Towards Reliable Generation of Clinical Chart Items: A Counterfactual Reasoning Approach with Large Language Models

Jiaxuan Li[1], Saed Rezayi[2], Peter Baldwin[2], Polina Harik[2], Victoria Yaneva[2]

[1]University of California Irvine, [2]NBME

jiaxul19@uci.edu (srezayidemne,pbaldwin,pharik,vyaneva)@nbme.org

Abstract

Educational assessment organizations continuously need new test items. This paper presents an exploratory study on the use of large language models (LLMs) for generating item drafts in medical education, focusing specifically on patient chart items. Using GPT-4, we developed and compared three prompting strategies—Chain-of-Thought, counterfactual reasoning, and information-theoretic sample selection—on the quality of the generated drafts. Our prompts include clinical vignettes from existing multiple-choice questions. Evaluation by two clinical experts showed that at least a quarter of the items were free from major flaws at first assessment, and half were considered useful starting points compared to creating items from scratch. We found our proposed counterfactual framework could generate novel items while maintaining the overall quality and accuracy of generated items. The quality of generated items was sensitive to the information-theoretic properties of examples in few-shot learning settings, where example questions with higher surprisal of the correct answers enhanced the quality of generated items. To the best of our knowledge, this is the first study to explore the potential of LLMs for automatic generation of clinical chart items.

1 Introduction

To ensure the relevance and integrity of examinations, educational assessment organizations must continuously develop new, high-quality test items. This is especially critical in the context of high-stakes assessments[1], where test items must not only cover necessary subject material but also conform to rigorous psychometric standards to ensure fairness, validity, and reliability. The process of crafting such test items is inherently complex and resource-intensive, requiring substantial expertise and time investment from subject matter experts. This is particularly challenging for medical education, where the test items need to accurately capture complex real-world problems and reflect highly specialized and rapidly changing knowledge.

Efforts to automate the full or partial creation of test items have long been explored as a means to address the need for scalable and efficient assessment development. Rule-based approaches and cognitive modeling have been widely applied in automated item generation (AIG) (Gierl and Lai, 2016; Lai et al., 2016a; Falcão et al., 2022; Circi et al., 2023). For instance, rule-based methods have been used to enhance distractor quality in MCQs through the integration of knowledge graphs (Lai et al., 2016b). More recently, LLMs have been profitably used for item generation across a range of domains including STEM education, cognitive assessments, as well as language proficiency testing (Attali et al., 2022; Prasetyo et al., 2020; Laverghetta Jr and Licato, 2023; Lee et al., 2023; Chan et al., 2024; Belzak et al., 2023). For example, LLMs in zero- or few-shot learning settings have successfully generated items that have achieved acceptable validity and reliability for various STEM subjects (Chan et al., 2024).

LLMs have demonstrated impressive performance with various medical tasks (Zhou et al., 2023). These include discriminative tasks like question answering (Jin et al., 2019; Yaneva et al., 2023; Naseem et al., 2021; Romanov and Shivade, 2018) as well as generative tasks such as clinical report generation (Johnson et al., 2016; Zhang et al., 2024b). However, most medical LLMs involve pretraining (Zhang et al., 2024a; Jin et al., 2023; Luo et al., 2022; Gu et al., 2021) or fine-tuning (Christophe et al., 2024; Gururajan et al., 2024; Luo et al., 2023), which may require expensive computation resources. The adoption of pretrained LLMs for AI-assisted item creation in the medical domain remains a challenge (Karabacak et al.,

[1]Examinations with significant consequences for the test-taker, such as professional certification or licensure.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 142–153

2023). A systematic survey suggests that off-the-shelf generative language models such as ChatGPT struggle to generate high-quality multiple-choice medical questions, even with advanced prompting strategies (Kıyak and Emekli, 2024).

In this paper, we perform an initial investigation of the potential of LLMs to assist with creating comprehensive documents of patient's medical record (clinical charts) and multiple choice questions for medical education exams. We prompt off-the-shelf pretrained language models with clinical vignettes from a publicly available dataset (MedQA; Jin et al., 2021) and develop three different approaches for item generation in a few-shot learning setting, including *Chain-of-Thought Generation*, *Counterfactual Generation*, and *Principled few-shot learning sample selection*. The generated items are evaluated by two licensed medical doctors who are medical school faculty. We found that our proposed counterfactual generation framework produces items with greater lexical and semantic distance from source material while maintaining overall quality, and that information-theoretic properties of samples in few-shot learning settings influence the quality of generated items. To the best of our knowledge, this is the first study to explore the potential of a counterfactual generation framework with principled learning sample selection for generating clinical chart items.

2 Method

2.1 Data

MedQA This study uses data from two distinct sources. The first source is MedQA (Jin et al., 2021), a publicly available dataset containing ≈ 60K clinical MCQs in English, simplified Chinese, and traditional Chinese. These MCQs were collected from various test preparation materials available online. In our study, we use the English-language subset, which contains 12,723 items.

