Toward Automated Evaluation of AI-Generated Item Drafts in Clinical Assessment

Tazin Afrin[1], Le An Ha[2], Victoria Yaneva[1], Keelan Evanini[1], Steven Go[1], Michael Heilig[1], Kristine DeRuchie[1]
[1]National Board of Medical Examiners, Philadelphia, USA
{tafrin, vyaneva, kevanini, sgo, mheilig, kderuchie}@nbme.org
[2]Ho Chi Minh City University of Foreign Languages, Vietnam
anhl@huflit.edu.vn

Abstract

This study examines the classification of AI-generated clinical multiple-choice question drafts as "helpful" or "non-helpful" starting points. Expert judgments were analyzed, and multiple classifiers were evaluated—including feature-based models, fine-tuned transformers, and few-shot prompting with GPT-4. Our findings highlight the challenges and considerations for evaluation methods of AI-generated items in clinical test development.

1 Introduction

The development of high-quality standardized assessments fundamentally depends on the availability of well-crafted test items. As the demand for more efficient and scalable item development grows, many organizations are turning to large language models (LLMs) to meet this need (LaFlair et al., 2023; Song et al., 2025). LLMs offer the promise of aiding the creation of items at scale – increasing diversity and improving security by eliminating item reuse. These benefits make LLMs an attractive solution for organizations seeking to streamline the assessment development process.

However, as LLMs become more widely used for generating content across various domains, evaluating the quality of the generated output has become increasingly critical to the usability of these models. Without a reliable and scalable method for assessing quality, there is a risk of replacing one bottleneck — manual content creation—with another: sorting through a vast amount of content that varies in quality. This issue is especially challenging in fields that require specialized expertise, such as medical educational assessment, and in contexts where there is no universal agreement among experts due to the nuanced and inherently subjective nature of the criteria used to define high-quality output. To fully harness the potential of LLMs in generating exam items, it is essential to address this evaluation bottleneck.

In this study, we present one of the first explorations of automated evaluation of AI-generated items in the clinical domain, using a dataset of 512 clinical multiple-choice questions (MCQs), each rated by two experts. This work presents the following original contributions:

- We collect and analyze expert ratings of AI-generated MCQs in the context of medical education assessment.
- We evaluate a range of automated classification metrics to determine how well they predict expert judgments, identifying which metrics align most closely with human assessments; an error analysis aims to identify areas where these automated metrics fall short.
- While primarily aimed at providing practical insights in assessment development, we also discuss the implications of these findings for the broader challenge of evaluating AI-generated expert text, highlighting the need for nuanced evaluation frameworks as generative AI becomes increasingly integrated into professional workflows.

2 Related Work

Since the advent of LLMs, the medical community has had a keen interest in exploring the medical knowledge of LLMs (He et al., 2025; Singhal et al., 2023; Tang et al., 2023; Yaneva et al., 2024; Zhou et al., 2023) and generating MCQs that can be used in medical education and assessments (Artsi et al., 2024; Al Shuraiqi et al., 2024). The quality of automatically generated MCQs has been evaluated using a range of methods across multiple studies (see Table 1 for a comparison).

Cheung et al. (2023) conducted a multinational prospective study evaluating the quality of MCQs produced by ChatGPT for graduate medical examinations across Hong Kong, Singapore, Ireland, and

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 172–182

Study	Evaluation Dimensions	Findings
Cheung et al. (2023)	Appropriateness, Clarity, Relevance, Discrimination, Exam Suitability	No significant difference between AI- and human-generated MCQs; AI scored slightly lower in relevance (p = 0.04)
Klang et al. (2023)	Accuracy, Terminology, Sensitivity	0.5% of questions were false; 15% required revisions due to various inaccuracies
Agarwal et al. (2023)	Validity, Difficulty, Reasoning	ChatGPT produced the least difficult questions; strong inter-rater reliability ($\kappa \geq 0.8$)
Ayub et al. (2023)	Accuracy, Complexity, Clarity	Only 40% of AI-generated questions were suitable for ABD-AE preparation
Bedi et al. (2025)	Distinguishability, Validity, Reviewer Consensus	64% of questions deemed valid; 51.8% distinguishability (random chance); reviewers took 3.2 min/question

Table 1: Summary of studies evaluating AI-generated medical MCQs

the United Kingdom. Five independent international assessors evaluated questions based on five domains: appropriateness, clarity and specificity, relevance, discriminative power of alternatives, and suitability for medical graduate examinations. The study found no significant difference in overall question quality between AI-generated and human-authored questions, except in the relevance domain, where AI-generated questions scored slightly lower (AI: 7.56 ± 0.94 vs. human: 7.88 ± 0.52; p = 0.04).

