医学教育 2026, **57**(3)：152～162

特集　医学教育における生成 AI 2026 ―「試行」から「効果検証」へ

原　著　Original Research Papers

生成 AI を用いた試験問題作成支援アプリケーションの有用性と受容に関する教員へのアンケート調査

橋本　恵太郎[*1,2,3]　前野　貴美[*4]　吉原　雅大[*1,3,5]　古屋　欽司[*1,3,6]
高屋敷　明由美[*1,2,3]　鈴木　將玄[*1,3,7]　前野　哲博[*1,2,3]

要旨：
背景：良質な多肢選択式問題の作成は教員の大きな負担であり，生成 AI による作問支援が注目されているが，教員視点の検証は不十分である．本研究は AI 作問支援アプリケーションが教員の作問効率および心理的負担に及ぼす影響を検討した．
方法：筑波大学医学類において本ツールを導入し，2025 年度作問担当教員 283 名に匿名 Web アンケートを実施した．
結果：ツールを使用したと回答した 87 名を解析対象とした．1 問あたり 30 分未満で作成できた割合は，一般問題で 29%から 64%，症例問題で 20%から 60%へ増加した（いずれも $p < 0.001$）．96%が心理的負担は軽減したと回答した．最大の問題点は過去問参照時のハルシネーションであった．
考察：ハルシネーションへの対処は今後の課題であるものの，AI 作問支援ツールは，形式調整などの作業負担を軽減することで，教員の作問効率向上に寄与する可能性が示唆された．
キーワード：生成 AI，試験問題作成，多肢選択式問題

Perceptions Usefulness and Acceptance of a Generative AI-Based Application for Medical Examination Question Development：A Faculty Survey

Etaro Hashimoto[*1,2,3]　Takami Maeno[*4]　Masaharu Yoshihara[*1,3,5]　Kinji Furuya[*1,3,6]
Ayumi Takayashiki[*1,2,3]　Masatsune Suzuki[*1,3,7]　Tetsuhiro Maeno[*1,2,3]

Abstract:
Background：Creating high-quality multiple-choice questions (MCQs) places a substantial burden on faculty. While generative AI-assisted question development has gained attention, evaluation from the faculty perspective remains insufficient. This study examined the impact of an AI-based question development support application on faculty question

[*1] 筑波大学医学類医学教育センター，Medical Education Center, College of Medicine, University of Tsukuba
[*2] 筑波大学医学医療系地域医療教育学，
Department of Primary Care and Medical Education, Institute of Medicine, University of Tsukuba
[*3] 筑波大学医学群医学群医学教育企画評価室，Department of Planning and Coordination for Medical Education, School of Medicine and Health Sciences, University of Tsukuba
[*4] 東京慈恵会医科大学教育センター，Center for Medical Education, The Jikei University School of Medicine
[*5] 筑波大学医学医療系解剖学発生学，
Department of Anatomy and Embryology, Institute of Medicine, University of Tsukuba
[*6] 筑波大学医学医療系消化器外科，
Department of Gastrointestinal and Hepato - Biliary - Pancreatic Surgery, Institute of Medicine, University of Tsukuba
[*7] 筑波大学医学医療系地域総合診療医学，
Department of General Medicine and Primary Care, Institute of Medicine, University of Tsukuba
受付：2026 年 3 月 16 日，受理：2026 年 4 月 20 日

development efficiency and psychological burden.
Methods：The application was introduced at the University of Tsukuba School of Medicine, and an anonymous web survey was administered to 283 faculty members responsible for exam question development in the 2025 academic year.
Results：Eighty-seven respondents who reported using the application were analyzed. The proportion of questions completed in less than 30 minutes per item increased from 29% to 64% for general questions and from 20% to 60% for clinical case questions (both $p < 0.001$). Ninety-six percent reported reduced psychological burden. The primary concern was hallucination during past exam question retrieval.
Conclusions：Although addressing hallucination remains a challenge, the results suggest that the AI-based question development support tool may contribute to improving faculty question development efficiency by reducing workload associated with tasks such as formatting.
Keywords: generative artificial intelligence, test item development, multiple-choice questions

