

ChatGPT's Accuracy, Agreement, and Reliability in Arabic Grammar Assessment: A Comparative Study with Human Raters

دقة ChatGPT واتفاقه وموثوقيته في تقويم القواعد النحوية العربية: دراسة مقارنة مع المقيمين البشريين

Farika Riskiyah^{1*}, Muhammad Ahsan Thoriq²

^{1,2} Universitas Negeri Malang, Indonesia

* farika.riskiyah.2602318@students.um.ac.id of the corresponding author

Abstract Large language models (LLMs) such as ChatGPT are increasingly used to assist in scoring open-ended responses, including Arabic grammatical analysis, yet their reliability relative to trained human raters remains empirically underexamined for morphologically complex languages like Arabic. This study evaluated ChatGPT's accuracy, agreement, and reliability in scoring an Arabic grammar test against human raters, using a comparative quantitative, psychometric design. The instrument comprised 20 essay items covering four aspects (i'rab, tasrif, tarkib, vocabulary/spelling), administered to 20 eleventh-grade students at MAN 2 Kota Malang, yielding 400 scoring units rated by an expert key, a second independent rater, and ChatGPT (twice, for test-retest). Analyses used exact-match percentage, Weighted Cohen's Kappa, and the Intraclass Correlation Coefficient (ICC). ChatGPT's raw accuracy was 65.2%, but after correcting for chance agreement, kappa was only 0.008 (not different from zero), far below the near-perfect agreement between the two human raters (kappa = 0.891). Its test-retest reliability was perfect (identical scores across all 400 units), yet single-rater ICC was poor (0.371). These findings indicate ChatGPT is not yet suitable as an independent evaluator for Arabic grammar assessment: it is internally consistent yet systematically lenient, awarding full marks to 27 of 33 items the expert scored as entirely wrong (81.8%), and failing to distinguish weak from strong students at the total-score level ($r = 0.055$, versus $r = 0.995$ between human raters) a "reliable but not valid" pattern showing high reliability alone does not guarantee suitability as an evaluator.

Keywords ChatGPT; Arabic grammar assessment; inter-rater reliability

ملخص البحث بدأت نماذج اللغة الكبيرة مثل ChatGPT تُستخدم كمساعد في تصحيح الإجابات المفتوحة، ومنها التحليل النحوي للغة العربية، إلا أن موثوقيتها مقارنة بالمقيمين البشريين لم تُختبر تجريبيًا بما يكفي في لغة ذات بنية صرفية معقدة كاللغة العربية. قُيِّمت هذه الدراسة دقة ChatGPT واتفاقه وموثوقيته في تقييم اختبار قواعد اللغة العربية مقارنة بالمقيمين البشريين، باستخدام تصميم كمي مقارنة بمقارنة سيكومترية. تكوّنت الأداة من 20 بندًا مقالًا تغطي أربعة جوانب (الإعراب، والتصرف، والتركيب، والمفردات/الإملاء)، طُبِّقت على 20 طالبًا من الصف الحادي عشر بمدرسة MAN 2 مدينة مالانج، فأنتجت 400 وحدة تقييم قُيِّمتها مفتاح الخبير، ومقيم ثانٍ مستقل، وChatGPT (مرتين لاختبار الثبات). استُخدم في التحليل نسبة التطابق التام، ومعامل كابا الموزون لكوهين، ومعامل الارتباط داخل الصنف (ICC). بلغت الدقة الخام لـ ChatGPT 65.2%، لكن بعد تصحيح احتمال التطابق العرضي، بلغت قيمة كابا 0.008 فقط (لا تختلف عن الصفر)، أدنى بكثير من التوافق شبه التام بين المقيمين البشريين (كابا = 0.891). وكانت موثوقية الاختبار وإعادة الاختبار لدى ChatGPT تامة (درجات مطابقة في جميع الوحدات)، إلا أن ICC للمقيم الواحد كان ضعيفًا 0.371. وتشير هذه النتائج إلى أن ChatGPT لا يصلح بعد ليكون مقيمًا مستقلًا لقواعد اللغة العربية: فهو متسق مع ذاته لكنه متساهل بشكل منهجي، إذ منح الدرجة الكاملة

لـ27 من أصل 33 إجابة اعتبرها الخبير خاطئة كلياً (81.8%)، وفشل في التمييز بين الطلاب الضعفاء والأقوياء عند مستوى الدرجة الإجمالية $r = 0.055$ ، مقابل $r = 0.995$ بين المقيمين البشريين) نمط "موثوق لكنه غير صالح" يؤكد أن الموثوقية العالية وحدها لا تضمن صلاحية النموذج كمقيّم.

كلمات أساسية ChatGPT؛ تقييم قواعد اللغة العربية؛ الموثوقية بين المقيمين

لقد هزّ حضور نماذج اللغة الكبيرة (LLM) مثل ChatGPT واحداً من أكثر الميادين التربوية محافظةً: ميدان التقييم. فلغفود طويلة، كان تصحيح الإجابات المفتوحة سواء في المقالات الإنجليزية أو في التحليل النحوي لأي لغة كانت يُعدّ مجاًلاً يتطلب بطبيعته مقيماً بشرياً، نظراً لما ينطوي عليه من تفكير لغوي دقيق ومتشعب (Kasneci et al., 2023). غير أن القدرات التوليدية التي يمتلكها ChatGPT في فهم البنية اللغوية وشرحها أثارت تساؤلاً لم يعد بإمكان الباحثين في تعليم اللغات تجاهله: هل يستطيع هذا النموذج أن يقيم إجابات الطلاب بالدقة والاتساق والموثوقية ذاتها التي يتحلى بها المقيم البشري المدرب؟ وهذا التساؤل ليس مجرد فضول أكاديمي؛ فملايين المعلمين حول العالم، ومن بينهم آلاف من مدرّسي اللغة العربية في المدارس الدينية (المدارس والمعاهد) في إندونيسيا، أخذوا يجربون بالفعل استخدام ChatGPT كمساعد في التصحيح، دون أن تتوافر لديهم إرشادات تجريبية كافية توضح متى وإلى أي مدى يمكن الوثوق بهذه الأداة.

لقد شهدت الأبحاث المتعلقة بالتصحيح الآلي للمقالات (Automated Essay Scoring – AES) المعتمد على نماذج اللغة الكبيرة تطوراً سريعاً خلال السنوات الثلاث الأخيرة، غير أنها تركّزت بشكل شبه كامل على اللغة الإنجليزية. وتُظهر الدراسات التي أُجريت على المقالات الحجاجية والكتابة بالإنجليزية كلغة أجنبية صورة متباينة: فبعضها يسجّل ارتباطاً قوياً بين ChatGPT والمقيمين البشريين (Kim, 2025; Quah et al., 2024)، في حين تخلص دراسات أخرى إلى أن ChatGPT يميل إلى التساهل المفرط أو، على العكس، إلى التشدد المفرط مقارنةً بالإنسان، مع اتفاق بعيد عن الكمال حين يُقاس باستخدام إحصاءات تصحّح أثر الصدفة كمعامل كبا لكوهين (Bui & Barrot, 2025b; Manning et al., 2025; Shin & Lee, 2024). أما الجانب الأقل تناوُلاً في الأدبيات فهو كيفية تغيّر هذا الأداء حين يكون موضوع التقييم ليس مقالاً سردياً بالإنجليزية، بل تحليلاً نحويّاً رسمياً للغة ذات بنية صرفية أكثر تعقيداً بكثير كاللغة العربية، بما تتضمنه من نظام الإعراب (وسم الوظائف النحوية)، والتصرّيف (تصرّيف الأفعال واشتقاقها)، والتركيب (بنية العبارات والجملة) وهي جوانب تستلزم استدلالاً متعدد الطبقات، لا مجرد الحكم الشمولي على جودة الكتابة.

