مجلة جامعة تبوك للعلوم الإنسانية والاجتماعية
University of Tabuk Journal for Humanities and Social Sciences
دورية علمية محكمة
ردمد ١٦٥٨-٨٧٩٧
١٤٢٧ 2006 جامعة تبوك University of Tabuk

دقة التقييم والأداء التفاضلي للمقيم الآلي في النماذج اللغوية الكبيرة (LLMs) لتقييم مقالات الطلبة القصيرة في أوضاع مختلفة: دراسة محاكاة

Scoring Accuracy and Differential Rater Functioning of Large Language Models (LLMs) in Evaluating Short Student Essays under Different Conditions: A Simulation Study

ماجد محمود الجوده

أستاذ مشارك، القياس والتقويم، التربية وعلم النفس، التربية والآداب، جامعة تبوك

Majed Mahmoud Ajoudeh

Associate Professor, Evaluation and Measurement, Education and Psychology, College of ,Education and Arts, University of Tabuk

m_aljodeh@ut.edu.sa

https://orcid.org/0009-0003-1530-930X

الملخص: سعت هذه الدراسة إلى استكشاف الدقة والأداء التفاضلي للمقيم الآلي في النماذج اللغوية الكبيرة في تقييم المقالات القصيرة، بالاعتماد على بيانات مولدة اصطناعيًا شملت (270) مقالة قصيرة متفاوتة الجودة. قُدِّمت المقالات للمقيم الآلي باستخدام النماذج اللغوية الكبيرة في ثلاثة أوضاع من التوجيه: من دون استخدام مصفوفة تصحيح (Rubric)، مصفوفة تصحيح مختصرة، ومصفوفة تصحيح مفصلة. كما تم توليد تقييمات بشرية معيارية لمقارنتها بالمقيم الآلي. أظهرت النتائج أن استخدام مصفوفة التصحيح المفصلة يزيد من الدقة والتوافق مع المقيمين البشريين. وكشف تحليل الانحدار عن أداء تفاضلي للمقيم الآلي؛ منها ميل المقيم الآلي لمكافأة المقالات الطويلة، ومحاباة لأسلوب الكتابة غير المباشر، وكذلك أداء تفاضلي تبعًا للمجموعة التي ينتمي إليها الطالب. وتؤكد هذه النتائج أن النماذج اللغوية الكبيرة تمثل أداة واعدة في مجال التقييم التربوي، شريطة توظيفها في إطار معايير واضحة وضوابط تقلل من الأداء التفاضلي، مع ضرورة دمجها بالتقييم البشري في السياقات التربوية ذات الخطورة العالية والحساسة.

الكلمات المفتاحية: النماذج اللغوية الكبيرة؛ التقييم الآلي؛ مصفوفة التصحيح؛ الأداء التفاضلي للمقيم؛ العدالة في التقييم؛ الذكاء الاصطناعي في التعليم.

Abstract: This study aimed to evaluate the accuracy and differential rater functioning (DRF) of LLMs in scoring short essays, based on synthetically generated data of 270 essays with varying quality. The LLMs scored the essays under three prompting conditions: without a rubric, with a brief rubric, and with a detailed rubric. Human benchmark ratings were also generated in comparison with the automated scores. The results showed that using detailed rubric improved accuracy and consistency with human raters. Regression analysis further revealed differential rater functioning: the LLMs were more likely to award higher scores to longer essays, favored indirect writing styles, and displayed different outcomes depending on student group membership. These findings suggest that LLMs could serve as useful assessment tools in educational settings, provided they are used within well-defined frameworks and safeguards that reduce DRF while maintaining human judgment in high-stakes or sensitive situations.

Keywords: Large Language Models LLMs); Automated Scoring; Rubric; Differential Rater Functioning (DRF); Fairness in Assessment; Artificial Intelligence in Education

المقدمة:

يعتبر قطاع التعليم من أكثر قطاعات الحياة تأثراً بالتطورات الحديثة والتحديات الراهنة، حيث يتحمل مسؤولية إعداد أجيال تتسلح بالمعرفة والمهارات اللازمة للتعامل مع هذه التطورات ومواجهة التحديات. لذلك، كان من الضروري توفير بيئات تعلمية مرنة وملائمة لمواجهة التطورات التعليمية، وعلى رأسها البيئات التي تتعامل مع الذكاء الاصطناعي (Artificial Intelligence - AI).

لقد شهد توظيف الذكاء الاصطناعي (AI) في التعليم نقاشًا واسعًا، خصوصًا فيما يتعلق بمزاياه المحتملة في التدريس والتقويم في المدارس والجامعات. وقد أثارت التطبيقات المختلفة للذكاء الاصطناعي نقاشات كثيرة حول آثارها على حياة الإنسان ومستقبل المجتمعات. إذ يسعى الكثيرون إلى فهم القدرات التي يمكن أن يطورها الذكاء الاصطناعي، والفرص التي سيوفرها، وتأثيره في القطاعات والمهن المختلفة.

ويبدو أن تطبيقات الذكاء الاصطناعي ستُصبح من أبرز قضايا تكنولوجيا التعليم على مستوى العالم خلال العشرين عامًا القادمة؛ حيث ستعتمد الأدوات والخدمات والتطبيقات على الذكاء الاصطناعي بما تمتلكه من إمكانات وقدرات عالية لدعم العملية التعليمية وإحداث تحولات جذرية في مسارها (Zawacki-Richter et al., 2019).

ومن التطبيقات التفاعلية ذات الصلة بالعملية التعليمية، والتي تُعد محور هذه الدراسة، الروبوتات الحوارية الذكية (Intelligent Chatbots)، وهي برامج متقدمة مصممة لمحاكاة المحادثة البشرية، بما يتيح التفاعل بين النظام والمتعلم عبر الصوت أو النص أو كليهما. ويمكن أن تتخذ هذه الروبوتات أشكالًا مختلفة مثل تطبيقات الأجهزة الذكية، أو مواقع الإنترنت، أو عبر منصات الهاتف والرسائل. ويستطيع الطلبة التفاعل معها بطرح أسئلة مرتبطة بموضوعات أو مجالات دراسية محددة.