背　景

医学教育において評価は，教育プロセスの中心に位置付けられており，信頼性と妥当性の高い評価の設計と実施は，教育の質を担保する上で不可欠である[1]．とりわけ，多肢選択式問題（Multiple Choice Questions，MCQ）は医学教育において広く用いられている評価手法であり，適切に設計された MCQ は，学生の知識のみならず，臨床推論能力など高次の認知能力を評価することが可能である[2]．しかし，その前提となる良質な試験問題の作成は，教員にとって大きな負担となっている[3]．先行研究では，質の高い MCQ1 問の作成に多大な時間的・経済的コストを要する可能性が指摘されており，この負担は教育の質にも影響を及ぼしうる[4, 5]．また，良質な MCQ を作成するためには，いわゆる問題作成上の技術的不備である Item Writing Flaws（IWF）を回避するなど，体系的な作問技法が求められるが，多くの教員は必ずしも十分な訓練を受けていない[6, 7]．その結果，作成が比較的容易な知識想起型問題への偏重や過去問の安易な流用が生じ，学生の表面的学習を助長する悪循環につながる可能性がある[8, 9]．

近年，この課題に対する解決策として，生成 AI を活用した作問支援が注目されている．大規模言語モデルを用いた研究では，AI が生成した MCQ が一定水準の品質を示すことや，作問時間の短縮に寄与する可能性が報告されている[10, 11]．一方で，生成 AI の出力にはハルシネーション（事実に基づかない情報の生成）が含まれる可能性があり，生成された問題の信頼性には課題が残されている．そのため，生成 AI は教員の専門的判断を支援する補助的ツールとして位置づけられており，最終的な内容確認は教員が行うことが不可欠とされている[12]．しかし，これらの問題の品質や作問効率に関する知見は主に実験的環境下で AI 出力の品質を評価したものであり[10–12]，教員が日常の作問業務のなかでツールを使用した際の使いやすさや作業負担の変化，ツールの受容といったユーザー中心の視点からの検証は限定的である．

我々は医師国家試験過去問と連携した多肢選択式問題作成支援 AI アプリケーションを開発し，2025 年度に筑波大学医学類における試験問題作成に導入・運営した．本研究では，教員の作問業務への影響を，作問効率，作業負担，ツールの受容，および教員が認識する問題の質の変化の 4 側面から主観的評価に基づいて検証することを目的とした．

方　法

1. 研究デザイン

本研究は横断研究である．報告は STROBE 声明に準拠した．

2. 対象

2025 年度に行われた筑波大学医学類 6 年生（合計 142 名）向けの総合試験（国家試験の形式・内容に準じた多肢選択式問題からなる 1 回 400 問全 3 回の試験）の作問を担当した教員 283 名を対象とした．各教員は令和 6 年版医師国家試

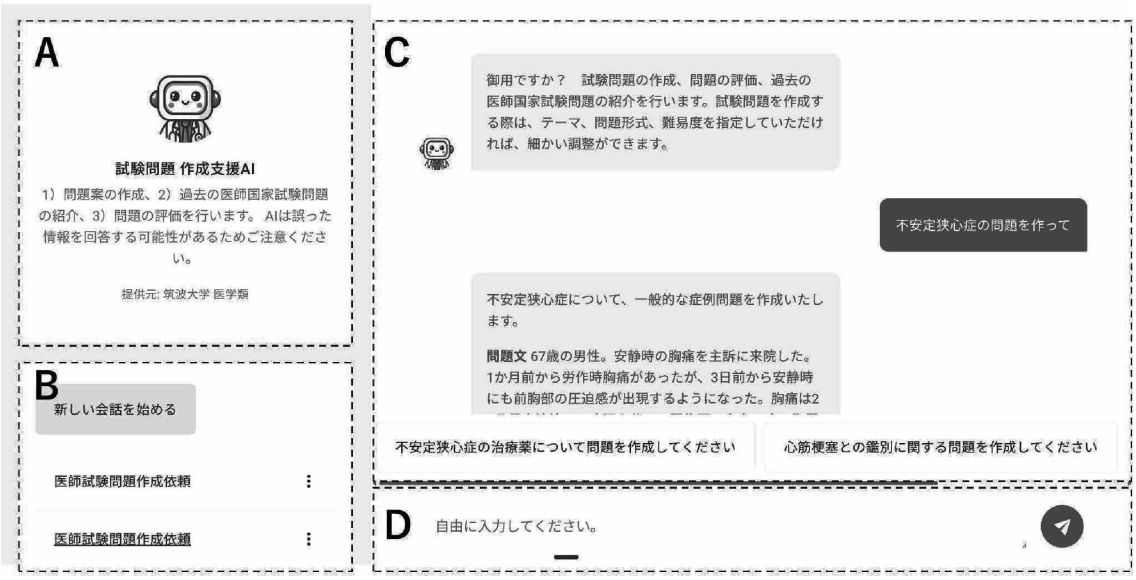

図1　アプリケーションの画面
A：アプリケーション説明欄，B：過去のチャットログ一覧，C：チャット内容表示欄，D：指示文入力欄

験出題基準・ブループリントに基づき指定された出題テーマおよび問題形式（一般問題または症例問題）について作問を行った．教員が作成した問題は，AI導入前と同様に，診療科・グループ内での確認を経て提出され，技術職員によるチェックおよび必要に応じた内容確認・修正を経て登録された．