وليست هذه الفجوة مجرد ثغرة أكاديمية هامشية. فاللغة العربية تُصنّف ضمن اللغات محدودة الموارد في تطوير نماذج اللغة الكبيرة مقارنةً بالإنجليزية؛ إذ إن مدوّنات التدريب المتاحة أقل بكثير، كما أن المهام النحوية الرسمية كالإعراب تكاد لا تجد نظيراً مباشراً لها في التقاليد اللسانية الغربية التي تهيمن على بيانات تدريب هذه النماذج. كذلك فإن الدراسات القائمة حول أتمتة تقييم اللغة العربية ركّزت في معظمها على المقالات الحرة باستخدام مقاربات تعلّم آلي كلاسيكية قائمة على تشابه النصوص (Du et al., 2024; Gomaa & Fahmy, 2014; Lotfy et al., 2023)، لا على تقييم نماذج

توليدية كبيرة مثل ChatGPT في بنود نحوية بنيوية محددة. وهذا التفاوت البحثي يحمل في طياته خطر نشوء فجوة في الممارسة: ففي حين بدأت المدارس الناطقة بالإنجليزية تمتلك أدلة تجريبية كافية لتقدير متى يمكن الوثوق بالذكاء الاصطناعي في التقييم، فإن مؤسسات تعليم اللغة العربية ومنها آلاف المدارس الدينية في إندونيسيا والشرق الأوسط وجنوب آسيا معرضة لاعتماد التقنية ذاتها دون تحقق مماثل من صلاحيتها.

وتأتي هذه الدراسة لسدّ هذه الفجوة من خلال دراسة استطلاعية تُقيّم ثلاثة أبعاد سيكومترية كثيرًا ما تُخلط بينها في الخطاب العام حول الذكاء الاصطناعي داخل الفصول الدراسية: الدقة أي مدى توافق درجات ChatGPT مع مفتاح الإجابة الذي وضعه الخبير، والاتفاق أي مدى انسجام درجات ChatGPT مع درجات مقيمين بشريين مستقلين، والموثوقية أي مدى ثبات ChatGPT في تقييم الإجابة نفسها عند تكرار عرضها عليه. وهذه الأبعاد الثلاثة مختلفة من الناحية المفاهيمية، وكما سيتبين في هذا المقال، فقد تُفضي إلى استنتاجات متعارضة إذا اقتصر القياس على بُعد واحد منها فقط؛ إذ قد يكون نموذج الذكاء الاصطناعي شديد الاتساق مع ذاته (موثوقًا) بينما ينحرف انحرافًا منهجيًا عن معيار الخبير (غير دقيق)، وهو نمط بدأ يحظى مؤخرًا باهتمام جاد في أدبيات تقييم نماذج اللغة الكبيرة بوصفها حكمًا في مجالات عدة خارج نطاق تعليم اللغات.

وللإجابة عن هذا التساؤل، صمّمنا أداة اختبار لقواعد اللغة العربية تغطي أربعة جوانب لغوية أساسية الإعراب، والتصريف، والتركييب، والمفردات/الإملاء يمثل كل جانب منها خمسة بنود اختبارية تُقيّم وفق مقياس تحليلي من 0 إلى 4. وفي المرحلة الاستطلاعية التي يعرضها هذا المقال، أجاب عشرون طالبًا من مدرسة الثانوية الحكومية الثانية بمدينة مالانج عن جميع البنود العشرين، مما أسفر عن 400 وحدة تقييم قُيِّمت بالتوازي من قبل مفتاح إجابة الخبير، ومقيمين بشريين مستقلين، وChatGPT. وقد أتاح هذا التصميم لنا ليس فقط حساب المؤشرات الإجمالية، بل أيضًا تحليل أداء ChatGPT بنديًا وبندًا، وجانبًا جانبيًا، بل وحتى طالبًا طالبًا وهو مستوى من الدقة التفصيلية نادرًا ما يتوافر في دراسات التصحيح الآلي واسعة النطاق التي تميل إلى الاكتفاء بعرض أرقام إجمالية دون تدقيق حالة بحالة.

وتتسم إسهامات هذا المقال بطابع مزدوج. فمن الناحية المنهجية، نبين كيف يمكن لمعامل كبا الموزون لكوهين أن يكشف عن فجوة خفية وراء أرقام اتفاق التطابق التام التي تبدو مطمئنة للوهلة الأولى - وهي ظاهرة معروفة في أدبيات القياس النفسي الكلاسيكية باسم "مفارقة كبا" (Feinstein & Cicchetti, 1990)، لكنها نادرًا ما تُطبّق صراحةً في تقييم أنظمة التصحيح الآلي القائمة على نماذج اللغة الكبيرة للغات غير الإنجليزية. أما من الناحية الموضوعية، فنقدّم واحدًا من أوائل الأدلة التجريبية حول أداء ChatGPT في المهام النحوية الرسمية للغة العربية - وهو مجال يستلزم استدلالًا صرفيًا نحويًا بالغ الدقة، بدلًا من الحكم الشمولي على جودة الكتابة الذي هيمن حتى الآن على أدبيات التصحيح الآلي.

1. طريقة البحث

اعتمدت هذه الدراسة تصميمًا كمياً مقارناً بمقاربة سيكومترية (قياس نفسي). وتكوّنت أداة الاختبار من 20 بنداً مقالياً بنيوياً موزعة بالتساوي على أربعة أقسام لغوية: القسم أ - الإعراب (البندود 1-5)، والقسم ب - التصريف (البندود 6-10)، والقسم ج - التركيب (البندود 11-15)، والقسم د - المفردات/الإملاء (البندود 16-20). ويقيس كل بند جانباً لغوياً واحداً فقط تجنباً لالتباس مصدر تباين الدرجات. وفي هذه المرحلة الاستطلاعية، أجاب عشرون طالباً من الصف الحادي عشر بمدرسة MAN 2 مدينة مالانج عن جميع البنود العشرين، مما أنتج 400 وحدة بيانات (طالب × بند).

وقُيِّمت كل وحدة بيانات بالتوازي من قبل ثلاثة مصادر باستخدام مقياس تقييم تحليلي على مقياس من 0 إلى 4 (0 = خطأ كلي، 4 = إجابة كاملة الصحة): (1) مفتاح إجابة الخبير، الذي قدّمه في هذه المرحلة الاستطلاعية المقيّم الأول؛ (2) المقيّم الثاني، وهو مقيّم بشري مستقل لم يطلع على درجات المقيّم الأول ولا على درجات ChatGPT أثناء عملية التقييم؛ (3) ChatGPT، الذي شُغل باستخدام موجّه تقييمي موحد يستند إلى المقياس ومفتاح الإجابة نفسيهما، وقُيِّم مرتين في وقتين مختلفين لاختبار مدى اتساقه (الموثوقية عبر الاختبار وإعادة الاختبار).

وحُسبت دقة ChatGPT كنسبة تطابق درجاته مع مفتاح الخبير (التطابق التام). وحُلّل الاتفاق بين المقيمين باستخدام معامل كابا الموزون لكوهين بالوزن التربيعي (Cohen, 1968)، وهو الأنسب للبيانات الترتيبية على مقياس من 0 إلى 4 مقارنةً بمعامل كابا غير الموزون، لأنه يفرض عقوبة تتناسب مع حجم الفرق بين الدرجات. وحُسبت فترات الثقة بنسبة 95% لكل معامل كابا باستخدام أسلوب المعاينة المكررة العنقودية (1000-2000 تكرار) مع وحدة إعادة المعاينة على مستوى الطالب لمراعاة عدم استقلالية البنود ضمن الطالب الواحد. وأُتُبعت في تصنيف قوة الاتفاق معايير Landis (1977) وKoch. واستُكمل ذلك بحساب موثوقية المقيمين الثلاثة مجتمعين باستخدام معامل الارتباط داخل الصنف من النموذج العشوائي ثنائي الاتجاه بالاتفاق المطلق (ICC(A,1) للمقيّم الواحد وICC(A,k) للمقيمين الثلاثة مجتمعين (Shrout & Fleiss, 1979; Koo & Li, 2016). وأُجريت جميع التحليلات الإحصائية باستخدام لغة بايثون.

