



ORIGINAL ARTICLE

Received: 11.10.2024

Accepted: 23.11.2024

Published: 23.12.2024

CITE THIS ARTICLE AS:

Kiyak YS, Kononowicz AA, Gorski S, "Multilingual Template-based Automatic Item Generation for Medical Education Supported by Generative Artificial Intelligence Models ChatGPT and Claude," Bio-Algorithms and Med-Systems vol. 20, no. 1, pp. 81-89, 2024, DOI: 10.5604/01.3001.0054.9192

AUTHORS' CONTRIBUTION:

- A. Conceptualization
- B. Data Curation
- C. Formal Analysis
- D. Funding Acquisition
- E. Investigation
- F. Methodology
- G. Project Administration
- H. Resources
- I. Software
- J. Supervision
- K. Validation
- L. Visualization
- M. Writing – Original Draft
- N. Writing – Review & Editing

CORRESPONDING AUTHOR:

Selim Kiyak; Gazi Üniversitesi Hastanesi; E Blok 9. Kat 06500 Beşevler, Ankara, Turkey; Phone: +90 312 202 7445; E-mail: yskiyak@gazi.edu.tr

COPYRIGHT:

Some rights reserved: Jagiellonian University Medical College. Published by Index Copernicus Sp. z o. o.

OPEN ACCESS:

The content of the journal „Bio-Algorithms and Med-Systems” is circulated on the basis of the Open Access which means free and limitless access to scientific data.

CREATIVE COMMONS CC BY:

Attribution. It is free to copy, distribute, present and perform the copyrighted work and derivative works developed from it, provided that the name of the original author is cited.

Multilingual Template-based Automatic Item Generation for Medical Education Supported by Generative Artificial Intelligence Models ChatGPT and Claude

Yavuz Selim Kiyak^{1,2}[ABCDGFMN](#) , Andrzej A. Kononowicz²[ABDFGN](#) ,
Stanisław Górski³[BFKN](#) 

¹Department of Medical Education and Informatics, Faculty of Medicine, Gazi University, Ankara, Turkey

²Department of Bioinformatics and Telemedicine, Jagiellonian University Medical College, Kraków, Poland

³Department of Medical Education, Jagiellonian University Medical College, Kraków, Poland

ABSTRACT

Objective: This study has two main aims. (1) To generate multiple-choice questions (MCQs) using template-based automatic item generation (AIG) in Polish and to evaluate the appropriateness of these MCQs in terms of assessing clinical reasoning skills in medical education; (2) to present a method for using artificial intelligence (AI) to generate new item models based on existing models for template-based AIG in medical education.

Methods: This was a methodological study. For the first aim, we followed Gierl's three-step template-based AIG method to generate MCQ items in Polish. The quality of the generated MCQs were evaluated by two experts using a structured form. For the second aim, we proposed a four-step process for using a parent template in English to transform it into new templates. We implemented this method in ChatGPT and Claude by using two medical MCQ item models.

Results: Both experts found the automatically generated Polish questions clear, clinically sound, and suitable for assessing clinical reasoning. Regarding the template transformation, our findings showed that ChatGPT and Claude are able to transform item models into new models.

Conclusions: We demonstrated the successful implementation of template-based AIG in Polish for generating case-based MCQs to assess clinical reasoning skills in medical education. We also presented an AI-based method to transform item models for enhancing diversity in template-based AIG. Future research should integrate AI-generated models into AIG, evaluate their exam performance, and explore their use in various fields.

KEYWORDS

automatic item generation, large language models, artificial intelligence, clinical reasoning, multiple-choice questions, medical education

LIST OF ABBREVIATIONS

- AI** – artificial intelligence
AIG – Automatic Item Generation
LLMs – large language models
MCQs – multiple-choice questions

INTRODUCTION

Clinical reasoning is a critical skill that requires particular attention to its teaching and assessment in medical education. This skill has been assessed through various methods including multiple-choice questions (MCQs), key-feature questions, script concordance tests, concept maps, oral examinations, objective structured clinical examinations, direct observations, and more [1]. These methods vary significantly in their efficiency, scalability, and the depth of reasoning they assess.

Case-based MCQs have been used for a long time as an efficient and scalable method to assess clinical reasoning due to their text-based content for simulating real-life scenarios that health professionals encounter [2]. However, the development of high-quality, context-rich MCQs that assess clinical reasoning is a labour-intensive process that necessitates significant resources. Not only does it require the input of experienced subject matter experts, but it also entails a considerable time commitment and thus financial resources. This is particularly challenging for medical schools that require a large number of MCQs to cover the diverse range of topics necessary for medical education.