ويمكن توظيف مثل هذه التطبيقات في العديد من مكونات العملية التعليمية حيث تعدُّ عملية التقييم جزءًا أساسيًا من العملية التعليمية بشكل عام، إذ توفر أدلة يمكن الاعتماد عليها لتحسين عمليتي التعليم والتعلم. فهي تزود المعلم بمعلومات تساعده على التدريس بشكل أفضل، وتمنح الطلبة تغذية راجعة تساعدهم على التعلم بفاعلية أكبر (Zawacki-Richter et al., 2019). وتتعدد ممارسات التقييم من التقييمات غير الرسمية إلى الاختبارات عالية المخاطر، ويمكن تصنيفها وفق معايير مختلفة مثل: توقيت تطبيقها، الجهة المشرفة عليها، آلية الحصول على النتائج، وأهدافها (Yang et al., 2021).

وقد أبرزت عدة دراسات الدور الفاعل للذكاء الاصطناعي في عمليتي التعليم والتعلّم (Vazquez et al., 2021؛ Pereira et al., 2019؛ Neto & Fernandes, 2019). وكذلك توصلت بعض الدراسات إلى نتائج إيجابية في توليد تقييمات جيدة للطلبة، مثل الاختبارات (Aljodeh, 2025).

النماذج اللغوية الكبيرة Large Language Models (LLMs)

ومن ضمن أدوات الذكاء الاصطناعي المستخدمة في حقل تعليم وتقييم الطلبة هي ما يُطلق عليه اسم النماذج اللغوية الكبيرة (Large Language Models أو LLMs)، وهي أنظمة ذكاء اصطناعي مبرمجة على كميات ضخمة من البيانات النصية من أجل تعلم الأنماط الإحصائية للغة الطبيعية.

وتتميز النماذج اللغوية الكبيرة بقدرتها على توليد نصوص مترابطة تشبه النصوص البشرية، والإجابة عن الأسئلة، وإتمام الجمل، والترجمة، والتلخيص، وأداء مجموعة واسعة من مهام معالجة اللغة الطبيعية دون الحاجة إلى تدريب مخصص لكل مهمة. ويُنظر إليها اليوم كأحد أبرز التطورات في مجال الذكاء الاصطناعي نظرًا لقدراتها الناشئة في التعلّم من السياق (In-context Learning) والتعميم على مهام جديدة. (Zaho et al., 2023))

وقد اعتبر بعض الباحثين أن هذه التطورات تمثل نقطة تحول في تاريخ معالجة اللغة الطبيعية (Chernyavskiy, Ilvovsky, & Nakov, 2021). ورغم الإمكانات الكبيرة لـ LLMs، إلا أن الأبحاث الحديثة نبهت إلى ضرورة وضع آليات واضحة للتدقيق (Auditing) كآلية حوكمة تضمن أن تكون النماذج مصممة ومستخدمة بطريقة أخلاقية وقانونية وتقنية آمنة (Arcas, 2022)

عند مراجعة الأدبيات الحديثة حول استخدام النماذج اللغوية الكبيرة (LLMs) في التقييم التعليمي، يظهر اتّجاهٌ إيجابيٌّ غالب لجدواها في التقييم الكتابي متعدّد الأبعاد، مع توافقٍ ملحوظ مع أحكام المقيمين البشريين متى أُحسن ضبط التنبيهات ومعايير التشغيل؛ فقد أظهرت دراسات تطبيقية قدرة LLMs على تقدير أبعادٍ مثل جودة الأفكار والتنظيم والأسلوب، وتقديم تقديرات متّسقة نسبيًا مع البشر. (Tang et al., 2024; Li & Liu, 2024)

في المقابل، تُبرز بحوث العدالة والتحيّز ضرورة الحذر المنهجي، إذ وثّقت فروقًا غير محمودة بين الفئات السكانية في سياق التقييم الآلي للمقالات، وكذلك إشارات لتحيزات عند تصحيح الأجوبة القصيرة، بما يستلزم فحصًا ممنهجاً للإنصاف عبر المجموعات واعتماد ضوابط تخفيف التحيّز (Schaller et al., 2024;

(Andersen et al., 2025) . ونبهت بعض الدراسات إلى ضرورة فحص العدالة الاجتماعية والتحيّزات الكامنة في هذه النماذج. فقد قدّم (Gallegos et al, 2023) مراجعةً شاملة أوضحت أن النماذج اللغوية الكبيرة قد تتعلّم وتُعيد إنتاج التحيّزات الاجتماعية السائدة في بيانات التدريب، بما في ذلك التحيّزات القائمة على النوع الاجتماعي أو الخلفية الثقافية أو العرقية، ودعت إلى تبنّي استراتيجياتٍ متعددة للحدّ من هذه التحيّزات على مستويات ما قبل التدريب وأثناءه وبعده.

كما طوّر (Chen et al, 2025) إطارًا تجريبيًا لقياس التحيّز الجندري عبر أسلوب السرد المفتوح، ووجدوا أن النماذج اللغوية تميل إلى إفراط تمثيل النساء في مهنٍ معينة وتكريس أنماطٍ جندرية نمطية عند توليد القصص، ما يعكس استمرار التأثير الاجتماعي للبيانات التي بُنيت عليها.

وتشير هذه النتائج إلى أنّ التحيّز في مخرجات النماذج اللغوية ليس عارضًا، بل هو نتيجة مباشرة لطبيعة البيانات وأساليب الضبط المستخدمة، الأمر الذي يجعل من الضروري تطوير آليات تدقيقٍ ومعاييرَ إنصافٍ أكثر دقة لضمان حيادية أنظمة التقييم الآلي والتطبيقات التعليمية التي تعتمد على الذكاء الاصطناعي.

