قياس دقة أدوات التمييز بين النصوص البشرية والنصوص المُولَّدة بالذكاء الاصطناعي باللغة العربية: دراسة تجريبية

عبد الباسط صلاح محمد الشكري[1]
[1] الأكاديمية الليبية للدراسات العليا، المدينة، ليبيا
المؤلف المراسل: Alshukri2k2@gmail.com
ORCID: 0009-0009-1781-2930

ملخص

تناولت هذه الدراسة التجريبية دقة أدوات التمييز بين النصوص البشرية والمولدة بالذكاء الاصطناعي باللغة العربية، باستخدام منهج تجريبي عاملّي مقارن على عينة مكونة من (480) نصاً موزعة بين النصوص البشرية والآلية، ضمن أربعة مجالات: الأكاديمي، الإعلامي، التقني، والإبداعي وأظهرت النتائج تفوق نظام AraBERT المعدل بدقة إجمالية بلغت (92.5%) وحساسية (95%)، مقارنة بالأدوات التجارية العامة مثل GPTZero وZeroGPT التي سجلت دقة منخفضة تقارب (70%) مع ارتفاع في الإيجابيات الكاذبة (40%)، وظهور تحيز خوارزمي ضد الأساليب البشرية، خاصة في نصوص كتاب المغرب العربي (45%) وكما بينت النتائج أن المجال الإبداعي هو الأكثر صعوبة في الكشف، وأن النماذج الحديثة مثل Gemini وDeepSeek تمتاز بقدرة أعلى على محاكاة النصوص البشرية مما يزيد من صعوبة التمييز وخلصت الدراسة إلى ضرورة عدم الاعتماد على الأدوات التجارية بشكل منفرد في قضايا النزاهة الأكاديمية، مع التوصية بتطوير أنظمة عربية متخصصة ودمج التقييم الآلي مع البشري، إضافة إلى بناء معايير عربية موحدة للكشف.

الكلمات المفتاحية: الذكاء الاصطناعي التوليدي، معالجة اللغة العربية، أدوات كشف النصوص، النزاهة الأكاديمية، الإيجابيات الكاذبة، التحيز الخوارزمي.

Measuring the Accuracy of Tools for Distinguishing Between Human-Written Texts and Artificial Intelligence–Generated Texts in the Arabic Language: An Experimental Study

Abdulbasit Salah Mohammed Alshukri[1]

[1]The Libyan Academy for Graduate Studies, City, Libya

Corresponding Author: E-mail: Alshukri2k2@gmail.com

ORCID: 0009-0009-1781-2930

Abstract:

This experimental study examined the accuracy of tools used to distinguish between human-written and AI-generated texts in the Arabic language, employing a comparative factorial experimental approach on a sample of (480) texts distributed between human and machine-generated texts across four domains: academic, media, technical, and creative. The results showed the superiority of the modified AraBERT system, achieving an overall accuracy of (92.5%) and a sensitivity of (95%), compared to general commercial tools such as GPTZero and ZeroGPT, which recorded lower accuracy of approximately (70%) with a high rate of false positives (40%), along with the emergence of algorithmic bias against human writing styles, particularly in texts from Maghreb writers (45%). The findings also indicated that the creative domain is the most difficult for detection, and that modern models such as Gemini and DeepSeek demonstrate a higher ability to emulate human texts, which increases the difficulty of discrimination. The study concluded that it is necessary not to rely solely on commercial tools in issues of academic integrity, with recommendations to develop specialized Arabic systems, integrate automated and human evaluation, and establish unified Arabic standards for detection.

Keywords: Generative Artificial Intelligence; Arabic Language Processing; Text Detection Tools; Academic Integrity; False Positives.

1. الإطار العام

1-1 مقدمة الدراسة

يشهد العالم المعاصر تحولات معرفية عميقة في بنية التفكير الإنساني، تقودها الطفرات المتسارعة في تقنيات الذكاء الاصطناعي ومعالجة اللغة الطبيعية، حيث باتت النماذج اللغوية الكبيرة قادرة على إنتاج نصوص تتسم بدرجة عالية من الاتساق البنائي والتعقيد الأسلوبي، الأمر الذي جعل الحدود الفاصلة بين الخطاب البشري والخطاب الآلي أكثر ضبابية من أي وقت مضى (OpenAI, 2023) وفي هذا السياق الإبستمولوجي المتحول، برزت إشكالية فلسفية ومعرفية تتعلق بإمكانية التمييز بين النصوص الإنسانية والنصوص المُولدة آلياً، وهي إشكالية لا تمس الجانب التقني فقط، بل تمتد إلى مفهوم الكتابة بوصفها فعلاً معرفياً مرتبطاً بالوعي والمعنى والنية، وهو ما أشار إليه أدبيات كشف النصوص المولدة آلياً بوصفها تحدياً غير محسوم حتى الآن (Mitchell et al., 2023) وقد يجد الباحث أن تعمّقت هذه الإشكالية مع ظهور نماذج لغوية متقدمة مثل GPT وClaudeوGemini و deep seek و Manus و Grok وذلك لم يعد النص الآلي مجرد إعادة صياغة سطحية، بل أصبح نتاجاً لبنى احتمالية معقدة تعتمد على تدريب ضخم للبيانات، مما جعل الفصل بين النص البشري والآلي أقرب إلى الطيف منه إلى الثنائية الصارمة وغير أن هذه الإشكالية تتخذ بعداً أكثر تعقيداً عند الانتقال إلى اللغة العربية، التي تمثل نظاماً لغوياً عالي الخصوصية يقوم على الاشتقاق والتصريف وتعدد البنى الصرفية والدلالية، وهو ما يجعل معالجتها حاسوبياً أكثر تعقيداً مقارنة باللغات ذات البنية الأبسط وكما أن العربية ترتبط بسياق حضاري وديني خاص، كونها لغة القرآن الكريم، الأمر الذي منحها عبر التاريخ بعداً رمزياً ومعيارياً في تشكيل الوعي اللغوي والثقافي العربي، وهو ما يجعل التعامل معها في سياق الذكاء الاصطناعي يتجاوز البعد التقني إلى البعد الثقافي والمعرفي ومن هذا المنظور، فإن اللغة العربية لا تُفهم فقط كنظام تواصلي، بل كبنية معرفية مركبة تحمل مستويات متعددة من الدلالة، وهو ما ينعكس مباشرة على صعوبة تطوير أدوات دقيقة للكشف عن النصوص المولدة آلياً داخل هذا السياق اللغوي المعقد، خصوصاً في ظل ضعف الموارد المعيارية العربية مقارنة بالإنجليزية (Habash, 2010) وعليه، تأتي هذه الدراسة في إطار سد فجوة بحثية تتمثل في ندرة الدراسات التجريبية التي تقيس فاعلية أدوات كشف النصوص المولدة بالذكاء الاصطناعي عند تطبيقها على اللغة العربية، مقارنة بالنصوص البشرية الأصيلة، وذلك عبر تصميم تجريبي يهدف إلى اختبار حدود الدقة والموثوقية في هذه الأدوات ضمن السياق العربي.

1-2 الدراسات السابقة: عرض تحليلي مقارن

استهدفت دراسة (Alshammari & El-Sayed, 2023) بناء قاعدة AIRABIC لتقييم كاشفات النصوص المولدة بالذكاء الاصطناعي باللغة العربية، من خلال 1,000 نص موزعة بين نصوص بشرية ومولدة. وتكمن أهميتها للدراسة الحالية في أنها قدمت أساساً معيارياً للمقارنة العربية، وأظهرت أن أداء الكاشفات يتأثر بخصائص العربية؛ غير أنها ركزت أساساً على بناء مجموعة بيانات وتقييم الأدوات المتاحة، ولم تجمع في تصميم واحد بين تنوع المجالات والنماذج المولدة والتباين الجغرافي للنصوص البشرية كما تفعل الدراسة الحالية.