3. アプリケーションの概要

2025年4月，作問依頼と同時に全教員に試験問題作成支援アプリケーションを公開した．本アプリケーションはチャットボット構築プラットフォームmiibo（miibo社，東京）上に構築したチャットボット型アプリケーションである．大規模言語モデルGemini 2.5 Pro（Google社，マウンテンビュー，カリフォルニア，米国）を基盤とし，1問1テーマ，二重否定の使用を避けるといったIWFの回避などの作問ノウハウと第109回から第119回までの医師国家試験問題をナレッジベースとして参照可能とした．過去問データベースへのアクセスにはRAG（Retrieval-Augmented Generation，検索拡張生成）を採用し，入力されたキーワードや問題番号に基づいて関連過去問を動的に検索・提示する構成とした．

主な機能は①医師国家試験過去問の参照，②多肢選択式問題案の作成，③問題案の評価の3つである．過去問参照機能では出題テーマに関連する国試過去問が提示される．問題案作成機能ではテーマと問題形式に応じた問題案（臨床シナリオ，選択肢，正答，解説等）が生成される．問題案評価機能では入力された問題の妥当性が作問ノウハウに基づき評価され改善案が提示される．これらの機能により，過去問検索から作問・評価・修正までの一連の工程を支援する（**図1**）．

本研究対象者にメーリングリストでアプリについて周知し，機密性保持のために，学内専用サイトからのアプリへのアクセスを求めた．アプリケーションの使用方法を解説する動画やPDF資料，プロンプト例を併せて提供した．

なお，本アプリケーションは2025年4月の公開後，機能改善のため，7月に1回のバージョンアップを行った．

4. 調査手続き

全総合試験の作問期間の終了後，2025年10月24日から11月24日までの1カ月間にWeb匿名

アンケートを実施した．アンケートは総合試験作問担当教員 283 名に電子メールで配布し，リマインダーを計 2 回送信した．アンケートの冒頭で研究の趣旨を説明し，回答をもって同意とした．なお，アンケートデータに加え，教員の問題提出時に使用するシートにアプリケーション使用チェック欄を設け，その記録を補助的データとして用いた．

5. 調査項目

調査票は，MCQ 作成における生成 AI 活用に関する先行研究[10]の評価項目および本アプリケーションの機能特性を踏まえ，研究者間で協議の上，独自に設計した．調査項目は以下の 5 領域で構成した．

なお，本アプリケーションは筑波大学医学類 6 年生の総合試験向けに開発したものではあるが，同じ 6 年生向けの領域別試験（総括試験）にも使用される可能性があったため，アンケートで利用状況等を調査した．

（1）基本属性（5 問）：年代（10 歳刻みの単一選択式），性別（男性 / 女性 / その他 / 答えたくない），所属領域（臨床医学 / 基礎医学 / 社会医学 / その他），医学類 6 年生向け試験における MCQ 作問経験年数（「1 年未満（今年度に初めて作問した）」～「21 年以上」および「作問経験は無い」の 7 件），普段の生成 AI 使用頻度（「日常的に使う（ほぼ毎日）」～「使ったことがない」の 5 件）をそれぞれ単一選択式で尋ねた．

（2）利用状況（3 問）：利用した試験（第 1～3 回総合試験・総括試験・いずれにも使用していない，複数選択可），作成した問題の種類（一般問題・症例問題・計算問題，複数選択可），使用した機能（過去問参照・問題案の作成・問題案の評価の 3 機能，複数選択可）を尋ねた．

（3）効率の評価（5 問）：アプリケーション使用前後の 1 問あたりの作問所要時間について，一般問題・症例問題それぞれに「15 分未満」「15～30 分」「30 分～1 時間」「1～2 時間」「2～4 時間」「4 時間以上」の 6 段階で回答を求めた（使用前のみ「作成したことはない」を加えた 7 段階）．使用前の時間は回顧的に回答を求め，「作成したことはない」を選択した回答は当該比較から除外した．また，アプリケーション利用による作問労力（心理的・身体的負担）の変化を「1：大幅に増加した」～「5：大幅に軽減された」の 5 件法リッカート尺度で尋ねた．