وبشكل عام فإن مجمل الدراسات في هذا الصدد نجد أنها غالباً ما تناولت بعداً واحداً فقط من التقييم (مثل الطول أو أسلوب الكتابة)، دون تصميم مقارنات منهجية شاملة بين أوضاع مختلفة للتوجيهات التي تقدم للتقييم الآلي في النماذج اللغوية الأخرى، ومن جانب أخر فإن الدراسات التي تناولت دراسة تأثير مصفوفات التصحيح (Rubrics) على عدالة التقييم الآلي وأدائه التفاضلي لا تزال محدودة، وغالباً ما تناولت المصفوفة بشكل عام دون التفصيل في مستوياتها في أوضاع مختلفة.

وعادة ما يتم تصحيح مقالات الطلبة القصيرة باستخدام مصفوفة التصحيح أو سلم التقدير (Rubric) والتي تتناول معايير مختلفة من التقييم مثل المحتوى، تنظيم الأفكار، اللغة والأسلوب، والملاءمة للموضوع، ويقوم المقيم بمقارنة مقال الطالب ضمن تلك المعايير في مستويات مختلفة للتقييم مثل (ممتاز، جيد جداً، جيد مقبول، ضعيف، ضعيف جداً) وتكميم المقال برقم يعكس جودته، مما يجعل التقييم أكثر عدالةً واتساقاً، فيما لو تم التقييم من دون وجود مثل تلك الإجراءات، ولزيادة دقة التقييمات قد تقيم المقالة من أكثر من مقيم، ويتم أخذ المتوسطات الحسابية للمقيمين (Brookhart, S. M, 2013)

وقد تتأثر التقييمات البشرية بعوامل مختلفة أو تنحاز إلى مجموعة معينة، مما قد يؤثر سلباً على عدالة تلك التقييمات، إلى جانب أن هذه الإجراءات طويلة قد ترهق المعلمين وأعضاء هيئة التدريس في العمل الأكاديمي، لذلك ومع التطور في تطبيقات الذكاء الاصطناعي قد تعمل بشكل فعلي على حل تلك المشكلات، إلا أن تلك التطبيقات لا تزال قيد الدراسة لإمكانية استغلالها في تحقيق تقييمات عادلة ومتسقة لتقييمات الطلبة.

ومن هنا وفي ضوء ما تقدم فإن الدراسة الحالية تأتي لتقديم نتائج عملية يمكن أن تسهم في تصميم سياسات تربوية تستفيد من إمكانات النماذج اللغوية الكبيرة مع الحد من مخاطرها، بما يحقق التوازن بين الكفاءة والعدالة في التقييم التربوي.

مشكلة الدراسة، وأهميتها:

في ظل التقدم التكنولوجي الحاصل على تطبيقات الذكاء الاصطناعي، وتزايد استخدامها في حقل التعليم، ومن أبرزها استخدام النماذج اللغوية الكبيرة في تقييم مقالات الطلبة نظراً لقدرتها الفائقة على تقديم أحكام سريعة، وبأقل التكاليف من الوقت والجهد فيما لو تم تقييمها باستخدام المقيمين البشريين (Zaho et al., 2023) ورغم الإمكانات الفائقة لهذه النماذج، إلا أنه يثار حولها بعض القضايا المتعلقة في العدالة والتحيز لمجموعة معينة تبعاً لعوامل مختلفة. تستقصي هذه الدراسة فاعلية النماذج اللغوية الكبيرة في إنتاج تقويمٍ منصفٍ ودقيقٍ للكتابة الأكاديمية القصيرة ضمن سياقات تعليمية. وبالتحديد فإن هذه الدراسة ستجيب عن مجموعة من الأسئلة هي:

1. ما مدى دقة تقييمات المقيم الآلي في النماذج اللغوية الكبيرة في أوضاع مختلفة من التوجيه (مصفوفة تصحيح مفصلة، مصفوفة تصحيح مختصرة، عدم وجود مصفوفة تصحيح)؟
2. هل يظهر المقيم الآلي في النماذج اللغوية الكبيرة أداءً تفاضلي Differential Rater (DRF) Functioning مرتبطة بعوامل غير بنائية مثل طول الفقرة، الأسلوب الكتابي، المجموعة التي ينتمي إليها الطالب في التقييمات؟

الأهمية النظرية للدراسة:

ومن المتوقع أن هذه الدراسة ستسهم في إثراء أدبيات القياس والتقويم فيما يتعلق باستخدام التقييم الآلي وتطبيقات الذكاء الاصطناعي في تقييم الطلبة، وكذلك إبراز دور التوجيهات المقدمة للنماذج اللغوية

الكبيرة في تقليل التحيز وتحقيق عدالة التقييمات، وبالتالي إضافة عاملاً جديداً يثير النقاش العلمي حول إمكانية استخدام هذه التقنيات.

الأهمية التطبيقية للدراسة:

وتبرز الأهمية التطبيقية لنتائج هذه الدراسة من خلال مساعدة المعلمين وأعضاء هيئة التدريس في الجامعات في كيفية استخدام تطبيقات الذكاء الاصطناعي في عمليات تقييم الطلبة، وخصوصاً تلك التي تستنزف الوقت والجهد لدي المقيمين، إلى جانب أن هذه الدراسة تستخدم بيانات محاكاة اصطناعية لتقليل القيود الأخلاقية المرتبطة بجمع البيانات من الطلبة.

هدف الدراسة:

تهدف الدراسة الحالية إلى تقصي وتحليل دقة النماذج اللغوية الكبيرة في تقييم المقالات الأكاديمية القصيرة مقارنة بالتقييم البشري ضمن مستويات مختلفة من التوجيهات المقدمة للمقيم الآلي (وجود مصفوفة تصحيح مفصلة، مصفوفة تصحيح مختصرة، من دون وجود مصفوفة تصحيح)، والكشف عن الأداء التفاضلي لها تبعاً لمجموعة من العوامل المتعلقة بطول المقالة، وأسلوبها، وتبعاً للمجموعة التي ينتمي لها الطالب (ذكور، إناث).