The use of Automatic Item Generation (AIG) presents an efficient solution to this challenge. AIG is a method that uses computer algorithms to generate test items based on predefined models and parameters [3]. AIG can be classified into two main categories: template-based and non-template-based. Both methods are more efficient than traditional item writing. While non-template-based methods are based on generative artificial intelligence (AI), such as ChatGPT [4–9], and susceptible to inaccuracies (e.g., biases, “hallucinations”) that stem from the data and the underlying model, template-based AIG can significantly streamline the process of question development and still possesses some advantages in comparison to non-template based methods, owing to accuracy not being sacrificed and to consistency in the questions.

Open access medical education and formative assessment resources are often limited by their availability in English language only. While it has value in reaching goals related to internationalization of study programs, the fact that the students use resources in their learning and assessment that are not in their mother-tongue or the language they will mainly use in their subsequent clinical practice may additionally increase the cognitive load, which may not be desirable in many educational scenarios. Even students who profess to have good English skills are affected

by this in their education, as a study on English and Romanian virtual patients has demonstrated [10]. To address this challenge, international initiatives have established repositories of educational content that are deliberately created in multiple languages. For instance, the iCoViP project has delivered a collection of more than 200 virtual patients for clinical reasoning education, each available in six languages [11]. Likewise, template-based AIG in the context of medical education has been successfully implemented in six languages: English, French, Chinese, Spanish, Korean, and most recently Turkish [3, 12, 13]. However, template-based AIG in medical education has not yet been implemented in Polish. In this study we aim to demonstrate how, in the context of a bilateral Polish-Turkish project, we added a new language version to the pool of template-based AIG methods.

The second challenge related to AIG methods that we aim to address was the limited diversity in the appearance of template-based generated MCQ questions. Since items are generated based on assigning specific values to components on pre-determined templates according to pre-determined rules, the resulting presentation of the questions may be limited [3]. This can affect the safety of tests as students may memorize templates, and it may also lower their acceptance among students as questions similar in appearance may be regarded as less challenging and inauthentic. This limitation can be addressed by generating more templates. Generative AI, such as in the form of ChatGPT or Claude, can be used to increase the heterogeneity of templates. A study conducted by Sayin et al. [14] used AI to increase the creativity and variety of templates in AIG. In their successful implementation, GPT was able to generate appropriate and unique sentences for use in reading comprehension item models. However, the use of AI for a similar purposes in medical education has not been demonstrated yet.

OBJECTIVE

Our study aims to show a workflow of translation and enhanced diversity of MCQs from template-based AIG. In particular, our goals were:

1. to generate MCQs using template-based AIG in Polish and to evaluate the appropriateness of these MCQs in terms of assessing clinical reasoning skills in medical education;
2. to present a method for using AI to generate new item models based on existing models for template-based AIG in medical education.

We conducted two consecutive study processes to address the respective research questions.

METHODS

This is a methodological research study.

Item Generation Process for Polish

The item (question) generation process consisted of these three main steps, as template-based AIG recommends [15]: development of a cognitive model, development of an item model, and generation of items using software. We followed these steps by using an existing cognitive model and an item template on hypertension in Turkish [12].

Stage 1 – Cognitive Model Development: the Turkish cognitive model was adapted to the Polish context collaboratively by the

research team, communicating in English to bridge the language barrier.

Stage 2 – Item Model Development: the foundation for this model was the item model used in the Turkish AIG study. It has been adapted by a Polish cardiologist. Necessary changes were made to tailor the content to the Polish healthcare context and to medical education based on medical guidelines and clinical/educational experience. This model provided the template for the MCQs. Tab. I. shows the item model (the English version can be found in the Supplementary Material).

Tab. I. The item model.

<AGE>-LETNI MĘŻCZYŻNA ZGŁASZA SIĘ DO PRZYCHODNI PODSTAWOWEJ OPIEKI ZDROWOTNEJ NA RUTYNOWE BADANIE PROFILAKTYCZNE. U PACJENTA AUTOMA <BLOOD PRESSURE>. PACJENT, U KTÓREGO WCZEŚNIEJ NIE ZDIAGNOZOWANO NADCIŚNIENIA TĘTNICZEGO, <COMPLAINT>. PACJENT <SMOKING>. W BADANIU FIZYKALNYM WSKAŹNIK MASY CIAŁA PACJENTA WYNOŚI <BMI VALUE> KG/M².

W przypadku tego pacjenta, który nie ma dodatkowych chorób współistniejących, oprócz zalecenia zmian w stylu życia, która z poniższych opcji jest najbardziej odpowiednia w porównaniu z innymi?

