. Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment*, 2:1–9.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Donald Hedeker and Robert D. Gibbons. 2006. *Longitudinal Data Analysis*. Wiley, Hoboken, NJ.

Yavuz Selim Kıyak and Emre Emekli. 2024. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgraduate Medical Journal*, 100(1189):858–865.

Geoff LaFlair, Kevin Yancey, Burr Settles, and Alina A von Davier. 2023. Computational psychometrics for digital-first assessments: a blend of ml and psychometrics for item generation and scoring. In *Advancing natural language processing in educational assessment*, pages 107–123. Routledge.

Zhiqing Lin and Huilin Chen. 2024. Investigating the capability of chatgpt for generating multiple-choice reading comprehension items. *System*, 123:103344.

Kheizaran Miri, Tahere Sarboozi Hoseinabadi, Ali Yaghobi, Sadaf Kholosi, and Mehdi Miri. 2024. Cost management analysis of objective structured clinical examination (OSCE): guide to the universities of medical sciences. *BMC Medical Education*, 24:1241.

Sanghamitra M. Misra and Srinivasan Suresh. 2024. Artificial intelligence and objective structured clinical examinations: Using ChatGPT to revolutionize clinical skills assessment in medical education. *Journal of Medical Education and Curricular Development*, 11:1–6.

OpenAI. 2025. ChatGPT-5. `https://chat.openai.com/`. Large language model.

Adriana Quiroga Velasquez. 2026. Medical education case generator GPT workflow. Association of American Medical Colleges (AAMC), Advancing AI Resource Collection. Last updated February 2026.

Andreas Säuberli and Simon Clematide. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)@ LREC-COLING 2024*, pages 22–37.

Tjorven Stamer, Jost Steinhäuser, and Kristina Flägel. 2023. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. *Journal of Medical Internet Research*, 25:e43311.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 119–129.

Imran Zafar, Hadeel Aboueisha, Ibrahim Elhassan, Fouzia Shersad, Suleyman Ayhan Caliskan, Asma Fatima Syeda, Mohammed Al-Houqani, and Mohi Eldin Magzoub. 2026. Ten tips for utilizing AI to generate high quality OSCE stations in medical education. *Frontiers in Medicine*, 12:1744657.

A Appendix

A.1 Generated CLA Example Item

Field	Generated Content
Gender	cisgender woman assigned female sex at birth
Race	WHITE White or Caucasian
Affect	fear (fear, anxiety, insecurity, inferiority, panic)
SiteOfCare	Emergency: Urgent treatment center
Diagnosis	asthma
Ctask	Benefits and risks of a medication
Vignette	Megan Porter, a 35-year-old woman, comes to the urgent treatment center with a 1-week history of chest tightness and nighttime dry cough. Over the past 3 months, she had two emergency visits for acute shortness of breath treated with bronchodilators. She reports frequent upper respiratory infections, takes no medications, and has a 10-year one-pack-per-day smoking history. After confirming normal vital signs, a clear chest x-ray, and bilateral end-expiratory wheezes, you explain that the findings are consistent with asthma, advise smoking cessation, and recommend a daily controller inhaler to reduce airway inflammation, outlining expected benefits and potential side effects. She fidgets, appearing anxious, and asks,
Prompt	"I'm nervous about starting this inhaler. I smoke; how will the inhaler help breathing at night and how can you help me avoid side effects?"
CaseExemplar1	I hear this makes you anxious. A daily controller inhaler, a preventer medicine, reduces swelling in your airways. It helps calm nighttime cough and tightness. It also lowers your risk of flare-ups, meaning asthma attacks. It can still improve your breathing at night even if you smoke. To keep side effects low, we will check your inhaler technique. You will use a spacer, a tube that helps more medicine reach your lungs, and rinse your mouth after each dose. We will start with the lowest dose that keeps you well, and follow up to adjust. It is okay to feel nervous. Your reaction makes sense, and we can manage this safely together so you feel comfortable starting. What worries you most about starting the inhaler?
CaseExemplar2	Thanks for telling me that. Starting a new inhaler can feel unsettling. A daily controller inhaler calms airway swelling, so it should ease your nighttime chest tightness and dry cough, and lower your risk of flare-ups. Even if you smoke, it should still help your breathing. We can keep side effects low by going over your technique together, using a spacer (a small attachable tube), rinsing your mouth after each use, starting at the lowest dose that works, and checking in to adjust based on how you feel. We can take it step by step. What worries you most about starting this inhaler?
LO1	Solicit the patient's perspective