In Klang et al. (2023), GPT-4 was utilized to generate MCQs for medical examinations. Specialist physicians, blinded to the source of the questions, evaluated them for mistakes and inaccuracies. The study reported that only 0.5% of AI-generated questions required replacement, while 15% required revisions due to issues like outdated terminology and demographic sensitivities.

Agarwal et al. (2023) assessed the applicability of ChatGPT, Bard, and Bing in generating reasoning-based MCQs in medical physiology. Two physiologists rated the AI-generated questions on validity, difficulty, and reasoning ability using a 0-3 scale. ChatGPT produced the least difficult questions, and all AI models showed limitations in generating high-level reasoning questions. Inter-rater reliability was high, with Cohen's kappa (κ) values ≥ 0.8 across all parameters.

In dermatology, Ayub et al. (2023) explored ChatGPT's potential in generating board-style questions. Two board-certified dermatologists conducted a qualitative analysis of 40 AI-generated questions, assessing accuracy, complexity, and clarity. Only 40% of the questions were deemed accurate and appropriate for American Board of Dermatology Applied Exam (ABD-AE) preparation, highlighting the need for expert oversight.

QUEST-AI (Bedi et al., 2025) is an AI system for generating, verifying, and refining USMLE-style items. Three physicians and two medical students participated in a twofold assessment: distinguishing between AI- and human-generated items and evaluating the validity of AI-generated content. Participants could only distinguish between the two at a rate of 51.8%, suggesting indistinguishability of AI-generated items. Furthermore, 64% of AI-generated items were unanimously deemed correct by reviewers, while 36% were flagged for issues like multiple correct answers or incorrect AI-selected answers. The average review time per item was 3.21 minutes, indicating efficiency advantages over traditional question drafting.

Among these studies, only Bedi et al.'s (2025) included automatic evaluation, in the form of an ensemble of language models to automatically flag flawed questions. None of the studies directly evaluated the AI-generated items in the context of operational assessments.

The literature so far provides important insights across a range of use cases, highlighting both the promise and current challenges of AI-assisted item development. However, the diversity in study designs, evaluation rubrics, expert backgrounds, and question types makes direct comparison across studies difficult. Most studies rely on expert judgment, while automated evaluation remains underexplored, with only preliminary use by Bedi et al. (2025). Our study is among the first to investigate automated methods for evaluating AI-generated medical questions, aiming to complement expert review with scalable and consistent quality checks.

3 Data

The dataset used in this study comprises 512 clinical MCQs generated by GPT-4-0314, aiming to cover 26 topics across various clinical domains. These include, but are not limited to, the respira-

tory system, renal and urinary systems, obstetrics and gynecology, behavioral disorders, and gastroenterology. The items were evaluated by ten subject matter experts (SMEs), who were physicians with extensive experience in writing clinical MCQs for high-stakes standardized assessments. The evaluation was organized such that each item was annotated by two SMEs and each SME saw ≈100 items. Additionally, the annotation was organized so that 5 pairs of SMEs that each shared the same domain of expertise annotated the same set of items that were grouped by topic. This paired assignment of SMEs to topics was necessary due to the highly technical nature of the MCQ content and a need to ensure, to the extent possible, that the SMEs had the right background to evaluate the items.

The SMEs were first shown an item stem (the clinical scenario that presents the problem to be solved) along with the key (the correct answer). They were then given up to 12 distractors (incorrect answer options) and asked to select those that, collectively, could form a partial or complete option set for the item. Following this selection, the SMEs evaluated several aspects of the item drafts, including their usefulness as starting points for developing items for a high-stakes clinical assessment. Each draft was rated as either a *Helpful* starting point (requiring relatively minor changes) or a *Non-Helpful* starting point (requiring substantive revisions). Optionally, SMEs also provided rationales for their selections.

