# نموذج قائم على الذكاء الاصطناعي التوليدي لتجويد مقياس تقييم مصفوفة الاختبارات

# A Generative AI-Based Framework for the Evaluation and Optimization of Educational Test Blueprints

**د. رنا فاروق محبوب، أستاذ مشارك، جامعة الملك عبد العزيز، المملكة العربية السعودية.**
**E-mail: rfmahbub@kau.edu.sa**

**المستخلص:**

الاستفادة من تطبيقات الذكاء الاصطناعي التوليدي له الأثر الكبير الذي يساهم في رفع كفاءة عمل عضو هيئة التدريس وأيضاً توفير الوقت والجهد في انشاء التقارير الخاصة بالاختبارات. تهدف هذه الدراسة قياس كفاءة نموذج مخطط تقييم الاختبارات (Full Exam Management GPT) القائم على الذكاء الاصطناعي التوليديChat GPT. يساهم النموذج في اعداد التقرير النهائي لمصفوفة الاختبار (Blueprint) وتحليل نتائج مؤشر التمييز DI ومن ثم تقديم مقترح بالتوصيات التي يجب أن تتم لتجويد أسئلة الاختبار. تم استخدام المنهج الوصفي لدراسة الحالة (الفردية) من خلال أربع مراحل متتالية طُبقت على النموذج شملت: ١- مواءمة المخرجات التعليمية وفق توصيف المقرر، ٢- اعداد الاختبار التحريري في ضوء مخرجات المقرر، ٣- استخلاص نتائج تقرير (DIA)، ٤- تنفيذ خطط التحسين المقترحة للأسئلة وإعادة الاختبار. أظهرت النتائج بعد المقارنة بين الاختبارين تحسُّن إيجابي في مؤشر التمييز للأسئلة ذات المؤشّر المنخفض في الاختبار الأول. كما تم تقييم جودة الأسئلة سواء باستبقائها أو استبعادها وفق مؤشر التمييز، بما يساهم في تجويد الاختبار التحريري. تؤكد نتائج دراسة الحالة أن نموذج (FEMGPT) قادر على اعداد تقرير مصفوفة الاختبار بما يضمن تجويد أسئلة الاختبار وتسهيل مهمة المعلم.

**الكلمات المفتاحية:** تقرير مصفوفة الاختبارات، الذكاء الاصطناعي التوليدي، مخرجات التعلم، مؤشر التمييز.

**Abstract:**

Leveraging generative AI applications significantly enhances faculty members' work efficiency, saving time and effort in the preparation of exam reports. This study aims to evaluate the effectiveness of the Full Exam Management GPT (FEMGPT) model, which is based on generative AI (ChatGPT). The model supports the preparation of the final blueprint report, analyses discrimination index (DI) results, and provides proposals and recommendations for improving exam questions. A descriptive case study approach was adopted, applied through four successive stages: 1- Aligning course learning outcomes according to the course specification. 2- Preparing the written exam based on the aligned outcomes. 3- Extracting and analysing the Discrimination Index Analysis (DIA) report results. 4- Implementing proposed improvement plans for exam questions and conducting a retest. The comparison between the two tests demonstrated a positive improvement in the discrimination index, particularly for questions that initially exhibited a low index. Furthermore, the quality of exam questions was evaluated based on whether items were retained or excluded according to their discrimination indices, contributing to the overall enhancement of the written

exam. The results of the case study affirm that the FEMGPT model is an effective tool for generating comprehensive test matrix reports, thereby facilitating and streamlining the examination preparation process for faculty members.

**Key words:** Test matrix, Full Exam Management GPT, learning outcomes, Assessment improvement.

**المقدمة:**

تنوعت ثورة الذكاء الاصطناعي وتطبيقاته في دعم العملية التعليمية، فاليوم لا يقتصر استخدامه على تعلم المعارف والمهارات، بل يسهُم في تقديم حلول تطويرية لأهم المشكلات التي تواجه المتعلم أو المستخدم بصورة ذكية في وقت وجهد أقل (عبد الصمد & أحمد، ٢٠٢٠). من هذا المنطلق سعى الباحثون لدراسة تطبيقات الذكاء الاصطناعي التوليدي في تطوير التعليم والتي تشمل استخدامه في توليد الصور والفيديوهات التعليمية والمحتوى المرئي للعرض وغيرها. ولا يقتصر ذلك على الطالب فقط، بل يشمل أيضاً المُعَلم، وذلك من خلال أداوت تساعد في تحسين العملية التعليمية وأساليب التقييم والتغذية الراجعة عن المقرر. قام Weng et al (2024) بمراجعة الاتجاهات والتحديات الحالية في تقييم تعلم الطلاب للمهارات المطلوبة في المقرر وذلك من خلال مراجعة منهجية تشمل: التعليم التقليدي، والتقييم المبتكر المعاد تقييمه، والتقييم المدمج بالذكاء الاصطناعي (بواسطة Chat GPT) وذلك ضمن أنشطة التقييم. أظهرت النتائج أن الأساليب التقليدية أقل فعالية عن البيئة التعليمية المدمجة بالذكاء الاصطناعي، كما أن تبني التقنية الذكية وأدواتها يساهم في تنمية مهارات التفكير العليا مثل الذاتي والإبداعي والنقدي. وذلك يتفق مع (Chiu, Xia, Zhou, Chai, & Cheng, 2023) إلى أن الذكاء الاصطناعي يلعب أدوارًا متعددة في التعليم، تشمل التعلّم، والتدريس، والتقويم، والإدارة. وهدفت الدراسة الى تحليل أدوار التعليم لفهم نتائج التعلم للطلاب والمعلمين الناتجة عن استخدام تقنيات الذكاء الاصطناعي، وتحديد التحديات واقتراح اتجاهات مستقبلية للبحث العلمي من خلال معايير PRISMA لاختيار الدراسات وتحليل المحتوى والترميز الاستقرائي لتصنيف الأدوار والنتائج والتحديات. وأظهرت النتائج في **التعليم** زيادة التكيّف والتفاعلية في البيئات الرقمية، وفي **التدريس** تقديم استراتيجيات تدريس متكيّفة، وفي **التقييم** سهولة التصحيح الآلي للأسئلة وتحسين جودتها، وفي **الإدارة** تم اتخاذ القرار التعليمي باستخدام البيانات الضخمة وتحسين منصات الإدارة التعليمية. من هذا المنطلق يتضح أن دور الذكاء الاصطناعي في العملية التعليمية وتفعيله لم يعد خيارًا، بل ضرورة حتمية لتحقيق تعليم فعّال ومواكب للتطورات الحديثة.

تُظهر مراجعة الأدبيات الحديثة اهتمامًا متزايدًا بتفعيل دور تطبيقات الذكاء الاصطناعي في البيئات التعليمية، لا سيما في مجالي التقييم والقياس. أجرى Santos وJunior (2024) دراسة تحليلية لكيفية استخدام الذكاء الاصطناعي (AI) في دعم تقييم الطلاب في تعليم الحوسبة. اعتمدت الدراسة منهج المراجعة المنهجية للأدبيات (SLR) وفق نموذج Kitchenham. وضحت الدراسة الفرص الرئيسية لاستخدام الذكاء الاصطناعي في التقييم وهي: ١- تقييم المهارات الشخصية والأكاديمية، ٢- تقييم المهام والمشروعات، ٣- تقييم المعرفة الحالية، ٤- تنبؤ بالأداء المستقبلي، ٥- تحليل جودة أسئلة الاختبارات حسب تصنيف بلوم لتحسين جودة التقييمات (Santos & Junior, 2024).