Chart items The second source is a dataset of 35 chart items (see Fig. 1 for an example item). These items were developed as part of a research project on assessing clinical reasoning and the specific items used in this study are referred to as SHARP items (SHort Answer, Rationale Provision; see Runyon et al. (2023) for a full description of the item format).

Clinical charts, also known as patient records, are comprehensive documents that typically include a patient's medical and social history, presenting symptoms, chief complaints, physical examination findings, and test results. They may also contain physician notes documenting patient visits, differential diagnoses, and treatment plans. In medical education, clinical charts serve as a structured and effective tool for training future physicians (Deschênes et al., 2025; Goulet et al., 2007), bridging the gap between theoretical knowledge and real-world medical practice (Al-Wassia et al., 2015).

One of the primary benefits of using patient charts in medical education is the enhancement of clinical reasoning and decision-making skills (Daniel et al., 2019). By reviewing and analyzing patient charts, medical students can practice prioritizing information, identifying key features, formulating differential diagnoses, developing treatment plans, and making informed clinical decisions.

2.2 Setup

Our primary goal is to develop a scalable pipeline to generate chart items by prompting language models with detailed instruction and medical scenarios. Each prompt comprises a medical vignette presented as a multiple-choice medical question from the MedQA dataset along with three examples of chart items from the SHARP dataset.

We implement three generation frameworks using GPT-4: *Chain-of-Thought* generation, which transforms a medical vignette from MedQA into a chart item by creating a medical record for a hypothetical patient (Section 2.3); *Counterfactual Generation*, which incorporates counterfactual reasoning to explore alternative outcomes and generate novel items while leveraging an agent-based self-prompting strategy to create a knowledge base for accuracy (Section 2.4); and an information-theoretic framework where the sample items in few-shot learning settings are selected based on information-theoretic properties, finding that LLMs perform better with "difficult" examples (Section 2.5). For each generation method, we produced 80 items that were evaluated by two licensed medical experts (Section 3).

2.3 Experiment 1: Chain-of-Thought

The first experiment uses Chain-of-Thought (CoT) prompting as a baseline. The approach was designed to be a robust framework for systematically generating high-quality medical assessment items that works by dividing the creation process into a sequence of cognitively manageable steps (Saparov and He, 2022).

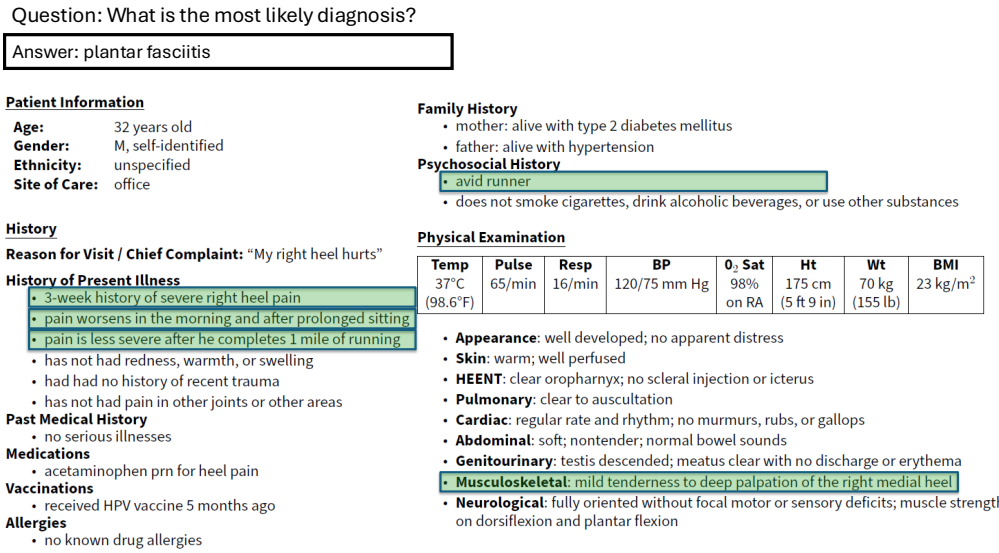


Figure 1: An example chart-type item from Runyon et al. (2023). A chart item includes a chart with patient information, medical history, chief complaint, and physical examination findings, as well as an associated question and answer. Not all chart information is equally relevant for correctly diagnosing and test-takers must determine relevancy as part of the task. The green boxes highlight the most relevant information for diagnosis in this example.