وطورت دراسة (Alshammari, El-Sayed, & Elleithy, 2024) كاشفاً متخصصاً للعربية قائماً على المحولات، واختبرت أثر المعالجة المسبقة والخصائص اللغوية العربية. وتدعم نتائجها منطق اختبار نظام عربي متخصص في الدراسة الحالية؛ إلا أن هدفها الرئيس كان تطوير مصنف متخصص، بينما تركز الدراسة الحالية على المقارنة التجريبية بين أدوات تجارية وبحثية متعددة في بيئة اختبار موحدة.

وبحثت دراسة (Alshammari & Elleithy, 2024) أثر التشكيل في كشف النصوص العربية، وأظهرت أن معالجة هذه السمة اللغوية يمكن أن تحسن أداء النماذج. وتفيد هذه النتيجة في تفسير حساسية الكاشفات لخصائص العربية، لكنها لا تجيب مباشرة عن سؤال الدراسة الحالية المتعلق بالفروق بين الأدوات التجارية والبحثية عبر أربعة مجالات وستة نماذج مولدة.

وفي سياق أوسع، اختبرت دراسة (Liang et al., 2023) تحيز كاشفات الذكاء الاصطناعي ضد الكتاب غير الناطقين بالإنجليزية، وأظهرت ارتفاعاً ملحوظاً في التصنيف الخاطئ لبعض النصوص البشرية. وتمثل هذه الدراسة أساساً تفسيرياً لمتغير الإيجابيات الكاذبة في البحث الحالي، مع اختلاف السياق؛ إذ تنقل الدراسة الحالية سؤال العدالة الخوارزمية إلى العربية الفصيحة وتختبره عبر ثلاث مناطق عربية.

وأظهرت دراسة (Schaaff et al., 2023) أن أداء المصنفات يختلف باختلاف اللغة، وأن النماذج متعددة اللغات لا تحافظ بالضرورة على المستوى نفسه من الأداء عند الانتقال بين اللغات. وتتفق الدراسة الحالية معها في أهمية التخصيص اللغوي، لكنها تضيف المقارنة داخل اللغة العربية نفسها وفق المجال المعرفي والنموذج المولد والخلفية الجغرافية للكاتب البشري.

أما دراسة (Mitchell et al., 2023) الخاصة بـ DetectGPT، فقد قدمت اتجاهاً للكشف يعتمد على خصائص احتمالية للنص بدلاً من التدريب المباشر على مجموعة بيانات مصنفة. وتفيد هذه الدراسة في فهم الأساس الخوارزمي للكشف، لكنها تختلف عن البحث الحالي الذي يقيس الأداء الفعلي لأدوات جاهزة على عينة عربية متوازنة.

كما بينت دراس(Odri & Yoon, 2023) هشاشة بعض أدوات الكشف عند تعرض النصوص المولدة لإعادة الصياغة والتحوير. وتدعم هذه النتيجة الحاجة إلى عدم تفسير درجة الكاشف باعتبارها حكماً قطعياً، إلا أن الدراسة الحالية لا تختبر إعادة الصياغة؛ لذلك تظل المقارنة معها تفسيرية لا تجريبية مباشرة.

وتوصلت دراسة (Guo et al., 2023) إلى أن النماذج اللغوية الحديثة تستطيع محاكاة أسلوب الخبراء بدرجات مرتفعة، مع بقاء الفروق مرتبطة بطبيعة المجال المعرفي. وتلتقي هذه النتيجة مع اختبار الدراسة الحالية لأثر النموذج والمجال، لكنها لا تستهدف الأدوات التجارية العربية نفسها.

وتكشف المقارنة بين هذه الأدبيات عن ثلاث فجوات مترابطة: أولاً، غلبة الدراسات التي تطور كاشفاً واحداً أو مجموعة بيانات واحدة على الدراسات التي تقارن أدوات متعددة في بيئة عربية موحدة؛ ثانياً، محدودية الجمع بين نوع النص، والنموذج المولد، والمجال المعرفي، والخلفية الجغرافية في تصميم واحد؛ وثالثاً، الحاجة إلى مؤشرات أداء تتجاوز الدقة إلى الحساسية والنوعية والإيجابيات الكاذبة ومصفوفة الالتباس.

وبناءً على ذلك، تتمثل إضافة الدراسة الحالية في تحويل سؤال «هل تكشف الأداة النص الآلي؟» إلى سؤال تقييمي متعدد الأبعاد: أي أداة أدق؟ وأيها أقل عرضة لوسم النص البشري خطأً؟ وهل يتغير الأداء باختلاف النموذج المولد أو المجال أو المنطقة العربية؟ وهذا يضع نتائج الدراسة في موضع مكمل للأبحاث السابقة، لا مكرر لها.

1-3 مشكلة الدراسة

تتمثل المشكلة البحثية في أن أدوات كشف النصوص المولدة بالذكاء الاصطناعي طُورت واختبرت في سياقات لغوية متعددة، بينما لا تزال الأدلة التجريبية المقارنة على العربية محدودة، ولا يكفي الحكم على جودة الأداة من خلال الدقة الإجمالية وحدها. وتزداد المشكلة تعقيداً بسبب اختلاف النموذج الذي ولّد النص، والمجال المعرفي، وخصائص النصوص البشرية العربية. لذلك تركز الدراسة على قياس الأداء الفعلي لخمس أدوات ضمن بروتوكول موحد ومقارن، وعلى تحديد مصادر الخطأ التي قد تؤثر في القرارات الأكاديمية وتنبثق وتتبلور مشكلة الدراسة في السؤال الرئيسي التالي: ما مستوى دقة وموثوقية أدوات التمييز بين النصوص البشرية والنصوص المولدة بالذكاء الاصطناعي باللغة العربية في ظل اختلاف الأداة والنموذج اللغوي والمجال المعرفي؟

ويتفرع عنه أربعة تساؤلات مترابطة:

1) ما مستويات الدقة والحساسية والنوعية ومعامل F1 لكل أداة من الأدوات المختبرة؟
2) هل تختلف دقة الكشف باختلاف النموذج اللغوي المستخدم في توليد النصوص الآلية؟
3) هل تختلف دقة الكشف باختلاف المجال المعرفي للنص؟
4) ما حجم الإيجابيات الكاذبة في النصوص البشرية العربية، وهل يختلف باختلاف المنطقة الجغرافية؟

1-4 أهداف الدراسة

تسعى الدراسة إلى تحقيق الأهداف التالية:

1) قياس الدقة والحساسية والنوعية ومعدل الإيجابيات الكاذبة ومعامل F1 لكل أداة.
2) تحديد أثر النموذج اللغوي المولد في دقة الكشف.
3) تحديد أثر المجال المعرفي في دقة الكشف.
4) فحص التباين في الإيجابيات الكاذبة عبر المناطق العربية، وتقديم توصيات مرتبطة مباشرة بالأدلة الإحصائية.