- Wystarczające jest zalecenie mierzenia ciśnienia krwi raz w roku.
- Pacjent powinien zgłosić się ponownie po co najmniej pięciu dniach od pomiaru ciśnienia krwi, bez żadnych interwencji.
- Zalecenie zmian w stylu życia jest wystarczające, a wizyta kontrolna powinna odbyć się po miesiącu.
- Należy przepisać leki hipotensyjne i umówić wizytę kontrolną po miesiącu.
- Należy zapewnić niezbędną pierwszą interwencję i natychmiast wysłać pacjenta na oddział ratunkowy.

Elements and their values to be assigned to corresponding parts in the question above:

<AGE>: 32, 44, 56

<BLOOD PRESSURE>:

- ciśnienie krwi wynosi 110/70 mmHg na podstawie pomiaru wykonanego raz w przychodni;
- średnie ciśnienie krwi w okresie 24 godzin jest rejestrowane jako 130/82 mmHg;
- pomiary uzyskane podczas dwóch oddzielnych wizyt w przychodni są rejestrowane jako 150/95 mmHg. Jednak po uwzględnieniu pomiarów wykonanych w domu średnia wynosi 145/93 mmHg;
- ciśnienie krwi wynosi 165/100 mmHg na podstawie dwóch pomiarów podczas dwóch oddzielnych wizyt w przychodni;
- ciśnienie krwi wynosi 190/130 mmHg na podstawie jednego pomiaru w przychodni.

<COMPLAINT>:

- nie występują dodatkowe dolegliwości;
- nie występują dodatkowe dolegliwości poza tym, że ma czasami łagodne bóle głowy.

<SMOKING>:

- nie pali;
- pali jedną paczkę papierosów dziennie.

<BMI VALUE>: 20, 23, 25, 29

Stage 3 – Item Generation: the MCQs were generated using the software that was employed in the Turkish AIG research. The software was specifically adapted for this study to accommodate the Polish template. This adaptation involved the same programming language and technologies (Python-based code and MySQL database) to ensure that the item generation process remained consistent, while being tailored for single-use in the Polish context.

This three-stage process has been carried out two times. In the second iteration, we updated the item model by taking into

account expert feedback and then generated the questions based on the updated model. Tab. I. shows the final version of the template.

Evaluation of the Polish MCQs

In the evaluation of Polish items, we adhered to a similar approach to that of previous studies [16, 17], wherein subject-matter experts were asked to appraise the quality of questions generated through the use of an AIG method. A random selection of the generated questions was evaluated by two subject

matter experts – one cardiologist and one internal medicine specialist – both of whom possessed over ten years of experience in teaching and assessment in medical education. They evaluated ten questions, which had been chosen randomly among 240 questions generated, using a standardized form to assess the clarity, clinical soundness, and relevance of the questions to clinical reasoning. The Supplementary Material section presents the form that includes the evaluation criteria and the ten MCQs generated by using the Polish item model (presented in Tab. I.). In the results section, we report the expert input in two rounds as a narrative.

Generating New Item Models Using AI

To demonstrate that AI can help us in generating new item models based on existing item models, we used two “parent” item models, which consist of existing templates built by human experts. The first item model was the English version of the Polish template presented in Tab. I. The second one was the item model used by Gierl et al. [3].

We developed a prompt (Tab. II.) to explain what we need from AI models. We designed it in such a way that the transformed templates remain compatible with the cognitive models underlying them.

Tab. II. The prompt for transforming item models using AI.

You are the Item Model Transformer, a specialized tool for transforming adaptable medical multiple-choice question (MCQ) item models into a different form by paraphrasing, mixing up the sentences, and/or changing their structures. This is to make them LOOK COMPLETELY DIFFERENT from the parent item model but have EXACTLY THE SAME MEANING, and ensuring they flow well and have no grammatical errors and maintaining compatibility with medical education and medicine terminology. The transformed item model should also strictly maintain grammatical correctness and logical coherence to make it versatile and robust, capable of handling various scenarios based on the values accurately.

You will ask the user to provide a parent item model. The parent item model consists of three main components:

1. A question 2. Answer options 3. Elements and their values.

The question-and-answer options could include some parts marked with special characters to show the parts that correspond to the elements that can be filled with the values.

You should ensure that, in the transformed version, when the corresponding values are assigned for the marked parts, the sentence is meaningful and free from any language and grammatical errors. This is crucial because, using different software, the item model will be used for generating new items (MCQs) by assigning values to the marked parts. Their combinations will be generated and used in MCQ exams in medical schools, so they must be meaningful and free from any language and grammatical errors. Your efforts will indirectly affect people's health, so be careful.