الطريقة والإجراءات:

عينة الدراسة:

أُستخدم الذكاء الاصطناعي في هذه الدراسة كأداة منهجية وليس بديلاً عن الباحث، حيث تم توليد المقالات الافتراضية باستخدام ChatGPT-4 لتجاوز القيود الأخلاقية المرتبطة بجمع بيانات الطلبة، كما تم توظيف GPT-4 لإجراء التقييمات الآلية وفق أوضاع مختلفة لمصفوفات التصحيح. وقد اقتصر دور الذكاء الاصطناعي على محاكاة البيانات والإجراءات، بينما تولّى الباحث مسؤولية التصميم التجريبي والتحليل الإحصائي وتفسير النتائج، بما ينسجم مع الاستخدام المسؤول والشفاف للذكاء الاصطناعي في البحوث التربوية.

والتطبيق منشور على الرابط الآتي: www.ChatGPT.com

حيث طلب منه توليد 270 مقالة أكاديمية قصيرة، بحيث طُلب من كل طالب افتراضي كتابة فقرة قصيرة (5–7 جمل) يعبّر فيها عن رأيه في أهمية التعلم التعاوني في تحسين التحصيل الدراسي، وهذه تعتبر مهمة تربوية

بسيطة لا تحتاج إلى متخصص لكتابتها، وبالتالي فهي مناسبة لمستويات مختلفة من الطلبة، وتم وضع خصائص للعينة بحيث تراعي مستويات مختلفة من الجودة والطول وأسلوب الكتابة، والمجموعة التي ينتمي إليها الطالب، وقد توزعت العينة بشكل متوازن على هذه المستويات، والجدول رقم 1 يبين خصائص العينة المولدة تبعا لمستويات مختلفة من الجودة، الطول، والأسلوب، موزعة بين مجموعتين (ذكور/إناث).

الجدول رقم 1: خصائص العينة المولدة تبعا لمستويات مختلفة من الجودة، الطول، والأسلوب، موزعة بين مجموعتين (ذكور/إناث).

الخاصية	المستوى	العدد
الجودة	منخفض	90
	متوسط	90
	مرتفع	90
	المجموع	270
الطول	قصير	90
	متوسط	90
	طويل	90
	المجموع	270
الأسلوب	مباشر	90
	بلاغي	90
	أكاديمي	90
	المجموع	270
المجموعة	A ذكور	135
	B إناث	135
	المجموع	270

تم تطبيق المقالات على النموذج اللغوي الكبير GPT-4 (OpenAI, 2023) عبر منصة ChatGPT لإنتاج التقييمات الآلية في أوضاع التوجيه الثلاثة (وجود مصفوفة تصحيح مفصلة، مصفوفة تصحيح مختصرة، من دون وجود مصفوفة تصحيح).

وعبر نفس المنصة تمت محاكاة ثلاثة مقيمين بشريين (Human Raters) ليمثلوا الخبرة التربوية في تقييم المقالات الأكاديمية القصيرة، وتم اختيار العدد 3 وذلك لغايات تقليل التحيز وتحقيق العدالة، وطلب من التطبيق بأن يتم توليد الدرجات التي يقدمها المقيمون البشريون تبعاً لجودة الفقرة، بحيث تحصل المقالة على درجة تتراوح بين 1 إلى 2 إذا كانت ذات جودة منخفضة، وبين 3 إلى 4 إذا كانت جودة الفقرة متوسطة، وبين 5 إلى 6 إذا كانت جودة الفقرة مرتفعة، بحيث تغطي الدرجات الفترة بين أقل درجة هي 1 وأعلى درجة هي 6، ومن ثم حسب الوسط الحسابي لكل طالب من المقيمين البشريين الثلاثة لتمثل الدرجة المرجعية للطالب والتي يتم مقارنتها مع درجة المقيم الآلي في النماذج اللغوية الكبيرة.

تُحاكي الإجراءات المعتمدة في هذه الدراسة الممارسات الشائعة في البحوث التربوية التي تتطلّب تقويماً متعددَ المصادر؛ إذ يُتوقَّع أن ينعكس ذلك إيجاباً على صدق التقديرات وموثوقيتها عندما تُستخدم كمقياسٍ مرجعي داخل الدراسة (Attali & Burstein, 2006).

مستويات التوجيه المقدمة للمقيم الآلي في النماذج اللغوية الكبيرة:

تم تقديم ثلاثة مستويات للمقيم الآلي في النماذج اللغوية الكبيرة، والتي على أساسها تم توليد الدرجات عبر تلك المستويات، والجدول رقم 2 يوضح هذه المستويات.

الجدول رقم 2: مستويات التوجيه (بدون مصفوفة تصحيح No Rubric، مصفوفة تصحيح مختصرة Brief Rubric، مصفوفة تصحيح مفصلة Detailed Rubric) المقدمة للمقيم الآلي في النماذج اللغوية الكبيرة.

مستوى التوجيه	المعايير/ الوصف	الدرجة
بدون مصفوفة تصحيح No Rubric	تقييم عام من دون معايير واضحة ومحددة	من 0 إلى 6
مصفوفة تصحيح مختصرة Brief Rubric	تقييم إجمالي حسب وضوح الأفكار، وسلامة اللغة، والأسلوب، وملاءمة الموضوع	من 0 إلى 6 بشكل إجمالي
	المحتوى والأفكار	0: الفقرة لا تحتوي على محتوى مناسب أو غير مرتبطة بالموضوع.

		1: الأفكار سطحية أو غير مكتملة.
		2: الأفكار واضحة، عميقة، ومترابطة.
مصفوفة تصحيح مفصلة Detailed Rubric	التنظيم وتسلسل الأفكار	0: لا يوجد تسلسل أو ترابط.
		1: الأفكار منظمة بتسلسل منطقي.
	اللغة والأسلوب	0: أخطاء لغوية كثيرة تعيق الفهم.
		1: لغة مفهومة مع بعض الأخطاء.
		2: لغة صحيحة وواضحة ومناسبة.
	الإبداع والتعبير	0: أسلوب بسيط جدًا أو مكرر.
		1: أسلوب متنوع ومعبّر.