When providing their ratings, SMEs were instructed to label drafts as *Helpful* starting points if only minimal revisions were needed. This included small edits to the stem—such as adding, modifying, or removing up to three minor history or exam details for accuracy, realism, or appropriateness—or minor changes to the answer options, like adding a distractor to complete a 4–5 option set with appropriate difficulty. Drafts requiring more substantive revisions to the stem or answer options were to be labeled *Non-Helpful* starting points. These guidelines were intended as reference points, with SMEs encouraged to use their judgment in assessing the overall effort required to finalize a draft. The instructions were presented both in writing and verbally during a training session, where SMEs also rated three sample items together. The discussion with the SMEs revealed the limitations of rigid criteria based on a specific number of item edits, as SMEs noted that a single change can sometimes require significant effort, while multiple superficial edits may be quick and easy to implement.

Data	Helpful	Non-Helpful
Set 1	280	232
Set 2	280	68

Table 2: Distribution of helpful and non-helpful items with (Set 1) and without (Set 2) SME disagreements.

This study focuses on developing an automated evaluation of the quality of AI-generated drafts by using the draft item as input and predicting the labels assigned by the SMEs. These labels are defined as follows: if both SMEs agreed that a draft was a helpful starting point, it is labeled *Helpful*. If at least one of the SMEs rated the draft as not helpful, it is labeled *Non-Helpful*, because the system should preferably reject any item draft that could be labeled as Non-Helpful by human annotators in order to streamline the review process. The data distribution following this labeling method is shown in Table 2 as Set 1. In a follow-up analysis, we refine the label distribution by removing the cases where the SMEs disagreed and consider only item drafts that were labeled by both annotators as either Helpful or Non-Helpful (Set 2 in Table 2). As shown in the following sections, this reduces labeling noise caused by rater disagreement but introduces class imbalance, making the classification task more challenging.

4 Analysis of Human Annotations

Overall, 70.8% of the individual annotations provided by the SMEs characterized the item as Helpful. However, the distribution of Helpful vs. Non-Helpful annotations varied substantially across raters with 47.1% Helpful ratings for the most rigorous annotator and 88.8% Helpful ratings for the most lenient annotator. These results suggest that the SMEs had different subjective interpretations of the definitions of Helpful and Non-Helpful provided in the annotation guidelines. The inter-annotator agreement statistics provide additional evidence for the challenging nature of the annotation task. The overall inter-annotator exact agreement was 67.1% and Cohen's κ was 0.223. Across the five pairs of raters that annotated the same sets of items, exact agreement ranged from 52.0% to 73.3% and Cohen's κ ranged from 0.078 to 0.376.

5 Automated Classification of Helpfulness

We conducted three types of automated classification experiments to predict the helpfulness of gen-

erated items based on the expert judgments. These included: (1) a feature-based approach utilizing interpretable features, (2) fine-tuned transformer models, and (3) an LLM judge utilizing few-shot learning with a focus on prompt engineering to explore creative prompting strategies.

5.1 Feature-Based Classification

In our experiment with hand-crafted interpretable features, we conducted a 5-fold cross-validation experiment utilizing a Random Forest (RF) classifier implemented via the `scikit-learn` Python library (Pedregosa et al., 2011). The models were trained on two types of manually engineered features: word count-based features and readability metrics. The word count features captured surface-level textual patterns such as the total word count in the item stem, the number of words in the key, the average word count between distractors, and the maximum word count between distractors. Readability was assessed using the Flesch-Kincaid Grade Level and Flesch Reading Ease scores (Kincaid et al., 1975), which estimate the linguistic complexity of the item content. These features serve as an interpretable baseline intended to quantify the extent to which surface characteristics such as item length are predictive of the two classes.

5.2 Transformer Models

We performed a 5-fold cross-validation experiment and fine-tuned three models: BERT-base-uncased, DeBERTa-v3-base, and DeBERTa-v3-large from HuggingFace (Wolf et al., 2020). The following parameters are used to fine-tune all models: batch size of 16, learning rate of $9e^{-6}$, 50 warmup steps, and a weight decay of 0.01. The input to the models consists of the stem, answer key, and distractor list, each separated with a [SEP] token.