على الصعيد التقليدي لأساليب التقييم والقياس غالباً ما يكون تحليل بيانات الاختبار قائم على مصفوفة الاختبارات (Blueprint) أو ما يُعرف أيضاً في بعض القطاعات التعليمية بمصطلح (مخطط التقييم). والذي يهدف الى بناء خريطة واضحة لربط أساليب التقييم وأسئلة الاختبارات في ضوء المحتوي الدراسي لتوصيف المقرر تشمل (المحتوى التعليمي، استراتيجيات التدريس، وأساليب التقييم)، وربطه بالمخرجات التعليمية. كما أنها تعمل بصورة تضمن توزيع الأوزان النسبية في عملية التقييم بما يتواءم مع مخرجات البرنامج (الهيئة الوطنية للاعتماد والتقويم الأكاديمي [NCAAA], 2024). وذكرت المراجعة المنهجية في هذا المجال لـ (عبد

اللطيف وآخرون، ٢٠٢٤) أن مصفوف الاختبار أداة فعّالة لتقليل التهديدات المرتبطة المصداقية (الصدق) والثبات (الاعتمادية)لأساليب التقييم الخاصة بالمقرر. وتم استخدام إطار المراجعة المنهجية لـ Arksey and O'Malley، وشملت أساليب كمية ونوعية ومختلطة، كما تم استخدام أطر مثل تصنيف بلوم المعدّل وهرم ميلر لقياس الكفاء. أظهرت النتائج أن مصفوفة الاختبار تساهم في تحسين مصداقية التقييم من خلال تغطية مناسبة لمخرجات التعلم. أيضاً دعم التقييم المستند إلى الأدلة والمساعدة في اتخاذ قرارات تعليمية عادلة وتعزيز الشفافية والقبول بين الطلاب والمعلمين بما يساهم في رفع مستوي جودة التصميم التربوي.

من هذا المنطلق يتضح أهمية مصفوفة الاختبارات في تجويد أساليب التقييم والاختبارات وفق توصيف المقرر بما يتماشى مع توجُّهات الاعتماد الأكاديمي للبرامج التعليمية وخاصة التعليم الجامعي. ولضمان رفع جودة الاختبارات وتحسين الأسئلة وتحقيق مصداقية التقييم تعمل البرامج الأكاديمية بالجامعات الى تفعيل مخطط التقييم في مقرراتها. وغالباً ما يتم العمل بهذا الأسلوب التقليدي المتعارف عليه عن طريقة نماذج متعددة ومتنوعة بواسطة برنامج الاكسل (Excel). ويقوم النموذج المصمم ببرنامج Excel على تحليل نتائج درجات الطلاب للاختبار وفق خطة القياس المضمّنة، ومن ثم تظهر النتائج التحليلية لدرجة صعوبة أسئلة الاختبار (Discriminatory Item Analysis) وتقيم جودتها وفق مؤشر التمييز (Discrimination Index) (Chiavaroli & Familari, 2011; Dela Peña et al., 2011; Matoba, 2020).

**مشكلة البحث:**

استخلاصاً مما سبق، يتبين أهمية تفعيل دور الذكاء الاصطناعي التوليدي في توليد نموذج عمل يسهل على أعضاء الهيئة التعليمية مهمة تحليل نتائج الاختبارات وتجويد الأسئلة بهدف تحسين الإنتاجية والحصول على نتائج دقيقة. طوّر مركز تطوير التعليم الجامعي بجامعة الملك عبد العزيز نموذج عمل لمخطط تقييم الاختبارات (Full Exam Management GPT) بواسطة تطبيق الذكاء الاصطناعي التوليديChat GPT. تم تصميم النموذج من خلال انشاء Prompt Template وتلقينه الأوامر والتعليمات بتسلسل كتابة واضح يشمل: ١- التحديد والوضوح (Specificity and Clarity)، ٢- هيكلة المدخلات والمخرجات (Structured Inputs and Outputs)، ٣- استخدام المحددات (Delimiters)، ٤- تقسيم المهام للسياقات المعقدة وتضمين التعليمات (Task Decomposition for Complex Operations) (Saravia, 2024). يعمل النموذج على توفير الوقت والجهد في إدارة الاختبارات وتطوير ومراجعة مصفوفة الاختبار وتحليل البيانات والمدخلات. كما يساهم في تحسين أسئلة الاختبارات في ضوء مخرجات المقرر ونتائج الطلاب، واعداد تقرير تحليل العناصر. وأخيراً تقديم تغذية راجعة وتقييم الإنجاز لتحقيق مخرجات المقرر وتقديم مقترح خطط التحسين للاختبار.

ولقياس فاعلية نموذج مخطط التقييم (FEMGPT) المعتمد على الذكاء الاصطناعي التوليدي، تم تطبيقه في تجويد الاختبارات. وذلك لدراسة وتتبع التحليلات والنتائج وخطط التحسين التي يقدمها من خلال الإجابة على التساؤلات:

١. هل يمكن لنموذج مخطط التقييم (FEMGPT) المعد بواسطة الذكاء الاصطناعي التوليدي Chat GPT أن يربط أساليب التقييم وأسئلة الاختبارات في ضوء المحتوي الدراسي لتوصيف المقرر بالمخرجات التعليمية؟

٢. هل يمكن أن يحلل نموذج مخطط التقييم (FEMGPT) المعد بواسطة الذكاء الاصطناعي التوليدي Chat GPT نتائج درجة صعوبة أسئلة الاختبارات وفق مؤشر التمييز DI؟

٣. هل يمكن أن يقدم نموذج مخطط التقييم (FEMGPT) المعد بواسطة الذكاء الاصطناعي التوليدي Chat GPT تقديم تغذية راجعة تسهم في تحسين الاختبارات من خلال تحليل مصفوفة الاختبارات؟

**هدف البحث وأهميته:**

يهدف البحث الى قياس فاعلية نموذج مخطط التقييم (FEMGPT) المعد بواسطة الذكاء الاصطناعي التوليدي Chat GPT وذلك من خلال:

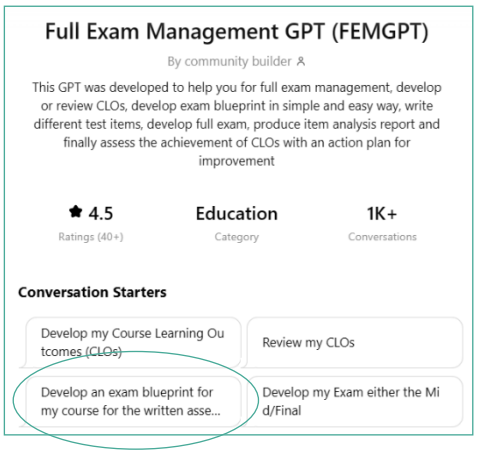
١. توضيح كيفية ربط المحتوى التعليمي بأساليب تقييم الاختبارات ورسم خريطة عمل لعملية تقييم الطلاب وفق نموذج (FEMGPT) تضمن تحقيق مخرجات المقرر بجودة وبأقل جهد ووقت.

٢. قياس فاعلية الاختبارات والتحقق من جودتها وفق مؤشر DI بما يحقق تحليل لبيانات الاختبار وخطط تحسين. تضمن التقليل من أخطاء صياغة الأسئلة و/أو تقديم درجة صعوبة لا تناسب مستوى مهارات المقرر.

٣. تقديم التغذية الراجعة للهيئة التعليمية وخطط التحسين المقترحة، بما يساهم في تحقيق مصداقية وثبات أكبر للاختبار.