Field	Generated Content
PCB1	Ask what worries her most about starting the inhaler.
LO2	Accurately explain the nature of the problem and/or its management
PCB2	Explain that daily controller medicine reduces airway swelling, nighttime symptoms, and flare-ups.
LO3	Address patient emotion
PCB3	Provide empathic reassurance that her fear is understandable and manageable.
SME comments about medical accuracy	"Might be nice to confirm her vitals as stable. It's unclear how sick she is."
SME comments about learning objective alignment	"The vignette and prompt alludes to addressing how to minimize side effects. This is not clearly addressed in LO/PCB. I actually think this aspect of the case may be too complex anyway so would modify the prompt and exemplar from asking this for this level of learner." "It would be more appropriate for a student to practice smoking cessation counseling in this scenario than teach about inhaler use, which is beyond their scope of knowledge."
SME comments about pcb alignment	None.
SME comments about exemplar alignment	"Might be a bit much to expect a student at this level to know how to minimize side effects." "Ex 2 better than Ex 1, especially with the mention of being anxious which feels a touch judgmental." "Exemplar 2 could provide more direct empathic reassurance to the patient."
SME comments about appropriate difficulty	"Not sure– I think knowing how to minimize side effects could be too advanced." "Would not expect the student to know how to counsel on inhaler use, but other than that it's fine." "I think this one finally gets it all correct!"

Note. LO = Learning Objective; PCB = Patient-Centered Behavior; SME = Subject Matter Expert.

A.2 Prompts for Workflow 1a

Initial Prompt: You are an expert clinical communication content designer. You are building content for a communication tool that provides medical professionals the opportunity to practice their communication skills. The content build is occurring in a step-by-step process, and this is just one step in the process. Do not generate any other case content beyond what is requested here. Do not modify, expand, or infer additional conditions, as those will be taken care of in subsequent steps.

I will provide you with a patient diagnosis and a communication task. The communication task is the purpose of the physician-patient communication (e.g., explaining the purpose of a vaccination, information about a low-risk procedure or medication, speak with the patient about a common or easily treatable diagnosis).

In this step you will review the diagnosis and communication task, and your task is to review this information and generate 3 fields: - AGE : The plausible age range of a patient with this particular diagnosis that is appropriate for the communication task, drawn from a lower bound of 18 years old and upper bound of 90 years old. The age range you should return should reflect the most common ages where this diagnosis and communication task might occur. - GENDER : The possible sex of the patient, either MAN, WOMAN, or EITHER. For this exercise we can assume the patient is cisgendered. - Site of Care : The possible locations where this communication task is taking place. Choose only those locations that are most common given the diagnosis and the communication task. The possible options are [Specific Case Guidelines]. Choose all sites of care that are applicable.

The structure of your output should be as follows: DIAGNOSIS: the exact diagnosis provided CTASK: the communication task as provided AGE: The plausible age range of the patient GENDER: The plausible gender(s) for the patient SOC: The plausible site(s) of care where the communication may be taking place Here are the inputs you need: DIAGNOSIS: {diag} CTASK: {comtask}

—

LO prompt (Step 2): "You are an expert clinical communication content designer. You are building content for a communication tool that provides medical professionals the opportunity to practice their communication skills. The content build is occurring in a step-by-step process, and this is just one step in the process. Do not generate any other case content beyond what is requested here. Do not modify, expand, or infer additional conditions, as those will be taken care of in subsequent steps. I will provide you with a patient diagnosis and a communication task. The communication task is the purpose of the physician-patient communication (e.g., explaining the purpose of a vaccination, information about a low-risk procedure or medication, speak with the patient about a common or easily treatable diagnosis). I will also provide you with: a plausible age range (AGE) for the patient; the plausible gender for the patient (GENDER); and the plausible site(s) of care where the communication task is taking place (SOC). Here is that base case information: {step1output}