（4）質の評価（11 問）：各機能の利用者を対象に，過去問参照機能では「キーワード検索だけではなく，自然な言語が使えることで見たい過去問を探しやすかった」「過去問を探す時間が短縮された」の 2 項目，問題案作成機能では「最初に AI が生成した問題案の質は高かった」「ゼロから作る心理的負担が減った」の 2 項目，問題案評価機能では「AI の評価に納得できた」「評価内容は見直しのきっかけになった」の 2 項目を，それぞれ「1：全くそう思わない」～「5：非常にそう思う」の 5 件法リッカート尺度で評価を求めた．AI 生成症例問題の質については，「症例の経過や病態は医学的に自然だった」「提示された情報（身体所見，検査結果など）の整合性がとれていた」の 2 項目を同様の 5 件法で尋ねた．難易度の適切さは「難しすぎた」～「簡単すぎた」の 5 段階で尋ねた．さらに，最終的に完成した問題の質の変化として，「学習目標・テーマとの整合性」「問題の妥当性」「ご自身の専門外の領域などでの，多様な観点からの出題」の 3 項目を「1：大幅に低下した」～「5：大幅に向上した」の 5 件法で尋ねた．

（5）満足度・意向（4 問）：操作性を「1：非常に使いにくい」～「5：非常に使いやすい」，総合満足度を「1：非常に不満」～「5：非常に満足」，継続利用意向を「1：利用したくない」～「5：ぜひ利用したい」の各 5 件法リッカート尺度で尋ねた．同僚への推奨度は 0（全く勧めない）～10（非常に勧めたい）の 11 件法の数値評価尺度で尋ねた．推奨度は Net Promoter Score に基づき推奨者・中立者・批判者に分類した．

加えて，未使用理由，メリット，改善要望，その他意見について自由記述欄を設けた．なお，各機能の評価項目は当該機能の利用者のみが回答する条件分岐を設定した．

6. 分析方法

(1) 解析対象

アンケート回答者のうち，本アプリケーションを使用したと回答した者を解析対象とした．Webアンケートでは回収率向上のため，基本属性など一部の項目以外は回答必須としなかったため，有効回答数が項目により異なる．

(2) 分析

リッカート尺度，数値評価尺度を用いた項目については，回答数および項目別割合を算出した．作問時間はAI使用前後の分布を比較するため，2つの検定を実施した．得られた時間カテゴリを順序尺度とみなし，カテゴリの増減の方向を評価する符号検定を行った．さらに，実用上の基準として30分未満を設定し，「30分未満」と「30分以上」に二値化したうえでMcNemar検定を実施した．分析にはMicrosoft Excel（Microsoft社，レドモンド，ワシントン，米国）およびR version 4.5.2（R Foundation for Statistical Computing，ウィーン，オーストリア）を用いた．

自由記述回答に対しては，研究責任者が全回答を精読し，内容の類似性に基づきカテゴリに分類した．

7. 倫理的配慮

本研究は筑波大学医の倫理委員会の承認を得た（承認番号2181）．匿名調査であり，個人の特定は不可能である．また，回答は任意とし，調査冒頭の説明に同意した者のみが回答した．

8. AI支援ツール，技術の利用

本研究では，Claude Opus 4.6およびClaude Sonnet 4.6（共にAnthropic社，サンフランシスコ，カリフォルニア，米国）を用いて論文構成の整理，表現の微修正，英語抄録の校正を行った．これらの用途は文章表現の補助に限定しており，研究内容の解釈，結果の判断および最終的な原稿の内容については著者が責任を持って行った．

結　果

1. 回答者の属性と利用状況

調査対象283名中101名より回答を得た（回答率36%）．うち本アプリケーション使用者87名を解析対象とした．解析対象者の属性と本アプリケーションの利用状況は**表1**の通りである．生成AIを週1回以上使用する者が83%を占めた．

なお，総合試験の提出問題1,200問中367問（31%）で本アプリケーションの使用が記録されていた．採点除外問題は，本アプリケーションを利用した367問中1問，利用していない833問中1問であった．

2. 効率の評価

2.1. 作問時間

アプリケーションの使用により，一般問題（n=69）・症例問題（n=65）ともに1問あたりの作問時間の分布は有意に短時間側へ変化した（符号検定，ともに$p < 0.001$）．なお，本検定は使用前後の双方に回答のあったものを対象とした．アプリケーション使用前後の作問時間分布を**図2**に示す．作問時間30分未満の割合がアプリケーションの使用前後で，一般問題では29%から64%に，症例問題では20%から60%となった（McNemar検定，ともに$p < 0.001$）．