**منهج البحث:**

يتبع منهج البحث دراسة الحالة (دراسة الحالة الفردية)، وهو منهج وصفي يقوم على جمع البيانات الكمية والكيفية ودراستها على عينة متعددة أو مفردة، ويساعد في تحليل المتغيرات والمعطيات وتحديد الروابط وعوامل نجاح أو فشل الحالة المدروسة (قندلجي، ٢٠١٩). ولدراسة فاعلية نموذج (FEMGPT) تم اختيار مقرر مبادئ علم الالياف الذي تم تطوير واعتماد توصيفه من قبل عمادة الجودة والاعتماد الأكاديمي، ضمن برنامج بكالوريوس العلوم (تكنولوجيا الأزياء والنسيج) بقسم الأزياء والنسيج – كلية علوم الانسان والتصاميم – بجامعة الملك عبد العزيز. واستخدام نموذج FEMGPT المعد بواسطة الذكاء الاصطناعي التوليدي من خلال انشاء مشروع جديد اختيار تطوير مصفوفة الاختبار التحريري للمقرر كما هي موضحة بالشكل رقم (١).




**الشكل رقم ١ صورة توضيحية لنموذج FEMGPT المعد بواسطة الذكاء الاصطناعي التوليدي**

تم تصميم إطار دراسة الحالة على أربعة مراحل من خلال تلقين الأوامر لنموذج FEMGPT:

**المرحلة الأولى:** مواءمة المخرجات التعليمية وفق توصيف المقرر وتشمل تحديد عدد الأسئلة ومستوى صعوبة ونوع المهارة للاختبارات وأساليب التقييم للمقرر والتوزيع الموزون لنسبة الدرجة لكل مخرج (معرفي، مهاري، قيم). في هذه الدراسة تم التركيز على المخرجات المعرفية التي تقاس من خلال الاختبارات التحريرية للمقرر وهي:

١،١ تصنف أنواع الألياف الى طبيعية وتحويلية وصناعية وكيفية غزلها.

١،٢ تتعرف على المواصفات القياسية لاختبارات الألياف والنسيج.

تم استخدام نسبة موزون الدرجة المعتمد بالتوصيف بدون تعديل (٢٠٪) للاختبار النصفي التحريري.

**المرحلة الثانية:** اعداد الاختبار التحريري في ضوء مخرجات المقرر لتجربته على دفعة الفصل الدراسي الأول ٢٠٢٥، ويمكن تلقين الأمر بطلب اعداد أسئلة مقترحة في ضوء نتائج *المرحلة الأولى* (لذات المشروع) بعد ارفاق المحتوي التعليمي. ومن ثم مراجعتها أو إعادة صياغاتها اما بطلب ذلك من خلال أمر جديد أو بشكل شخصي. وأيضاً في هذه المرحلة تم استخدام منصة ZIPGRADE لإنشاء التصحيح الالكتروني للاختبار الورقي بهدف الحصول على تقرير تحليل بيانات الاختبار (Discriminatory Item Analysis DIA). وغالباً ما يمكن الحصول عليه من خلال عمل الاختبارات الالكترونية على نظام الفصول الافتراضية Blackboardالخاص بالجامعة أو الجهة التعليمية.

**المرحلة الثالثة:** استخلاص نتائج تقرير (DIA) بعد تنفيذ الاختبار للحصول على التغذية الراجعة وخطط التحسين المقترحة، ووضعها في تقرير مصفوفة الاختبار. وفي هذه المرحلة نقوم بتلقين الأوامر للنموذج بالتفاصيل التي نريد التعمق فيها على حسب النتائج مثل: تفصيل أوضح لعمليات التحسين في الأسئلة كحذف أو تعديل في جذر السؤال أو في المشتتات وهكذا.

**المرحلة الرابعة:** تنفيذ خطط التحسين المقترحة للأسئلة وإعادة الاختبار على دفعة الفصل الدراسي الثاني ٢٠٢٥. ومن ثم العمل على نفس المشروع وإعادة المرحلة الثالثة ورفع تقرير (DIA) الجديد. من ثم تلقين الأمر بطلب التغذية الراجعة وعمل مقارنة بين نتائج التقريرين للدفعتين، والحصول على خطة تقييم نهائي لجودة الأسئلة من حيث قياس مستواها (صعبة، متوسطة، سهلة) أو مستبعدة لثبوت عدم كفاءتها لقياس مخرجات المقرر.

**النتائج والمناقشة:**

**نتائج المرحلة الأولى:**

تم تلقين الأمر للنموذج بالنص الآتي: "من خلال توصيف المقرر المرفق بإنشاء مخطط الامتحان النصفي، للمخرجين المعرفية (١,١، ١,٢) بحيث يكون مجموع موزون الدرجة ١٠٪ لكل مخرج، ومجموع الأسئلة عشرون سؤال موزعة بنسبة ٥٠٪ من الدرجة لكل مخرج بموزون درجة لكل مخرج". يوضح الجدول رقم (١) تفاصيل نتائج التوزيع النسبي لموزون الدرجة على مخرجات المقرر.

**الجدول رقم ١ نتائج التوزيع النسبي لموزون الدرجة على مخرجات المقرر.**

| رقم المخرج | نوع الأسئلة | عدد الأسئلة | المستوى المعرفي | النسبة % | المحتوى التعليمي |
|---|---|---|---|---|---|
| ١,١ | اختيار من متعدد MCQ | ١٠ | التذكر | ١٠ | تعريف وتصنيف أنواع الألياف وطرق غزلها |
| ١,٢ | اختيار من متعدد MCQ | ١٠ | التذكر | ١٠ | فهم واستخدام المواصفات القياسية لاختبارات الألياف والنسيج |

يمكن في هذه المرحلة أيضاً، تعديل موزون الدرجة على المخرجات بنسب مختلفة إذا حصل أي تعديلات طارئة خلال الفصل الدراسي على سبيل المثال ظروف معيّنة كحذف أحد أساليب التقييم أو تغيير في سير العملية التعليمة أو تقليص عدد الأسابيع الدراسية... الخ. لهذا تم تجربة تلقين النموذج الأمر الآتي للنموذج: "مرفق المحتوى التعليمي المحدد للطلاب للتقييم بالاختبار النصفي، في ضوء توصيف المقرر قدم مقترح تعديل نسبة الدرجة". تم التفاعل من قبل FEMGPT بالتعديل الموضح بالشكل رقم (٢). حيث جعل للمخرج (١,١) النسبة الأكبر من الدرجة بما يشكل (٦٠٪) و (٤٠٪) على المخرج (١,٢) وأوضح GPT أن سبب توزيع النسب هو أن في **الاختبار النصفي**: التركيز الأكبر على المخرج ١,١ وذلك لأنهاء الطلاب معظم الدروس التعليمية التي تغطيه، وتم تقديمه مبكرًا في الخطة الدراسية. بينما يرجح أن يكون في **الاختبار النهائي**: بتوزيع متوازن على كلا المخرجين (١,١ و١,٢)، وبذلك يمكن أن نؤكد على أن FEMGPT يقدم تفسيرات ومقارنات

منطقية مدروسة من خلال ما يتم تلقينه من أوامر وما تم تعليمه للذكاء الاصطناعي مستندا على مرجعية الاعدادات التي يعتمد عليها المستخدم (Saravia, 2024).