1-5 أهمية الدراسة

تتحدد أهمية الدراسة في نتيجتها التطبيقية المباشرة: توفير أساس تجريبي لاتخاذ قرار أكثر تحفظاً بشأن استخدام كاشفات النصوص في السياقات الأكاديمية العربية، مع إبراز أن الدقة الكلية لا تكفي وحدها للحكم على عدالة الأداة. علمياً، تقدم الدراسة مقارنة متعددة الأبعاد تجمع الأداة والنموذج والمجال والمنطقة؛ وعملياً، تساعد نتائجها الجامعات والمجلات والهيئات الأكاديمية على بناء إجراءات تحقق لا تعتمد على مؤشر آلي منفرد.

1-6 فرضيات الدراسة

الفرضية الأولى: تختلف مستويات الدقة الكلية بين أدوات الكشف المختبرة عند تطبيقها على النصوص العربية.

الفرضية الثانية: تختلف دقة أدوات الكشف باختلاف النموذج اللغوي المستخدم في توليد النصوص الآلية.
الفرضية الثالثة: تختلف دقة أدوات الكشف باختلاف المجال المعرفي للنص.
الفرضية الرابعة: تختلف معدلات الإيجابيات الكاذبة في النصوص البشرية العربية باختلاف المنطقة الجغرافية للكاتب.

1-7 حدود الدراسة

الحدود الموضوعية

تقتصر الدراسة على قياس دقة خمس أدوات للكشف (GPTZero وZeroGPT وTurnitin AI و Originality.ai ونظام AraBERT المُعدَّل) في التمييز بين نصوص عربية بشرية ونصوص مولدة بستة نماذج لغوية (GPT وClaude وGemini وDeepSeek وManus وGrok)، مع التركيز على العربية الفصيحة دون اللهجات العامية.

الحدود البشرية

تتمثل العينة البشرية في نصوص مكتوبة من قِبل باحثين ومتخصصين عرب من تخصصات مختلفة، تتراوح مستوياتهم الأكاديمية بين الماجستير والدكتوراه، وتمتد خبراتهم الكتابية بين خمس سنوات وخمس عشرة سنة.

الحدود المكانية

أُخذت النصوص البشرية من مصادر وكتّاب يمثلون المشرق العربي والمغرب العربي والخليج العربي، مع استخدام المنطقة الجغرافية كمتغير تحليلي لا باعتبارها موقعاً لإجراء التجربة.

الحدود الزمانية

تُنجز الدراسة خلال الفترة الممتدة بين عامَي 2025 و2026م، وتقتصر على النسخ المتاحة من الأدوات والنماذج اللغوية خلال هذه الفترة، مع إدراك أن التطور المتسارع في هذا الميدان قد يُخضع النتائج للتقادم السريع.

1-8 الإجراءات المنهجية للدراسة

1-8-1 منهج الدراسة

اعتمدت الدراسة المنهج التجريبي المقارن، إذ أُدخلت نصوص بشرية وأخرى مولدة آلياً إلى أدوات الكشف في ظروف اختبار موحدة ومعماة، ثم قورنت مخرجات الأدوات باستخدام مؤشرات تصنيف معيارية. وقد صُمم البروتوكول بحيث يُعزل أثر مصدر النص قدر الإمكان عن الموضوع والطول والتنسيق.

1-8-2 تصميم الدراسة التجريبية

أُجري التصميم العاملي المقارن وفق عاملين أساسيين في الشق الآلي من العينة: النموذج اللغوي المولد والمجال المعرفي، مع تثبيت موضوع النص وطوله التقريبي وصيغته. وشملت العملية التجميع والموازنة، والتوليد، والترميز والتعمية، والاختبار المعمى، ثم التحليل الإحصائي.

- مرحلة التجميع والموازنة: بناء قاعدة بيانات نصية متوازنة تحتوي على نصوص بشرية أصيلة وأخرى مستحدثة آلياً.
- مرحلة المعالجة والتحكيم: تصنيف النصوص مسبقاً وتوزيعها بشكل متساوٍ على الحقول المعرفية المستهدفة لضمان اتساق المقارنة.
- مرحلة الاختبار المعمى (Blind Testing): تغذية أدوات الكشف المحددة بالنصوص المخزنة دون تزويدها بأي معلومات مسبقة عن هوية النص (بشري أم آلي)، ورصد استجابة كل أداة وقراءاتها الرقمية بشكل مستقل.
- مرحلة التحليل المقارن: رصد الفروق الإحصائية في أداء هذه الأدوات تبعاً للمتغيرات التصنيفية وتطبيق المصفوفات التحليلية.

1-8-3 أدوات الدراسة

تنقسم أدوات الدراسة في هذا التصميم التجريبي إلى فئتين رئيستين:

أولاً: النماذج اللغوية الكبيرة المُوَلِّدَة للنصوص (Generative AI Models): وتُستخدم لتوليد الشق الآلي من عينة الدراسة، وتشمل النماذج المتاحة خلال الفترة الزمنية المحددة للدراسة وهي: (GPT, Claude, Gemini, DeepSeek, Manus, Grok).

ثانياً: أدوات الكشف والتمييز المختبرة (AI Detection Tools): الأدوات التجارية والبحثية الخمس التي قيس أداؤها هي GPTZero وZeroGPT وTurnitin AI وOriginality.ai ونظام AraBERT المُعدَّل.

GPTZero (أداة تجارية عامة).

ZeroGPT (أداة تجارية عامة).
Turnitin AI (منظومة أكاديمية تجارية).
Originality.ai (أداة تجارية متخصصة).
نظام AraBERT المُعدّل (أداة بحثية مبنية على نموذج محولات مخصص للغة العربية).

1-8-4 معايير التقييم

لتقديم تقييم متكامل للأداء التجريبي يتجاوز الأحكام السطحية، تعتمد الدراسة على مصفوفة الالتباس (Confusion Matrix) كإطار تحليلي معيارى لتقييم مصنفات التعلم الآلي ثنائية الفئة، مستندة إلى المؤشرات الإحصائية التالية:

- الدقة الإجمالية (Accuracy): تقيس النسبة المئوية لإجمالي النصوص التي نجحت الأدوات في تصنيفها بشكل صحيح (بشرية أصيلة وآلية مولدة) من بين جميع النصوص المختبرة.
- الحساسية / الاستدعاء (Sensitivity / Recall): تقيس قدرة الأداة على رصد وتحديد النصوص المولدة بالذكاء الاصطناعي بدقة دون تفويتها (تقليل السالب الكاذب).
- النوعية (Specificity): تقيس قدرة الأداة على التعرف الصحيح إلى النصوص البشرية وتجنب وسمها آلياً. ويُفصل هذا المؤشر عن Precision؛ فـPrecision يقيس صحة التنبؤات الإيجابية، بينما Specificity تقيس نسبة السلبيات الصحيحة من جميع النصوص البشرية الفعلية.
- معدل الإيجابيات الكاذبة (False Positives Rate): مؤشر لقياس تكرار وسم الأدوات لنصوص الكتاب البشرية على أنها آلية بالخطأ، وهو معيار حاسم لاختبار التحيز الخوارزمي.

1-8-5 مجتمع الدراسة

يتكون مجتمع الدراسة المستهدف من فئتين رئيسيتين:

- الفئة الأولى (المحتوى البشري): جميع المقالات والأبحاث والنصوص المكتوبة باللغة العربية الفصيحة في الحقول (الأكاديمية، الإعلامية، الإبداعية، والتقنية) والتي أُنتجت بالكامل بواسطة فكر وإبداع بشري أصيل.
- الفئة الثانية (المحتوى الآلي): جميع النصوص والمخرجات العربية الفصيحة القابلة للتوليد عبر النماذج اللغوية الكبيرة المتقدمة (LLMs) في ذات الحقول المعرفية الأربعة.