وتجدر الإشارة، بوصف ذلك محدودية تصميمية متعمّدة يُفصح عنها بشفافية، إلى أنه لما كان مفتاح إجابة الخبير في هذه المرحلة الاستطلاعية صادراً عن المقيّم الأول، فإن مؤشر "دقة ChatGPT مقارنةً بمفتاح الخبير" يتطابق بنيوياً مع مؤشر "اتفاق ChatGPT مع المقيّم الأول". ولذلك يُعدّ المقيّم الثاني المقارن الوحيد المستقل تماماً عن عملية وضع مفتاح الإجابة، وقد أُعطيت نتائجه وزناً تفسيرياً أكبر في مناقشة الاتفاق.

2. نتائج البحث

يعرض هذا القسم نتائج الدراسة بالترتيب وفقاً للأبعاد السيكمترية الثلاثة التي جرى فحصها الإحصاء الوصفي، والدقة، والتوافق بين المقيمين، والموثوقية، والتباين حسب الجانب اللغوي، وتحليل مستوى الطالب، وأنماط الأخطاء المنهجية دون تفسير معمق. أما المناقشة النقدية والتوليف بين هذه النتائج فتعرض بشكل منفصل في القسم الرابع.

الإحصاء الوصفي

يعرض الجدول ١ المتوسطات والانحرافات المعيارية للدرجات حسب كل مقيم وكل جانب لغوي. وبشكل عام، حقق ChatGPT أعلى متوسط للدرجات (م = ٣,٦٢؛ ع = ٠,٨٨)، يليه الخبير/المقيم الأول (م = ٣,٤٢؛ ع = ١,٢٢) ثم المقيم الثاني (م = ٣,٠٨؛ ع = ١,٢٦). وقد اتسق هذا النمط من الدرجات الأعلى التي يمنحها ChatGPT في معظم الجوانب اللغوية، باستثناء جانب التركيب.

الجانب	الخبير م (ع)	المقيم الثاني م (ع)	ChatGPT م (ع)
الإعراب	3.68 (0.78)	2.71 (0.77)	3.33 (1.12)
التصريف	2.71 (1.65)	2.66 (1.66)	3.75 (0.69)
التركيب	3.79 (0.74)	3.65 (0.82)	3.54 (1.02)
المفردات/الإملاء	3.48 (1.19)	3.31 (1.32)	3.85 (0.46)
المجموع	3.42 (1.22)	3.08 (1.26)	3.62 (0.88)

الجدول ١. الإحصاء الوصفي لدرجات كل مقيم وكل جانب لغوي (على مقياس ٠-٤؛ ن=١٠٠ لكل جانب، ن=٤٠٠ إجمالاً)

دقة ChatGPT مقارنةً بمفتاح الإجابة الخبير

بلغت الدقة الخام (التطابق التام) لـ ChatGPT مقارنةً بمفتاح الإجابة الخبير ٦٥,٢٪ إجمالاً، مع تفاوت تراوح بين ٥٠,٠٪ في جانب التصريف و ٧٧,٠٪ في جانب المفردات/الإملاء. وبعد تصحيح احتمال التطابق العرضي باستخدام معامل كابا الموزون لكوهين (Weighted Cohen's Kappa) بأوزان تربيعية، بلغت قيمة كابا = ٠,٠٠٨ (فاصل ثقة ٩٥٪: [-٠,٠٥٥، ٠,٠٦٦]) للبيانات الإجمالية وهي قيمة لا تختلف إحصائياً عن الصفر.

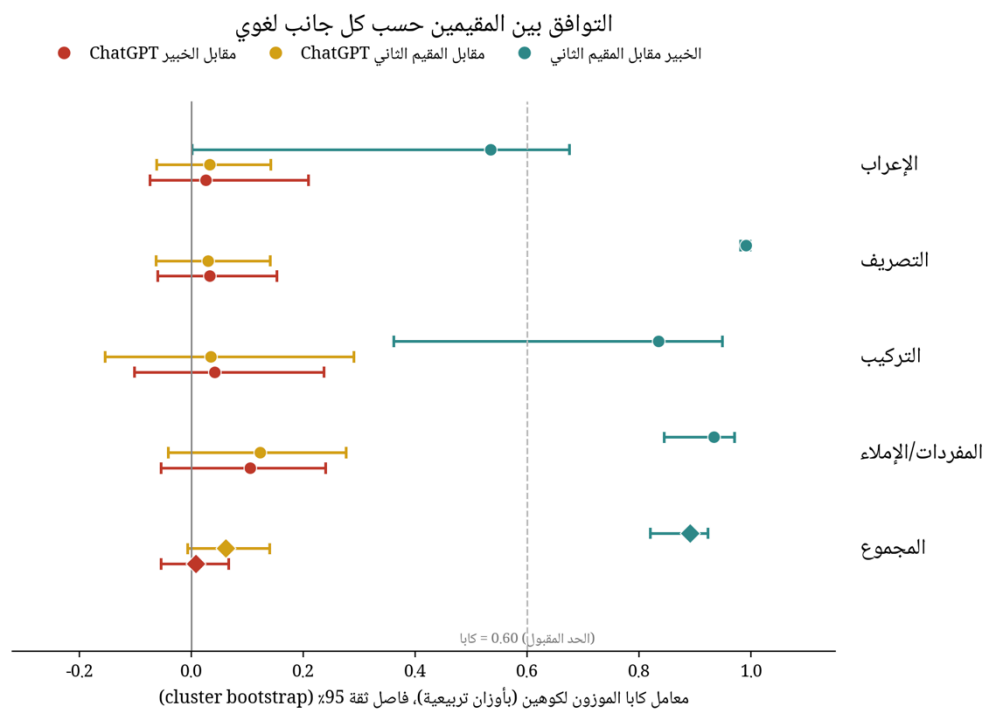
التوافق بين المقيمين

يعرض الجدول ٢ والشكل ١ (مخطط الغابة/forest plot) قيم كابا الموزون وفواصل الثقة ٩٥٪ (باستخدام أسلوب cluster bootstrap على مستوى الطالب) لثلاثة أزواج من المقيمين في كل جانب لغوي. وقد تراوح التوافق بين الخبير (المقيم الأول) والمقيم الثاني من كابا = ٠,٥٣٤ (الإعراب) إلى كابا = ٠,٩٩١ (التصريف)، بإجمالي كابا = ٠,٨٩١ (فئة "شبه تام" وفق معايير Landis و Koch). في المقابل، لم يتجاوز توافق ChatGPT مع أي من المقيمين البشريين سواء مع الخبير أو مع المقيم الثاني فئة "ضئيل" (كابا > ٠,٢٠) في أي جانب من الجوانب، بإجمالي كابا بين ChatGPT والمقيم الثاني = ٠,٠٦١، وبين ChatGPT والخبير = ٠,٠٠٨.

الجانب	زوج المقيمين	(كابا) κ	%فاصل الثقة 95	الفئة
--------	--------------	----------	----------------	-------

متوسط

ضئيل



مقابل المقيم الثاني			ChatGPT	
مقابل الخبير الثاني			ChatGPT	
ضئيل	شبه تام		[-0.074, 0.209]	0.026
التصريف			[0.981, 0.998]	0.991
مقابل المقيم الثاني			[-0.064, 0.140]	0.030
مقابل الخبير الثاني			[-0.061, 0.152]	0.032
ضئيل	شبه تام		[0.361, 0.948]	0.834
التركيب			[-0.155, 0.290]	0.034
مقابل الخبير الثاني			[-0.102, 0.236]	0.041
ضئيل	شبه تام		[0.844, 0.970]	0.933
المفردات/الإملاء			[-0.042, 0.276]	0.123
مقابل المقيم الثاني			[-0.055, 0.239]	0.105
ضئيل	شبه تام		[0.820, 0.923]	0.891
المجموع			[-0.007, 0.139]	0.061
مقابل الخبير الثاني			[-0.055, 0.066]	0.008

الجدول ٢. معامل كابا الموزون لكوهين بين المقيمين حسب كل جانب لغوي مع فاصل ثقة ٩٥% (باستخدام cluster bootstrap على مستوى الطالب).