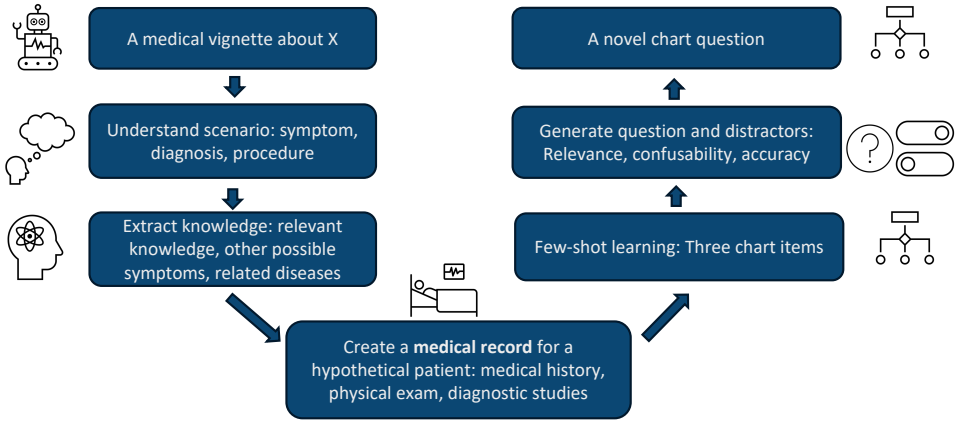


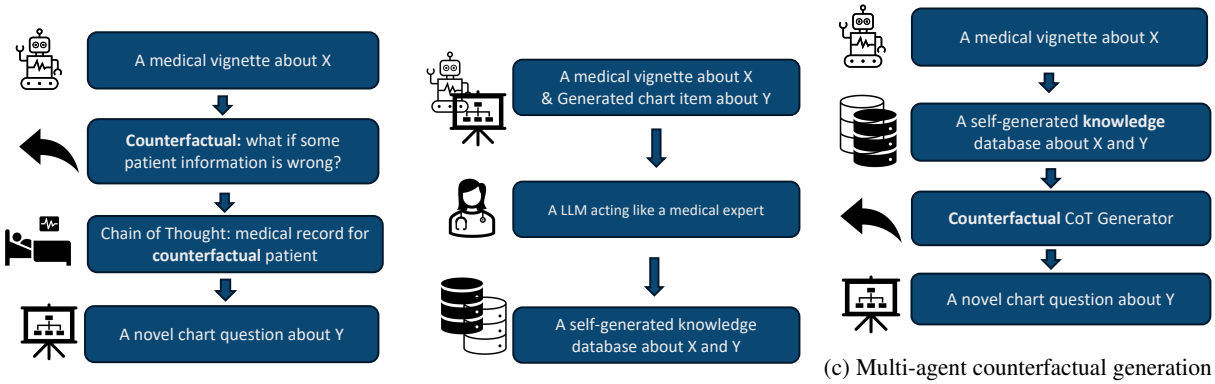
Figure 2: An illustration of Chain-of-Thought generation for chart-type items. The model is instructed to transform a simple medical scenario drawn from the MedQA dataset into a novel chart question step by step.

The CoT generator transforms a medical vignette extracted from the MedQA dataset into a chart question step by step (see Fig. 2). First, the model is instructed to identify the symptoms, diagnosis, and procedures described in the medical vignette to ensure that the model captures the parent medical scenario. Next, the model generates key knowledge relevant to the parent medical scenario, including key symptoms, potential differential diagnoses, and related diseases. The model then creates a detailed medical record for a hypothetical patient incorporating incorporating information from parent medical vignette and relevant information generated by the model. The model is further guided by referencing three sample SHARP chart items as examples of the desired chart-format output. The final output includes a clinical chart, a question with a correct answer and ten distractors. We instruct the model to adhere some general principles for question and distractor generation (see Appendix A).

2.4 Experiment 2: counterfactual generation

A counterfactual chart item is one whose key diagnostic findings intentionally contradict the parent vignette's findings such that the correct diagnosis changes. It leverages a three-step process that integrates CoT prompting, counterfactual reasoning, and self-generated knowledge infusion (see Fig.3).

The first step focuses on generating content that differs from the source material by transforming



(a) Step 1: Counterfactual generation (b) Step 2: Knowledge generation (c) Multi-agent counterfactual generation with self-generated knowledge

Figure 3: An illustration of three-step knowledge-infused counterfactual generation for chart items. In Step 1, a chart item related to a novel medical scenario is generated by instructing the model to identify misinformation from the parent vignette. In Step 2, a knowledge database is generated for the parent and generated medical scenarios. In Step 3, the database is integrated with the counterfactual generator from Step 1 to regenerate the chart item.

the parent medical vignette into a different medical scenario through counterfactual reasoning. Counterfactual reasoning has been widely applied in various settings to explore alternative scenarios or causal inference in LLM performance (Qin et al., 2019; Zellers et al., 2019; Mostafazadeh et al., 2016; Meng et al., 2022; Rajani et al., 2019; Saparov and He, 2022; Frohberg and Binder, 2022; Elazar et al., 2021; Rudinger et al., 2020; Li et al., 2023). The model is instructed to conduct counterfactual reasoning in a Chain-of-Thought framework. We set up a counterfactual premise where the model is informed that certain elements of the parent vignette are transcribed incorrectly. Based on this counterfactual premise, the model needs to creatively "recover" the clinical chart, leading to a hypothetical patient record based on a new medical scenario. The model reasons based on the generated counterfactual record to develop a clinical assessment. The goal of this process is to generate content that deviates from the parent vignette while maintaining clinical plausibility.