**الاختبار النصفي**

| المخرج التعليمي (CLO) | عدد الأسئلة | نوع السؤال | النسبة المئوية (%) | الدرجة |
|---|---|---|---|---|
| CLO 1.1 | 12 | MCQs، مطابقة | 60% | 12 |
| CLO 1.2 | 8 | MCQs، أسئلة قصيرة | 40% | 8 |
| الإجمالي | 20 | | 100% | **20** |

**الشكل رقم ٢ تعديل FEMGPT على التوزيع النسبي للدرجة في ضوء المحتوي التعليمي**

**نتائج المرحلة الثانية:**

إعداد أسئلة الاختبار التحريري وفق مخرجات المقرر من المراحل المهمة في عملية تقويم مخرجات المقرر المعرفية، لأنها حجر الأساس الذي يقوم عليه تحليل البيانات وتجويد عملية القياس (Weng et al.,2024). تم تلقين الأمر للنموذج بالنص الآتي "ضع أسئلة اختيار من متعدد عددها عشرون مستعيناً بالمحتوي التعليمي المرفق، وفي ضوء المصفوفة المقدمة سابقاً". يوضح الشكل (٣) المصفوفة المقترحة لربط المحتوي التعليمي بتوزيع الأسئلة، وهي تعد كنمط استفهامي يستخدمه الذكاء الاصطناعي التوليدي للتأكد من فهم الأمر بالشكل المطلوب قبل استكمال المهمة.

**Midterm Exam Blueprint (Based on Course Topics)**

| Topic # | Topic Name | Recall (40%) | Cognitive (60%) | Total MCQs |
|---|---|---|---|---|
| 1 | مقدمة عن الألياف وخواصها | 2 | 3 | 5 |
| 2 | الفحص الميكروسكوبي للألياف وخلط الألياف | 1 | 3 | 4 |
| 3 | فحص الحرق للألياف الطبيعية والصناعية والتحويلية | 1 | 2 | 3 |
| 4 | الفحص الكيميائي للألياف الطبيعية والصناعية والتحويلية | 1 | 2 | 3 |
| 5 | مراحل الغزل وتكوين الخيط | 2 | 3 | 5 |
| Total | — | 7 | 13 | 20 MCQs |

**الشكل رقم ٣ مصفوفة ربط المحتوي التعليمي بتوزيع الأسئلة**

استكمالاً لما جاء أعلاه، يصيغ FEMGPT أسئلة الاختبار من متعدد وفق المصفوفة بعد تلقين الأوامر سواء بتعديل صيغة جذر السؤال وتحديد عدد المشتتات وفق مستوى الصعوبة المطلوب. توضيح لاستجابة النموذج لحوار الأوامر موضح أدناه:

"شكرًا لك على تزويدي بالمواد المرفقة!

استنادًا إلى مخطط الامتحان الخاص بك (الذي يركز على نواتج التعلم ١,١ و١,٢ مع ٢٠ سؤال اختيار من متعدد (MCQs))، سأقوم الآن بإنشاء ٢٠ سؤال اختيار من متعدد باستخدام سيناريوهات حالة مستمدة من محتوى الشرائح.


المجلة الدولية لتكنولوجيا التعليم والمعلومات ISSN Print: 1658-9556 / ISSN Online: 2961-4023

**سيتبع كل سؤال البنية التالية:**

✓ سيناريو حالة (قصة قصيرة مرتبطة بالموقف)

✓ سؤال واضح ( Lead-in)يطلب من الطالب اتخاذ قرار أو تحديد الإجابة

✓ خمس اختيارات للإجابة (إجابة صحيحة واحدة وأربعة مشتتات)

✓ توضيح للإجابة الصحيحة مع شرح السبب"

أكدت دراسة (Rivera-Rosas et al., 2024) التي استخدمت ChatGPT-3.5 لتوليد خمسون سؤال من نوع اختيار من متعدد (MCQs) لمقرر التشريح وعلم الحركة ( The Anatomy and Kinesiology Course)، أنه رغم جودة وكفاءة الأسئلة المولّدة الا أنه يجب على أستاذ المقرر أن يراجع صحة وصلاحية الأسئلة لتجنب مشكلات هلوسة الذكاء الاصطناعي. من هذا المنطلق تتفق هذه الدراسة مع (Rivera-Rosas et al., 2024)، أن نموذج FEMGPT أيضاً لا يخلو من الهلوسة ويجب على المستخدم البشري المراجعة والتدقيق. الاستجابة أدناه لأحد الأسئلة المولّدة باستخدام النموذج ويظهر بها خطأ شائع جداً في صياغة السؤال وغفل عنه ال AI، وهي ألا يكون في جذر السؤال أي تلميح عن الإجابة الصحيحة.

**اعداد التصحيح الالكتروني باستخدام منصة ZIPGRADE**

**Section 2: CLO 1.2 – المواصفات القياسية لاختبارات الألياف والنسيج (10 MCQs)**

**Q11: Standard Fiber Testing Methods**

**Scenario:**
أثناء العمل في مختبر تحليل النسيج، طُلب من هدى استخدام طريقة لتحديد نوع الألياف بناءً على طريقة احتراقها ورائحتها.

يظهر هنا التلميح

**Question:**
ما الاختبار المناسب لهذه الحالة؟

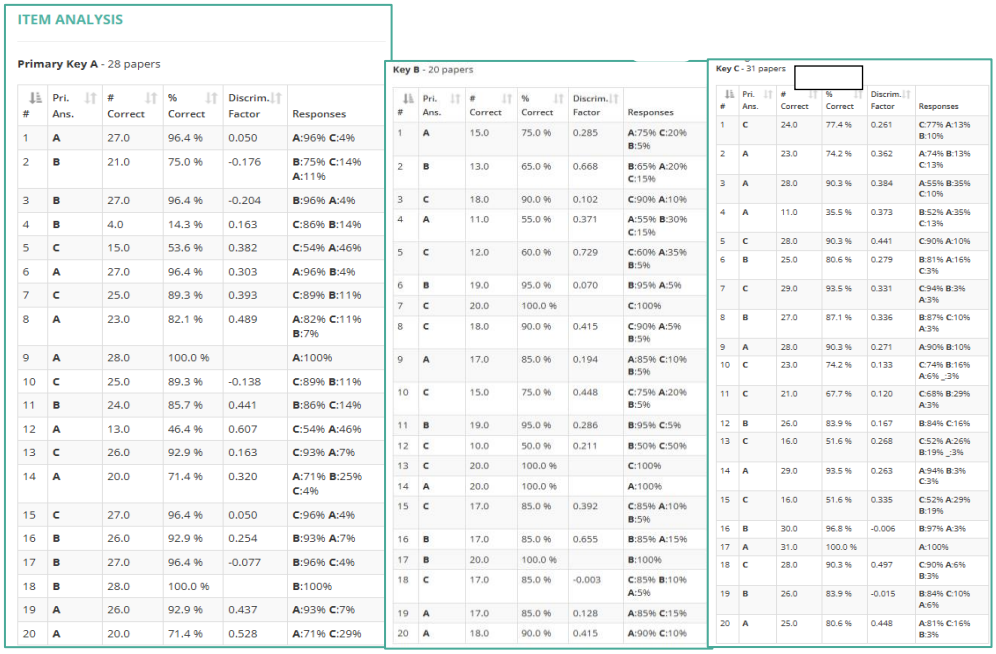
**Choices:**

A) الفحص الميكروسكوبي
B) اختبار الحرق
C) اختبار الامتصاصية
D) اختبار الوزن النوعي
E) اختبار الثباتية اللونية

**Correct Answer: ✓ B) اختبار الحرق**

**Explanation:** اختبار الحرق يساعد في تحديد نوع الألياف بناءً على لون اللهب والرائحة وبقايا الرماد، وهو طريقة شائعة للتمييز بين الألياف الطبيعية والصناعية.

ZipGrade LLC هي شركة ذات مسؤولية محدودة أسسها John Viebach، وأستلهم فكرتها من زوجته التي تعمل معلمة. تقدم المنصة خاصية الاختبار والتصحيح الالكتروني واعداد ملخص تقرير DIA لنتائج الاختبار. تم استخدام التصحيح الالكتروني في هذه الدراسة للحصول على تحليل بيانات الاختبار، شكل ذلك الاختبار النصفي للفصل الدراسي الأول والثاني بعد الأخذ بالتوصيات. تم تصميم نماذج الإجابة لثلاث نماذج للاختبار النصفي (A, B, C)، كما هي موضحة بالشكل رقم (٤).