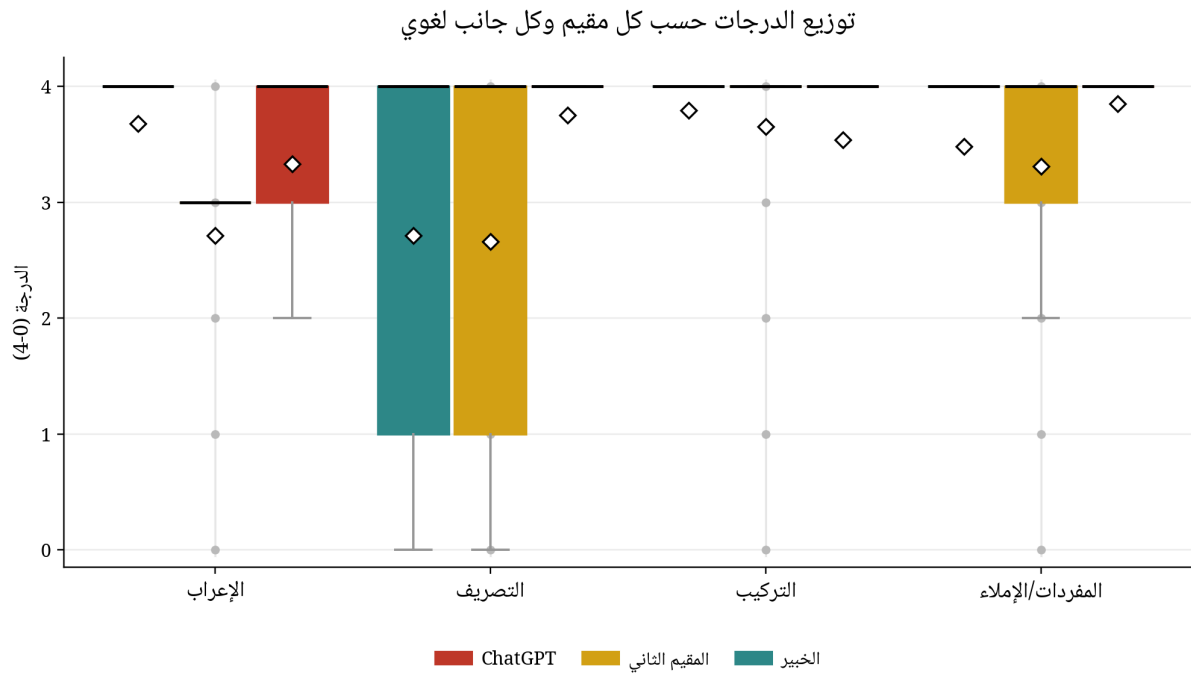
الشكل ١. مخطط الغابة (forest plot) لمعامل كابا الموزون بين المقيمين حسب كل جانب لغوي، مع خط عتبة كابا = ٠,٦٠ كمرجع للتوافق المقبول.

موثوقية الاختبار وإعادة الاختبار ومعامل ICC الإجمالي

في اختبار الموثوقية بأسلوب الاختبار وإعادة الاختبار، أنتج ChatGPT درجات متطابقة تمامًا في كلتا جولتي التقييم لجميع وحدات البيانات البالغ عددها ٤٠٠ وحدة. وأظهر معامل الارتباط داخل الفئة (ICC) الإجمالي للمقيمين الثلاثة أن $ICC(A,1)$ للمقيّم الواحد بلغ ٠,٣٧١ (فئة "ضعيف")، بينما بلغ $ICC(A,k)$ لمجموع المقيمين الثلاثة ٠,٦٤٠ (فئة "متوسط") (Koo & Li, 2016). ويعود هذا الارتفاع في $ICC(A,k)$ إلى إسهام التوافق المرتفع بين المقيمين البشريين الاثنين؛ إذ إن إدراج ChatGPT ضمن فريق المقيمين أدى إلى خفض الموثوقية الإجمالية بدلاً من رفعها مقارنةً بالاعتماد على المقيمين البشريين فقط.

تباين الدرجات حسب الجانب اللغوي

يعرض الشكل ٢ توزيع الدرجات (مخطط الصندوق) للمقيمين الثلاثة في كل جانب لغوي. وبالترتيب من الأكثر تحديًا لـ ChatGPT: يقع الإعراب والتصريف (كابا مقارنة بالخبير ٠,٠٢٦ و ٠,٠٣٢ على التوالي) في مستوى صعوبة شبه متماثل، يليهما التركيب (كابا = ٠,٠٤١)، ثم المفردات/الإملاء بوصفه الجانب ذا التوافق الأعلى نسبيًا رغم بقائه ضمن فئة "ضئيل" (كابا = ٠,١٠٥). وقد أظهر ChatGPT باستمرار توزيعًا أضيق للدرجات وميلًا نحو القيمة القصوى مقارنةً بالمقيمين البشريين، لا سيما في جانبي التصريف والمفردات/الإملاء.



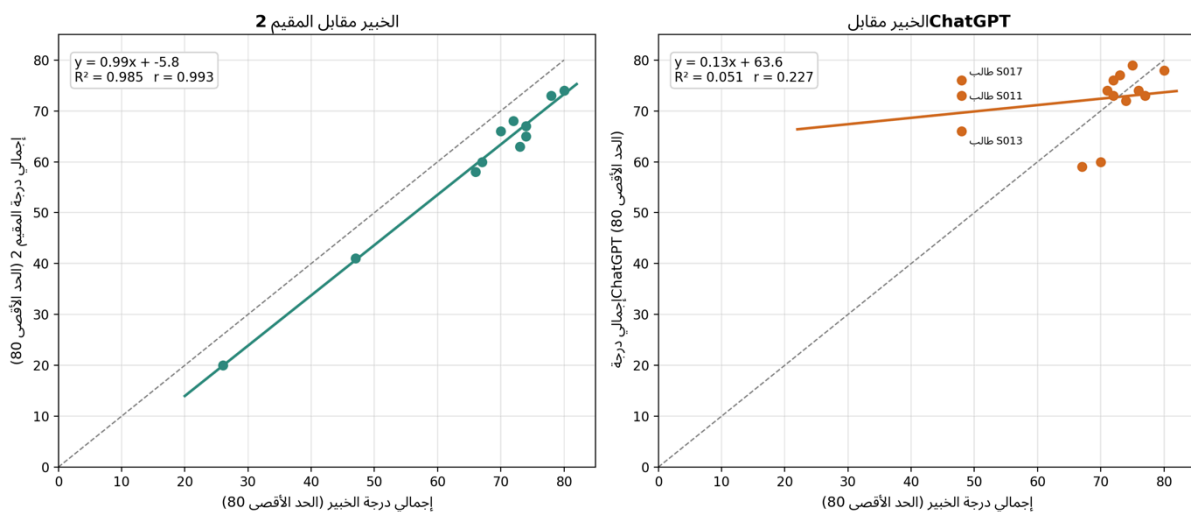
الشكل ٢. توزيع الدرجات حسب كل مقيم وكل جانب لغوي. تشير المعينات البيضاء إلى المتوسط الحسابي؛ وكثيرًا ما تتطابق الوسائط لكل من

ChatGPT

والخبير عند الدرجة ٤.