2.2. 心理的負担

各評価項目の回答分布を**表2**に示す．作問に要する労力については，「大幅に軽減」と「やや軽減」が合計で84名中74名（88%）を占めた．増加したとの回答は0%であった．ゼロから作問する負担の低減は，「非常にそう思う」と「そう思う」が合計で73名中70名（96%）を占めた．

3. 質の評価

3.1. 問題の質の変化

アプリケーションを利用して作成した問題の質について，学習目標・テーマとの整合性と問題の妥当性はともに「向上」が84名中47名（56%）を占めた（**表2**）．自身の専門外の領域などでの多様な観点からの出題は「大幅に向上」と「やや

表 1 回答者の属性と利用状況

項目	カテゴリ	n（%*）
年代（n=87）		
	30 代	23（26）
	40 代	44（51）
	50 代	17（20）
	60 代	3（3）
性別（n=86）		
	男性	61（71）
	女性	24（28）
	無回答	1（1）
所属領域（n=87）		
	臨床医学	79（91）
	社会医学	4（5）
	基礎医学	3（3）
	その他	1（1）
作問経験年数（n=86）		
	1 年未満	18（21）
	1 ～ 2 年	11（13）
	3 ～ 5 年	18（21）
	6 ～ 10 年	24（28）
	11 年以上	15（17）
生成 AI の使用頻度（n=87）		
	ほぼ毎日	29（33）
	週 3 ～ 4 回	23（26）
	週 1 ～ 2 回	20（23）
	週 1 回未満	12（14）
	使用経験なし	3（3）
利用した試験（n=87，複数選択可）		
	第 1 回総合試験	53（61）
	第 2 回総合試験	54（62）
	第 3 回総合試験	57（66）
	総括試験	57（66）
作成した問題の種類（n=84，複数選択可）		
	一般問題	74（88）
	症例問題	70（83）
使用した機能（n=83，複数選択可）		
	問題案の作成	73（88）
	過去問参照	38（46）
	問題案の評価	28（34）

* % は小数点第 1 位で四捨五入しており，合計が 100 とならない場合がある．

向上」が合計で 84 名中 36 名（43%）を占めた．いずれの項目においても，「変わらない」～「大幅に向上」という維持以上の回答が 84 名中 83 名（99%）以上を占めた．

3.2. AI 生成症例問題の質

アプリケーションで教員が作問を指示した際に AI が最初に生成した症例問題の質について，症例の自然さは「非常にそう思う」と「そう思う」の肯定的回答が 70 名中 53 名（76%）を占め，提