This is a three-step process that starts with receiving the parent item from the user.

Step 1: Generate the transformed item model, which includes the transformed versions of the question, the answer options, and elements and their values.

Step 2: Ask the user if there are any problems with the transformed item model.

Step 3: Based on the user's input, make necessary changes.

Repeat this process until the user approves the transformed item model.

We defined and followed a four-step process for using a parent template to transform it into new templates:

1. Choose an AI model: We chose ChatGPT-4 and Claude 3, large language models (LLMs), for our demonstration as they are the most advanced AI models as of April 2024. Claude 3 has three models: Opus, Sonnet, and Haiku. Opus is the most advanced, while Haiku is the least advanced. The cost of using them follows the opposite order: Haiku is the cheapest with Opus being the most expensive. We used Sonnet as an optimal choice.

2. Enter the prompt or use a customized AI tool: We used the prompt in Tab. II. Since ChatGPT allows us to develop custom GPTs for specific purposes [18] and it has been used in medical education [7] and medicine [19], we developed a custom GPT named “Item Model Transformer” (accessible via this link: <https://chat.openai.com/g/g-tYcMdrxl-item-model-transformer>). We used

the prompt in the instruction of this custom GPT. Since Claude 3 does not allow creating custom tools without coding as of April 2024, we directly used the prompt in Claude 3 Sonnet.

3. Prepare and provide an item model to the AI, and provide feedback in the conversation if necessary: We prepared the existing item models for AI because the original presentations of the item models could be difficult for AI tools to properly interpret. For example, when an element is represented as a proper word or phrase, such as “<BLOOD PRESSURE>”, it is more prone to misunderstand the sentence structure. For this purpose, we used a standard (such as “<E1>”) to show the parts (elements) that will be manipulated. Additionally, we presented the values of the elements in a bulleted point form.

4. Make remaining necessary changes outside of artificial intelligence environments: Sometimes it is difficult for humans to

express their needs regarding AI support. Moreover, as of April 2024, expert review is still essential before using the content generated by AI, especially in fields like medicine. Therefore, it is important to review the content generated and make any remaining necessary changes, preferably together with the support of a group of experts.

In this four-step process, we kept the human involvement as low as possible in order to show the capabilities of AI tools.

Ethical Approval

The study did not require ethical approval, as it did not involve human participants.

RESULTS

Evaluation of Polish Items

Round 1: Expert 1 critiqued the artificial nature of some questions, such as phrases implying non-smoking as a symptom, which could detract from their clinical relevance. Additionally, the expert noted a possible improvement of clarity regarding how medium blood pressure was obtained, whether through patient self-monitoring or automatic blood pressure monitoring. Despite these issues, the questions were deemed clear, clinically sound, and capable of assessing clinical reasoning. Expert 2 identified issues with answer choices related to lifestyle changes and unspecified interventions,

suggesting that these could lead to confusion among students. We made the necessary changes based on these comments.

Round 2: Expert 1 expressed general satisfaction with the questions, finding no significant issues to report. The clarity and clinical relevance of the questions were upheld. Expert 2 confirmed that all questions were clear, clinically sound, and well-suited for assessing clinical reasoning skills. The expert also observed that the similarity of answer options across different questions could potentially reduce how engaging the exam is to students. While noting the potential for monotony due to similar answer choices, Expert 2 also acknowledged that this consistency might aid students in understanding the questions better and encourage deeper clinical reasoning. Both experts stated that the items were suitable for assessing clinical reasoning rather than only factual recall.

Generating New Item Models Using AI

We presented the parent templates and the transformed templates by Claude 3 (Sonnet) and ChatGPT-4 in Tab. III. for our parent template and in Tab. IV. for Gierl et al.'s template. While Claude 3 does not have a feature to share conversations via a link (as of April 2024), the conversation with ChatGPT is publicly accessible from these links:

Our template: <https://chat.openai.com/share/7d885d61-09ee-4cd1-9843-82faa7ec49b3>,

Gierl et al.'s template: <https://chat.openai.com/share/ca7d89b6-e608-4724-8be9-3b924fca8afc>.

Tab. III. Our parent item model and the transformed item models by Claude and ChatGPT.