التحليلات الإحصائية ومؤشرات الدقة والأداء التفاضلي:

بعد الحصول على ملفات البيانات المولدة من منصة ChatGPT 4 تم إجراء التحليلات الإحصائية المناسبة باستخدام برامج التحليل الإحصائي SPSS وبرنامج R ولغايات التحقق من دقة التقييمات للمقيم الآلي في النماذج اللغوية الكبيرة، ومقارنتها مع متوسط تقييم المقيمين البشريين، فلقد تم استخدام المؤشرات الآتية:

1. متوسط الخطأ المطلق (Mean Absolute Error – MAE):

واستخدم هذا المؤشر لإيجاد الفروق المطلقة بين المقيمين البشريين والمقيم الآلي، ويتم حساب ذلك من خلال المعادلة الآتية:

$$\text{MAE}=|\hat{y}_i - y_i| \sum_{i=1}^{n} \frac{1}{n}$$

حيث إن الرموز المستخدمة لهذا المؤشر في سياق هذه الدراسة تعني ما يلي:

$\hat{y}_i$: القيمة المقدرة من المقيم الآلي للطالب رقم i

y_i : المتوسط الحسابي لتقييمات المقيمين البشريين للطالب رقم i

n: عدد الطلبة

وكلما كان هذا المتوسط قليلاً، كانت الدقة عالية لتقييمات المقيم الآلي مقارنة بالمقيم البشري.

(Willmott & Matsuura,2005)

2. الجذر التربيعي لمتوسط الخطأ (Root Mean Square Error – RMSE)

يقيس هذا المؤشر متوسط حجم الأخطاء بين القيم المتوقعة والقيم المرجعية، ويمثل مقياسًا لحجم الانحرافات عن القيم الحقيقية، بحيث يُعطي صورة واضحة عن مدى دقة النموذج في التقدير (Chai & Draxler, 2014). وفي هذه الدراسة، يساعد RMSE في الكشف عن مدى تقارب أو تباعد تقييمات النماذج اللغوية الكبيرة عن متوسطات تقييمات المقيمين البشريين للمقالات القصيرة (Chai & Draxler, 2014).

ويعطى هذا المؤشر من خلال المعادلة الآتية:

$$\text{RMSE}= \sqrt{\sum_{i=1}^{n} \frac{1}{n}(y_i - \hat{y}_i)^2}$$

حيث إن الرموز المستخدمة لهذا المؤشر في سياق هذه الدراسة تعني ما يلي:

y_i: المتوسط الحسابي لتقييمات المقيمين البشريين للطالب رقم i

$\hat{y}_i$: القيمة المقدرة من المقيم الآلي للطالب رقم i

n: عدد الطلبة

وكلما قلت قيمة هذا المؤشر دل على أن التقارب بين المقيم الآلي ومتوسط المقيمين البشريين أعلى، بمعنى اتفاق أكبر بينهما.

2. معامل كابا الموزون (Quadratic Weighted Kappa – QWK)

يستخدم هذا المؤشر لقياس درجة الاتفاق بين مقيمين أو أكثر عند تصنيف الأفراد على مقياس رتبي، ولتطبيق هذا المؤشر يتم بناء ثلاث مصفوفات، بحيث أن المصفوفة تمثل تقييمات المقيمين الملاحظة، والمصفوفة الثانية تمثل

التقييمات المتوقعة وفقاً للتوزيعات الاحتمالية، والمصفوفة الثالثة تمثل مصفوفة الأوزان وهي عبارة عن مربعات الفرق بين تقييم المقيم الأول والمقيم الثاني عند كل حالة، بحيث تأخذ الفروق الكبيرة وزناً أكبر من الفروق الكبيرة. ويستخدم هذا المؤشر على نطاق واسع في دراسات القياس النفسي، وتقويم الكتابة، وتصحيح الاختبارات، وتقييم النماذج التنبؤية، حيث يُعد مقياسًا أكثر دقة من مجرد معاملات الارتباط (Ben-David, 2008). ويعطى هذا المؤشر من خلال المعادلة الآتية:

$$QWK = 1 - \frac{\sum_{i,j} O_{i,j} w_{i,j}}{\sum_{i,j} E_{i,j} w_{i,j}}$$

حيث إن الرموز المستخدمة لهذا المؤشر في سياق هذه الدراسة تعني ما يلي:

$O_{i,j}$: المصفوفة الملاحظ للاتفاق بين المقيم الآلي في النماذج اللغوية الكبيرة والمقيمين البشريين.

$E_{i,j}$: مصفوفة التقييمات المتوقعة وفقاً للتوزيعات الاحتمالية.

$w_{i,j}$: الأوزان التربيعية التي تعكس حجم الفروق بين التقييمات.

وكلما كان هذا المؤشر قريباً من $+1$ كلما دل على اتفاق أعلى بين المقيم الآلي والمقيمين البشريين.

أما بالنسبة لدراسة الأداء التفاضلي للمقيم الآلي فقد تم استخدام تحليل الانحدار الخطي المتعدد (Multiple Linear Regression) للكشف عن العوامل المؤثرة في تقييمات المقيم الآلي في النماذج اللغوية الكبيرة (LLMs)، وبالتالي فإن الصيغة العامة لمعادلة الانحدار المتعدد في سياق هذه الدراسة على الشكل الآتي:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

حيث إن الرموز المستخدمة في سياق هذه الدراسة تمثل ما يلي:

Y: تمثل التقييم الصادر عن المقيم الآلي في النماذج اللغوية الكبيرة.

β_0: الحد الثابت (ثابت الانحدار)

β_1 ، β_2، β_3 : معاملات الانحدار للمتغيرات المستقلة: طول المقالة، أسلوب المقالة، المجموعة (ذكور، إناث)

X_1: المتغير المستقل الأول في الدراسة والذي يعكس طول مقالة الطالب.

X_2: المتغير المستقل الثاني، والذي يعكس أسلوب مقالة الطالب.

X_3: المتغير المستقل الثالث، والذي يعكس المجموعة التي ينتمي إليها الطالب (ذكور، إناث)

ε: الخطأ العشوائي.

ويساعد تحليل الانحدار على تفسير ما إذا كانت الفروق في التقديرات تعكس جودة الكتابة فعلًا، أم أنها ناجمة عن تحيزات مرتبطة بخصائص النصوص أو الطلبة (Field, 2018).