ITEM ANALYSIS

Primary Key A - 28 papers

| # | Pri. Ans. | # Correct | % Correct | Discrim. Factor | Responses |
|---|---|---|---|---|---|
| 1 | A | 27.0 | 96.4 % | 0.050 | A:96% C:4% |
| 2 | B | 21.0 | 75.0 % | -0.176 | B:75% C:14% A:11% |
| 3 | B | 27.0 | 96.4 % | -0.204 | B:96% A:4% |
| 4 | B | 4.0 | 14.3 % | 0.163 | C:86% B:14% |
| 5 | C | 15.0 | 53.6 % | 0.382 | C:54% A:46% |
| 6 | A | 27.0 | 96.4 % | 0.303 | A:96% B:4% |
| 7 | C | 25.0 | 89.3 % | 0.393 | C:89% B:11% |
| 8 | A | 23.0 | 82.1 % | 0.489 | A:82% C:11% B:7% |
| 9 | A | 28.0 | 100.0 % | | A:100% |
| 10 | C | 25.0 | 89.3 % | -0.138 | C:89% B:11% |
| 11 | B | 24.0 | 85.7 % | 0.441 | B:86% C:14% |
| 12 | A | 13.0 | 46.4 % | 0.607 | C:54% A:46% |
| 13 | C | 26.0 | 92.9 % | 0.163 | C:93% A:7% |
| 14 | A | 20.0 | 71.4 % | 0.320 | A:71% B:25% C:4% |
| 15 | C | 27.0 | 96.4 % | 0.050 | C:96% A:4% |
| 16 | B | 26.0 | 92.9 % | 0.254 | B:93% A:7% |
| 17 | B | 27.0 | 96.4 % | -0.077 | B:96% C:4% |
| 18 | B | 28.0 | 100.0 % | | B:100% |
| 19 | A | 26.0 | 92.9 % | 0.437 | A:93% C:7% |
| 20 | A | 20.0 | 71.4 % | 0.528 | A:71% C:29% |

Key B - 20 papers

| # | Pri. Ans. | # Correct | % Correct | Discrim. Factor | Responses |
|---|---|---|---|---|---|
| 1 | A | 15.0 | 75.0 % | 0.285 | A:75% C:20% B:5% |
| 2 | B | 13.0 | 65.0 % | 0.668 | B:65% A:20% C:15% |
| 3 | C | 18.0 | 90.0 % | 0.102 | C:90% A:10% |
| 4 | A | 11.0 | 55.0 % | 0.371 | A:55% B:30% C:15% |
| 5 | C | 12.0 | 60.0 % | 0.729 | C:60% A:35% B:5% |
| 6 | B | 19.0 | 95.0 % | 0.070 | B:95% A:5% |
| 7 | C | 20.0 | 100.0 % | | C:100% |
| 8 | C | 18.0 | 90.0 % | 0.415 | C:90% A:5% B:5% |
| 9 | A | 17.0 | 85.0 % | 0.194 | A:85% C:10% B:5% |
| 10 | C | 15.0 | 75.0 % | 0.448 | C:75% A:20% B:5% |
| 11 | B | 19.0 | 95.0 % | 0.286 | B:95% C:5% |
| 12 | C | 10.0 | 50.0 % | 0.211 | B:50% C:50% |
| 13 | C | 20.0 | 100.0 % | | C:100% |
| 14 | A | 20.0 | 100.0 % | | A:100% |
| 15 | C | 17.0 | 85.0 % | 0.392 | C:85% A:10% B:5% |
| 16 | B | 17.0 | 85.0 % | 0.655 | B:85% A:15% |
| 17 | B | 20.0 | 100.0 % | | B:100% |
| 18 | C | 17.0 | 85.0 % | -0.003 | C:85% B:10% A:5% |
| 19 | A | 17.0 | 85.0 % | 0.128 | A:85% C:15% |
| 20 | A | 18.0 | 90.0 % | 0.415 | A:90% C:10% |

Key C - 31 papers

| # | Pri. Ans. | # Correct | % Correct | Discrim. Factor | Responses |
|---|---|---|---|---|---|
| 1 | C | 24.0 | 77.4 % | 0.261 | C:77% A:13% B:10% |
| 2 | A | 23.0 | 74.2 % | 0.362 | A:74% B:13% C:13% |
| 3 | A | 28.0 | 90.3 % | 0.384 | A:55% B:35% C:10% |
| 4 | A | 11.0 | 35.5 % | 0.373 | B:52% A:35% C:13% |
| 5 | C | 28.0 | 90.3 % | 0.441 | C:90% A:10% |
| 6 | B | 25.0 | 80.6 % | 0.279 | B:81% A:16% C:3% |
| 7 | C | 29.0 | 93.5 % | 0.331 | C:94% B:3% A:3% |
| 8 | B | 27.0 | 87.1 % | 0.336 | B:87% C:10% A:3% |
| 9 | A | 28.0 | 90.3 % | 0.271 | A:90% B:10% |
| 10 | C | 23.0 | 74.2 % | 0.133 | C:74% B:16% A:6% _3% |
| 11 | C | 21.0 | 67.7 % | 0.120 | C:68% B:29% A:3% |
| 12 | B | 26.0 | 83.9 % | 0.167 | B:84% C:16% |
| 13 | C | 16.0 | 51.6 % | 0.268 | C:52% A:26% B:19% _3% |
| 14 | A | 29.0 | 93.5 % | 0.263 | A:94% B:3% C:3% |
| 15 | C | 16.0 | 51.6 % | 0.335 | C:52% A:29% B:19% |
| 16 | B | 30.0 | 96.8 % | -0.006 | B:97% A:3% |
| 17 | A | 31.0 | 100.0 % | | A:100% |
| 18 | C | 28.0 | 90.3 % | 0.497 | C:90% A:6% B:3% |
| 19 | B | 26.0 | 83.9 % | -0.015 | B:84% C:10% A:6% |
| 20 | A | 25.0 | 80.6 % | 0.448 | A:81% C:16% B:3% |

**الشكل رقم 1 نموذج تقرير DIA لنتائج الاختبار النصفي للفصل الدراسي الأول بعد التصحيح الالكتروني**

**نتائج المرحلة الثالثة:**

بعد تنزيل تقرير تحليل البيانات DIA، يتم رفع المرفق على تطبيق الذكاء الاصطناعي FEMGPT لإكمال العمل على نفس المشروع. أُستُخدم النص التالي لتلقين الأمر "استخدم تقرير تحليل بيانات الاختبار المرفق، لإتمام عرض نتائج مصفوفة الاختبار في ضوء ما سبق". إضافةً، في حال طبق الاختبار عدة نماذج، مثلما تم في هذه الدراسة فيمكن رفعها بشكل منفرد أو جميعاً في آن واحد. نفذت هذه الدراسة كلا الطريقتين في **الفصل الدراسي الأول ٢٠٢٥ تم رفع الملفات منفصلة** ومن ثم تلقين FEMGPT الأمر بالمقارنة بينهم، وذلك بهدف حسم ما إذا كان الذكاء الاصطناعي قادر على تمييز نتيجة السؤال ودرجة الصعوبة. علماً أن الأسئلة رتبت عشوائياً في الثلاث نماذج مع تغيير بعض الأسئلة بين النماذج بمستوى صعوبة متساوية. ومن هنا يمكننا التأكد مما إذا كانت هناك هلوسة ناتجة عن الذكاء الاصطناعي خلال عملية التحليل.