ITEM MODEL TRANSFORMED BY CLAUDE 3	OUR ITEM MODEL WRITTEN BY HUMANS	ITEM MODEL TRANSFORMED BY ChatGPT-4
<p>During a routine visit to a primary care facility, a male patient aged <E1> years old is found to have <E2>. The individual, who has not previously been diagnosed with hypertension, reports <E3>. Additionally, it is noted that the patient <E4>, and their body mass index is calculated to be <E5> kg/m².</p> <p>Considering this patient without any other comorbidities, which of the following approaches would be most appropriate in addition to recommending lifestyle modifications?</p> <p>Answer Options:</p> <p>A. Suggesting an annual blood pressure measurement would suffice.</p> <p>B. Advising the patient to return after at least five days of blood pressure monitoring, without any intervention.</p> <p>C. Recommending lifestyle changes and scheduling a follow-up visit in one month's time.</p> <p>D. Prescribing antihypertensive medication and arranging a follow-up appointment in one month.</p> <p>E. Providing necessary immediate intervention and referring the patient to the emergency room promptly.</p>	<p>A <E1> years-old male patient applies to a primary care clinic for a routine check-up. The patient's <E2>. The patient, who has not been diagnosed with hypertension before, has <E3>. The patient <E4>. On physical examination, the body mass index of the patient is <E5> kg/m².</p> <p>In this patient without any additional comorbidities, in addition to recommend lifestyle changes, which one of the following options is the most appropriate approach compared to others?</p> <p>A. Blood pressure measurement once a year is sufficient to recommend.</p> <p>B. The patient should re-apply after at least five days of blood pressure measurement, without any intervention.</p> <p>C. Recommendation of lifestyle changes is sufficient and a follow-up visit should be made after one month.</p> <p>D. Antihypertensive medication should be prescribed and a follow-up visit should be made after one month.</p> <p>E. Necessary first intervention should be provided and the patient should be sent to the emergency room immediately.</p>	<p>A male patient aged <E1> years visits a primary healthcare provider for a standard check-up. The patient's blood pressure <E2>. He has no history of hypertension and currently has <E3>. He <E4>. Upon physical examination, his body mass index (BMI) is determined to be <E5> kg/m².</p> <p>Considering the absence of additional comorbidities, which of the following is the most suitable course of action?</p> <p>A. Annual monitoring of blood pressure is sufficient.</p> <p>B. The patient should return for another blood pressure assessment after at least five days, without any medical intervention prior.</p> <p>C. Advising dietary and lifestyle adjustments, with a follow-up appointment in one month, is adequate.</p> <p>D. Initiate antihypertensive medication and schedule a follow-up visit in one month.</p> <p>E. Provide the necessary immediate treatment and refer the patient to the emergency department immediately.</p>

Tab. III. cd. Our parent item model and the transformed item models by Claude and ChatGPT.

ITEM MODEL TRANSFORMED BY CLAUDE 3	OUR ITEM MODEL WRITTEN BY HUMANS	ITEM MODEL TRANSFORMED BY ChatGPT-4
<p>Elements and their values:</p> <p><E1>:</p> <ul style="list-style-type: none"> • 32; • 44; • 56. <p><E2>:</p> <ul style="list-style-type: none"> • a blood pressure reading of 110/70 mmHg, based on a single measurement taken in the clinic; • an average 24-hour blood pressure recording of 130/82 mmHg; • blood pressure measurements of 150/95 mmHg obtained during two separate clinic visits, while the average home readings are 145/93 mmHg; • a blood pressure of 165/100 mmHg, based on two measurements taken during two separate clinic visits; • a blood pressure of 190/130 mmHg, based on a single measurement in the clinic. <p><E3>:</p> <ul style="list-style-type: none"> • no additional complaints; • occasional mild headaches, but no other complaints. <p><E4>:</p> <ul style="list-style-type: none"> • is a non-smoker; • smokes one pack of cigarettes per day. <p><E5>:</p> <ul style="list-style-type: none"> • 20; • 23; • 25; • 29. 	<p><E1>:</p> <ul style="list-style-type: none"> • 32; • 44; • 56. <p><E2>:</p> <ul style="list-style-type: none"> • blood pressure is 110/70 mmHg based on a measurement taken once in the clinic; • average blood pressure over a 24-hour period is recorded as 130/82 mmHg; • measurements, obtained during two separate visits in the clinic, are recorded as 150/95 mmHg. However, when the measurements taken at home are considered, the average is 145/93 mmHg; • blood pressure is 165/100 mmHg based on two measurements during two separate visits in the clinic; • blood pressure is 190/130 mmHg based on one measurement in the clinic. <p><E3>:</p> <ul style="list-style-type: none"> • no additional complaint; • no additional complaint except the patient has sometimes a mild headache. <p><E4>:</p> <ul style="list-style-type: none"> • does not smoke; • smokes one package of cigarette each day. <p><E5>:</p> <ul style="list-style-type: none"> • 20; • 23; • 25; • 29. 	<p><E1> Options:</p> <ul style="list-style-type: none"> • 32; • 44; • 56. <p><E2> Options:</p> <ul style="list-style-type: none"> • was recorded at 110/70 mmHg from a single clinic measurement; • was averaged at 130/82 mmHg over a 24-hour period; • included measurements of 150/95 mmHg from two separate clinic visits and an average of 145/93 mmHg from home measurements; • was measured at 165/100 mmHg from two separate clinic visits; • was recorded at 190/130 mmHg from a single clinic measurement. <p><E3> Options:</p> <ul style="list-style-type: none"> • no additional symptoms; • no additional symptoms, although he occasionally experiences mild headaches. <p><E4> Options:</p> <ul style="list-style-type: none"> • does not smoke; • smokes one pack of cigarettes daily. <p><E5> Options:</p> <ul style="list-style-type: none"> • 20; • 23; • 25; • 29.