تحليل مستوى الطالب

يعزز التحليل على مستوى إجمالي درجة كل طالب (الحد الأقصى ٨٠ درجة من ٢٠ بنّاءً بمقياس ٤-٠) النتائج التي ظهرت على مستوى البنود. ويعرض الشكل ٣ مخطط الانتشار لإجمالي الدرجات مع خط الانحدار، مقارنةً بخط التوافق ($y = x$). وقد بلغ الارتباط بين الخبير والمقيم الثاني على مستوى الطلاب $r = 0.995$ (معامل التحديد $R^2 = 0.990$)، في حين بلغ الارتباط بين الخبير و ChatGPT $r = 0.055$ فقط ($R^2 = 0.003$)، مع خط انحدار يكاد يكون مسطحًا. وقد برزت حالات الطلاب S011 و S013 و S017: إذ قِيمهم كلا المقيمين البشريين بدرجات منخفضة إجمالي الدرجات لكل طالب (ن=20): خط الانحدار مقابل خط التوافق (متقطع)



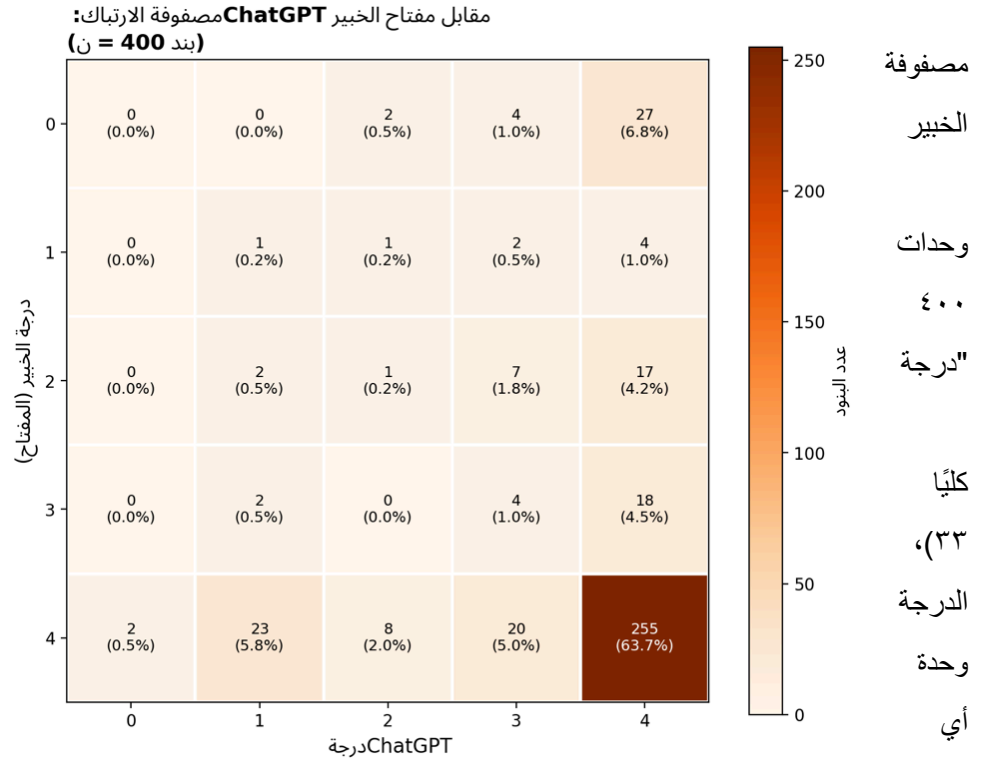
جداً (إجمالي درجات تراوح بين ٢٠ و ٤٧ من ٨٠)، بينما منحهم ChatGPT درجات تراوحت بين ٧١ و ٧٥ من ٨٠ وهو ما يعادل أداء أفضل الطلاب وفق تقييم البشر.

الشكل ٣. إجمالي درجة كل طالب (ن = ٢٠): الخبير مقابل المقيّم الثاني (يساراً) يُظهر تطابقاً شبه تام مع خط التطابق؛ الخبير مقابل ChatGPT (يميناً) يُظهر خط انحدار يكاد يكون مسطحاً.

أنماط الأخطاء المنهجية: مصفوفة الالتباس

يعرض الشكل ٤ الالتباس بين درجات ودرجات

ChatGPT لجميع البيانات البالغ عددها وحدة. ففي صف الخبير = صفر (الإجابات الخاطئة وفق الخبير، ن = منح ChatGPT الكاملة (٤) ٢٧، ولم يمنح (٨١،٨٪)، ولم يمنح وحدة درجة صفر. وإجمالاً، منح



ChatGPT الدرجة الكاملة (٤) ٣٢١ وحدة من أصل ٤٠٠ (٨٠،٣٪)، مقارنةً بـ ٣٠٨ وحدات (٧٧،٠٪) لدى الخبير، و ٢٠٢ وحدة فقط (٥٠،٥٪) لدى المقيّم الثاني.

الشكل ٤. مصفوفة الالتباس بين ChatGPT ومفتاح الإجابة الخبير (ن = ٤٠٠ بند). يُظهر صف "الخبير = صفر" أن ٢٧ من أصل ٣٣ إجابة خاطئة كلياً قد مُنحت مع ذلك الدرجة القصوى من قبل ChatGPT.

المناقشة

تناقش نتائج القسم الثالث نقدياً من خلال ثلاث عدسات سيكومترية متكاملة، فُصلت عمداً لتجنب الخلط بينها كما يحدث كثيراً في الخطاب العام حول الذكاء الاصطناعي في الفصول الدراسية: (٤.١) دقة ChatGPT مقارنةً بمفتاح الإجابة الخبير بما يتجاوز مجرد نسبة التطابق التام؛ (٤.٢) توافق ChatGPT مع المقيمين البشريين وكيفية تقارب هذا التوافق وتباعده تبعاً لتعقيد البنية اللغوية؛ و(٤.٣) نمط "الموثوقية دون الصلاحية" الذي يشير إلى تحيز منهجي في تقييم ChatGPT.

دقة ChatGPT في تقييم قواعد اللغة العربية: ما وراء التطابق التام

قد تُوحي الدقة الخام (٦٥,٢٪) للوهلة الأولى بأن ChatGPT موثوق به بقدر معقول لتقييم قواعد اللغة العربية. غير أن الفجوة الحادة بين هذه الدقة الخام ومعامل كابا الموزون (٠,٠٠٨) تمثل تجلياً لظاهرة تُعرف في أدبيات القياس النفسي الكلاسيكية بـ"مفارقة كابا": فعندما يكون توزيع الدرجات شديد الانحياز نحو فئة واحدة وفي هذه الحالة، بلغت نسبة الدرجات التي منحها الخبير القيمة ٤ من إجمالي الدرجات ٧٧٪ يرتفع احتمال التطابق العرضي ارتفاعاً كبيراً، بحيث تنهار الإحصاءات التي تصحح أثر التوافق العرضي مثل كابا لتقترب من الصفر رغم أن نسبة التطابق الخام تبدو مرتفعة (Cicchetti & Feinstein, 1990). وهذا يعني أن نسبة التطابق التام في مجموعة بيانات شديدة الانحياز نحو القيمة القصوى ليست دليلاً صالحاً على جودة تقييم ChatGPT وهو استنتاج بالغ الأهمية للباحثين والممارسين الذين يقيمون أنظمة التصحيح الآلي المعتمدة على النماذج اللغوية الكبيرة دون مراعاة شكل توزيع الدرجات.

وليس هذا النمط ظاهرة فريدة بهذه الدراسة. فباستخدام معامل كابا لكوهين في سياق تقييم امتحانات قائمة على معايير تقييم تحليلية في التعليم العالي، أفاد (Flodén 2025) أيضاً بتوافق ضئيل فقط (كابا = ٠,١٣٩) بين ChatGPT والمقيمين البشريين، أدنى بكثير من الانطباع الذي توحى به مقارنة الدرجات/العلامات مقارنة سطحية وهو ما يعزز الحجة القائلة بأن مقاييس الدقة البسيطة تبالغ منهجياً في تقدير كفاءة ChatGPT التقييمية عندما لا يُؤخذ توزيع الدرجات بعين الاعتبار. ويظهر نمط مماثل في دراسات التصحيح الآلي لمقالات اللغة العربية: فقد أفاد (Ghazawi و Simpson 2025)، اللذان قارنا أداء ChatGPT بعدد من النماذج اللغوية الكبيرة الأخرى على مجموعة بيانات مقالات عربية، بأن أفضل نموذج لغوي توليدي (ACEGPT، بمعامل كابا موزون تربيعياً = ٠,٦٧) لا يزال أدائه أدنى من نموذج BERT المدرب خصيصاً رغم صغر حجمه النسبي (QWK = ٠,٨٨). وتؤكد هذه النتيجة أن محدودية دقة ChatGPT في تقييم المهام باللغة العربية ليست مجرد مسألة إجرائية خاصة بهذه الدراسة، بل

نمط ثابت عبر الدراسات: إذ يبدو أن النماذج التوليدية العامة لم تُضاه بعد أداء المقاربات المصممة أو المدربة خصيصًا لمعالجة صرف اللغة العربية بدقة.