يوضح الجدول رقم (٢) نتائج تحليل مؤشر التمييز ومستوى الصعوبة لنماذج الاختبار الثلاثة، نلاحظ أن الاختبار أظهر نسب متقاربة بين المخرجين في مؤشر DI والذي وصف نطاق مستوى صعوبة الأسئلة بين منخفض الى مرتفع. Chiavaroli and Familari (2011) أشارا إلى أن حساب نتائج مؤشر التمييز DI تعطي قيمة تقع بين (-١,٠ و +١,٠) ، بحيث كلما اقتربت القيمة من (-١,٠) تشير الى قدرة تمييز ضعيفة، بمعني أنه سهل جدا أو صياغة السؤال تحتاج إلى تعديل أو حذف. من ناحية أخرى كلما اقتربت القيمة من (+١,٠) تصبح قدرة تمييز السؤال عالية وذات جودة ممتازة ودرجة صعوبة قد تكون عالية. الجدول رقم (٢) جمع نطاق DI كمتوسط لنتائج البيانات للاختبار (A, B, C) للمخرجات التعليمية (١,١ ، ١,٢) والذي أظهر أن هناك انخفاض في مؤشر التمييز وضعف في بعض الأسئلة وأيضاً، يوجد أسئلة ذات جودة عالية.

**الجدول رقم 1مخطط امتحان منتصف الفصل الدراسي الأول – تحليل DI لمفتاح التصحيح (A, B, C)**

| رقم المخرج | نوع الأسئلة | عدد الأسئلة | المستوى المعرفي | النسبة % | نطاق مستوى الصعوبة | نطاق مؤشر التمييز DI |
|---|---|---|---|---|---|---|
| ١,١ | اختيار من متعددMCQ | ١٠ | التذكر | ١٠ | من متوسطة إلى عالية (٣٥,٥٪–١٠٠٪) | من منخفض جدًا إلى مرتفع (-٠,٠٠٦ إلى ٠,٧٢٩) |
| ١,٢ | اختيار من متعددMCQ | ١٠ | التذكر | ١٠ | من متوسطة إلى عالية (٤٦,٤٪–٩٦,٨٪) | من منخفض جدًا إلى مرتفع (-٠,٠١٥ إلى ٠,٦٥٥) |


المجلة الدولية لتكنولوجيا التعليم والمعلومات ISSN Print: 1658-9556 / ISSN Online: 2961-4023

استكمالا لما جاء في نتائج تحليل بيانات الاختبار، أعطى النموذج FEMGPT تحليل منفصل لكل مفتاح تصحيح في جدول، ومن ثم تم تلقين الأمر بجمع البيانات في جدول واحد لتسهل المقارنة (جدول رقم ٣). علاوة على ذلك أعطي الذكاء الاصطناعي توصيات إضافية للتطّوير بشكل مختصر لنقاط الضعف كما يلي:

١. **زيادة صعوبة الأسئلة السهلة جدًا ذات التمييز المنخفض:** ينبغي زيادة صعوبة الأسئلة مثل السؤال Q1 في المفتاح A، والسؤال Q6 في المفتاح B، والسؤال Q16 في المفتاح C، أو تحسين اتساقها مع نواتج التعلم (CLOs).
٢. **مراجعة الأسئلة ذات التمييز السلبي:** يجب مراجعة الأسئلة مثل السؤال Q10 في المفتاح A، والسؤال Q18 في المفتاح B، والسؤال Q16 في المفتاح C، حيث أظهرت تمييزًا سلبيًا، مما يشير إلى الحاجة لمراجعتها لضمان فعاليتها في قياس معرفة الطلاب.
٣. **الاحتفاظ بالأسئلة ذات التمييز العالي والصعوبة المتوسطة:** يُوصى بالاحتفاظ بالأسئلة التي تقع ضمن نطاق الصعوبة المتوسطة والتمييز العالي (مثل السؤال Q5 في المفتاح A، والأسئلة Q2 وQ5 في المفتاح B، والسؤال Q18 في المفتاح C)، حيث إنها تُظهر أداءً جيدًا وتخدم الغرض التقييمي بشكل فعال.

بعد ذلك يمكن تلقين AI أوامر أخرى بناء على رغبة المستخدم مثل تقديم مقترح لتحسين الأسئلة وفق توصياته وهذا أمر اختياري، حيث إنه لا يوجد ضمان بأن ChatGPT لن يستخدم هذه المعلومات. على الرغم من ذلك أثبتت دراسة (Rivera-Rosas et al., 2024) أنه لم تظهر دلالات بأن الذكاء الاصطناعي استخدم الأسئلة ذاتها خلال عملية التعلم والمراجعة للطلاب. حيث تم تجربة توليد الأسئلة من خلال ChatGPT-3.5 في الدراسة، وتم اختبار الطلاب وكانت نتائج تحليل DI توضح أن مستويات الصعوبة أظهرت تفاوت في نتائج الاختبار كما كان متوقع من الذكاء الاصطناعي.

**الجدول رقم 2 نتائج تحليل DI لأسئلة الاختبار للفصل الدراسي الأول**

| المفتاح | رقم السؤال | نسبة الإجابة الصحيحة(%) | معامل التمييز | التقييم بالرمز | التوصية |
|---|---|---|---|---|---|
| A | Q1 | 96.40% | 0.05 | ⚠ | سؤال سهل جدًا مع تمييز منخفض. يوصى بزيادة الصعوبة أو تحسين الاتساق مع نواتج التعلم |
| A | Q4 | 14.30% | 0.163 | ✗ | سؤال صعب جدًا مع تمييز منخفض. يوصى بمراجعة صياغة السؤال وتحسين الوضوح |
| A | Q5 | 53.60% | 0.382 | ✓ | سؤال بصعوبة متوسطة وتمييز عالٍ. سؤال فعال، يُبقي كما هو |
| A | Q10 | 89.30% | -0.138 | ✗ | سؤال سهل مع تمييز سلبي. يوصى بمراجعة صياغة السؤال لرفع فعاليته |
| A | Q12 | 46.40% | 0.607 | ✓ | سؤال بصعوبة متوسطة وتمييز عالٍ. سؤال قوي، يُبقي كما هو |
| B | Q2 | 65% | 0.668 | ✓ | سؤال بصعوبة متوسطة وتمييز عالٍ. سؤال فعال، يُبقي كما هو |
| B | Q5 | 60% | 0.729 | ✓ | سؤال بصعوبة متوسطة وتمييز عالٍ. سؤال فعال، يُبقي كما هو |
| B | Q6 | 95% | 0.07 | ⚠ | سؤال سهل جدًا مع تمييز منخفض. يوصى بزيادة مستوى التفكير أو تحسين الصياغة |
| B | Q10 | 75% | 0.448 | ✓ | سؤال سهل مع تمييز عالٍ. سؤال فعال، يُبقي كما هو |
| B | Q18 | 85% | -0.003 | ✗ | سؤال سهل مع تمييز سلبي. يوصى بمراجعة السؤال لاحتمالية وجود غموض أو أخطاء |
| C | Q4 | 35.50% | 0.373 | ⚠ | سؤال صعب مع تمييز متوسط. يوصى بضمان وضوح الصياغة والاتساق مع نواتج التعلم |

| | | | | | |
|---|---|---|---|---|---|
| سؤال بصعوبة متوسطة وتمييز متوسط. سؤال فعال، يُبقي كما هو مع مراقبة لاحقة | ⚠ | 0.279 | 80.60% | Q6 | C |
| سؤال سهل مع تمييز منخفض. يوصى بمراجعة صياغة السؤال لزيادة فعاليته | ⚠ | 0.133 | 74.20% | Q10 | C |
| سؤال سهل جدًا مع تمييز سلبي. يحتاج إلى إعادة صياغة وزيادة في الصعوبة | ✗ | -0.006 | 96.80% | Q16 | C |
| سؤال سهل مع تمييز عالٍ. سؤال قوي، يُبقي كما هو | ✓ | 0.497 | 90.30% | Q18 | C |

**نتائج المرحلة الرابعة:**

تم إعادة صياغة الأسئلة ذات مؤشر تمييز ضعيف كما جاء في توصيات نموذج الذكاء الاصطناعي، وهي كما يلي: **مفتاح :A**

١. Q1، تم تحسين المشتتات في السؤال
٢. Q4، تم تعديل صياغة جذر السؤال
٣. Q10، تم تعديل السؤال بتعمق أكبر في جذر السؤال

**مفتاح: B**

١. Q10، سؤال مشترك مع نموذج A بذات التعديل
٢. Q18، السؤال لم يتم تعديله لأنه مشترك مع مفتاح C وحصل على فقيمة مؤشر تمييز عالية جدول رقم (٣).