You have 2 tasks: *TASK 1* Identify the learning objectives that would most naturally occur in the communication scenario given the diagnosis and plausible age, gender, and site of care. You are to select 2-5 learning objectives from the following list. The list is in the following format:[Specific Case Guidelines] REQUIREMENTS: *It is important that I am able to distinctly measure each learning objective. However, some of the learning objectives are similar to one another, and, if included for the same communication practice item, it will be difficult to distinguish between what behaviors are exemplifying the different learning objectives. *Similar to the previous requirement, ensure that the learning objectives are not all from the same learning objective CATEGORY. There will be situations when multiple learning objectives from the same category are appropriate, given they can be distinctly measured by different observable behaviors by the physician. After choosing the learning objectives from the following list, review them to ensure that they meet these criteria. Learning objectives: {thelos}

TASK 2 Identify the most appropriate patient affects that the patient might plausibly be exhibiting given the working diagnosis and the communication task. REQUIREMENT: *If the communication task is not about addressing the patient's emotion, only pick possible affects that are more subtle or neutral. The patient should not have a strong affect if the diagnosis and communication task do not indicate it.* Here are the possible patient emotions: {pt_emo}

OUTPUT STRUCTURE The structure of your output should be as follows: DIAGNOSIS: the

exact diagnosis provided CTASK: the communication task as provided AGE: the age range of the patient GENDER: the plausible genders of the patient SOC: the plausible site(s) of care of the patient LOS: the 2-4 identified learning objectives (* new information to be amended to the base case information) AFFECT: the possible affect(s) of the patient, given the other case information (* new information to be amended to the base case information)

Generate that output.

—

Base Case Details Generation (Step 3): You are an expert clinical communication content designer. You are building content for a communication tool that provides medical professionals the opportunity to practice their communication skills. The content build is occurring in a step-by-step process, and this is just one step in the process. Do not generate any other case content beyond what is requested here. Do not modify, expand, or infer additional conditions, as those will be taken care of in subsequent steps. I will provide you with a patient diagnosis and a communication task. The communication task is the purpose of the physician-patient communication (e.g., explaining the purpose of a vaccination, information about a low-risk procedure or medication, speak with the patient about a common or easily treatable diagnosis).

INPUTS I will provide you with: DIAGNOSIS : the patient's diagnosis; CTASK : the communication task; AGE : a plausible age range for the patient; GENDER : the plausible gender(s) for the patient; SOC : the plausible site(s) of care where the communication task can plausibly take place; LOS : the learning objectives that have been identified for the communication scenario; AFFECT : the plausible affect(s) for the communication scenario. Here is that base case information: {step2output} Here are additional descriptions of the selected learning objectives: {focallos}

YOUR TASK Your task is to generate the main 'case bundle'. This consists of 4 components. I will describe each of the components in more detail, and then provide examples of case bundles developed by subject matter experts. The components that you generate should adhere to the descriptions below and follow the same general structure (form, content) as the examples.

CASE BUNDLE COMPONENT DESCRIPTIONS

VIGNETTE: [Specific Case Guidelines]

BACKGROUND: [Specific Case Guidelines]

ENCOUNTER POINT: [Specific Case Guidelines]

PROMPT: [Specific Case Guidelines]

IMPORTANT DETAILS: [Specific Case Guidelines]

Length and style targets: [Specific Case Guidelines]

Flow and coherence checks: [Specific Case Guidelines]

Before finalizing each case, review the output for the following details: [Specific Case Guidelines]

EXAMPLES Here are some example case bundles that have been developed. Use these to help guide the content and structure of the case bundle components: {examples}

OUTPUT STRUCTURE The output should be the 4 case components as follows: VIGNETTE: the generated vignette BACKGROUND: the generated background ENCOUNTER_POINT: the generated encounter point PROMPT: the generated prompt

Do not include any additional case details except for the 4 case bundle components. Generate that output.

—

Patient-Centered Behaviors (Step 4): You are an expert clinical communication content designer. You are building content for a tool that helps medical professionals practice communication skills.