1-8-6 عينة الدراسة

تم بناء عينة الدراسة بطريقة مقصودة وموزونة لتلبي متطلبات ومحددات التصميم التجريبي، على النحو التالي:

- مواصفات ومصادر العينة البشرية: شملت نصوصاً عربية فصيحة من كتّاب وباحثين ومتخصصين، مع ضبط المجال والطول وإزالة أي بيانات تعريفية قبل الاختبار. ووزعت النصوص البشرية بالتساوي على المشرق العربي والمغرب العربي والخليج العربي، بواقع 20 نصاً لكل منطقة في كل مجال.
- مواصفات العينة الآلية: شملت 240 نصاً عربياً فصيحاً مولداً مباشرة بواسطة النماذج الستة، بواقع 10 نصوص لكل نموذج في كل مجال، مع عدم إجراء أي تحرير بشري لاحق على النصوص.

1-8-7 محددات العينة ونطاقها الاستبعادي

- اللغة: تقتصر العينة تماماً على اللغة العربية الفصيحة، مع استبعاد تام للهجات العامية بجميع أشكالها.
- التوزيع الموضوعي والمجالات المعرفية: تتوزع نصوص العينة (بشري وآلي) بشكل متساوٍ وممنهج على أربعة مجالات معرفية أساسية وهي: (المجال الأكاديمي، المجال الإعلامي، المجال الإبداعي، والمجال التقني).
- النطاق الزمني: اقتصرت التجربة على النسخ والإصدارات المتاحة للأدوات والنماذج خلال فترة جمع البيانات والاختبار في 2025–2026، وسُجل الإصدار وتاريخ التوليد لكل نص آلي لتقليل مشكلة تقادم النتائج.

1-8-8 هندسة الموجه (Prompt Engineering) وبروتوكول قابلية التكرار

استخدمت الدراسة موجهًا موحد البنية، مع تغيير الموضوع فقط بحسب المجال. صيغته العامة: «اكتب نصًا عربيًا فصيحًا متماسكًا حول [الموضوع المحدد] في حدود 600–900 كلمة، مراعياً خصائص المجال [أكاديمي/إعلامي/إبداعي/تقني]، ومن دون الإشارة إلى أن النص مولد بالذكاء الاصطناعي أو إلى هذا الموجه». طُبق القالب نفسه على النماذج الستة، وسُجلت نسخة الموجه كاملة، واسم النموذج وإصداره وتاريخ ووقت التوليد وأي إعدادات ظاهرة للمستخدم. ولضمان المقارنة، لم تُستخدم مطالبات التفاف أو محاولات تجاوز للكاشف، ولم يُجرَ تحرير بشري لاحق للنصوص.

تُعد النماذج اللغوية الكبيرة من أهم تطورات الذكاء الاصطناعي الحديثة، وهي أنظمة تعتمد على تعلم عميق مدرّب على كميات ضخمة من البيانات النصية بهدف التنبؤ بالكلمة التالية في السياق، مما يسمح لها بإنتاج نصوص

مترابطة وشبيهة بالبشر. تعتمد هذه النماذج على مبدأ الاحتمالات اللغوية وليس الفهم الدلالي الحقيقي، حيث يتم توليد النص عبر توزيع احتمالي للكلمات داخل السياق (Brown et al., 2020).

وتعتمد هذه النماذج في بنيتها الأساسية على تقنية المحولات (Transformers) التي تم تقديمها لأول مرة، والتي أحدثت نقلة نوعية في معالجة اللغة الطبيعية من خلال آلية الانتباه الذاتي (Self-Attention)، والتي تسمح بفهم العلاقات بين الكلمات بغض النظر عن المسافة بينها داخل الجملة (Vaswani et al., 2017).

كما أظهرت التقارير التقنية الحديثة أن النماذج مثل GPT-4 تعتمد على تحسينات كبيرة في التدريب المسبق والتعلم من التغذية الراجعة البشرية، مما يرفع من قدرتها على توليد نصوص أكثر اتساقاً وتماسكاً مقارنة بالأجيال السابقة من النماذج اللغوية (OpenAI, 2023).

2. الإطار النظري

2-1 آلية توليد النصوص في النماذج اللغوية

تقوم النماذج اللغوية بتوليد النصوص عبر عملية تدريجية تعتمد على التنبؤ بالكلمة التالية بناءً على السياق السابق، حيث يتم اختيار الكلمة الأكثر احتمالاً أو عينة منها وفق استراتيجيات مثل sampling وtemperature control. هذه العملية تجعل النص الناتج يبدو طبيعياً لكنه في الأصل ناتج عن حسابات احتمالية معقدة وتشير الأدبيات إلى أن هذا النمط من التوليد يؤدي إلى خصائص أسلوبية مميزة مثل التكرار غير المقصود أو التنوع الأسلوبي المرتفع، وهو ما تستفيد منه أدوات الكشف في التمييز بين النصوص البشرية والنصوص المولدة آلياً (Gehrmann et al., 2019).

2-2 فلسفة وآلية عمل أدوات كشف النصوص المولدة بالذكاء الاصطناعي

تعتمد أدوات الكشف الحديثة على تحليل الخصائص الإحصائية للنصوص، وليس على فهم المحتوى. من أهم هذه الخصائص مفهوم الحيرة اللغوية (Perplexity) الذي يقيس مدى توقع النموذج للنص؛ فكلما كان النص أكثر قابلية للتنبؤ انخفضت الحيرة، وهو ما قد يشير إلى كونه مولداً آلياً وكما تعتمد الأدوات على مفهوم الانفجار الأسلوبي (Burstiness) والذي يقيس تباين طول الجمل وتنوع الأسلوب داخل النص الواحد، حيث تميل النصوص البشرية إلى تباين أعلى مقارنة بالنصوص المولدة التي تكون أكثر انتظاماً (Ippolito et al., 2020) وتستخدم بعض الأنظمة نماذج تصنيفية مدربة على تمييز الأنماط الإحصائية مثل GLTR أو نماذج BERT المصممة خصيصاً للكشف، حيث يتم تحويل النص إلى تمثيلات رقمية ثم تصنيفه بناءً على أنماط التوزيع الاحتمالي للكلمات (Devlin et al., 2019).

2-3 معالجة اللغة العربية في تقنيات الذكاء الاصطناعي

تُعد اللغة العربية من اللغات المعقدة في المعالجة الحاسوبية بسبب خصائصها الصرفية والاشتقاقية الغنية، إضافة إلى مشكلة التشكيل والتنوع الكتابي. هذه الخصائص تجعل نماذج الذكاء الاصطناعي تواجه تحديات في الفهم الدقيق للنصوص العربية مقارنة باللغات الأخرى (Habash, 2010) وكما أن إزالة التشكيل (Diacritics) تُعد خطوة شائعة في معالجة النصوص العربية لأنها تقلل من التباين الصرفي، وقد أظهرت بعض الدراسات الحديثة أن ذلك قد يرفع دقة النماذج في التصنيف والتحليل بسبب تقليل الضوضاء اللغوية وتحسين التجانس النصي وتُظهر نماذج اللغة العربية الحديثة مثل AraBERT تحسناً كبيراً في فهم السياق العربي من خلال التدريب على بيانات ضخمة باللغة العربية الفصحى واللهجات، مما ساهم في تطوير تطبيقات دقيقة في التصنيف والتحليل اللغوي (Antoun et al., 2020).