**مفتاح: C**

١. Q4، تم تعديل صياغة جذر السؤال (سؤال مشترك مع مفتاح A)
٢. Q10، سؤال مشترك مع نموذج A&B بذات التعديل
٣. Q16، تم تغيير السؤال في جميع النماذج، حيث أن في المفتاحين A&B كان مؤشر التمييز منخفض
٤. Q18، السؤال لم يتم تعديله لأنه مشترك مع مفتاح C وحصل على فقيمة مؤشر تمييز عالية جدول رقم (٣).

في هذه المرحلة تم تنفيذ الاختبار على دفعة **الفصل الدراسي الثاني ٢٠٢٥** وتم استخدام تلقين الأمر بالطريقة الثانية حيث **تم رفع ملف واحد بنتائج تحليل البيانات للمفاتيح الثلاثة للاختبار.** تم استخدام الأمر التالي على ذات المشروع FEMGPT المنجز بالفصل الأول " تم الأخذ بالملاحظات أعلاه وتحسين الأسئلة، مرفق نتائج تحليل البيانات لاستكمال نتائج التحسين. قم بعرض النتائج ومن ثم مقارنتها مع الفصل السابق مع تلخيص أهم التحسينات المنجزة". كانت استجابة FEMGPT كالتالي:

رائع جدًا🖐 !
كونك قد قمت **بتحسين الأسئلة بناءً على الملاحظات أعلاه**، فهذه خطوة مهمة جدًا في رفع جودة الاختبار وضمان دقته في قياس نواتج التعلم المطلوبة.
**الخطوة التالية التي يمكننا العمل عليها معًا:**
١. **إعادة تحليل الأسئلة بعد التحسين:**
- إذا توفر لديك **تقرير تحليل البنود الجديد** (بعد تعديل الأسئلة)، يمكنني:
  - حساب **مستوى الصعوبة الجديد.**
  - قياس **معامل التمييز.**

- مقارنة التحسينات بنتائج النسخة السابقة.
- تحديث **الخطة العلاجية**.

٢. **تحديث خطة العمل بناءً على التحسينات:**

- يمكنني مساعدتك في:
- إعداد **تقرير متابعة جودة الأسئلة**.
- تقديم **نقاط القوة بعد التحسين**.
- اقتراح **مزيد من التحسينات إن لزم الأمر**.

٣. **تحليل مدى تحقيق نواتج التعلم:CLOs**

- بناءً على الأسئلة الجديدة وتحليلها.

**مناقشة توصيات Full Exam Management GPT**

يظهر جدول رقم (٤) نتائج التحليل العامة للمقارنة بين الاختبارين قبل وبعد تحسين الأسئلة. نجد أن مؤشر التمييز قد تحسن في معظم الأسئلة بما يدل على التحسينات الإيجابية للاختبار. حيث وصلت قيمة DI في المفتاح C الى (٠,٨٧٨) وتعد الأعلى مقارنة بالمفتاح A بقيمة (٠,٨٣٢) الذي أظهر تحسن واضح بعد التعديل. وبالرغم من ذلك قد أظهرت نتائج مفتاح C وجود أسئلة ضعيفة. بينما المفتاح B كانت قيمة DI متقاربة بـ (٠,٧٩٤) مع مؤشر التمييز للمفتاح A، ولم يظهر مستوى سلبي مقارنة (A, C).

**الجدول رقم 3 تحليل عام (مقارنة بين نتائج الاختبارين)**

| المفتاح | نطاق مستوى الصعوبة | نطاق مؤشر التمييز DI | ملاحظات |
|---|---|---|---|
| A | ١٤,٣٪ إلى ١٠٠٪ | -٠,٧٠٦ إلى ٠,٨٣٢ | تحسن واضح في عدة أسئلة من حيث التمييز |
| B | ١٤,٣٪ إلى ١٠٠٪ | ٠,٤٠٦ إلى ٠,٧٩٤ | معظم الأسئلة أصبحت ذات تمييز جيد جدًا |
| C | ٢٥٪ إلى ١٠٠٪ | -٠,٨٧٨ إلى ٠,٨٧٨ | يوجد تحسن، لكن ما زالت بعض الأسئلة تُظهر تمييزًا سلبيًا |

يظهر جدول رقم (٥) مقارنة تحليل مؤشر التمييز DI لاختبار الفصل الدراسي الثاني ٢٠٢٥ مع مقارنة لمعامل التمييز للفصل الدراسي الأول لمفاتيح (A, B, C)، وذلك بعد العمل بالتحسينات والتوصيات المقترحة بالاختبار السابق. يوضح الجدول أن هناك تباين كبير لمؤشر التمييز في Q4، حيث إن DI في المفتاح (A) أظهر مستوي منخفض جداً بـ (-0.706) وجاءت التوصية بتعديل السؤال مرة اخرى. بينما في مفتاح DI (C) ارتفع بشكل كبير بقيمة (0.878). قد يرجع ذلك الى حجم العينة أو لاستجابة الطلاب للاختبار بشكل مختلف. وللتأكد من جودة السؤال يمكن الإبقاء عليه مع عمل تحسينات وقياسه في الاختبار التالي. من جانب آخر نجد أن Q10 قد أعطي نتيجة متباينة في المفاتيح الثلاثة مع أن السؤال مشترك بينهم، في مفتاح A التوصية إيجابية بتحسن كبير والاحتفاظ بالسؤال بالصيغة المحسن، حيث كانت قيمة DI (-0.138) وارتفعت الى (0.385) ومازالت في المفتاح B مرتفعة بـ (0.263). بينما في المفتاح C لم تذكر القيمة وكانت التوصية بالتحسين. من منطلق أخر، كانت التوصية باستبعاد Q19 بالمفتاح B رغم أن السؤال في اختبار الفصل الأول كان DI بمستوى مرتفع ولم يكن هنا أي توصيات من FEMGPT. وعليه يمكن أيضاً أن يتم تجربة استخدام السؤال في اختبار اخر للتأكد من جودته. كما يوضح الجدول توصيات إيجابية على باقي الأسئلة ويقارن مستوى التحسن في أسئلة الاختبار.