In our item model's transformation (Tab. III.), ChatGPT's first output included grammatical errors that emerged due to incompatibility of some values with the body of the question. For example, it offered to use "It is noted that his <E2>." in the body of the question, but one of the values that it recommended to replace <E2> with was "a 24-hour average blood pressure shows 130/82 mmHg", which is incompatible. Therefore, it required us to add one more input: "Some of the values in E2 and E4 are not compatible with the corresponding part in the question. It would generate questions with grammatical errors." After this feedback, it made changes and corrected the structure. Tab. III. includes the final corrected version. However, Claude 3 (Sonnet) did not need any feedback. The first output was free from any significant language problems.

In the transformation of Gierl et al.'s item model (Tab. IV.), both Claude and ChatGPT generated item models without any significant errors.

DISCUSSION

Polish Items

The first aim of this study was to explore the possibility of implementing template-based AIG in Polish for case-based MCQs to assess clinical reasoning skills. Similar to the Turkish study which we sourced the models from, our research has demonstrated that AIG can be successfully adapted to the Polish language and context. Our

Tab. IV. Our parent item model and the transformed item models by Claude and ChatGPT.

ITEM MODEL TRANSFORMED BY CLAUDE 3	GIERL ET AL.'S ITEM MODEL WRITTEN BY HUMANS	ITEM MODEL TRANSFORMED BY ChatGPT-4
<p>A patient in their [E1] was recently discharged after undergoing a [E3] surgery. However, within [E2] days of the operation, they were readmitted due to abdominal discomfort. Upon examination, the patient had an elevated temperature of 38.5°C and exhibited [E4]. What would be the most appropriate next step in managing this patient's condition?</p> <p>A. Initiate antibiotic therapy. B. Encourage ambulation and mobilization. C. Surgically re-open the operative site. D. Administer anticoagulant medication. E. Perform percutaneous drainage.</p> <p>[E1]:</p> <ul style="list-style-type: none"> • Middle-aged adult (40–65 years old). <p>[E2]:</p> <ul style="list-style-type: none"> • Early postoperative period (1–6 days). <p>[E3]:</p> <ul style="list-style-type: none"> • Stomach resection (gastrectomy). • Partial colon removal (right or left hemicolectomy). • Appendix removal (appendectomy). • Gallbladder removal (laparoscopic cholecystectomy). <p>[E4]:</p> <ul style="list-style-type: none"> • Wound appearing red and tender. • Abdominal guarding and rebound tenderness. • Generalized abdominal tenderness. • Calf tenderness. 	<p>A [E1] was readmitted to hospital for pain in the abdominal area. He was on post operative day [E2] after recovering from a [E3]. The patient has a temperature of 38.5°C. Physical examination reveals [E4]. Which one of the following is the best next step for this patient?</p> <p>A. Antibiotics. B. Mobilise. C. Reopen Wound. D. Anticoagulation. E. Percutaneous Drainage.</p> <p>[E1]:</p> <ul style="list-style-type: none"> • Age 40 to 65 years. <p>[E2]:</p> <ul style="list-style-type: none"> • 1 to 6 days. <p>[E3]:</p> <ul style="list-style-type: none"> • Gastrectomy. • Right Hemicolectomy. • Left Hemicolectomy. • Appendectomy. • Laparoscopic Cholecystectomy. <p>[E4]:</p> <ul style="list-style-type: none"> • Red and Tender Wound. • Guarding and Rebound. • Abdominal Tenderness. • Calf Tenderness. 	<p>A patient aged between 40 and 65 years was rehospitalized due to abdominal pain and is currently on day [E2] following a [E3] procedure. The patient exhibits a fever of 38.5°C. Upon examination, [E4] is noted. What is the most appropriate subsequent action for this patient?</p> <p>A. Administer antibiotics. B. Encourage mobilization. C. Surgical wound reassessment. D. Start anticoagulation therapy. E. Implement percutaneous drainage.</p> <p>[E1]:</p> <ul style="list-style-type: none"> • Age range from 40 to 65 years. <p>[E2]:</p> <ul style="list-style-type: none"> • Postoperative days ranging from 1 to 6. <p>[E3]: Procedures:</p> <ul style="list-style-type: none"> • Gastrectomy. • Right Hemicolectomy. • Left Hemicolectomy. • Appendectomy. • Laparoscopic. • Cholecystectomy. <p>[E4]: Clinical findings:</p> <ul style="list-style-type: none"> • Red and Tender Wound. • Guarding and Rebound. • Abdominal Tenderness. • Calf Tenderness.