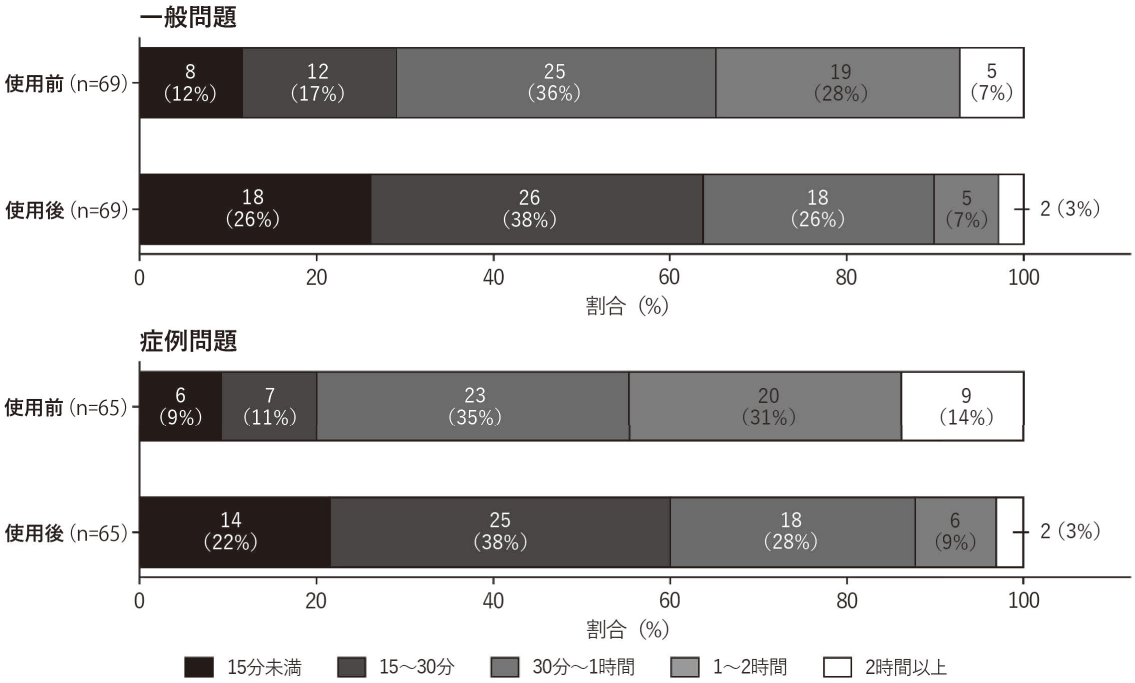

図2　アプリケーション使用前後の1問あたりの作問時間の変化（一般問題，症例問題）

示情報の整合性は「非常にそう思う」と「そう思う」が70名中57名（81%）を占めた（**表2**）．難易度は84名中57名（68%）が適切と回答した．

4. 満足度・利用意向

アプリケーションの操作性については，「非常に使いやすい」と「やや使いやすい」が合計で84名中72名（86%）を占めた．満足度については，「非常に満足」と「やや満足」が合計で84名中71名（85%）を占めた（**表2**）．継続利用については「ぜひ利用したい」と「やや利用したい」が合計で83名中74名（89 %）を占めた．アプリケーションの利用推奨度は平均8.2 ± 1.9点（n=83）で，推奨者（9〜10点）39名（47%），中立（7〜8点）32名（39%），批判者（0〜6点）12名（15%）であった．

5. 自由記述の内容

5.1. アプリケーション利用のメリット

本アプリケーションを利用するメリットに関する34件の自由記述から，単純な肯定などの一般的感想を除いた29件を5カテゴリに分類した（**表3**）．作問時間の短縮に関する記載が8件と最多であった．そのほか，国試水準への準拠・難易度の適正化，およびたたき台としての有用性がそれぞれ6件で続いた．後者では，AI出力をそのまま使用するのではなく修正・加筆して活用するという記述が複数みられた．

5.2. アプリケーションの課題

本アプリケーションの課題に関する42件の自由記述から，画像対応等の機能拡張要望等を除いた31件を4カテゴリに分類した（**表3**）．過去問参照時のハルシネーションに対する指摘が15件と最多であった．そのほか，医学知識更新の必要性，AI生成物に対する検証の負担，アプリケーションの教員への周知の強化などの記述が複数みられた．

考　察

本研究では，アンケートによる主観的評価ではあるが，作問時間の短縮や心理的負担の軽減といったAIによる作問作業の効率化の傾向が顕著

表 2　各評価項目の回答分布 n（%*）

項目	n	5	4	3	2	1
効率の評価						
		大幅に軽減	やや軽減	変わらない	やや増加	大幅に増加
作問に要する労力の変化	84	46（55）	28（33）	10（12）	0（0）	0（0）
		非常にそう思う	そう思う	どちらともいえない	そう思わない	全くそう思わない
ゼロから作問する心理的負担が減った	73	47（64）	23（32）	2（3）	1（1）	0（0）
問題の質の変化						
		大幅に向上	やや向上	変わらない	やや低下	大幅に低下
学習目標・テーマとの整合性	84	10（12）	37（44）	37（44）	0（0）	0（0）
問題の妥当性	84	12（14）	35（42）	36（43）	1（1）	0（0）
専門外を含む多様な観点からの出題	84	13（16）	23（27）	47（56）	1（1）	0（0）
AI 生成症例問題の質						
		非常にそう思う	そう思う	どちらともいえない	そう思わない	全くそう思わない
症例の経過や病態が医学的に自然だった	70	12（17）	41（59）	15（21）	2（3）	0（0）
提示情報（身体所見･検査結果等）の整合性がとれていた	70	9（13）	48（69）	13（19）	0（0）	0（0）
難易度		難しすぎた	やや難しかった	適切であった	やや簡単だった	簡単すぎた
難易度	84	0（0）	12（14）	57（68）	14（17）	1（1）
操作性						
		非常に使いやすい	やや使いやすい	どちらともいえない	やや使いにくい	非常に使いにくい
使いやすさ	84	30（36）	42（50）	9（11）	2（2）	1（1）
満足度						
		非常に満足	やや満足	どちらともいえない	やや不満	非常に不満
総合満足度	84	37（44）	34（41）	12（14）	0（0）	1（1）
継続利用意向						
		ぜひ利用したい	やや利用したい	どちらともいえない	あまり利用したくない	利用したくない
今後の継続利用意向	83	66（80）	8（10）	8（10）	0（0）	1（1）