ثم تم تحليل البيانات المولدة باستخدام مجموعة الحزم الإحصائية SPSS وبرنامج R وسجلت النتائج.

نتائج الدراسة:

نتائج الدراسة المتعلقة بالسؤال الأول:

ويتعلق السؤال الأول بمدى دقة تقديرات المقيم الآلي مقارنة مع متوسط التقييمات البشرية في أوضاع مختلفة من التوجيه (مصفوفة تصحيح مفصلة Detailed Rubric، مصفوفة تصحيح مختصرةBrief Rubric ، عدم وجود مصفوفة تصحيح No Rubric)

تم إيجاد المتوسطات الحسابية والانحرافات المعيارية لمتوسطات تقييمات المقيمين البشريين وتقييمات المقيم الآلي في النماذج اللغوية الكبيرة في جميع مستويات التوجيه، وسجلت النتائج في الجدول رقم 3 الآتي.

الجدول رقم 3: المتوسطات الحسابية والانحرافات المعيارية لمتوسطات تقييمات المقيمين البشريين وتقييمات المقيم الآلي في النماذج اللغوية الكبيرة في جميع مستويات التوجيه.

المقيم	المتوسط	الانحراف المعياري	عدد المقالات (الطلبة)
متوسط تقييمات المقيمين البشريين	3.95	1.02	270
المقيم الآلي من دون مصفوفة تصحيح	3.72	1.20	270
المقيم الآلي بمصفوفة تصحيح مختصرة	3.88	1.10	270
المقيم الآلي بمصفوفة تصحيح مفصلة	3.97	0.95	270

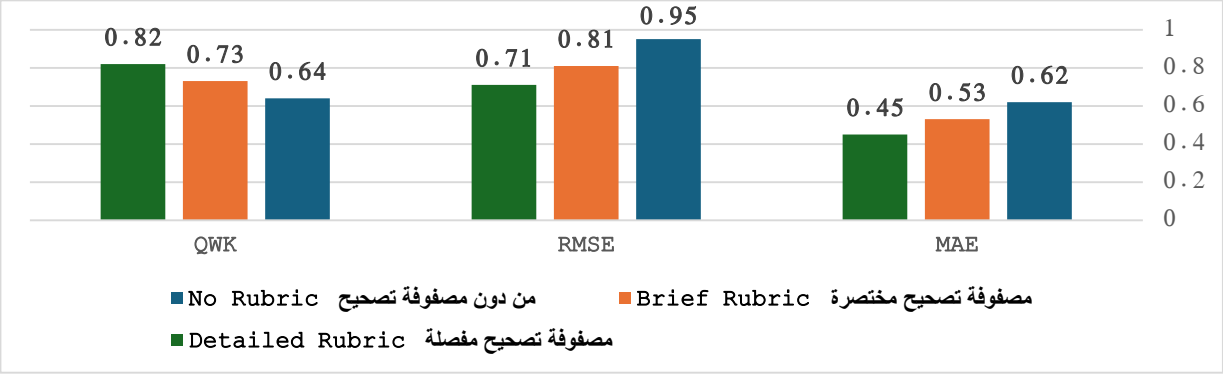
يلاحظ من الجدول رقم 3 أن متوسط تقييمات المقيم الآلي تقاربت من متوسطات التقييمات البشرية بدرجات متفاوتة، وكان أفضلها في الوضع الذي يكون فيه توجيه المقيم الآلي بمصفوفة تصحيح مفصلة، ويلاحظ من الانحرافات المعيارية وجود تفاوت في تقييمات المقيم الآلي في حالة عدم وجود مصفوفة تصحيح توجه هذا المقيم، لذلك يظهر أنه كلما كان التوجيه أكثر للمقيم الآلي كانت تقييمات المقيم الآلي قريبة من تقييمات البشر.

وتم استخراج المؤشرات الإحصائية المتعلقة بدقة تقييمات المقيم الآلي مقارنةً مع متوسطات تقييمات المقيمين البشريين وهي: MAE وRMSE وQWK في جميع مستويات التوجيه للمقيم الآلي، وسجلت النتائج في الجدول رقم 4.

الجدول رقم 4: مؤشرات دقة التقييمات للمقيم الآلي مقارنة مع متوسطات التقييمات البشرية في مستويات التوجيه المقدمة له.

QWK	RMSE	MAE	مستوى التوجيه
0.64	0.95	0.62	No Rubric من دون مصفوفة تصحيح
0.73	0.81	0.53	Brief Rubric مصفوفة تصحيح مختصرة
0.82	0.71	0.45	Detailed Rubric مصفوفة تصحيح مفصلة

من خلال ملاحظة قيم MAE وRMSE نجد أنها تقل كلما زاد مستوى التوجيه، فلقد وصلت إلى أقل قيمها عند وجود مصفوفة تصحيح مفصلة كتوجيه مقدم للمقيم الآلي، وهذا يعني أن الاتفاق بين المقيمين البشريين والمقيم الآلي أعلى في هذه الحالة، ومن جهة أخرى أن أفضل حالات مؤشر كابا الموزون QWK كانت 0.82 وهي قريبة من +1 عند وجود مصفوفة تصحيح مفصلة، وهذا يدعم أيضاً النتيجة التي توصلت إليها مؤشرات MAE وRMSE، والشكل رقم 1 يبين تمثيل بياني لهذه المؤشرات في الأوضاع المختلفة من التوجيه.



الشكل رقم 1: التمثيل البياني لمؤشرات الدقة في جميع أوضاع التوجيه

إن النتائج الظاهرة في الجدول رقم 4، والشكل رقم 2 تشير إلى أنه كلما كانت معايير التقييم المقدمة للمقيم الآلي واضحة ومفصلة، تكون تقييماته أقرب إلى تقييمات المحكمين البشريين.