5.3 LLM as Judge with Few-shot Learning

We used GPT-4 (OpenAI et al., 2024) as a judge to determine the helpfulness of the generated item drafts. We employed few-shot prompting and tested the following four distinct prompt designs (refer to Appendix A for the complete prompts):

Simple Prompt: In this approach, we did not provide detailed instructions to the LLM. We instructed the LLM to take the a role of a highly knowledgeable medical educator, provided it with two labeled examples (one Helpful item requiring few edits and one Non-Helpful item requiring major edits), and then asked it to classify a third item.

Criteria-Based Prompt: In this prompting strategy, the LLM was prompted to act as a highly knowledgeable medical educator and was given a set of review criteria, including clarity, relevance, validity, formatting, cognitive level, and statistical usability. Similar to the simple prompt, two labeled examples were followed by a third item to be classified. In this case, the model was explicitly instructed not to provide an explanation.

Criteria-Based Prompt with Rationale: This prompting strategy followed the structure of the criteria-based prompt. In addition, the SME rationales for the example items in the prompt were included, and the model was instructed to provide a clear rationale for its decision.

Similarity-Based Prompt with Rationale: Building on the third prompt, this version improved the example selection process by choosing examples most similar to the target item. Similarity was computed using cosine distance between sentence-level vector embeddings of the items. The sentence vectors were extracted from the sentence transformer embedding model (Zhang et al., 2025).

6 Results

We evaluated all models using two metrics: weighted F1-score and accuracy, with results presented in Table 3. For Set 1, which includes cases that SMEs disagreed upon, the majority baseline yielded an F1-score of 0.387 and an accuracy of 0.547. For Set 2, without SME disagreements, the corresponding scores were 0.718 and 0.805.

Comparative analysis indicates that, while the feature-based Random Forest classifier outperformed the baseline in terms of F1 score, it consistently underperformed on the accuracy metric across both sets. Notably, for all feature ablation combinations, the classifier's accuracy remained below the majority baseline. Among the feature sets, word count features achieved the best performance, suggesting that item length provides a predictive signal when modeling helpfulness. To further investigate this relationship, we computed the Pearson correlation between the helpfulness label and various hand-crafted features. Interestingly, word count features did not exhibit a statistically significant correlation with helpfulness. However, the number of words in the item stem and the readability measured via the Flesch Reading Ease score showed a negative correlation (shown in Table 4).

The fine-tuned transformer models outperformed

		Set 1		Set 2	
		F1-score	**Accuracy**	**F1-score**	**Accuracy**
Baseline	Majority class	0.387	0.547	0.718	0.805
RF	Word count features	**0.505**	0.506↓	**0.741**	0.779↓
	Readability features	0.479	0.482↓	0.709↓	0.747↓
	All features	0.486	0.492↓	0.739	0.796↓
Transformers	BERT	0.558	0.561	0.769	0.802↓
	DeBERTa-base	**0.589**	**0.604**	0.718	0.805
	DeBERTa-large	0.564	0.568	**0.771**	**0.810**
GPT-4	Simple Prompt	0.465	0.575	**0.768**	**0.815**
	Criteria-Based Prompt	0.482	0.573	0.755	0.792↓
	Criteria-Based Prompt w/ Rationale	**0.537**	**0.586**	0.742	0.754↓
	Similarity-Based Prompt w/ Rationale	0.534	0.574	0.732	0.732↓

Table 3: Comparison of model performance using weighted F1-score and accuracy. Models that did not improve over the baseline are marked with the ↓ symbol. The best performing models within each type are marked in bold.

Features	Correlation	P-val Range
Word Count		
Stem	-0.062	[0.093, 0.500]
Answer	0.015	[0.523, 0.798]
Avg. Distractor	0.022	[0.401, 0.974]
Max. Distractor	0.016	[0.481, 0.944]
Readability		
Grade Level	0.053	[0.051, 0.813]
Reading Ease	-0.062	[0.019*, 0.910]

Table 4: Average magnitudes and p-value ranges for correlations between the helpfulness label and handcrafted features over 5-fold cross validation. $*p < 0.05$

the majority baseline, with the exception of BERT-base-uncased on the accuracy metric. DeBERTa-v3-base consistently outperformed both BERT-base-uncased and DeBERTa-v3-large on Set 1, which includes item drafts with discrepant SME ratings. In contrast, DeBERTa-v3-large achieved the best performance on Set 2, where item drafts with discrepant ratings were removed.