التوافق بين ChatGPT والمقيمين البشريين: التقارب والتباعد والتعقيد اللغوي

يتجلى نمط التقارب والتباعد في هذه الدراسة بوضوح شديد: إذ بلغ التوافق بين المقيمين البشريين اللذين عملا بشكل مستقل فئة "شبه تام" (كابا = ٠,٨٩١ وفق معايير Koch و Landis)، في حين لم يتجاوز توافق ChatGPT مع كلا المقيمين البشريين فئة "ضئيل" في أي جانب. وتتسق هذه النتيجة مع جزء من أدبيات التصحيح الآلي باللغة الإنجليزية التي أفادت أيضًا بتوافق بعيد عن الكمال بين ChatGPT والبشر عند قياسه بمعامل كابا الموزون، وإن ظل عمومًا ضمن فئتي "مقبول" إلى "متوسط" (Shin & Lee, 2024; Uyar & Büyükahiska, 2025). لا فئة "ضئيل" كما في هذه الدراسة بينما أفادت دراسات أخرى طبقت أسلوب الضبط الدقيق بمعامل ICC مرتفع بلغ ٠,٩٧ أو ارتباط تراوح بين ٠,٧٥ و ٠,٨٥ (Bridgeman et al., 2012; Quah et al., 2024; Yavuz et al., 2025). ومن المرجح أن هذا التفاوت بين الدراسات يعود بشكل كبير إلى نوع المهمة: فالتقييم الشمولي للمقالات عادة ما يتسامح مع التباين الذاتي ويمنح درجات لطلاقة الكتابة السطحية، في حين يتطلب التصحيح التحليلي القواعدي الرسمي كما في هذه الدراسة منطقيًا صارم القواعد لا يترك مجالًا لمثل هذا التسامح. أما نمط التوافق المنخفض الأقرب إلى نتائج هذه الدراسة، فقد أفادت به على نطاق واسع دراسات استخدمت معايير تقييم صارمة ومهيكلية (Bucol & Manning et al., 2025; Sangkawong, 2025; Bui & Barrot, 2025a, 2025b)، ما يشير إلى أن تصميم معايير التقييم وشكل المهمة لا اللغة المستهدفة وحدها يشكّلان محددًا مهمًا لمدى تقارب ChatGPT مع المقيمين البشريين.

كما يتدرج نمط تباعد ChatGPT بشكل منهجي وفقًا لتعقيد البنية اللغوية للبناء: فالإعراب والتصريف اللذان يتطلبان منطقيًا صرفيًا نحوياً متعدد الخطوات حول الوظيفة النحوية وتصريف الكلمات هما الجانبان الأدنى توافقًا (كابا يتراوح حول ٠,٠٣)، يليهما التركيب (كابا = ٠,٠٤١)، في حين أظهرت المفردات/الإملاء، بوصفها مهمة مطابقة معجمية منفصلة نسبيًا، أعلى توافق رغم بقائه ضمن فئة "ضئيل" (كابا = ٠,١٠٥). ويتسق هذا التدرج في الصعوبة مع دراسات أخرى في تصحيح اللغة العربية آليًا: فقد أكد (Steiss, 2024) اللذان طوّروا نظامًا للتصحيح الآلي للإجابات القصيرة بالعربية قائمًا على التعلم العميق، أهمية القدرة على تفسير الدرجة في البنود القواعدية الرسمية نظرًا لارتفاع مخاطر الخطأ في البنى الصرفية المعقدة وهو ما يتسق مع كون الإعراب والتصريف يتطلبان مستويات من الدقة المنطقية أعلى بكثير من مجرد التعرف على الأنماط السطحية. كما أفادت دراسات التصحيح الآلي العربي القائمة على أساليب كلاسيكية باستمرار بتحديات مماثلة في المهام التي تتطلب فهمًا عميقًا للبنية الصرفية (Abu Guba & Abu Quba, 2025; Gomaa & Fahmy, 2014)، مما يعزز الافتراض بأن هذا القصور ليس مجرد ظاهرة خاصة بنموذج واحد أو دراسة واحدة، بل تحدٍّ بنيوي في معالجة اللغة العربية الرسمية آليًا بشكل عام.

موثوق ولكن غير صالح: فهم التحيز المنهجي في تقييم ChatGPT

يؤكد مبدأ أساسي في القياس النفسي أن الموثوقية شرط ضروري، لكنه ليس شرطاً كافياً، لتحقيق الصلاحية (Koo & Li, 2016; Shrout & Fleiss, 1979). وتشير موثوقية الاختبار وإعادة الاختبار التامة التي أظهرها ChatGPT في هذه الدراسة درجات متطابقة تماماً في جولتي التقييم لجميع وحدات البيانات الأربعمئة إلى أن النموذج لا يَحْمَنُ بشكل عشوائي؛ بل على العكس، فهو يطبّق نمط تقييم متسقاً لكنه متحيز بشكل منهجي. ويتعزز هذا الاستنتاج بفشل التمييز على مستوى الطالب (ارتباط الخبير بـ ChatGPT لم يتجاوز $r = 0.05$)، وبمصفوفة الالتباس التي أظهرت أن ٢٧ من أصل ٣٣ إجابة خاطئة كلياً وفق الخبير (٨١,٨٪) مُنحت مع ذلك درجة كاملة من قبل ChatGPT. وقد بدأ هذا الفصل الواضح بين الاتساق الذاتي (الموثوقية) والمطابقة لمعيار مرجعي (الصلاحية) يحظى باهتمام متزايد في أدبيات تقييم جودة مخرجات ChatGPT خارج سياق تصحيح المقالات: فقد فصل (Kooli & Yusuf, 2024; Küchemann et al., 2024; Zhao et al., 2023)، على سبيل المثال، صراحةً بين قياس الموثوقية وقياس الصلاحية عند تقييم جودة بنود الأسئلة التي يولدها ChatGPT، مؤكداً أن التوافق المتوسط بين المقيمين إزاء مخرجات ChatGPT لا يضمن بالضرورة أن تكون تلك المخرجات صالحة من الناحية الجوهرية وهو إطار فكري يتسق مع نمط "موثوق ولكن غير صالح" الذي كُشف عنه في هذه الدراسة.

ويتسم نمط التحيز المكتشف بالتساهل، ويتسق مع اتجاه التحيز الذي أفاد به عدد من دراسات التصحيح الآلي باللغة الإنجليزية (Alqurashi et al., 2025; Bouziane & Bouziane, 2024). وإن لم يكن هذا الاتجاه ثابتاً في جميع الأدبيات إذ أفادت دراسات أخرى بأن ChatGPT أكثر صرامة من البشر (Aydın et al., 2025; Bui & Barrot, 2025b). ومن المرجح أن ChatGPT يمنح درجات مرتفعة استناداً إلى إشارات سطحية مثل اكتمال بنية الإجابة الظاهري أو استخدام مصطلحات تقنية ذات صلة (كذكر مصطلح الإعراب الصحيح دون تطبيقه بدقة) وهو نمط استدلال سطحي يتسق مع مواطن ضعف النماذج اللغوية الكبيرة حين لا ترتبط الإشارات السطحية ارتباطاً كاملاً بصحة الإجابة الجوهرية. وتدعم نتائج (Abu Guba, 2025)، اللذين قيما موثوقية ChatGPT في ترجمة النصوص العربية إلى الإنجليزية ووجدوا تبايناً كبيراً في الاتساق تبعاً لتعقيد بنية الجملة المصدر، فكرة أن ضعف الموثوقية هذا لا يقتصر على سياق تقييم القواعد وحده، بل يعزز المؤشرات على أن التعقيد الصرفي النحوي للغة العربية يمثل مصدر تحدٍّ متكرر لـ ChatGPT عبر أنواع المهام اللغوية، وليس ظاهرة خاصة بمهمة تقييم القواعد فحسب.