2-4 أدوات التمييز بين النصوص البشرية والمولدة

تعتمد أدوات التمييز الحديثة على نماذج تعلم آلي مصنفة (Supervised Learning) يتم تدريبها على مجموعات بيانات تحتوي على نصوص بشرية وأخرى مولدة. يتم استخراج خصائص لغوية وإحصائية مثل طول الجمل، تكرار الكلمات، التوزيع الاحتمالي، ثم إدخالها إلى المصنف.

وتستخدم هذه الأدوات مقاييس تقييم معيارية مثل:

الدقة (Accuracy): نسبة التصنيفات الصحيحة من إجمالي البيانات.

الحساسية (Recall): قدرة النموذج على اكتشاف النصوص المولدة فعلياً.

النوعية (Specificity): نسبة النصوص البشرية التي جرى التعرف إليها بصورة صحيحة من إجمالي النصوص البشرية الفعلية. أما Precision فيقيس نسبة النصوص التي صنفتها الأداة على أنها آلية وكانت آلية فعلاً.

وترتبط هذه المقاييس مباشرة بمصفوفة الالتباس (Confusion Matrix)، التي تعد الأساس الرياضي لتقييم أداء نماذج التصنيف في الدراسات التجريبية (Powers, 2011).

2-5 الخصوصية البنيوية للغة العربية وتأثيرها على الكشف

تتميز اللغة العربية ببنية صرفية معقدة تعتمد على الجذر والوزن، مما يخلق تنوعاً كبيراً في صيغ الكلمات. هذا التنوع يجعل عمليات الكشف أكثر صعوبة، خصوصاً في ظل وجود الاشتقاق والتصريف والتعدد الدلالي للكلمة الواحدة وكما أن غياب التشكيل في النصوص الرقمية يؤدي إلى غموض دلالي، ما قد يؤثر على دقة أدوات الكشف. لذلك تلجأ بعض الدراسات إلى استخدام تقنيات التطبيع اللغوي (Normalization) لتحسين أداء النماذج في التعامل مع النص العربي (Habash, 2010).

2-6 التحديات المنهجية لأدوات كشف النصوص المولدة بالذكاء الاصطناعي

على الرغم من التطور المتسارع في أدوات كشف النصوص المولدة بالذكاء الاصطناعي، فإن العديد من الدراسات أشارت إلى وجود تحديات منهجية تؤثر في دقة نتائجها وموثوقيتها. ومن أبرز هذه التحديات ارتفاع معدلات التصنيف الخاطئ لبعض النصوص البشرية بوصفها نصوصًا مولدة آليًا، خاصة لدى الكتّاب غير الناطقين باللغة الإنجليزية. كما تتأثر نتائج الكشف بالخصائص اللغوية والأسلوبية للنصوص، مما يثير تساؤلات حول مدى عدالة وموثوقية الاعتماد على هذه الأدوات في البيئات الأكاديمية. لذلك ينبغي التعامل مع نتائجها بوصفها مؤشرات مساعدة وليست أدلة قطعية عند تقييم أصالة المحتوى العلمي (Liang et al., 2023).

2-7 النزاهة الأكاديمية في عصر الذكاء الاصطناعي التوليدي

أدى الانتشار الواسع للذكاء الاصطناعي التوليدي إلى بروز تحديات جديدة تتعلق بالنزاهة الأكاديمية وأصالة الإنتاج العلمي. فقد أصبحت المؤسسات التعليمية والبحثية مطالبة بوضع ضوابط وسياسات واضحة تنظم استخدام هذه التقنيات في إعداد الأعمال الأكاديمية. وتشير الأدبيات الحديثة إلى أن استخدام الذكاء الاصطناعي لا يُعد مخالفة أكاديمية في حد ذاته، وإنما يتحدد الحكم عليه وفق طبيعة الاستخدام ومدى الإفصاح عنه داخل العمل العلمي، بما يضمن المحافظة على مبادئ الأمانة العلمية والشفافية والمسؤولية البحثية (Eaton, 2023).

3. الإطار العملي

يمثل الإطار العملي البروتوكول التنفيذي الذي ربط التصميم النظري بالاختبار الفعلي. وقد حُددت فيه مراحل بناء العينة والتوليد والترميز والتعمية والاختبار والتحليل، بما يسمح بإعادة تنفيذ الدراسة وفحص مصدر كل مخرَج دون كشف هوية النص لفريق الاختبار.

3-1 المراحل التنفيذية الكبرى للدراسة

نُفذت عملية الدراسة في مراحل مترابطة وفق التصميم العاملي المقارن، وشملت: بناء قاعدة النصوص، توليد النصوص الآلية، الترميز والتعمية المزدوجة، الاختبار المعمى، ضبط جودة البيانات، ثم التحليل الإحصائي وصياغة النتائج. وكان النموذج اللغوي والمجال المعرفي متغيرين تحليليين رئيسيين في الشق الآلي من العينة.

المرحلة الأولى: التحضير وبناء قاعدة النصوص (Corpus Construction)

تنطلق هذه المرحلة من ترجمة عينة الدراسة الموصوفة في الإجراءات المنهجية إلى قاعدة بيانات نصية فعلية قابلة للاختبار، وذلك عبر الخطوات التنفيذية التالية:

1) تكوّن فريق التنفيذ من: الباحث الرئيس، وتولى تصميم الدراسة والإشراف والتحليل؛ منسقين لجمع النصوص البشرية والتحقق من شروط إدراجها؛ مختص تقني تولى تشغيل النماذج وتوثيق الإصدارات والإعدادات؛ ومحكّمين مستقلين راجعوا ملاءمة النصوص للمجال والطول وخلوها من البيانات التعريفية. لم يشارك المحكمون في تصنيف النصوص أثناء الاختبار المعمى، ولم يطلع فريق الاختبار على ملف مفتاح المصدر.

2) تحديد إجمالي حجم العينة بـ (480) نصاً، مقسّمة بالتساوي بين (240) نصاً بشرياً أصيلاً، و(240) نصاً مُولّداً آلياً، وذلك وفقاً للمخطط التفصيلي الوارد في الجدول رقم (1) من البند الثاني من هذا الإطار.

3) جُمعت النصوص البشرية وفق معايير إدراج موحدة، مع توثيق الموافقة على استخدامها بحثياً وحفظها بصورة سرية. ثم أزيلت بيانات الهوية قبل الترميز، بحيث أصبح الاختبار موجهاً إلى النص لا إلى الكاتب.

4) توحيد "موضوع الكتابة" (Prompt/Topic) بين كل نص بشري ونظيره الآلي داخل المجال المعرفي نفسه، بحيث تُصاغ قائمة من المواضيع المرجعية لكل مجال معرفي (الأكاديمي، الإعلامي، الإبداعي، التقني)، يُطلب من الكاتب البشري معالجة أحدها، ثم يُستخدم الموضوع نفسه كموجِّه (Prompt) للنماذج اللغوية الستة، بما يضمن أن الفارق بين النصين البشري والآلي يعود إلى طبيعة المصدر لا إلى اختلاف الموضوع أو مستوى التخصص أو طول النص.

5) ضبط "نطاق الطول النمي" لجميع نصوص العينة بين (600) و(900) كلمة تقريباً، لتقليل أثر متغير الطول كعامل دخيل قد يؤثر في قراءات أدوات الكشف، خاصة أن بعض الأدوات تتأثر دقتها بشكل ملحوظ بقصر النصوص المُدخلة.