The second step aims to improve the factual grounding of generated items by creating a self-generated knowledge base. This step addresses LLMs' tendency to hallucinate (Xu et al., 2024; Zhang et al., 2023) and is accomplished by initiating a new session in which the language model assumes the role of a medical expert. Agent-based prompting (Wu et al., 2024) enables the model to adapt to this role for generating medical knowledge. The correct answer from the parent vignette (X) and the generated correct answer from the counterfactual scenario (Y) are provided to the model. The task is to synthesize a detailed knowledge base about X and Y, including their symptoms, diagnostic criteria, and distinguishing features. This approach attempts to ground the generated counterfactual scenario in medical knowledge. In the final step, the knowledge base from Step 2 is integrated back into the counterfactual generation process. Combining the content variation from Step 1 with the knowledge grounding from Step 2, the model generates a refined chart item based on the counterfactual scenario.

2.5 Experiment 3: sample selection

We explore whether the performance of the language model is sensitive to the information-theoretic properties of the few-shot learning samples. Language model performance has been shown to depend on the quality of the samples in few-shot learning (Rasheed and Zarkoosh, 2024). Although all chart items used as examples were judged to be of high quality by human medical experts, certain information theoretic properties might make some examples better suited for the item generation task. In this experiment, we evaluate whether the quality of automatic generation is affected by the information content of the few-shot learning examples.

We hypothesize that the information-theoretic properties of example items are directly related to how challenging they are for the LLM to solve. Specifically, we use the surprisal of the correct answer given the question stem as a metric to assess an item's difficulty for the LLM. Surprisal is calculated as the negative logarithm of the probability that the LLM assigns to the correct answer

given the question stem ($-\log p(\text{answer} \mid \text{stem})$), thereby quantifying how unexpected the correct answer is. Consequently, an item with relatively low surprisal is considered relatively easy for the LLM to answer correctly.

Sample selection is guided by two complementary hypotheses. The first posits that easier questions lead to better performance because they are straightforward for LLMs to mimic and regenerate (*Easy Sample Hypothesis*). Conversely, the second hypothesis suggests that more challenging examples may compel LLMs to engage in deeper reasoning, improving their ability to generate complex items (*Hard Sample Hypothesis*). By evaluating the effect of the surprisal of selected examples, we can maximize the quality of the generated items.

We calculate the surprisal of the correct answer using GPT-2 (Radford et al., 2019), and select the three sample items with the *lowest* surprisal. These selected samples are used in the multi-agent counterfactual generator with self-generated knowledge (Fig. 3). We then compare the performance with that of the counterfactual generator with randomly selected examples described in Section 2.4. If the items generated in Experiment 3 are considered of higher quality than items in Experiment 2, *Easy Sample Hypothesis* is supported.

3 Evaluation

While there is no consensus on the evaluation protocol of generated items (Circi et al., 2023), we aim to evaluate various aspects related to their practical use in assessment. The generated items may contain various flaws that affect their suitability for assessment. These flaws include, but are not limited to, clinical inaccuracies, contradictions, or hallucinations; incorrect designation of the correct answer; distractors (incorrect answers) that may actually be correct; or content that is unsuitable for assessment due to overly high or low complexity. Evaluating these issues requires review by human experts, as they cannot currently be assessed automatically.

Since an exhaustive list of all potential flaws could not be constructed a priori due to the unknown nature of AI-generated items, we focused our evaluation on the general suitability of these items for use in high-stakes medical education assessment as perceived by experts with both clinical and educational backgrounds. We designed a rubric that covered the following questions, with the full list provided in Appendix B:

(1) Can the chart stem be used on a high-stakes assessment?

(2) Please select up to 5 distractors that would, as a group, constitute a partial or full option set. Do not select any that would not be suitable for this chart, or that are too similar to others that have been selected as suitable.

(3) Can the chart item as a whole be used on a high-stakes assessment as currently written?

(4) Is this draft a usable starting point for writing or updating a chart item?

Two licensed medical doctors who also served as faculty at accredited medical schools in the United States were recruited. Each expert was assigned the same set of 100 automatically generated items, of which 33-34 were generated using each of the three methods (see Appendix C).

The results from the expert evaluation are presented in Table 1. Responses to each of the four rubric questions were dichotomized: (1) *stem quality*: minor changes / substantive changes; (2) *distractor quality*: substantive changes *not* required / substantive changes required; (3) *chart quality*: minor changes / substantive changes; and (4) *helpfulness*: helpful / *not* helpful. For each question, we calculate two success metrics: *strict*, which requires two favorable expert judgments and *loose*, which only requires one favorable judgment.