study added Polish as the seventh language to the language list used for template-based AIG in medical education. This is an important improvement to the diversity and inclusivity of AIG applications in terms of covering non-English speaking countries.

The positive feedback from subject matter experts, including a cardiologist and an internal medicine specialist with teaching and healthcare experience, is valid evidence of the quality of the AIG-generated questions. Their agreement on all evaluation criteria indicates that the questions are clear, clinically sound, possess a single correct answer, and are sufficient to allow the identification of the correct answer. Most importantly, these questions are recognized as tools for assessing clinical reasoning rather than mere factual recall.

Nonetheless, similar to the limitations acknowledged in the Turkish study [12], our research is not without its limitations. The evaluation by a relatively small number of experts and the review of only a select sample of questions may affect the generalizability of our findings.

Additionally, while the AIG system was successfully adapted from the Turkish version, future research could explore the development of original item models specific to the Polish medical curriculum. Another possible future research direction could be the use of the generated questions in an exam setting to reveal other characteristics, such as their psychometric properties, as the Turkish study showed acceptable item difficulty and discrimination levels [13].

Generating New Item Models Using AI

The second aim was to present a new method for generating item models to use for AIG. We presented a four-step process for this purpose. A previous study successfully showed the use of AI for a similar purpose but for reading comprehension items [14]. In this current study, for the first time in a medical education context, the findings from an informal evaluation showed that ChatGPT and Claude can successfully transform existing item models into new item models, while preserving their underlying meaning and structure without significant errors, though further

changes for optimization could be made in the fourth step as we proposed in our method. The models generated by the AI tools were grammatically correct, logically coherent, and compatible with the terminology. This shows the potential of using AI to increase the diversity and creativity of item models used in AIG. By using our four-step process, template-based AIG developers could be able to generate multiple item models (and therefore more diverse items) based on the same cognitive model. Our approach addresses one of the important limitations of the template-based AIG.

An interesting aspect of our research is the demonstration of the idea of using two alternative large language models: ChatGPT and Claude. It was not our objective to benchmark the two AI models in generating MCQ but to illustrate the feasibility of the transformation process across different systems. However, our initial observations made in the context of the study suggest that Claude did not appear inferior to ChatGPT for the task, but in fact required fewer manual corrections. Considering the outcome of other research, which demonstrated that Claude-2 surpassed ChatGPT-3.5 in answering and explaining physiology MCQs [20], it indicates that the selection of tools for assisting in MCQ AIG should not be limited to models developed by OpenAI alone. Future research should investigate the quality of question generation by different AI models in greater depth.

The involvement of human experts was minimal in our demonstration, limited to providing the initial parent item models and reviewing the transformed outputs. This highlights the capability of AI systems to semi-autonomously generate new content with minimal human intervention, potentially streamlining the item model development process. However, it is important to note that the transformed models generated by the AI tools still require review and validation by subject matter experts before being implemented in real-world settings. While the AI outputs were largely appropriate for use, there were instances where human feedback was necessary to correct minor errors or incompatibilities. This points out the need for a collaborative

approach, where AI augments and supports human expertise rather than completely replacing it, at least in the near future.