＊ % は小数点第 1 位で四捨五入しており，合計が 100 とならない場合がある．

にみられた．テーマとの整合性や問題の妥当性など，問題の品質についても大多数が，維持あるいは向上したと回答した．最大の課題は過去問参照時のハルシネーションであった．

1. 認知負荷理論の観点からの効率化の機序

認知負荷理論では，外在的負荷は課題そのものの本質ではなく，課題の提示や作業手順などに起因する負荷とされている[13]．問題形式の標準化や問題文・選択肢の文案作成などの外在的負荷を AI が引き受けたことで，教員は妥当性判断や専門的内容の吟味など本質的負荷に集中でき，それが作問時間の短縮や心理的負担の軽減に寄与したと推察される．自由記述においても，AI 出力をそのまま使用するのではなく，たたき台として修正・加筆して活用するという記述が複数みられ，

表3　自由記述のカテゴリ分類

区分	カテゴリ	件数	代表的な回答例
メリット（34件）			
	作問時間の短縮	8	「問題作成時間が大幅に短縮されたのを実感した」
	国試水準への準拠・難易度の適正化	6	「国家試験のレベルに見合った問題を作成できる」，「過去問から大きく外れないように設計されている」
	たたき台の有用性	6	「たたき台を数分で作ることができる」，「疾患名からそれらしい臨床データを生成してくれる」
	問題の質向上・偏りの是正	5	「自作した問題の客観的評価が得られ，問題の質向上につながる」，「作問者の専門や好みに引っ張られない偏りの少ない問題ができる」
	心理的負担の軽減	4	「本当にゼロから作る負担がなくなり便利でした」，「心理的負担が大幅に軽減されます」
	小計	29	※一般的感想など5件を除く
課題（42件）			
	ハルシネーション	15	「参考の過去問ナンバーを調べたら，実際には存在しないものだった」，「確認作業が増えて，効率的にはならなかった」
	出力品質の課題	8	「現行のガイドラインでは不可とされるものを古いガイドラインに則って正答として扱ってしまう」，「仲間はずれ選択肢ができてしまう」
	AI出力の検証負担	4	「ダブルチェック機能がついているとよい」，「作問後の検証機能も充実させていただきたい」
	教員への周知・研修の必要性	4	「事前の講習会やe-learningなどを実施してもよいかもしれません」，「ベストプラクティスを共有してほしい」
	小計	31	※機能拡張要望（画像対応等）など11件を除く

こうした運用の実態を裏付けている可能性がある．

2. 生成AI活用に伴う新たな外在的認知負荷

本研究では過去問参照時のハルシネーションが最大の課題として挙げられた．事実や根拠に基づかない情報を生成するハルシネーションはAIの利用に伴う新たな外在的負荷といえる．生成AIの活用が作問の効率化に至るかどうかは，従来の外在的負荷の軽減と新たな外在的負荷の増大のバランスに強く影響されるだろう．しかし，ハルシネーションはプロンプトやナレッジデータの最適化といった技術的対処により低減可能と考えられる．

3. 本アプリケーションに対する教員の受容

技術受容モデル（Technology Acceptance Model, TAM）では，利用意図は「知覚された有用性」と「知覚された使いやすさ」によって規定される[14]．本研究では有用性と使いやすさの評価がいずれも高く，継続利用意向も高かったことから，TAMの想定に一致する結果といえる．

一方で，自由記述では教員への周知や研修の必要性が指摘された．ツールの性能に加え，周知や研修体制の整備が受容の向上の鍵である．

4. 教育現場への応用

本研究の結果は，生成AIを活用した作問支援が教員の教育活動に複数の側面で寄与しうることを示唆している．作問時間の短縮と心理的負担の軽減は，教員が問題の質の吟味や他の教育活動に時間と労力を振り向けることを可能にすると考えられる．問題の質が維持・向上したとの回答が大多数を占めたことは，効率化が評価の質を損なわ

ない可能性を示唆している．高い継続利用意向と推奨度は，他施設への展開可能性を支持しており，生成 AI を活用した作問支援の組織的導入は現実的な選択肢となりうる．

5. 限界

本研究にはいくつか限界がある．

第一に，作問時間の測定方法の主観性である．使用前の作問時間は回顧的自己報告であり，想起バイアスにより実時間より長く回答している可能性がある．ただし，生成 AI による作問の時間短縮効果は複数報告されており，一定の妥当性を有すると考えられる[10, 12, 15]．また，本研究で回答を求めた作問所要時間は作問開始から完成までの総所要時間であり，AI 出力の検証に要する時間も含まれていると考えられる．ただし，検証時間は明示的には分離されておらず，今後は AI 出力の検証と問題作成それぞれに要する時間を区別して調査することが，効率化の機序をより詳細に理解するうえで必要である．