Across the three methods, both experts agreed that over 24% of generated stems required only "minor changes." In addition, the quality of the generated disractors was perceived to be high, with both raters agreeing that the distractors for at least 79% of the items required only minor changes. Across three methods, at least 85% generated chart items were considered usable with minor changes by at least one annotator. Although the CoT framework's items were deemed usable most often, the counterfactual framework performed similarly. The information-theory-based framework using sample items with lowest surprisal has a reduced performance compared to other methods. This suggests that language models' generation performance benefits more from examples with higher item surprisal, supporting the *Hard Sample Hypothesis*. Moreover, both experts agreed that over half of the items (52%) were helpful starting points for writing a new item. Here, items generated using

	Stem	Suggested distractors	Whole item	Helpfulness
CoT	**91% (35%)**	**100% (82%)**	**94% (26%)**	97% (79%)
Counterfact	**91% (32%)**	**100% (82%)**	91% (24%)	**100% (53%)**
Info theory	85% (24%)	97% (79%)	85% (15%)	100% (52%)

Table 1: This table displays the proportion of items that were favorably judged for each of the four questions in the annotation rubric. Two evaluation criteria were used: the *loose* criterion, where an item is considered favorably judged if at least one of the two participating physicians judged it favorably; and the *strict* criterion, where an item is favorably judged only if both physicians agreed. Proportions are presented in the format: loose (strict).

the CoT method significantly outperformed other items on that criterion with 79% of the CoT items judged to be helpful starting points by both experts. Overall, the expert evaluation suggests that approximately a quarter of the generated items were free from major flaws *at first assessment*, and half were regarded as useful starting points for item development compared to creating items from scratch.

The variability in expert agreement underscores the subjective nature of evaluating item quality, particularly for stems and charts, where the experts exhibited the most disagreement. The two experts had inter-annotator agreement of $\kappa = 0.1$ on chart stem quality. A qualitative inspection of the annotators' comments suggests that the low inter-rater agreement might be due to different conceptual understanding of the rubric. For example, both annotators commented that one question "needs mother's prenatal history", but one annotator considered this critical and suggested substantial changes needed, whereas the other considered it a minor modification (see Appendix E and F for more discussion on limitations and ethical considerations).

We also evaluated whether counterfactual generation produces items with greater semantic distance from their source material. To quantify semantic distance, we computed cosine similarity between word embeddings of each generated item and its parent vignette, with lower similarity indicating greater lexical/semantic divergence. Results showed that methods based on counterfactual generation (Exp 2 & 3) produced items with significantly lower cosine similarity to their parent vignettes than CoT generation (Exp 1), suggesting greater variation from the source material (see Appendix D).

4 Discussion

This study demonstrated the potential of LLMs to be used as automated assistive tools when writing items for medical assessments. The findings highlight key insights into the quality of the items generated across the three methods. Notably, over 24% of the generated stems were rated as requiring "minor changes" by both experts, with 85% of the items judged to require minor changes by at least one expert. This suggests that a significant portion of the generated items lack what could initially be considered irreparable flaws, inaccuracies, or contradictions. While this cannot yet be considered evidence that the items can be profitably used on an assessment without significant review and modifications, it is an encouraging initial assessment.

The integration of counterfactual reasoning and agent-based knowledge infusion showed effectiveness in producing content that differs more from source material. This suggests that tasking the model with identifying misinformation and generating counterfactual scenarios helps prevent the model from simply replicating existing data.

Of particular interest are the findings on information-theoretic sample selection, which highlight the nuanced role of item surprisal in few-shot learning. The observed differences in item generation when challenging examples were used suggest that example difficulty may influence LLM generation patterns. This insight underscores the importance of principled sample selection in optimizing LLM performance for automated item generation.

Future research should focus on automating the evaluation process, expanding applicability to other domains, and reducing the computational overhead of LLM-based pipelines. Integrating external knowledge sources, such as medical databases, could potentially improve the factual grounding of generated chart items. Retrieval-Augmented Generation techniques could be explored to access and incorporate external data during the item generation process. This approach might allow the model to generate more contextually informed items and better adapt to specialized knowledge domains.

References

Heidi Al-Wassia, Rolina Al-Wassia, Shadi Shihata, Yoon Soo Park, and Ara Tekian. 2015. Using patients' charts to assess medical trainees in the workplace: a systematic review. *Medical teacher*, 37(sup1):S82–S87.

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.

Susan M Case and David B Swanson. 1998. *Constructing written test questions for the basic and clinical sciences*. National Board of Medical Examiners Philadelphia.

Kuang Wen Chan, Farhan Ali, Joonhyeong Park, Kah Shen Brandon Sham, Erdalyn Yeh Thong Tan, Francis Woon Chien Chong, Kun Qian, and Guan Kheng Sze. 2024. Automatic item generation in various stem subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, page 100344.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

Ruhan Circi, Juanita Hicks, and Emmanuel Sikali. 2023. Automatic item generation: foundations and machine learning-based approaches for assessments. In *Frontiers in Education*, volume 8, page 858273. Frontiers Media SA.