Future research could explore the seamless integration of AI-generated item models into existing template-based AIG systems, as well as their psychometric performance and effectiveness in assessing clinical reasoning skills when administered in exams. Our study was limited by the use of only two examples to demonstrate AI for transforming templates. Additionally, investigating the scalability and generalizability of this approach across different medical disciplines and educational contexts would be valuable. The proposed templates are relatively simple for demonstrative purposes and we do not support nesting of templates. Moreover, it would be interesting to investigate whether the current capabilities of AI technologies allow us to generate useful item models and even cognitive models from scratch without providing a parent item model. However, a previous study suggested that "it is neither possible nor desirable to create specifications and instructions using artificial intelligence" in the context of template-based AIG [14]. Therefore, this could be a difficult task for AI within the current capabilities.

CONCLUSIONS

We demonstrated the successful implementation of template-based AIG in Polish for generating case-based MCQs to assess clinical reasoning skills in medical education. We also presented a novel method of using AI models to transform existing item models into new models while preserving their underlying structure and meaning. The AI-generated item models showed potential in enhancing the diversity of item models used in template-based AIG. Nevertheless, human expert review remains crucial before utilizing the outputs from AI. Future research should explore integrating AI-generated models into template-based AIG processes, evaluating their performance in real exams, and investigating the use of this approach across different areas.

REFERENCES

- Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Acad Med.* 2019;94(6):902–12.
- Pugh D, De Champlain A, Touchie C. Plus ça change, plus c'est pareil: Making a continued case for the use of MCQs in medical education. *Med Teach.* 2019;41(5):569–77.
- Gierl MJ, Lai H, Tanygin V. *Advanced Methods in Automatic Item Generation.* 1st edition. New York: Routledge; 2021.
- Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. 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.* 2023;18(8):e0290691.
- Coşkun Ö, Kiyak YS, Budakoğlu İ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. *Med Teach.* 2024;13:1–7.
- Kiyak YS, Coşkun Ö, Budakoğlu İ, Uluoğlu C. ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *Eur J Clin Pharmacol.* 2024;80(5):729–35.
- Kiyak YS, Kononowicz AA. Case-based MCQ generator: A custom ChatGPT based on published prompts in the literature for automatic item generation. *Med Teach.* 2024;48(6):1018–20.
- Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions. *Acad Med.* 2023;99(5):508–12. <https://doi.org/10.1097/ACM.0000000000005626>.

9. Zuckerman M, Flood R, Tan RJB, Kelp N, Ecker DJ, Menke J, et al. ChatGPT for assessment writing. *Med Teach*. 2023;45(11):1224–7.
10. Fors UGH, Muntean V, Botezatu M, Zary N. Cross-cultural use and development of virtual patients. *Med. Teach*. 2009;31(8):732–8.
11. Mayer A, Da Silva Domingues V, Hege I, Kononowicz AA, Larrosa M, Martínez-Jarreta B, et al. Planning a Collection of Virtual Patients to Train Clinical Reasoning: A Blueprint Representative of the European Population. *IJERPH*. 2022;19(10):6175.
12. Kiyak YS, Budakoğlu İİ, Coşkun Ö, Koyun E. The First Automatic Item Generation in Turkish for Assessment of Clinical Reasoning in Medical Education. *Tıp Eğitimi Dünyası*. 2023;22(66):72–90.
13. Kiyak YS, Coşkun Ö, Budakoğlu İİ, Uluoğlu C. Psychometric Analysis of the First Turkish Multiple-Choice Questions Generated Using Automatic Item Generation Method in Medical Education. *Tıp Eğitimi Dünyası*. 2023;22(68):154–61.
14. Sayin A, Gierl M. Using OpenAI GPT to Generate Reading Comprehension Items. *Educ. Meas*. 2024;43(1):5–18.
15. Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. *Med Educ*. 2012;46(8):757–65.
16. Leo J, Kurdi G, Matentzoglou N, Parsia B, Sattler U, Forge S, et al. Ontology-Based Generation of Medical, Multi-term MCQs. *Int J Artif Intell Educ*. 2019;29:145–88.
17. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *RPTTEL*. 2020;15:12.
18. Masters K, Benjamin J, Agrawal A, MacNeill H, Pillow MT, Mehta N. Twelve tips on creating and using custom GPTs to enhance health professions education. *Med Teach*. 2024;46(6):752–56.
19. Fisher AD, Fisher G. Evaluating performance of custom GPT in anesthesia practice. *J. Clin. Anesth*. 2024;93:111371.
20. Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in Answering and Explaining Conceptual Medical Physiology Multiple-Choice Questions. *Cureus*. 2023; 15(9):e46222. <https://doi.org/10.7759/cureus.46222>.

Supplementary materials are available at Bio-Algorithms and Med-Systems (ISSN: 1896-530X) – online: <https://bamsjournal.com/api/myfiles/view/2700581>.