وتتسم الآثار العملية لنمط "موثوق ولكن غير صالح" هذا بطابع مزدوج. فمن جهة، يرفض هذا النمط السردية المتفائلة القائلة بأن ChatGPT "دقيق بما فيه الكفاية" استناداً إلى أن نسبة التطابق التام تبدو متوسطة إذ تفشل هذه السردية في مراعاة أثر انحياز توزيع الدرجات ومفارقة كابا. ومن جهة أخرى، يرفض هذا النمط أيضاً السردية المتشائمة القائلة بأن ChatGPT "يقيم بشكل عشوائي لا يمكن التنبؤ به": فالموثوقية التامة تشير بالأحرى إلى أن أخطاء ChatGPT ذات طابع منهجي، ويمكن من حيث المبدأ تتبعها وتصحيحها من خلال هندسة المطالبات أو ضبط معايير التقييم بشكل أكثر وضوحاً، بدلاً من كونها ضوضاء عشوائية يصعب معالجتها.

يفضي الجمع بين العدسات النقدية الثلاث السابقة إلى سردية متماسكة، لكنها تتعارض مع التفاؤل الذي كثيراً ما يسود الخطاب العام حول الذكاء الاصطناعي في التقييم التربوي. فدقة ChatGPT مقارنةً بمفتاح الإجابة الخبير

منخفضة جدًا (كابا = ٠,٠٠٨)؛ وتوافقه مع المقيّم البشري المستقل منخفض جدًا (كابا = ٠,٠٦١)؛ في حين أن موثوقيته الذاتية تامة. ومقارنةً بأدبيات التصحيح الآلي باللغة الإنجليزية، تُعدّ ضخامة الفجوة المكتشفة في هذه الدراسة متطرفة: إذ عادة ما تُبلغ الدراسات على المقالات الإنجليزية عن قيم كابا أو ChatGPT-ICC مقابل البشر في نطاق "متوسط" إلى "كبير"، وإن أفاد عدد غير قليل منها أيضًا بتوافق منخفض مماثل لنتائج هذه الدراسة (Bui & Barrot, 2025a; Chan et al., 2023; Jung et al., 2024; Quah et al., 2024). ومن المرجح أن هذا التفاوت بين الدراسات يعود إلى نوع المهمة (التقييم الشمولي للمقالات مقابل التصحيح التحليلي القواعدي المهيكل)، وتصميم المطالبة، وهو الأكثر صلة بهذه الدراسة، أي تعقيد البنية الصرفية النحوية للغة المستهدفة. ونظرًا لأن أداة هذه الدراسة تتطلب تحديدًا منطقيًا قواعديًا بالغ الدقة في لغة ذات نظام إعرابي غني (نظام الحالات النحوية)، فإن أداء ChatGPT الأدنى بكثير من متوسط دراسات التصحيح الآلي باللغة الإنجليزية يقدم مؤشرًا مبدئيًا على أن قدرة النماذج اللغوية الكبيرة على تقييم القواعد الرسمية للغة العربية لا تزال متأخرة بشكل ملحوظ مقارنة بقدرتها على تقييم جودة الكتابة الإنجليزية بشكل شمولي.

تقدّم هذه الدراسة مساهمة على ثلاثة مستويات على الأقل من الخطاب العالمي. أولاً، على مستوى العدالة التقنية اللغوية، تضيف هذه النتائج دليلاً تجريبيًا على أن أداء النماذج اللغوية الكبيرة التوليدية لا يمكن افتراض تجانسه عبر اللغات المختلفة. فبينما تُبلغ أدبيات التصحيح الآلي المعتمدة أساسًا على اللغة الإنجليزية بتوافق يتراوح بين المتوسط والمرتفع، تُظهر النتائج على اللغة العربية إحدى اللغات التي يتجاوز عدد متحدثيها ٤٠٠ مليون نسمة، وهي لغة طقسية لأكثر من ١,٩ مليار مسلم حول العالم فجوة أداء أكبر بكثير. وتعزز أدلة إضافية من مجالات لغوية عربية أخرى كتحديات موثوقية ChatGPT في الترجمة من العربية إلى الإنجليزية (Abu Guba & Abu Quba, 2025)، وأداء النماذج التوليدية الذي لا يزال دون مستوى النماذج المدربة خصيصًا في تصحيح المقالات العربية آليًا (Ghazawi & Simpson, 2025) هذا النمط القائل بأن هذه الفجوة ذات طابع بنيوي، وليست خاصة بمهمة واحدة. وتتصل هذه النتائج بالنقاش العالمي حول الفجوة الرقمية اللغوية، الذي يسلط الضوء على خطر أن تؤدي أدوات الذكاء الاصطناعي المطوّرة والمقيّمة أساسًا على اللغات الغنية بالموارد إلى توسيع الفجوة في الوصول إلى تقنيات تعليمية موثوقة للمجتمعات الناطقة بغير الإنجليزية، بدلًا من تضيقها.

ثانيًا، على مستوى منهجية تقييم الذكاء الاصطناعي في التعليم، تُبرهن هذه الدراسة على أهمية الفصل الصريح بين ثلاثة أبعاد سيكومترية الدقة والتوافق والموثوقية بدلًا من الاكتفاء بالإبلاغ عن مقياس واحد كالارتباط أو نسبة التطابق الخام، وهو ما قد يكون مضللًا. كما يقدم أسلوب cluster bootstrap لحساب فواصل ثقة معامل كابا، إلى جانب تحليل النتائج على مستوى البند الفردي ومستوى الطالب الفردي، إطارًا منهجيًا قابلاً للتكرار من قبل باحثين آخرين يقيمون النماذج اللغوية الكبيرة بوصفها مقيّمة في لغات أو مجالات أخرى بما في ذلك اللغات الأخرى المحدودة الموارد في آسيا وأفريقيا والشرق الأوسط التي تواجه تحديات مماثلة.

ثالثًا، على مستوى السياسات التعليمية العملية، تحمل نتيجة "موثوق ولكن غير صالح" رسالة مهمة لصانعي السياسات ومطوّري المنصات التعليمية حول العالم الذين يدرسون حاليًا دمج الذكاء الاصطناعي التوليدي في أنظمة

التقييم التكويني والختامي: فالتساق النموذج ليس ضماناً لصلاحيته، ويجب اختبار كلا الأمرين بشكل منفصل قبل منح أي أداة ذكاء اصطناعي صلاحية تقييمية تؤثر في الدرجات الأكاديمية للطلاب. وفي سياق أوسع، تتصل هذه النتائج أيضاً بتطوير مناهج تعليم اللغة العربية في المؤسسات الإسلامية الدولية، التي تستكشف بشكل متزايد رقمنة تعليم النحو والصرف إذ تشير هذه الدراسة إلى أن الدور الأكثر واقعية للذكاء الاصطناعي حالياً هو أداة مساعدة للفرز الأولي تحت إشراف المعلم، وليس بديلاً كاملاً عن المقيم البشري، لا سيما في الجوانب الصرفية النحوية المعقدة.