Michelle Daniel, Joseph Rencic, Steven J Durning, Eric Holmboe, Sally A Santen, Valerie Lang, Temple Ratcliffe, David Gordon, Brian Heist, Stuart Lubarsky, et al. 2019. Clinical reasoning assessment methods: a scoping review and practical guidance. *Academic Medicine*, 94(6):902–912.

Marie-France Deschênes, Nicolas Fernandez, Kathleen Lechasseur, Marie-Ève Caty, Busra Meryem Uctu, Yasmine Bouzeghrane, and Patrick Lavoie. 2025. Transformation and articulation of clinical data to understand students' clinical reasoning: a scoping review. *BMC Medical Education*, 25(1):52.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Filipe Falcão, Patrício Costa, and José M Pêgo. 2022. Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, 27(2):405–425.

Jörg Frohberg and Frank Binder. 2022. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140.

Mark J Gierl and Hollis Lai. 2016. The role of cognitive models in automatic item generation. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, pages 124–145.

François Goulet, André Jacques, Robert Gagnon, Pierre Racette, and William Sieber. 2007. Assessment of family physicians' performance using patient charts: interrater reliability and concordance with chart-stimulated recall interview. *Evaluation & the health professions*, 30(4):376–392.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

B Guimarães, J Pais, E Coelho, A Silva, A Povo, I Lourinho, M Severo, and MA Ferreira. 2013. Assessing inter-rater agreement about item-writing flaws in multiple-choice questions of clinical anatomy. In *EDULEARN13 Proceedings*, pages 5921–5924. IATED.

Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrián Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, et al. 2024. Aloe: A family of fine-tuned open healthcare llms. *CoRR*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Mert Karabacak, Burak Berksu Ozkara, Konstantinos Margetis, Max Wintermark, and Sotirios Bisdas. 2023. The advent of generative language models in medical education. *JMIR Medical Education*, 9:e48163.

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*, page qgae065.

Hollis Lai, Mark J Gierl, B Ellen Byrne, Andrew I Spielman, and David M Waldschmidt. 2016a. Three modeling applications to promote automatic item generation for examinations in dentistry. *Journal of dental education*, 80(3):339–347.

Hollis Lai, Mark J Gierl, Claire Touchie, Debra Pugh, André-Philippe Boulais, and André De Champlain. 2016b. Using automatic item generation to improve the quality of mcq distractors. *Teaching and learning in medicine*, 28(2):166–173.

Antonio Laverghetta Jr and John Licato. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 414–428.

Philseok Lee, Shea Fyffe, Mina Son, Zihao Jia, and Ziyu Yao. 2023. A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, 38(1):163–190.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 804–815, Toronto, Canada. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2021. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Septian Eko Prasetyo, Teguh Bharata Adji, and Indriana Hidayah. 2020. Automated item generation: Model and development technique. In *2020 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pages 64–69. IEEE.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Areeg Fahad Rasheed and M Zarkoosh. 2024. Mashee at semeval-2024 task 8: The impact of samples quality on the performance of in-context learning for machine text classification. *Authorea Preprints*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675.

Christopher R Runyon, Miguel A Paniagua, Francine A Rosenthal, Andrea L Veneziano, Lauren McNaughton, Constance T Murray, and Polina Harik. 2023. Sharp (short answer, rationale provision): A new item format to assess clinical reasoning. *Academic Medicine*, pages 10–1097.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *CoRR*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. In *COLM*.

Ziwei Xu, Sanjay Jain, and Mohan S Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *CoRR*.

Victoria Yaneva, Peter Baldwin, Daniel P Jurich, Kimberly Swygert, and Brian E Clauser. 2023. Examining chatgpt performance on usmle sample items and implications for assessment. *Academic Medicine*, pages 10–1097.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *ACL*, pages 4791–4800.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. 2024a. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2024b. Rexrank: A public leaderboard for ai-powered radiology report generation. *arXiv preprint arXiv:2411.15122*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A General Principles

We instruct the model to adhere to the following principles during generation: (a) *Informativity*: the new item should contain all the necessary information for a chart item; (b) *Accuracy*: the generated answer should be consistent with all of the information from the generated chart; (c) *Novelty*: the generated chart item should be sufficiently different from the parent item; and (d) *Validity*: the generated chart must include sufficient information to unambiguously identify the correct answer.

According to the chart question stem and correct answer, the model then crafts ten distractors—plausible but incorrect answer choices that are meant to be attractive to examinees who do not know the correct answer. We instruct the model to focus on the following properties during the distractor generation process: (a) *Relevance*: the distractors should be relevant to the chart question stem; (b) *Dissimilarity*: the distractors should not be synonyms or very similar to the correct answer; (c) *Incorrectness*: the distractors cannot be plausible correct answers for the generated chart question. Distractors with these characteristics enhance items' discriminative power.

The model is instructed to use descriptive language about any physical exam findings that follows patient chart documentation standards, such as specifying *warm*, *dry*, or *no rashes or lesions* instead of vague terms like *normal*.