Task: (1) Identify Patient-Centered Behaviors (PCBs) for the provided case. PCBs are discrete, case-specific, observable clinician behaviors that demonstrate how each Learning Objective (LO) would be satisfied. (2) Usually provide one PCB per Learning Objective identified for the case; add multiple ONLY if distinct aspects of the same LO must be addressed.

Constraints: - Only produce PCBs; do not generate any other case content. - Do not modify, expand, or infer additional conditions beyond the inputs. - Keep each PCB to the main idea; avoid parenthetical details, rationales, or sub-examples. - Do not include PCBs for Learning Objectives that are not present in the [CASE_BUNDLE].

EXAMPLES: [Specific Case Guidelines]

Inputs: Core case components [CASE_BUNDLE] containing: VIGNETTE, BACKGROUND, ENCOUNTER POINT, PROMPT.

Descriptions of the inputs are as follows: VIGNETTE: A brief clinical overview that introduces the primary communication issue and orients the physician to the scenario. BACKGROUND: Clinical and contextual information leading up to the conversation so the communication moment is understandable and plausible. ENCOUNTER POINT: The moment in the physician-patient interaction that places the clinician at a precise conversational moment that naturally triggers the patient's response. PROMPT: The patient's authentic, direct quote that launches the communication challenge and creates opportunity to practice the specified learning objectives (LOS). LEARNING OBJECTIVES: The learning objectives that the physician should demonstrate during the case. The format of the learning objectives are: [Specific Case Guidelines] - Example cases with completed PCBs [PCB_EXAMPLES] to illustrate how PCBs map to LOs in context. - **IMPORTANT: When reviewing these examples, pay close attention to how the PATIENT CENTERED BEHAVIORS are examples of the LEARNING OBJECTIVES within the context of the other core case content.**

Output format: Return an array that lists the PCBs sequentially and pairs each with its LO name: PCB1: [LO_NAME] - PCB text PCB2: [LO_NAME] - PCB text ...continue until all LOs are covered. **IMPORTANT** : the [LO_NAME] should be in the EXACT same format as provided in the [CASE_BUNDLE].

Do not include any additional case information. Here are case details to guide your task: [CASE_BUNDLE] : {casebundle} [LEARNING OBJECTIVES] : {focallos} [PCB_EXAMPLES] : {pcbexamples}"

—

Exemplar 1 Generation Prompt (Step 5): Role: You are an expert clinical communication content designer. Your goal is to improve physician–patient communication for physicians by crafting an exemplar physician response that integrates the provided patient-centered behaviors (PCBs) and demonstrates the specified learning objectives (LOs).

Encounter Information: - Diagnosis: {ptdx} - Communication Purpose: {ctask} - Case Bundle (Vignette, Background, Encounter Point, Prompt): {casebundle} - Learning Objectives: {focallos} - Patient-Centered Behaviors (PCBs): {pcbs}

Definitions: - Vignette: A brief clinical overview introducing the primary communication issue and orienting the physician to the scenario. - Background: Clinical and contextual information leading up to the conversation. - Encounter Point: The precise conversational moment that triggers the patient's response. - Prompt: The patient's direct quote that launches the communication challenge and opportunity to practice the LOs. - Learning Objectives: [Specific Case Guidelines]

Your Task: - Write a single, natural-sounding physician reply that immediately follows the patient's Prompt in the Case Bundle. - The reply must exemplify optimal patient-centered communication for physicians to emulate and incorporate every PCB.

Content Requirements: [Specific Case Guidelines]

Technical Requirements: [Specific Case Guidelines]

Communication Style: [Specific Case Guidelines]

Reference Examples; use these to guide the length, content, and structure of the exemplar response: {exemplarexamples}

Quality Checkpoints (self-check before finalizing): 1. Directly addresses the patient's Prompt and immediate concern. 2. Integrates every PCB seamlessly (no labels, no meta-commentary). 3. Matches tone to the patient's emotional state. 4. Stays within 50–100 words in one paragraph. 5. Sounds authentic—what a skilled physician would actually say.