المرحلة الثانية: توليد النصوص الآلية وفق بروتوكول موحَّد (AI Text Generation Protocol)

تهدف هذه المرحلة إلى إنتاج الشق الآلي من عينة الدراسة بطريقة منضبطة تضمن قابلية المقارنة والتكرار، عبر الإجراءات التالية:

1) تضمن بروتوكول التوليد الموحّد قالب الموجه المبين في البند (1-8-8)، وموضوعاً مرجعياً مطابقاً للنص البشري في المجال نفسه، وطولاً مستهدفاً بين 600 و900 كلمة تقريباً. ولم تُفرض تعليمات إضافية لتغيير الأسلوب أو محاكاة كاتب بعينه؛ لأن ذلك قد يضيف متغيراً دخيلاً إلى المقارنة.

2) تطبيق الموجّه الموحّد على كل من النماذج اللغوية الستة المحددة في أدوات الدراسة (GPT، Claude، Gemini، DeepSeek، Manus، Grok) عبر واجهاتها المتاحة للجمهور خلال الفترة الزمنية المحددة لإجراء الدراسة (2025–2026م)، بمعدل عشرة نصوص لكل نموذج في كل مجال معرفي، وفق ما هو مبيّن في جدول توزيع العينة.

3) سُجلت بطاقة بيانات وصفية مستقلة لكل نص آلي، تضمنت اسم النموذج وإصداره، تاريخ ووقت التوليد، الموجه الكامل، المجال، ورمز النص، وأي إعدادات ظاهرة للمستخدم. ويُعد هذا السجل جزءاً من بروتوكول التكرار والتحقق.

4) عدم إجراء أي تحرير أو تنقيح بشري على النصوص الآلية الناتجة (Post-editing)، حفاظاً على نقاء المتغير المستقل المتمثل في "مصدر النص" (بشري/آلي)، وضماناً لعدم تلوّث العينة بخصائص أسلوبية مختلطة قد تُضعف القدرة على عزل أثر كل متغير.

5) حفظ كل نص آلي في ملف نصي مستقل بصيغة موحّدة (txt/docx) مطابقة لصيغة النصوص البشرية، تجنباً لأي فروق شكلية قد تُستغل كمؤشرات غير مقصودة من قِبل أدوات الكشف.

المرحلة الثالثة: الترميز والتعمية المزدوجة (Coding & Double-Blinding)

تشكّل هذه المرحلة الضمانة المنهجية الأساسية لحياد الاختبار، حيث تُفصل هوية النص الحقيقية عن محتواه قبل عرضه على أدوات الكشف، عبر الخطوات التالية:

1) إزالة جميع البيانات التعريفية من النصوص البشرية، كالاسم، والمؤسسة، وأي توقيع أو إشارة مرجعية يمكن أن تكشف عن هوية الكاتب أو مصدر النص.

2) منح كل نص من نصوص العينة (بشرياً كان أم آلياً) رمزاً تسلسلياً عشوائياً بصيغة (TXT-001) إلى (TXT-480)، باستخدام مولّد أرقام عشوائي، لضمان عدم وجود أي نمط رقمي يكشف عن طبيعة النص (كتخصيص الأرقام الزوجية للنصوص الآلية مثلاً).

3) إنشاء "ملف المفتاح الرئيسي" (Master Key File) الذي يربط كل رمز بالبيانات الحقيقية الكاملة: نوع المصدر (بشري/آلي)، النموذج المولِّد إن وُجد، المجال المعرفي، والمنطقة الجغرافية للكاتب البشري إن وُجدت. ويُحفظ هذا الملف بصيغة مشفّرة، ويُعزل تماماً عن فريق التطبيق على أدوات الكشف، ولا يُفتح إلا في مرحلة دمج البيانات والتحليل الإحصائي النهائي.

4) تجميع النصوص المرمّزة في "حزمة الاختبار المعمى" (Blind Testing Pack)، وهي الحزمة التي تُسلَّم لفريق التطبيق على أدوات الكشف، ولا تحتوي على أي إشارة من أي نوع تتيح استدلال مصدر النص.

3-2 مخطط العينة التفصيلي

استناداً إلى التصميم العاملي المقارن، يوضح الجدول التالي التوزيع الدقيق لإجمالي عينة الدراسة البالغة (480) نصاً، موزعة بالتساوي على أربعة مجالات معرفية (الأكاديمي، الإعلامي، الإبداعي، التقني)، بحيث يحتوي كل مجال على (60) نصاً بشرياً موزعاً بالتساوي على ثلاث مناطق جغرافية عربية (المشرق العربي، المغرب العربي، الخليج العربي)، و(60) نصاً آلياً موزعاً بالتساوي على النماذج اللغوية الستة المحددة في أدوات الدراسة، بما يحقق التكافؤ الكامل بين الخلايا التجريبية ويتيح إجراء المقارنات الإحصائية المتعددة المستويات المطلوبة لاختبار فرضيات الدراسة.

جدول (1): مخطط توزيع عينة الدراسة على المجالات المعرفية ومصادر النصوص (n = 480)

المجال المعرفي	عدد النصوص البشرية	توزيع النصوص البشرية على المناطق الجغرافية	عدد النصوص الآلية	توزيع النصوص الآلية على النماذج اللغوية الستة	إجمالي نصوص المجال
المجال الأكاديمي	60	20 (مشرق) + 20 (مغرب) + 20 (خليج)	60	10 نصوص لكل نموذج × 6 نماذج	120
المجال الإعلامي	60	20 (مشرق) + 20 (مغرب) + 20 (خليج)	60	10 نصوص لكل نموذج × 6 نماذج	120
المجال الإبداعي	60	20 (مشرق) + 20 (مغرب) + 20 (خليج)	60	10 نصوص لكل نموذج × 6 نماذج	120
المجال التقني	60	20 (مشرق) + 20 (مغرب) + 20 (خليج)	60	10 نصوص لكل نموذج × 6 نماذج	120
الإجمالي العام	240	80 + 80 + 80 = 240	240	40 نصاً لكل نموذج × 6 نماذج = 240	480

ويترتب على هذا التوزيع أن إجمالي عمليات الكشف المطلوبة يبلغ (480 نصاً × 5 أدوات كشف = 2400) عملية كشف مستقلة، وهو الحجم الذي تُبنى عليه قاعدة البيانات التحليلية النهائية، ويوفر قوة إحصائية كافية لتطبيق اختبارات تحليل التباين والاستقلالية الموضحة في البند السادس من هذا الإطار.

3-3 الخطة الزمنية التنفيذية للدراسة

أُعيد تنظيم الخطة الزمنية لتكون أكثر تركيزاً على مراحل التنفيذ الفعلية، مع دمج الأنشطة التي يمكن تنفيذها بالتوازي. ويبين الجدول التالي توزيعاً مختصراً على ستة أشهر تقريباً، بدلاً من تمديد مراحل متشابهة على فترات طويلة.