When using GPT-4 to assess helpfulness, we evaluated its performance both with and without rationale explanations. On Set 1, which includes item drafts with discrepant SME ratings, the *Criteria-Based Prompt with Rationale* outperformed all prompting strategies. While the *Similarity-Based Prompt with Rationale* yielded competitive results, it did not surpass the performance of the *Criteria-Based Prompt with Rationale*. In contrast, for Set 2, with no discrepant SME ratings, the *Simple Prompt* achieved the highest performance. The other three prompts did not exceed the baseline accuracy on Set 2, suggesting that in the absence of discrepancy, GPT-4 performs best with simple prompts.

7 Error Analysis

The experiments presented in Section 6 show that modeling draft helpfulness is a challenging task for various classifiers. Our findings identify DeBERTa-base and the *Criteria-Based Prompt with Rationale* as the most effective approaches, which is why we focus on these two models to further understand their error patterns.

The confusion matrices in Appendix B show that both the DeBERTa-base and GPT-4 models are effective at recalling Helpful items—correctly identifying many of them as Helpful. However, both models also exhibit a tendency to incorrectly classify Non-Helpful items as Helpful. We suspect our labeling choice of marking items with annotator disagreement as Non-Helpful may have contributed to the misidentification. To better understand this effect, we analyzed how the models handled the items marked by disagreement. Of the 164 such items, DeBERTa-base predicted 103 (63%) as Helpful and 61 (37%) as Non-Helpful. The GPT-4 model using the criteria-based prompt with rationale predicted 126 (77%) as Helpful and only 38 (23%) as Non-Helpful. These findings suggest that, although items with discrepant ratings were labeled as Non-helpful, the models exhibit a consistent bias toward predicting them as Helpful.

To better understand why a Non-Helpful item might have been labeled as Helpful, we examined the comments provided by both human annotators and the rationales generated by GPT-4 as part of the *Criteria-Based Prompt with Rationale*. On average, GPT-4's rationales were 4.6 times longer than those written by humans. A qualitative analysis reveals that human annotators tend to focus more on issues related to the stem of the question, whereas GPT-4 emphasizes aspects of the distractors and the correct answer. Additionally, GPT-4 often describes items as clear, concise, and easy to understand. Human rationales also contain highly technical medical knowledge about flaws in the

Human Rationale	GPT-4 Rationale
Draft item seems to hint at diagnosis (medication known to cause acute kidney injury, AKI), and yet answer relates to hematuria and proteinuria, which may or may not be seen in AKI. Also, shortness of breath is not further elaborated on in history or physical exam. If this is fluid overload in setting of AKI, would likely have crackles or lung findings. Draft item lacks focus and is not necessarily aligned with indicated answer.	The question stem is clear, concise, and relevant to the intended learning objectives. The options are unambiguous and easy to understand. However, the correct answer should be revised to match the format of the other options. Instead of "Proteinuria and/or hematuria," the correct answer should be written as "Protein and hematuria." This change will make the correct answer consistent with the other options and maintain the clarity of the question. Additionally, the question matches the cognitive level required for the audience and is free of bias and stereotypes.

Table 5: Human annotator rationale vs. GPT-4 rationale for an example where the item draft was labeled as Non-Helpful by the SMEs but predicted as Helpful by GPT-4 using the Criteria-Based Prompt with Rationale.

item draft that the GPT-4 may not pick up.

Given that the items were generated by GPT-4 and are also being evaluated by GPT-4, this may partly explain its tendency to find the items easier to understand. Table 5 presents an example where the item was labeled as Non-Helpful by human annotators but predicted as Helpful by GPT-4, along with their respective rationales.

8 Discussion and Conclusion

This study investigated human and automated evaluations of AI-generated item drafts, intended to serve as starting points for item development. The results indicate that this is a challenging task for both experts and machines. Despite our efforts to mitigate variability—through detailed guidelines, topic-to-rater matching, and a group calibration exercise—inter-rater agreement remained modest.

A likely explanation lies in the inherently subjective nature of the task. A given draft may evoke different ideas and interpretations depending on the item writer's experience, domain-specific preferences, or approaches to the item development process. Writers may also vary in their thresholds for what they perceive as a "substantive" effort required to revise a draft. While future research should further refine rating criteria and protocols, the subjective nature of evaluating helpfulness is unlikely to be eliminated entirely.