1. نتائج الدراسة المتعلقة بالسؤال الثاني من الدراسة:

ويبحث السؤال الثاني في تقصي الأداء التفاضلي للمقيم الآلي Differential Rater Functioning (DRF) في أوضاع مختلفة لخصائص مقالات الطلبة (الطول، الأسلوب، المجموعة) مع ثبات مستوى جودة المقالات، حيث تم توليد مقالات مختلفة في الجودة محددة مسبقًا (منخفضة، متوسطة، عالية)، لذلك تم استخدام الانحدار الخطي المتعدد لتفسير التباين في درجات المقيم الآلي في النماذج اللغوية الكبيرة بناءً على خصائص المقالات (الطول، الأسلوب) وخصائص المجموعات (ذكور/إناث). فإذا ظهر أن معامل أحد المتغيرات المستقلة دال إحصائيًا، فهذا يشير إلى أن المقيم الآلي يمنح درجات أعلى أو أقل بسبب هذه الخاصية، حتى مع ثبات جودة الفقرة. وبذلك، يوفر الانحدار أداة موضوعية للكشف عن التفاضل الممنهج للمقيم الآلي والذي يؤثر سلباً على عدالة التقييم (Zumbo, 2007).

وأظهرت نتائج تحليل الانحدار، بعد أن تم إدخال متغيرات خصائص الفقرة كمتغيرات مستقلة حيث ادخل متغير الطول كمتغير منفصل (0= قصير، 1= متوسط، 2= طويل) ومتغير الأسلوب كمتغير منفصل ثنائي (0 = أسلوب مباشر، 1= أسلوب غير مباشر(بلاغي، أكاديمي)) ومتغير المجموعة أيضاً متغير ثنائي(0 = مجموعة الأساس ذكور، 1 = مجموعة الإناث) أن درجات المقيم الآلي تتأثر بعدد من الخصائص للمقالات القصيرة للطلبة، والجدول رقم 5 يوضح معاملات الانحدار المعيارية لخصائص المقالات القصيرة.

الجدول رقم 5: معاملات الانحدار المعيارية لخصائص المقالات القصيرة.

p	t	SE	β	المتغير المستقل
<0.001	64.2	0.06	3.85	الثابت (Intercept)
<0.01	4.00	0.08	0.32	الطول (Length)
<0.05	1.96	0.07	0.13	الأسلوب (Style)
<0.05	-2.57	0.07	-0.18	المجموعة (Group)

من الجدول رقم 5 يلاحظ أن في كل زيادة في فئة الطول ترتفع تقديرات المقيم الآلي بما نسبته 32%، لذلك هذا يظهر تفاضلاً واضحاً للمقيم الآلي لصالح المقالات الأطول على الرغم من ثبات جودة المقالة، ومن

خلال معامل الانحدار لمتغير الأسلوب يظهر أن المقيم الآلي يميل لصالح أسلوب المقالة غير المباشر، حيث حصلت في المتوسط على درجة أعلى بمقدار 0.13 نقطة مقارنة بالمقالات المباشرة رغم ثبات جودة المقالة أيضاً.

وبالنسبة لمتغير المجموعة فإن القيمة السالبة (–0.18) لمعامل الانحدار المعياري تعني أن المجموعة الإناث حصلت في المتوسط على درجات أقل بمقدار 0.18 نقطة مقارنة بالمجموعة المرجعية ذكور، بمعنى أن المقيم الآلي يميل إلى تقليل درجات مجموعة معينة كالإناث بشكل طفيف، حتى مع ثبات جودة المقالات الأكاديمية القصيرة للطلبة.

مناقشة النتائج والاستنتاجات والتوصيات:

فيما يتعلق بدقة تقييم المقيم الآلي في النماذج اللغوية الكبيرة، مقارنة مع متوسط تقييمات المقيمين البشريين عند مستويات التوجيه المختلفة، فإن النتائج التي ظهرت في الجداول رقم 3، و4 تشير إلى أداء المقيم الآلي دون مصفوفة تصحيح (No Rubric) الأقل دقة (MAE=0.62, RMSE=0.95, QWK=0.64)، بينما تحسنت الدقة مع وجود مصفوفة مختصرة، ووصلت إلى أفضل حالاتها عند استخدام مصفوفة مفصلة (MAE=0.45, RMSE=0.71, QWK=0.82)، وهذا يعني أن النماذج اللغوية الكبيرة قادرة على محاكاة التقييم متعدد الأبعاد عند وجود معايير واضحة وأن مصفوفات التصحيح المفصلة تقلل التباين وتحسن العدالة في التقييم، وهذا ما توصلت إليه بعض الدراسات مثل (Tang et al., 2024; Li & Liu, 2024; Brookhart, 2013)

وفيما يتعلق بنتائج الدراسة المعروضة في الجدول رقم 5 (جدول تحليل الانحدار) للكشف عن الأداء التفاضلي للمقيم DRF تبين أن هناك 3 نماذج من الأداء التفاضلي للمقيم:

أولاً: الطول، حيث تبين أن النماذج اللغوية الكبيرة تميل إلى مكافأة الفقرات الأطول حتى مع ثبات الجودة، وهذا يتفق مع دراسة (Schaller et al., 2024) التي وجدت أن أنظمة التصحيح الآلي المبنية على النماذج اللغوية الكبيرة تُظهر تحيزًا للطول، مما يؤثر سلباً على عدالة التقييم.

ثانياً: الأسلوب أظهرت نتائجنا وجود تفاضلٍ طفيفٍ لصالح الأسلوب غير المباشر (بلاغي/أكاديمي) مقارنةً بالمباشر، ما يشير إلى حساسية المقيم الآلي لخصائصٍ نصيّةٍ تتجاوز جودة المحتوى. وبخلاف ما توثّقه الأدبيات حول تحيّز الطول تحديدًا في أنظمة التقييم الآلي (Schaller et al., 2024)، لا تُجمِع الدراسات المنشورة على اتجاهٍ أسلوبيٍّ ثابت؛ وعليه تُعدّ هذه النتيجة إسهامًا تجريبيًا يستدعي فحصًا منهجيًا لاحقًا إلى جانب مؤشّرات العدالة الأوسع التي أبرزتها المراجعات الحديثة (Gallegos et al., 2023).

ثالثاً: المجموعة، حيث ظهر تفاضلاً للمقيم الآلي لصالح مجموعة دون غيرها وهي في سياق هذه الدراسة لمجموعة الذكور المفترضة، وتتفق مع ما أشارت إليه دراسة (Gallegos et al., 2023) حول التحيزات الديموغرافية في النماذج اللغوية الكبيرة.