الإستنتاج

تظهر هذه الدراسة التجريبية أن ChatGPT، عند تقييمه في مهمة تقييم قواعد اللغة العربية التي تشمل الإعراب والتصريف والتركيب والمفردات/الإملاء، ينتج نمطاً سيكومترياً معقداً يمكن تلخيصه في ثلاث نتائج رئيسية. أولاً، فيما يتعلق بالدقة، تبين أن نسبة التطابق التام التي بدت متوسطة (٦٥,٢٪) مضللة بعد تصحيحها باستخدام معامل كابا الموزون لكوهين (كابا = ٠,٠٠٨) وهي مفارقة كابا أبلغ عنها أيضاً في دراسات أخرى للتصحيح القائم على كابا (Flodén, 2025)، وتتسق مع أداء النماذج اللغوية التوليدية الذي لا يزال دون مستوى المقاربات المدربة خصيصاً في مهام التصحيح الآلي العربي (Ghazawi & Simpson, 2025). ثانياً، فيما يتعلق بالتوافق، لم يتجاوز ChatGPT فئة "ضئيل" في توافقه مع المقيمين البشريين (كابا التوافق = ٠,٠٦١)، وهو أدنى بكثير من التوافق بين المقيمين البشريين أنفسهم (كابا = ٠,٨٩١)، مع تبين أن الإعراب والتصريف هما الجانبان الأكثر تحدياً بما يتسق مع ما يتطلبه من منطق صرفي نحوي أعلى مقارنة بالمفردات/الإملاء ذات الطابع المعجمي المنفصل نسبياً، وينسجم مع التحديات البنيوية التي أفادت بها دراسات أخرى للتصحيح الآلي للغة العربية). ثالثاً، فيما يتعلق بالموثوقية، أظهر ChatGPT اتساقاً تاماً لكن مع صلاحية منخفضة جداً مقارنة بالخبير وهو نمط "موثوق ولكن غير صالح" يتسق مع أطر تقييم جودة مخرجات ChatGPT في مجالات أخرى (Küchemann et al., 2024)، ويتعزز بمؤشرات على تحديات مماثلة في مهام لغوية عربية أخرى كالترجمة (Abu Guba & Abu Quba, 2025).

وتحمل نتيجة أن ChatGPT متسق ولكنه غير صالح دلالات عملية واضحة: فاستخدام ChatGPT كمقيّم مستقل لتقييم قواعد اللغة العربية، لا سيما لأغراض التقييم الختامي المؤثر في الدرجات الأكاديمية للطلاب، لا يمكن التوصية به في الوقت الراهن دون ضبط إضافي لمعايير التقييم وإشراف بشري. وفي المقابل، يفتح الاتساق المرتفع لـ ChatGPT فرصة للاستفادة منه كأداة مساعدة تكوينية واسعة النطاق كتنظيم تغذية راجعة أولية سريعة للطلاب طالما ظل المعلم هو صاحب القرار النهائي في التقييمات ذات الأثر المعنوي الكبير. وستعزز الدراسات اللاحقة على النطاق الكامل (٤٠٠ طالب)، باستخدام مفتاح إجابة خبير مستقل تماماً، وتنويع استراتيجيات صياغة المطالبات، والتدقيق النوعي لنصوص إجابات الطلاب، تعميم هذه النتائج، كما ستختبر ما إذا كان بالإمكان سد فجوة الصلاحية المكتشفة في هذه الدراسة من خلال ضبط معايير التقييم وهندسة المطالبات

المراجع

- Abu Guba, M. N., & Abu Quba, A. (2025). AI Translation: Evaluating ChatGPT's Reliability in Translating Arabic to English. *Journal of Artificial Intelligence and Technology*. <https://doi.org/10.37965/jait.2025.0831>
- Alqurashi, A., Alharbi, B., & Sabbeh, S. (2025). An Automatic Grading System for Arabic Language Short-Answer Questions Using Deep Learning. *Engineering, Technology & Applied Science Research*, 15(5), 26665–26675. <https://doi.org/10.48084/etasr.10917>
- Aydın, B., Kışla, T., Elmas, N. T., & Bulut, O. (2025). Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System*, 133, 103784. <https://doi.org/10.1016/j.system.2025.103784>
- Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education*, 3(1), 201. <https://doi.org/10.1007/s44217-024-00320-6>
- Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education*, 25(1), 27–40. <https://doi.org/10.1080/08957347.2012.635502>
- Bucol, J. L., & Sangkawong, N. (2025). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, 62(3), 867–882. <https://doi.org/10.1080/14703297.2024.2363901>
- Bui, N. M., & Barrot, J. (2025a). Using generative artificial intelligence as an automated essay scoring tool: a comparative study. *Innovation in Language Learning and Teaching*, 1–16. <https://doi.org/10.1080/17501229.2025.2521003>
- Bui, N. M., & Barrot, J. S. (2025b). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*, 30(2), 2041–2058. <https://doi.org/10.1007/s10639-024-12891-w>
- Chan, K. K. Y., Bond, T., & Yan, Z. (2023). Application of an Automated Essay Scoring engine to English writing assessment using Many-Facet Rasch Measurement. *Language Testing*, 40(1), 61–85. <https://doi.org/10.1177/02655322221076025>
- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6), 551–558. [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M)
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. <https://doi.org/10.1037/h0026256>
- Du, H., Jia, Q., Gehringer, E., & Wang, X. (2024). Harnessing large language models to auto-evaluate the student project reports. *Computers and Education: Artificial Intelligence*, 7, 100268. <https://doi.org/10.1016/j.caeai.2024.100268>
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
- Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, 51(1), 201–224. <https://doi.org/10.1002/berj.4069>

- Ghazawi, R., & Simpson, E. (2025). How well can LLMs grade essays in Arabic? *Computers and Education: Artificial Intelligence*, 9, 100449. <https://doi.org/10.1016/j.caeai.2025.100449>
- Gomaa, W. H., & Fahmy, A. A. (2014). Automatic scoring for answers to Arabic test questions. *Computer Speech & Language*, 28(4), 833–857. <https://doi.org/10.1016/j.csl.2013.10.005>
- Jung, J. Y., Tyack, L., & von Davier, M. (2024). Combining machine translation and automated scoring in international large-scale assessments. *Large-Scale Assessments in Education*, 12(1), 10. <https://doi.org/10.1186/s40536-024-00199-7>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kim, Y. (2025). Automated Essay Scoring With <scp>GPT</scp> -4 for a Local Placement Test: Investigating Prompting Strategies, Intra-Rater Reliability, and Alignment With Human Scores. *TESOL Quarterly*, 59(S1). <https://doi.org/10.1002/tesq.3405>
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kooli, C., & Yusuf, N. (2024). Transforming Educational Assessment: Insights Into the Use of ChatGPT and Large Language Models in Grading. *International Journal of Human–Computer Interaction*, 1–12. <https://doi.org/10.1080/10447318.2024.2338330>
- Küchemann, S., Rau, M., Schmidt, A., & Kuhn, J. (2024). ChatGPT's quality: Reliability and validity of concept inventory items. *Frontiers in Psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1426209>
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>
- Lotfy, N., Shehab, A., Elhoseny, M., & Abu-Elfetouh, A. (2023). An Enhanced Automatic Arabic Essay Scoring System Based on Machine Learning Algorithms. *Computers, Materials & Continua*, 77(1), 1227–1249. <https://doi.org/10.32604/cmc.2023.039185>
- Manning, J., Baldwin, J., & Powell, N. (2025). Human versus machine: The effectiveness of ChatGPT in automated essay scoring. *Innovations in Education and Teaching International*, 62(5), 1500–1513. <https://doi.org/10.1080/14703297.2025.2469089>
- Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education*, 24(1), 962. <https://doi.org/10.1186/s12909-024-05881-6>
- Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies*, 29(18), 24735–24757. <https://doi.org/10.1007/s10639-024-12817-6>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>

- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education*, 12(1), 20–32. <https://doi.org/10.21449/ijate.1517994>
- Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56(1), 150–166. <https://doi.org/10.1111/bjet.13494>
- Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P. L. H. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies*, 28(6), 7031–7063. <https://doi.org/10.1007/s10639-022-11473-y>