第二に，複数のバイアスの存在である．ツールを高く評価する者ほど回答に応じやすい選択バイアス，組織として導入したツールに対する権威バイアス，研究責任者が開発者でもあることによる社会的望ましさバイアスにより，肯定的回答に偏った可能性がある．今後は問題提出時にアンケートへの回答を求めるなど，回答率を高める工夫により選択バイアスの低減を図り，多施設での検証によりこれらのバイアスの影響を低減する必要がある．

第三に，問題の品質に対する客観的評価の欠如である．本研究の品質評価は教員の主観に基づいており，採点除外問題が発生していることからも，検証プロセスの重要性が示唆された．識別指数・難易度指数による項目分析等を用いた客観的な品質検証は今後の課題である．

第四に，運用期間中にバージョン更新を行ったため，回答者間で評価条件が一様でなかった可能性がある．今後は使用時期の記録によりバージョンを特定し，層別の評価を行う必要がある．

第五に，単施設で実施された点であり，組織文化や教員特性の影響を受けている可能性がある．多施設での再現性の検討が必要である．

6. 今後の展望

第一に，学生の回答データに基づく問題の質の客観的評価である．今後は識別指数や難易度指数などの項目分析を実施し，生成された問題の測定特性を定量的に検証する必要がある．これにより，問題の測定特性に与えた影響を明らかにすることが可能となる．

第二に，多施設への展開である．異なる環境下での検証を通じて，結果の一般化可能性を検討することが必要である．

第三に，ハルシネーションへの対処と検証である．本研究で最大の課題として抽出された過去問参照時のハルシネーションについては，本調査実施後にプロンプトの制約条件の強化やナレッジデータの再整理などの技術的対処により，大幅に改良できる見通しがついている．今後はその効果を定量的に検証するとともに，改善後のアプリケーションが教員の検証負担をどの程度軽減するかを評価する必要がある．

認知負荷理論の観点から，生成 AI の導入効果は従来の外在的負荷の軽減と新たな負荷とのバランスに規定されうる．今後は，教員の主観的評価と学生の回答データに基づく客観的評価を継続し，生成 AI による作問支援の正味の効果を多面的に検証していく必要がある．

謝　辞

アンケートにご協力いただいた筑波大学医学類教職員の皆様に感謝申し上げます．

文　献

1）Schuwirth LWT, van der Vleuten CPM. General overview of the theories used in assessment：AMEE Guide No. 57. *Med Teach* 2011；**33**(10)：783-97.
2）Javaeed A. Assessment of Higher Ordered Thinking in Medical Education：Multiple Choice Questions and Modified Essay Questions. *MedEdPublish* 2018；**7**：128.
3）Karthikeyan S, O'Connor E, Hu W. Barriers and facilitators to writing quality items for medical

school assessments-a scoping review. *BMC Med Educ* 2019；**19**(1)：123.
4) Magzoub ME, Zafar I, Munshi F, et al. Ten tips to harnessing generative AI for high-quality MCQS in medical education assessment. *Med Educ Online* 2025；**30**(1)：2532682.
5) Royal KD, Hedgpeth M-W, Jeon T, et al. Automated item generation：the future of medical education assessment. *EMJ Innov* 2018；**2**(1)：88-93.
6) Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Appl Meas Educ* 2002；**15**(3)：309-34.
7) Downing SM. The effects of violating standard item writing principles on tests and students：the consequences of using flawed test items on achievement examinations in medical education. *Adv Health Sci Educ Theory Prac*t 2005；**10**(2)：133-43.
8) Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ* 1983；**17**(3)：165-71.
9) Tarrant M, Knierim A, Hayes SK, et al. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Educ Today* 2006；**26**(8)：662-71.
10) Cheung BHH, Lau GKK, Wong GTC, 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.
11) Laupichler MC, Rother JF, Grunwald Kadow IC, et al. Large Language Models in Medical Education：Comparing ChatGPT-to Human-Generated Exam Questions. *Acad Med* 2024；**99**(5)：508-12.
12) Artsi Y, Sorin V, Konen E, et al. Large language models for generating medical examinations：systematic review. *BMC Med Educ* 2024；**24**(1)：354.
13) van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education：design principles and strategies. *Med Educ* 2010；**44**(1)：85-93.
14) Davis FD. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q* 1989；**13**(3)：319-40.
15) Schneid SD, Armour C, Evans S, et al. Alexa, write my exam：ChatGPT for MCQ creation. *Med Educ* 2024；**58**(11)：1373-4.