Output Format: - Provide ONLY the exemplar response in this format: [EXEMPLAR]: Your generated physician response here

Generate the exemplar response now.

—

Exemplar 2 Generation Prompt (Step 6):

Role: You are an expert clinical communication content designer. Your goal is to improve physician–patient communication for physicians by crafting an exemplar physician response that integrates the provided patient-centered behaviors (PCBs) and demonstrates the specified learning objectives (LOs). Some exemplars for this clinical case have already been developed, and an additional exemplar is needed to provide an example of another way that the physician could include the PCBs while demonstrating the LOs for the specific communication task. The additional exemplar should incorporate the same PCBs, but should be

structurally different than previously generated exemplars - the new exemplar should use different wording and sentence construction.

(rest of prompt copied from Step 5)

—

Patient Ethnicity Selection (Step 7): You are an expert clinical communication content designer. Given the case below, select all patient race codes that could appropriately be represented for this scenario. Do not choose only the most common association; include any race code that is plausible and does not risk reinforcing stereotypes. Only exclude a race code if there is a clear, compelling clinical or contextual reason making it implausible, or if using it in this context would likely reinforce common racial stereotypes.

Inputs: - Diagnosis: {ptdx} - Communication Purpose: {ctask} - Case Bundle (Vignette, Background, Encounter Point, Patient Prompt): {casebundle}

Key definitions: - Vignette: Brief clinical overview introducing the primary communication issue and orienting the clinician to the scenario. - Background: Clinical and contextual information that makes the communication moment understandable and plausible. - Encounter Point: The precise moment in the interaction that naturally triggers the patient's response. - Patient Prompt: The patient's direct quote that launches the communication challenge and enables practice of the specified learning objectives.

Decision rules: - Clinically plausible means the race code does not contradict the diagnosis or case details; do not use epidemiologic prevalence alone as a reason to exclude. - Avoid reinforcing stereotypes; if a race association is commonly stereotyped in the context of this diagnosis or scenario, exclude that code. - If the case does not provide a compelling exclusion reason, include the race code. - If the case explicitly specifies the patient's race, include only that specified code.

Valid race codes: [Specific Case Guidelines]

Output: - Return only the selected race codes, each wrapped in parentheses, in all caps, separated by a single space. - Do not include any explanations or additional text. - If no compelling exclusions are found, include all race codes.

A.3 Item Quality Dimensions

Items were evaluated with respect to six different dimensions:

- **Medical Accuracy** It was important that the generated item present accurate and plausible medical scenarios for the learners to practice patient-centered behaviors.
- **Learning Objective Alignment** The extent to which the vignette background naturally elicits, depends upon, or provides sufficient context for what the learning objectives are intended to measure.
- **Patient-Centered Behavior Alignment** The degree to which the generated PCBs accurately operationalize the learning objectives within the context of the specific clinical vignette.
- **Exemplar Alignment** The degree to which the exemplar responses accurately and naturally instantiate the PCBs (and, by extension, the LOs) in the vignette context. Here "naturally" means in a manner consistent with how an experienced physician would communicate with a patient (appropriate tone, sequencing, level of detail, etc.).
- **No Stereotypes** The generated item does not include or suggest harmful stereotypes pertaining to medical conditions, demographic characteristics (e.g., race, gender, age), or the association between specific groups and particular illnesses or behaviors.
- **Appropriate Difficulty** The item presents a clinical communication scenario that is suitable for a pre-clerkship medical student, requiring only the level of medical knowledge, reasoning, and communication skills typically expected for learners at this level of training.

A.4 Evaluation Questions for Subject Matter Experts

1. What level of change is needed to make this vignette medically accurate?
    a. No change (factually correct)
    b. Minimal (modification of one element (e.g., medical condition incorrectly described/ diagnostic test incorrect)
    c. Moderate (multiple elements within the vignette need to be revised for medical accuracy)
    d. Major-Not salvageable (changes would cascade throughout the vignette and exemplars)

2. What level of changes are needed for the learning objectives (LO) to align with the vignette?
    a. N/A- vignette was not salvageable
    b. No change (leave as is)
    c. Minimal (LO not aligned with the case.)
    d. Moderate (2 LOs not aligned with the case.)
    e. Major (more than 2 LOs not aligned with the case)