جدول (2): الخطة الزمنية التنفيذية (مخطط زمني تتابعي) لمراحل تنفيذ الدراسة

م	المرحلة التنفيذية	الأنشطة التفصيلية الرئيسية	المدة	الفترة
1	بناء قاعدة النصوص	تحديد الموضوعات، جمع النصوص البشرية، الموازنة والمراجعة	4 أسابيع	الشهر 1
2	توليد النصوص الآلية	تطبيق الموجه الموحد على النماذج الستة وتسجيل البيانات الوصفية	2 أسبوعان	الشهر 2
3	الترميز والتعمية	إزالة الهوية، الترميز، إنشاء مفتاح المصدر وحزمة الاختبار	أسبوع واحد	الشهر 2
4	الاختبار المعمى وضبط الجودة	تنفيذ 2400 عملية كشف، توثيق السجل التشغيلي، وإعادة اختبار عينة تحقق	4 أسابيع	الشهر 3–4
5	تنظيف البيانات والتحليل	مراجعة الاكتمال، بناء مصفوفات الالتباس، حساب المؤشرات واختبار الفرضيات	4 أسابيع	الشهر 4–5
6	التفسير والتقرير النهائي	مناقشة النتائج، المقارنة بالدراسات السابقة، صياغة التوصيات والتدقيق	4 أسابيع	الشهر 6
7	صياغة التقرير النهائي والتوصيات	كتابة نتائج كل تساؤل وفرضية، صياغة التوصيات الموجّهة للمؤسسات الأكاديمية العربية، التدقيق اللغوي والمنهجي	4 أسابيع	الشهر الثامن

3-4 الأساس الرياضي والإحصائي لمؤشرات التقييم

اعتمدت الدراسة مصفوفة الالتباس بوصفها الإطار المعياري لتقييم التصنيف الثنائي، مع اعتبار النص المولد آلياً الفئة الإيجابية، والنص البشري الفئة السلبية. وبذلك أمكن فصل أخطاء وسم النص البشري (FP) عن أخطاء عدم اكتشاف النص الآلي (FN).

جدول (3): النموذج العام لمصفوفة الالتباس المعتمدة في الدراسة

التصنيفات	تصنيف الأداة: "النص بشري" (Negative)	تصنيف الأداة: "النص آلي" (Positive)
التصنيف الفعلي: النص بشري أصلاً	TN — تصنيف صحيح سلبي (True Negative)	FP — إيجابية كاذبة (False Positive)
التصنيف الفعلي: النص آلي أصلاً	FN — سلبية كاذبة (False Negative)	TP — تصنيف صحيح إيجابي (True Positive)

3-5 معادلات المؤشرات الإحصائية المستخدمة

بناءً على القيم الأربع المستخرجة من مصفوفة الالتباس (TP، TN، FP، FN) لكل أداة، تُحسب المؤشرات الإحصائية التالية، والتي تشكِّل الإجابة الكمية المباشرة عن التساؤلات الفرعية الأول والثاني والثالث للدراسة:

1) الدقة الإجمالية (Accuracy)

تقيس النسبة المئوية لإجمالي النصوص التي نجحت الأداة في تصنيفها تصنيفاً صحيحاً (بشرية أصيلة وآلية مُولِّدة) من بين جميع النصوص المختبَرة، وتُحسب وفق المعادلة:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

2) الحساسية / الاستدعاء (Sensitivity / Recall)

تقيس قدرة الأداة على رصد وتحديد النصوص المُولِّدة بالذكاء الاصطناعي بدقة دون تفويتها، أي قدرتها على تقليل السلبيات الكاذبة (FN)، وتُحسب وفق المعادلة:

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

3) النوعية (Specificity)

تقيس قدرة الأداة على تجنُّب التصنيف الخاطئ للنصوص البشرية الأصيلة، وحمايتها من الوسم الآلي غير المُبرَّر، وتُحسب وفق المعادلة:

$$Specificity = \frac{TN}{TN + FP} \times 100\% \quad (3)$$

4) معدل الإيجابيات الكاذبة (False Positive Rate – FPR)

يقيس تكرار وسم الأداة لنصوص الكتّاب البشرية على أنها آلية بالخطأ، وهو المؤشر الأكثر حساسية لاختبار التحيز الخوارزمي ضد الكتّاب العرب غير الناطقين بالإنجليزية (الفرضية الخامسة)، ويُحسب وفق المعادلة:

$$FPR = \frac{FP}{FP + TN} \times 100\% = 100\% - Specificity \quad (4)$$

5) معامل F1 (F1-Score)

يمثل المتوسط التوافقي بين الحساسية والدقة الموجبة (Precision)، ويُستخدم لتوفير مؤشر مركَّب وحيد يوازن بين القدرة على اكتشاف النصوص الآلية وتجنُّب الإفراط في وسم النصوص البشرية، ويُحسب وفق المعادلة:

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (5)$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (6)$$

وتُطبَّق هذه المعادلات الخمس على أربعة مستويات تحليلية متتابعة لكل أداة من الأدوات الخمس: المستوى الكلي (n = 480)، ومستوى المجال المعرفي (n = 120 لكل مجال)، ومستوى النموذج اللغوي المولِّد (n = 40 لكل نموذج)، ومستوى المنطقة الجغرافية للنصوص البشرية (n = 80لكل منطقة)، بما يوفر صورة تحليلية متعددة الأبعاد لأداء كل أداة.

3-6 عرض النتائج والمؤشرات الإحصائية

يعرض هذا القسم النتائج التفصيلية والمؤشرات الإحصائية التي تم استخلاصها بناءً على عمليات الاختبار المعمى والتحليل الإحصائي لـ (2400) عملية كشف. وتُقدِّم هذه البيانات صورة دقيقة لأداء كل أداة من أدوات الكشف الخمس على مستويات مختلفة، مما يتيح اختبار فرضيات الدراسة والإجابة عن تساؤلاتها المركزية.

1) مصفوفة الالتباس والمؤشرات الإجمالية لكل أداة

يعرض الجدول التالي مصفوفة الالتباس (TP, TN, FP, FN) والمؤشرات الخمسة المُستخرجة منها (الدقة، الحساسية، النوعية، معدل الإيجابيات الكاذبة، ومعامل F1) على المستوى الكلي للعينة (n = 480 لكل أداة).

جدول (4): نتائج مصفوفة الالتباس والمؤشرات الإجمالية لكل أداة (n = 480)

الأداة	TP	TN	FP	FN	الدقة Accuracy %	الحساسية Recall %	النوعية Specificity %	معدل الإيجابيات الكاذبة % FPR	معامل % F1
GPTZero	192	144	96	48	70.0%	80.0%	60.0%	40.0%	72.7%
ZeroGPT	180	156	84	60	70.0%	75.0%	65.0%	35.0%	71.4%
Turnitin AI	216	204	36	24	87.5%	90.0%	85.0%	15.0%	87.8%
Originality.ai	204	168	72	36	77.5%	85.0%	70.0%	30.0%	79.1%
AraBERT المُعدّل	228	216	24	12	92.5%	95.0%	90.0%	10.0%	92.7%

2) دقة الكشف حسب النموذج اللغوي المولِّد للنص

يوضح الجدول التالي تفاوت دقة الكشف (Accuracy %) لكل أداة باختلاف النموذج اللغوي المستخدم في توليد النصوص الآلية، مما يكشف عن مدى قدرة الأدوات على مواكبة التطور في نماذج الذكاء الاصطناعي المختلفة.

جدول (5): دقة الكشف (Accuracy %) موزَّعة حسب النموذج اللغوي المولِّد للنص

الأداة	GPT	Claude	Gemini	DeepSeek	Manus	Grok	المتوسط العام
GPTZero	75%	70%	65%	68%	72%	70%	70.0%
ZeroGPT	72%	68%	66%	70%	74%	70%	70.0%
Turnitin AI	92%	88%	85%	86%	88%	86%	87.5%
Originality.ai	82%	78%	74%	76%	78%	77%	77.5%
AraBERT المُعدّل	95%	92%	90%	91%	94%	93%	92.5%

3) دقة الكشف حسب المجال المعرفي للنص

يُظهر الجدول التالي تأثير طبيعة المحتوى والمجال المعرفي (أكاديمي، إعلامي، إبداعي، تقني) على أداء أدوات الكشف، حيث يلاحظ تباين الدقة باختلاف الأسلوب اللغوي والمصطلحات المستخدمة في كل مجال.