Turning to the automated evaluation, classification models exhibited modest success. Even when analysis was limited to instances where both raters agreed—a subset of examples with arguably clearer ground truth—the performance of the classifiers remained moderate. One contributing factor was class imbalance within the dataset (while this skew affects supervised models, the GPT-4-based few-shot prompting approach used a balanced set of examples, mitigating this issue during inference). Notably, rationale-augmented prompts improved GPT-4's performance in Set 1, suggesting that structured reasoning can help guide the model's decisions in more complex cases. However, in Set 2, with no disputed labels, the simple prompt outperformed more elaborate versions—highlighting that, in low-ambiguity scenarios, additional reasoning may introduce unnecessary "cognitive" load and reduce accuracy.

Several limitations warrant consideration. First, the relatively small sample size constrains the generalizability of our findings. Model performance may differ across item formats or subject areas not represented in our dataset. There could also be potential misalignments between item content and rater expertise. For example, an item involving pediatric trauma may have been assigned to a general pediatrician, whereas the underlying clinical focus would have been more suitable for an emergency physician. In addition, the choice of particular specialists in this study were limited by the item writer's availability and addressing these limitations in future work is important for obtaining a more robust evaluation.

The practical aim of this study was to explore whether automated evaluation methods could help streamline the human review process. The extent to which this goal was achieved remains open to interpretation. On the one hand, several of the classifiers outperformed baseline models and demonstrated reasonable recall for identifying helpful items (in other words, there is lower risk of discarding helpful drafts). However, the relatively low precision in identifying unhelpful drafts—when combined with subjective preferences— limits the utility of such methods in practice. Further precision refinement is needed before automated triaging can be considered a dependable aid in clinical test development.

References

Mayank Agarwal, Priyanka Sharma, and Ayan Goswami. 2023. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. *Cureus*, 15(6):e40977.

Somaiya Al Shuraiqi, Abdulrahman Aal Abdulsalam, Ken Masters, Hamza Zidoum, and Adhari AlZaabi. 2024. Automatic generation of medical case-based multiple-choice questions (mcqs): A review of methodologies, applications, evaluation, and future directions. *Big Data and Cognitive Computing*, 8(10).

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(1):354.

Ibraheim Ayub, Dathan Hamann, Carsten R. Hamann, and Matthew J. Davis. 2023. Exploring the potential and limitations of chat generative pre-trained transformer (chatgpt) in generating board-style dermatology questions: A qualitative analysis. *Cureus*, 15(8):e43717.

Suhana Bedi, Scott L. Fleming, Chia-Chun Chiang, Keith Morse, Ankit Kumar, Bhavik Patel, Jindal A. Jindal, Christopher Davenport, Christina Yamaguchi, and Nigam H. Shah. 2025. QUEST-AI: A System for Question Generation, Verification, and Refinement using AI for USMLE-Style Exams. In *Proceedings of the Pacific Symposium on Biocomputing 2025*, pages 54–69.

Billy H. H. Cheung, Gary K. K. Lau, Gordon T. C. Wong, Elaine Y. P. Lee, Dhananjay Kulkarni, Choon S. Seow, Ruby Wong, and Michael Co. 2023. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLOS ONE*, 18(8):e0290691.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 118:102963.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated readability index. *Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel: Inst Sim Trng*.

E. Klang et al. 2023. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Medical Education*, 23(1):1–9.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Poko-

rny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yishen Song, Junlei Du, and Qinhua Zheng. 2025. Automatic item generation for educational assessments: a systematic literature review. *Interactive Learning Environments*, pages 1–20.

Liyan Tang, Ziyue Sun, Betina R. Idnay, Gongbo Zhang, Yufan Zhang, Chen Wang, Yanshan Zhang, and Hong Yu. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6:158.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Victoria Yaneva, Peter Baldwin, Daniel P. Jurich, Kimberly Swygert, and Brian E. Clauser. 2024. Examining chatgpt performance on usmle sample items and implications for assessment. *Academic Medicine*, 99(2):192–197.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models. *Preprint*, arXiv:2412.19048.