B Evaluation Protocol

1) Evaluation of the Chart: Evaluate the Chart's suitability for use on a high stakes assessment. Minor changes are defined as the necessity to make minor changes to the chart including but not limited to: the addition, modification, or deletion of three or fewer minor history/physical exam details to make the chart more correct, realistic, or at a more appropriate difficulty level. Substantive changes entail an extensive rewrite of the chart and include but are not limited to: the addition, modification, or deletion of four or more substantive history/physical exam details to make the chart more correct, realistic, or at a more appropriate difficulty level.

Question: Can the chart be used on a high stakes assessment?

i. Yes, with some minor changes

ii. Substantive changes required, or the chart is too flawed to be useful

Note that if the expert selected "Substantive changes required", they would skip the next two questions and go directly to the fourth question on Helpfulness.

2) Selection of Appropriate Option Set: Please select up to 5 distractors that would, as a group, constitute a partial or full option set. Do not select any that would not be suitable for this chart, or that are too similar to others that have been selected as suitable. The N/A option should be used if you selected "Substantive changes required, or the chart is too flawed to be useful" in response to the above question about the associated chart.

i. N/A – Substantive changes required, or the chart is too flawed to be useful

ii. Distractor Suggestion 1

...

xi. Distractor Suggestion 10

3) Evaluation of the Chart Item as a whole (chart plus the option Set): To what extent can the chart item as a whole (i.e., chart plus the options set) be used on a high stakes assessment? Minor changes to the chart item as a whole is defined as the necessity to make minor changes to EITHER the chart (i.e., requires a minor rewrite of the chart including but not limited to: the addition, modification, or deletion of three or fewer minor history/physical exam details to make the chart more correct, realistic, or at a more appropriate difficulty level OR minor changes to the option set (i.e., the need to create one additional option to complete a sufficient option set of at least 4 options (preferably 5) with appropriate difficulty for a high stakes assessment). Substantive changes to the chart item as a whole is defined as the necessity to make substantive changes to EITHER the chart (i.e., requires an extensive rewrite of the chart including but not limited to: the addition, modification, or deletion of four or more substantive history/physical exam details to make the item more correct, realistic, or at a more appropriate difficulty level (i.e., suitable for high stakes assessment) OR substantive changes to the option set (i.e., the need to create three or more options to complete a sufficient option set of at least 4 options (preferably 5) with appropriate difficulty for a high stakes assessment). If EITHER the chart OR the option set need substantive changes, then this is considered as the need for substantive changes to the chart item as a whole. If BOTH the chart and the option set require minor changes, this is considered as the need for minor changes to the chart item as a whole.

Question: Can the chart item as a whole be used on a high stakes assessment as currently written?

i. Yes, with some minor changes

ii. Substantive changes required, or the chart is too flawed to be useful

4) Evaluation of helpfulness: Is this draft a usable starting point for writing or updating a chart item?

i. Yes, this draft would be helpful

ii. No, it would be easier for me to write an item from scratch

It is important to clarify that *we do not consider the "minor changes" category as suggesting an item is ready for assessment without significant additional work* (see Section F for discussion on ethical considerations). Instead, the distinction between minor and substantive changes serves as a simple way to differentiate items with major flaws from those with flaws that may be fixable.

C Recruitment

To perform this evaluation, two licensed medical doctors who also served as faculty at accredited medical schools in the United States were recruited. The recruitment was performed by ANONYMIZED INSTITUTION's Assessment Alliance, which engages with educators, learners, and other members of the health profession's education community to identify how to best prepare medical professionals to safely care for a diverse patient population.

Once recruited, the human experts were invited to a kickoff meeting, where they were briefed on the purpose of the experiment and the evaluation rubric, instructed on the use of the annotation platform (items were displayed using the John Snow Labs annotation system), and given an opportunity to ask questions. Following this meeting, the experts were given two weeks to complete their annotations. Each expert was assigned the same set of 100 automatically generated items, of which 33-34 were generated using each of the three methods described in Section 2.

D Automated evaluation of item variation

An important consideration for newly generated items is the extent to which they differ from their source material. Understanding these differences can help identify which generation methods produce more varied content and potentially guide selection of items for further development by human item writers. To this end, we explore the use of cosine similarity between word embeddings of generated items and their parent medical vignettes as one measure of content variation.

Cosine similarity between word embeddings quantifies lexical and semantic overlap between generated and parent items, with lower values indicating less overlap—i.e., greater textual divergence. We define an experimental group where cosine similarity is calculated between each generated item and its corresponding parent vignette. This is com-

pared against a baseline group, where cosine similarity is computed between each generated item and a random non-parent vignette from the same set of parent vignettes. Figure 4 shows the average cosine similarity in experimental and baseline groups across the three generation methods.