3. Do patient-centered behaviors (PCBs) map to Learning Objectives?
    a. N/A- vignette was not salvageable
    b. Yes – no change
    c. No- PCB requires minor revision(s)
    d. No- Do not use PCB

4. Do the exemplars align with the PCBs? If not, what level of change is needed for better alignment?
    a. N/A- vignette was not salvageable
    b. No change
    c. Minimal (change to one aspect of the exemplar)
    d. Moderate (Multiple elements of the exemplar need to be revised)
    e. Major-Not salvageable (several elements of the exemplars would need to be revised)

5. Does this vignette perpetuate any harmful stereotypes?
    a. N/A- vignette was not salvageable
    b. Yes (explain in comments)
    c. No

6. Is this vignette (background, exemplars, and PCBs) at the appropriate level of difficulty for a learner ready to start clerkships?
    a. N/A- vignette was not salvageable
    b. Too easy for student
    c. Appropriate for a student ready to start clerkships
    d. Above the level of the student

A.5 Descriptive Statistics by Generation Method by All Item Dimensions

Category	Method 1a	Method 1b	Method 2a	Method 2b	Total
Medical Accuracy					
None	146 (73.0%)	134 (67.0%)	114 (57.0%)	105 (52.5%)	499 (62.4%)
Minimal	28 (14.0%)	44 (22.0%)	53 (26.5%)	46 (23.0%)	171 (21.4%)
Moderate	9 (4.5%)	9 (4.5%)	13 (6.5%)	19 (9.5%)	50 (6.2%)
Major	17 (8.5%)	13 (6.5%)	20 (10.0%)	30 (15.0%)	80 (10.0%)
LO Alignment					
None	174 (94.6%)	174 (93.5%)	123 (67.6%)	142 (82.1%)	613 (84.6%)
Minimal	8 (4.3%)	10 (5.4%)	33 (18.1%)	19 (11.0%)	70 (9.7%)
Moderate	1 (0.5%)	1 (0.5%)	24 (13.2%)	9 (5.2%)	35 (4.8%)
Major	1 (0.5%)	1 (0.5%)	2 (1.1%)	3 (1.7%)	7 (1.0%)
PCB Alignment					
None	151 (82.5%)	165 (89.2%)	95 (52.8%)	107 (62.2%)	518 (71.9%)
Minimal	32 (17.5%)	20 (10.8%)	79 (43.9%)	56 (32.6%)	187 (26.0%)
Unusable	0 (0.0%)	0 (0.0%)	6 (3.3%)	9 (5.2%)	15 (2.1%)
Exemplar Alignment					
None	113 (61.7%)	102 (55.1%)	121 (67.2%)	120 (69.8%)	456 (63.3%)
Minimal	54 (29.5%)	63 (34.1%)	45 (25.0%)	36 (20.9%)	198 (27.5%)
Moderate	16 (8.7%)	19 (10.3%)	14 (7.8%)	13 (7.6%)	62 (8.6%)
Major	0 (0.0%)	1 (0.5%)	0 (0.0%)	3 (1.7%)	4 (0.6%)
Appropriate Difficulty					
Too Easy	3 (1.6%)	4 (2.2%)	4 (2.2%)	1 (0.6%)	12 (1.7%)
Appropriate	155 (84.7%)	153 (82.7%)	143 (79.4%)	127 (75.6%)	578 (80.7%)
Too Hard	25 (13.7%)	28 (15.1%)	33 (18.3%)	40 (23.8%)	126 (17.6%)
Stereotype Inclusion					
No	178 (97.3%)	180 (97.3%)	178 (98.9%)	167 (99.4%)	703 (98.2%)
Yes	5 (2.7%)	5 (2.7%)	2 (1.1%)	1 (0.6%)	13 (1.8%)

A.6 Full Rater Alignment Agreement by Item and Method

Figure 2: Number of raters indicating that the six item dimensions needed no edits for each generation method and item pair. For example, for Case 17 and Generation Method 1a, only 4 of the 10 raters indicated that no changes were necessary across all six item dimensions.