بشكل عام إن جل ما كشفته هذه الدراسة أن المقيم الآلي يحتاج إلى معايير واضحة ومفصلة لكي يقدم تقييمات دقيقة، فكلما زادت تفاصيل مصفوفة التصحيح المقدمة للمقيم الآلي زادت دقته في تقييم مقالات الطلبة، وهذا ما دعا إليه (Zumbo, 2007) وأن هناك تحديات ومشاكل متعلقة بخصائص المقالات والتي من الممكن أن تلعب دوراً مهماً ومؤثراً على عدالة التقييمات التي يقدمها المقيم الآلي في النماذج اللغوية الكبيرة، لذلك هناك دعوات لحوكمة وتدقيق استخدامات المقيم الآلي في النماذج اللغوية الكبيرة في الحقل التربوي قبل استخدامه (Arcas, 2022)

الاستنتاجات والتوصيات:

من خلال نتائج هذا الدراسة يمكن التوصل إلى ما يلي:

- النماذج اللغوية الكبيرة جيدة ومفيدة عند استخدامها في تقييم المقالات، ولكن أمر دقتها مشروط بوجود مصفوفة تصحيح مفصلة ذات معايير واضحة ومحددة.
- في حال اعتماد النماذج اللغوية الكبيرة من دون ضوابط قد تظهر تفاضلاً تبعاً لمتغيرات مختلفة كخصائص المقالات وخصائص الطلبة الديموغرافية، وهذا يهدد عدالة التقييم.
- نتائج هذه الدراسة تثري الأدب النظري نظراً لما اعتمدته من تحليل شامل لمتغيرات كخصائص المقالات والمجموعة للطالب.

وتوصي الدراسة بما يلي:

- ضرورة تقديم معايير واضحة ومحددة للمقيم الآلي في النماذج اللغوية الكبيرة، ومصفوفات تصحيح مفصلة عند تقييم المقالات الأكاديمية للطلبة.
- دمج التقييم الآلي مع التقييم البشري عند وجود ضرورة وخطورة من نتائج التقييم.
- إجراء دراسات مستقبلية على عينات فعلية من الطلبة لتأكيد النتائج المستخلصة من المحاكاة.
- تطوير خوارزميات تصحيحية داخل النماذج اللغوية الكبيرة لتقليل التحيز للطول والأسلوب.

References:

Aljodeh, M. M. (2025). "Assessing AI-generated math items: Evidence from eighth-grade triangle congruence". *TPM – Testing, Psychometrics, Methodology in Applied Psychology, 32 (S2)*, 251–265. https://tpmap.org/submission/index.php/tpm/article/view/219

Andersen, N., Mang, J., & Zehner, F. (2025). Algorithmic fairness in automatic short answer scoring. International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-025-00495-5

Arcas, B. (2022). "Do large language models understand us?". Daedalus, 151 (2), 183–197. https://doi.org/10.1162/daed_a_01909

Attali, Y., & Burstein, J. (2006). "Automated essay scoring with e-rater® V.2". *The Journal of Technology, Learning and Assessment, 4 (3)*, 1–30.

Ben-David, A. (2008)." Comparison of classification accuracy using Cohen's weighted kappa". *Expert Systems with Applications, 34 (2)*, 825–832. https://doi.org/10.1016/j.eswa.2006.10.022

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. *ASCD.*

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7 (3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, E., Zhan, R.-J., Lin, Y.-B., & Chen, H.-H. (2025). From structured prompts to open narratives: Measuring gender bias in LLMs through open-ended storytelling. arXiv preprint arXiv:2503.15904. https://arxiv.org/abs/2503.15904

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). "Transformers: The end of history for natural language processing*?" In ECML PKDD 2021 (pp. 523–540). Springer*. https://doi.org/10.1007/978-3-030-86523-8_41

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics (5th ed.)*. Sage Publications.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2023). Bias and fairness in large

language models: A survey. arXiv preprint arXiv:2309.00770. https://arxiv.org/abs/2309.00770

Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. Humanities and Social Sciences Communications, 11, 126. https://doi.org/10.1057/s41599-024-03209-9

Neto, A. J. M., & Fernandes, M. A. (2019, July). Chatbot and conversational analysis to promote collaborative learning in distance education. In 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT) (pp. 324–326). IEEE. https://doi.org/10.1109/ICALT.2019.00102

OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Pereira, J., Fernández-Raga, M., Osuna-Acedo, S., Roura Redondo, M., Almazán-López, O., & Buldón-Olalla, A. (2019). "Promoting learners' voice productions using chatbots as a tool for improving the learning process in a MOOC". *Technology, Knowledge and Learning, 24 (3), 609–627*. https://doi.org/10.1007/s10758-018-9380-9

Schaller, N.-J., Ding, Y., Horbach, A., Meyer, J., & Jansen, T. (2024). Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 210–221). Association for Computational Linguistics.

Tang, X., Chen, H., Lin, D., & Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. Heliyon, 10(14), e34262. https://doi.org/10.1016/j.heliyon.2024.e34262

Vázquez-Cano, E., Mengual-Andrés, S., & López-Meneses, E. (2021)." Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments". *International Journal of Educational Technology in Higher Education, 18 (1)*, 1–20. https://doi.org/10.1186/s41239-021-00273-0

Willmott, C. J., & Matsuura, K. (2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average

model performance". *Climate Research, 30 (1)*, 79–82. https://doi.org/10.3354/cr030079

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). "Human-centered artificial intelligence in education: Seeing the invisible through the visible". *Computers and Education: Artificial Intelligence, 2, 100008*. https://doi.org/10.1016/j.caeai.2021.100008

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., & Wen, J. (2023)." A survey of large language models." *arXiv preprint arXiv:2303.18223*. https://doi.org/10.48550/arXiv.2303.18223

Zumbo, B. D. (2007). "Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going". *Language Assessment Quarterly, 4 (2),* 223–233. https://doi.org/10.1080/15434300701375832