Paired t-tests between experimental and baseline groups within each method did not reveal statistically significant results (all $p > 0.25$), possibly due to high variability in similarity values or limited sample size. Nevertheless, we observed consistent trends across all conditions, where similarities between generated items and their parent vignettes were not significantly different from similarities with unrelated vignettes. To quantify relative content variation across methods, we used the difference in cosine similarity between experimental and baseline groups as an index of textual divergence. A second set of paired t-tests with Bonferroni correction was conducted to compare this divergence index across generation methods. The results revealed that CoT generation produced significantly smaller divergence from source material than both counterfactual ($t = 4.64, p < 0.001$) and information-theory-based generation ($t = 4.41$, $p < 0.001$). No significant difference was found between counterfactual and information-theory-based methods ($t = -0.1, p = 0.91$).

These findings suggest that counterfactual and information-theoretic approaches produce content with greater lexical and semantic distance from their source vignettes compared to CoT generation. However, it is important to note that cosine similarity captures only surface-level textual differences and does not necessarily reflect clinically meaningful variation or educational value of the generated items.

E Limitations

A key limitation for this research is the fact that the evaluation relied on only two human raters. These raters had not undergone specific training in item writing for high-stakes clinical exams, and this was their first time evaluating AI-generated items. These factors may have contributed to the observed variability in their judgments while limiting their generalizability. Additionally, given the well-documented variability in how human experts write clinical MCQs (e.g., Guimarães et al., 2013), judgments about the need for "minor" vs "substantive" changes may reflect subjective differences in

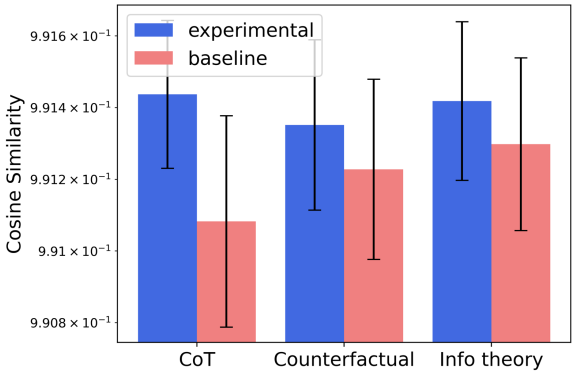


Figure 4: Average cosine similarity in experimental and baseline groups across three generation methods. Blue bars represent cosine similarity between generated item and its corresponding parent vignette. Red bars represent cosine similarity between generated item and a random vignette from a set of parent vignette.

opinion rather than a definitive standard of quality.

Rater performance may have been further influenced by biases such as social desirability or confirmation bias. Social desirability bias could lead raters to align their evaluations with perceived research goals or provide overly favorable feedback due to the novelty of AI in clinical item generation. Confirmation bias might cause raters to focus on strengths or weaknesses based on their pre-existing beliefs about AI's capabilities. Measuring attitudes toward AI as part of the recruitment process is an area for improvement in future research.

In terms of evaluation design, the rubric was purposefully broad given the stage of this research and did not account for specific flaws that might arise in clinical MCQs. Examples of such flaws include susceptibility to "testwiseness," which refers to an examinee's familiarity with general test-taking strategies, and "construct-irrelevant difficulty," which refers to item features that increase an item's difficulty for reasons unrelated to the trait that is the intended target of the assessment (Case and Swanson, 1998). Future research should endeavor to better understand and identify specific flaws that may be prevalent within AI generated items, and facilitate their evaluation through more granular rubrics.

Similar to the human evaluation, the automated evaluation also suffered limitations stemming from the preliminary nature of this study. While a useful approximation of the differences that exist between items, cosine similarity focus only on relative item variation and do not guarantee that items are sufficiently novel for a given application.

Last but not least, the performance of the generated items in practical settings is currently unknown. Key metrics such as the extent to which examinees find an item difficult, the power of an item to discriminate between examinees of different proficiency levels, and examinee perceptions of clarity require pretesting with an examinee sample and remain untested at this stage.

In summary, future research should not only focus on improving the technical components of item generation, but also include larger-scale evaluations, enhanced rubrics, qualitative analyses, the utilization of raters trained in item writing, and the collection of examinee response data in real-world assessment settings.

F Ethical considerations

As AI continues to evolve and its application is extended to more domains, its integration into item development raises important ethical considerations. A key concern is ensuring that AI-generated items meet the necessary quality standards for a given type of assessment. While AI can generate item drafts, these items must be thoroughly reviewed by expert item writers to ensure that they are appropriate, clinically accurate, and meet the intended learning or assessment objectives. Human oversight remains essential to finalize each item, and AI-generated content should undergo the same rigorous review processes as items that are written without AI assistance.

The use of AI also requires clear accountability and transparency in the development process and avoidance of over-reliance on technology. While AI can assist in generating drafts, the final responsibility for ensuring the quality, fairness, and ethical use of any test item remains with human experts. It is crucial to maintain transparency about how AI is used and to ensure that stakeholders are aware of both the capabilities and limitations of AI in this context.

By ensuring that human expertise remains central to the item development process, establishing rigorous review procedures, and maintaining transparency and accountability, AI can be used ethically and responsibly to support the creation of high-quality assessment items.