**الجدول رقم 4 نتائج تحليل DI مع مقارنة بين الاختبارين**

| التوصية | مستوى التحسين | معامل التمييز بعد التحسين | معامل التمييز السابق | رقم السؤال | المفتاح |
|---|---|---|---|---|---|
| زيادة الصعوبة أو ضبط الاتساق مع نواتج التعلم | ⚠ تراجع طفيف | -0.026 | 0.05 | Q1 | A |
| إعادة صياغة السؤال أو استبعاده | ❌ تدهور كبير | -0.706 | 0.163 | Q4 | A |
| الاحتفاظ بالسؤال كما هو | ✅ تحسن ممتاز | 0.810 | 0.382 | Q5 | A |
| الاحتفاظ بالسؤال كنموذج | ✅ حسن جذري | 0.385 | -0.138 | Q10 | A |
| الاحتفاظ بالسؤال كما هو | ✅ حسن إضافي | 0.832 | 0.607 | Q12 | A |
| الاستفادة منه كنموذج | ✅ سؤال جديد متميز | 0.702 | — | Q15 | A |
| لا يزال مقبولًا – يُحتفظ به | ⚠ تراجع طفيف | 0.407 | 0.668 | Q2 | B |
| يُستخدم مع الحذر | ⚠ تراجع بسيط | 0.430 | 0.729 | Q5 | B |
| يوصى بمراجعة عمق السؤال | ⚠ لم يتحسن | ≈0.00 | 0.07 | Q6 | B |
| قد يحتاج تعديلًا طفيفًا | ⚠ أقل أداءً | 0.263 | 0.448 | Q10 | B |
| الاحتفاظ به كمثال ناجح | ✅ إضافة ممتازة | 0.794 | — | Q12 | B |
| ضرورة مراجعته أو حذفه | ❌ تدهور واضح | -0.406 | -0.003 | Q19 | B |
| الاحتفاظ به كمثال مميز | ✅ تحسن كبير جدًا | 0.878 | 0.373 | Q4 | C |
| يُفضل تحسينه | ⚠ غير محسّن | غير مذكور | 0.133 | Q10 | C |
| إعادة صياغته بالكامل ضرورية | ❌ تراجع شديد | -0.878 | -0.006 | Q16 | C |
| يُستخدم كنموذج قوي | ✅ تحسن ممتاز | 0.845 | 0.335 | Q15 | C |
| يُعتمد كنموذج للاختبارات القادمة | ✅ سؤال جديد متميز | 0.845 | — | Q19 | C |

**الملخص:**

تطبيقات الذكاء الاصطناعي التوليدي تساهم بشكل فعّال في تجويد أساليب التقييم للمقرر الدراسي، مما يضمن اعداد أسئلة اختبار ذات جودة وكفاءة عالية تحقق مخرجات المقرر. استعرض البحث دراسة حالة لمقرر (مبادئ علم الألياف) - ببرنامج تكنولوجيا الأزياء والنسيج - بقسم الأزياء والنسيج- بجامعة الملك عبد العزيز. وطبقت الدراسة نموذج FEMGPT المعد بواسطة الذكاء الاصطناعي التوليدي على المخرجات المعرفية فقط، بهدف قياس مؤشر التمييز DI لأسئلة الاختبار التحريري (MCQ)، وذلك من خلال أربع مراحل تسلسلية لدراسة الحالة الفردية تشمل الفصل الدراسي الأول والثاني للعام الجامعي ٢٠٢٥. حيث تم تلقين الأوامر لنموذج FEMGPT في كل مرحلة لأداء مهمة محدد، بداية من اطلاع النموذج على توصيف المقرر ومن ثم انشاء مصفوفة للاختبار بتوزيع نسبة الدرجة وتحديد مستوى صعوبة الأسئلة ونوعها بما يتوافق مع مخرجات المقرر. كما تم تلقين الأمر بتوليد أسئلة مقترحة تواءم بين المحتوي والمخرج التعليمي، وتؤكد الدراسة ضرورة المراجعة للأسئلة من قبل المعلم وعدم الاعتماد الكامل في صياغة السؤال على AI بسبب الهلوسة التي قد تظهر وتؤدي الى ضعف في جذر ومشتتات السؤال.

إتماماً لذلك، تظهر نتائج الدراسة قدرة FEMGPT على تحليل نتائج بيانات الاختبار المستخلصة من خلال التصحيح الالكتروني، وتقديم توصيات تحسين لأسئلة الاختبار ذات مؤشر التمييز المنخفض. كما أثبت تحليل المقارن بين اختبار الفصل الأول والثاني أن العمل وفق توصيات AI ساهمت في ارتفاع مؤشر التمييز DI

لمعظم الأسئلة. كما أن الدراسة تحث على استخدام النموذج في مشاريع متعددة للاختبارات سواء النصفية أو النهائية لكل مقرر، ومتابعة العمل على المشروع ذاته في كل فصل دراسي لتتبع عملية تجويد أسئلة الاختبارات واستبقاء الأفضل لبنك الأسئلة.

**الشكر والتقدير:**

أتقد بجزيل الشكر والعرفان لمركز تطوير التعليم الجامعي – بجامعة الملك عبد العزيز، والذي قام بإنشاء نموذج Full Exam Management GPT باستخدام الذكاء الاصطناعي التوليدي (Chat GPT). من خلال برنامج زمالة تطبيقات الذكاء الاصطناعي في التعليم والبحث العلمي.

**قائمة المراجع:**

الهيئة الوطنية للاعتماد والتقويم الأكاديمي. (2024). *معايير وإرشادات الاعتماد*. استرجع من https://www.ncaaa.edu.sa/

عبد الصمد، أ. الس. م.، &أحمد، ك. م. م. (2020). *تطبيقات الذكاء الاصطناعي ومستقبل تكنولوجيا التعليم = Artificial Intelligence Applications and Future of the Instructional Technology*. القاهرة: المجموعة العربية للتدريب والنشر.

قندلجي، عامر. (2019). *منهجية البحث العلمي* (كتاب إلكتروني). مجموعة اليازوري للنشر والتوزيع.

**Bibliography:**

Abdellatif, H., Alsemeh, A. E., Khamis, T., & Boulassel, M.-R. (2024). Exam blueprinting as a tool to overcome principal validity threats: A scoping review. *Educación Médica, 25*, 100906. https://doi.org/10.1016/j.edumed.2024.100906

Chiavaroli, N., & Familari, M. (2011). When majority doesn't rule: The use of discrimination indices to improve the quality of MCQs. *Bioscience Education, 17*(1), 1–7. https://doi.org/10.3108/beej.17.8

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, 100118. https://doi.org/10.1016/j.caeai.2022.100118

Dela Peña, R. A., Musa, L. G., Soco, L., Sta. Maria, M. B., Politico, V., Avelino, D., Cruz, V. A., Martin, E., Guno, E., De Guzman, M., & Dimitui, E. A. (2011). Post-hoc analysis of departmental final examinations for selected general education subjects for Arellano University, AY 2010–2011. Arellano University, Research and Publications Department. https://www.researchgate.net/publication/316491151

Matoba, B. (2020, August 5). Discriminatory item analysis: A test administrator's best friend. *EMS1*.

https://www.ems1.com/ems-products/education/articles/discriminatory-item-analysis-a-test-administrators-best-friend

Rivera-Rosas, C. N., Calleja-López, J. R. T., Ruibal-Tavares, E., Villanueva-Neri, A., Flores-Felix, C. M., & Trujillo-López, S. (2024). Exploring the potential of ChatGPT to

create multiple-choice question exams. *Educación Médica, 25*, 100930. https://doi.org/10.1016/j.edumed.2024.100930

Santos, S. C. dos, & Junior, G. A. S. (2024). Opportunities and challenges of AI to support student assessment in computing education: A systematic literature review. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), Volume 2* (pp. 15–26). SCITEPRESS – Science and Technology Publications. https://doi.org/10.5220/0012552500003693

Saravia, E. (2024). *Prompt engineering guide*. DAIR.AI. https://www.promptingguide.ai/

Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. F. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology, 40*(6), 37–55. https://doi.org/10.14742/ajet.14145