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 LLM Prompts

Simple Prompt

System:
Here are 2 examples of medical MCQ questions where the first example is *Non-helpful* and the second example is *Helpful*. Given a third example, your job is to answer if it is Helpful or Non-helpful.

User:
Example 1: [example 1 question]
Answer: [example 1 answer]
Options: [example 1 options]
Label: Non-helpful

Example 2: [example 2 question]
Answer: [example 2 answer]
Options: [example 2 options]
Label: Helpful

Example 3: [test example question]
Answer: [test example answer]
Options: [test example options]

Is the third example *Helpful* or *Non-helpful* ?"

Criteria-Based Prompt

System:
You are a highly knowledgeable medical educator and expert in medical exam question design. Your task is to review a set of Multiple Choice Questions (MCQs) intended for a medical education platform.

Criteria:

- Clarity and Conciseness: Is the question stem clear and concise, avoiding unnecessary complexity? Are the options unambiguous and easy to understand?

- Relevance and Focus: Does the question align with the intended learning objectives or topic? Is it free of irrelevant or extraneous details that might confuse the respondent?

Criteria-Based Prompt (Cont..)

- Answer Key Validity: Is the correct answer clearly supported by the question and defensible? Are distractors (incorrect options) plausible but clearly incorrect?

- Formatting and Grammar: Is the question grammatically correct, free of typos, and formatted appropriately?

- Cognitive Level: Does the question match the cognitive level (e.g., recall, application, analysis) required for the audience or context?

- Bias and Sensitivity: Is the question free of bias, stereotypes, or language that might disadvantage certain groups?

- Statistical Usability (Optional): Does the question have characteristics likely to yield good discrimination and difficulty levels if data is available?

User:
Here are two examples of well-structured MCQs where the first example is *Non-helpful* and the second example is *Helpful*:

Example 1:
- Question: [example 1 question]
- Options: [example 1 options]
- Correct Answer: [example 1 answer]
- Label: Non-helpful

Example 2:
- Question: [example 2 question]
- Options: [example 2 options]
- Correct Answer: [example 2 answer]
- Label: Helpful

Now, classify the following question:

- Question: [test example question]
- Options: [test example options]
- Correct Answer: [test example answer]

Criteria-Based Prompt (Cont..)

Instruction:
ONLY return one of the following labels:
- *Non-helpful*
- *Helpful*

Do **NOT** provide any additional explanation.

Criteria-Based Prompt with Rationale

System:
You are a highly knowledgeable medical educator and expert in medical exam question design. Your task is to review a set of Multiple Choice Questions (MCQs) intended for a medical education platform.

Criteria:

- Clarity and Conciseness: Is the question stem clear and concise, avoiding unnecessary complexity? Are the options unambiguous and easy to understand?

- Relevance and Focus: Does the question align with the intended learning objectives or topic? Is it free of irrelevant or extraneous details that might confuse the respondent?

- Answer Key Validity: Is the correct answer clearly supported by the question and defensible? Are distractors (incorrect options) plausible but clearly incorrect?

- Formatting and Grammar: Is the question grammatically correct, free of typos, and formatted appropriately?

- Cognitive Level: Does the question match the cognitive level (e.g., recall, application, analysis) required for the audience or context?

- Bias and Sensitivity: Is the question free of bias, stereotypes, or language that might disadvantage certain groups?

Criteria-Based Prompt with Rationale (Cont..)

- Statistical Usability (Optional): Does the question have characteristics likely to yield good discrimination and difficulty levels if data is available?

User:
Here are two examples of well-structured MCQs where the first example is *Non-helpful* and the second example is *Helpful*:

Example 1:
- Question: [example 1 question]
- Options: [example 1 options]
- Correct Answer: [example 1 answer]
- Label: Non-helpful
- Rationale: [example 1 rationale]

Example 2:
- Question: [example 2 question]
- Options: [example 2 options]
- Correct Answer: [example 2 answer]
- Label: Helpful
- Rationale: [example 2 rationale]

Now, classify the following question:

- Question: [test example question]
- Options: [test example options]
- Correct Answer: [test example answer]

Instruction:
ONLY return one of the following labels:
- *Non-helpful*
- *Helpful*

Provide a clear Rationale for your assessment, highlighting any issues related to the system criteria.

B Error Analysis Confusion Matrices