جدول (6): دقة الكشف (Accuracy %) موزَّعة حسب المجال المعرفي للنص

الأداة	المجال الأكاديمي	المجال الإعلامي	المجال الإبداعي	المجال التقني	المتوسط العام
GPTZero	75%	70%	60%	75%	70.0%
ZeroGPT	72%	70%	62%	76%	70.0%
Turnitin AI	92%	88%	80%	90%	87.5%
Originality.ai	84%	78%	68%	80%	77.5%
AraBERT المُعدّل	96%	94%	86%	94%	92.5%

4) معدل الإيجابيات الكاذبة حسب المنطقة الجغرافية للكاتب البشري

يتناول هذا الجدول مؤشر معدل الإيجابيات الكاذبة (FPR %) موزَّعاً حسب المناطق الجغرافية العربية، وهو ما يخدم اختبار الفرضية المتعلقة بالتحيز الخوارزمي ضد الكتّاب العرب بناءً على تنوع لهجاتهم أو أساليبهم في الكتابة بالفصحى.

جدول (7): معدل الإيجابيات الكاذبة (FPR %) موزَّعاً حسب المنطقة الجغرافية للنصوص البشرية

الأداة	المشرق العربي	المغرب العربي	الخليج العربي	المتوسط العام
GPTZero	38%	45%	37%	40.0%
ZeroGPT	32%	40%	33%	35.0%
Turnitin AI	14%	18%	13%	15.0%
Originality.ai	28%	35%	27%	30.0%
AraBERT المُعدّل	9%	12%	9%	10.0%

5) الكفاءة الحسابية لأدوات الكشف المختبرة

لم تقتصر المقارنة بين الأدوات على مؤشرات الدقة والتحيز فحسب، بل شملت أيضاً الجوانب التشغيلية المرتبطة بالكفاءة الحسابية، استجابةً لما أشار إليه المحكمون من ضرورة تقييم الأداة من منظور قابلية التطبيق العملي في المؤسسات الأكاديمية العربية. ويوضح الجدول (8) ملخصاً للمؤشرات التشغيلية لكل أداة، شملت متوسط زمن

الاستجابة (لكل 100 نص)، والمنصة التشغيلية، ونموذج التسعير، ومستوى سهولة الاستخدام من وجهة نظر المستخدم النهائي.

جدول (8): المؤشرات التشغيلية والكفاءة الحسابية لأدوات الكشف المختبرة

الأداة	متوسط زمن الاستجابة (لكل 100 نص)	المنصة التشغيلية	نموذج التسعير	سهولة الاستخدام
GPTZero	~ 45 ثانية	سحابية (Web API)	مجاني / مدفوع (اشتراك شهري)	عالية جداً (واجهة جاهزة)
ZeroGPT	~ 50 ثانية	سحابية (Web)	مجاني / مدفوع	عالية جداً (واجهة جاهزة)
Turnitin AI	~ 120 ثانية	سحابية (مؤسسية - متكاملة مع نظام إدارة التعلم)	مدفوع (ترخيص مؤسسي)	متوسطة (تتطلب دمجاً مع LMS)
Originality.ai	~ 90 ثانية	سحابية (Web/ متصفح)	مدفوع (نظام وحدات/رصيد)	عالية (امتداد متصفح)
AraBERT المُعدَّل	~ 150 ثانية	محلية (On-Premise) / سحابية (API مخصص)	مفتوح المصدر (مجاني بالكامل)	متوسطة (تتطلب بيئة تشغيل Python وإعداداً مسبقاً)

تكشف المقارنة في الجدول (8) أن الأدوات التجارية العامة (GPTZero وZeroGPT) تتمتع بكفاءة حسابية عالية وسرعة فائقة في الاستجابة، مما يجعلها عملية في التطبيقات التي تتطلب معالجة كميات ضخمة من النصوص في زمن قصير جداً (كالتصفية الأولية للمقالات في المجلات الإلكترونية). غير أن هذه الميزة التشغيلية تُقابل بهبوط حاد في الموثوقية وارتفاع كبير في الإيجابيات الكاذبة، كما أظهرت النتائج في الجدول (4) وفي المقابل، يظهر نظام AraBERT المُعدَّل بطئاً نسبياً (نحو 150 ثانية لكل 100 نص)، يُعزى إلى طبيعة تشغيل النماذج المحلية الضخمة (Transformer-based) وتعقيد معالجة اللغة العربية الصرفية والاشتقاقية مقارنة بالاستدعاء السريع عبر واجهات برمجة التطبيقات السحابية الجاهزة. ومع ذلك، فإن هذا البطء النسبي ليس عاملاً حاسماً في السياقات الأكاديمية والبحثية التي تُعنى بالنزاهة العلمية، إذ تكون دقة التصنيف وعدالته أولوية قصوى تفوق عامل السرعة، خاصةً أن إجمالي زمن معالجة العينة الكاملة (480 نصاً) على نظام AraBERT لا يتجاوز 12 دقيقة، وهو وقت مقبول تماماً في سياق التحقيق في قضايا الانتحال الأكاديمي، حيث تكون الدقة مطلوبة على حساب السرعة الفائقة وكما أن طبيعة المصدر المفتوح (Open Source) لنظام AraBERT تمنح المؤسسات الأكاديمية العربية ميزة استراتيجية؛ فهي قادرة على تشغيله محلياً (On-Premise) دون الحاجة إلى رفع نصوصها الحساسة إلى خوادم خارجية، مما يحقق استقلالية أكبر عن الخدمات السحابية ويتيح ضبط معايير المعالجة بما يتلاءم مع سياسات الأمان والخصوصية الداخلية لكل جامعة. وتشير هذه المعطيات إلى أن المفاضلة بين الأدوات ينبغي أن تتم بناءً على سياق الاستخدام: في حالات الفحص الأولي السريع قد تكون الأدوات السريعة (مثل GPTZero) خياراً مساعداً، أما في حالات التحقيق العميق في قضايا النزاهة الأكاديمية والاتهامات الذاتية، فيُفضَّل الاستثمار في الأنظمة البطيئة نسبياً ذات الموثوقية والدقة الأعلى، مع مراعاة توفر الكوادر التقنية لتشغيلها محلياً عند استخدام الأنظمة مفتوحة المصدر.

3-7 خطة اختبار الفرضيات

جدول (9): نتائج اختبار فرضيات الدراسة

الفرضية	الاختبار الإحصائي	قيمة الاختبار	قيمة الدلالة Sig	القرار الإحصائي
الفرضية الأولى: تتفاوت مستويات دقة أدوات الكشف عند تطبيقها على النصوص العربية مقارنة بالمعدلات المعلنة لتلك الأدوات على النصوص الإنجليزية.	اختبار t لعينة واحدة (One-Sample t-Test)	4.821	0.013	رفض الفرضية الصفرية
الفرضية الثانية: تحقق الأدوات المصممة خصيصًا للغة العربية مستويات حساسية ونوعية أعلى من الأدوات العامة متعددة	اختبار t لعينتين مستقلتين (Independent Samples t-Test)	3.547	0.021	رفض الفرضية الصفرية

الفرضية	الاختبار الإحصائي	قيمة الاختبار	قيمة الدلالة Sig	القرار الإحصائي
اللغات.				
الفرضية الثالثة: تتباين دقة أدوات الكشف تبعًا للنموذج اللغوي المستخدم في توليد النصوص.	تحليل التباين الأحادي (One-Way ANOVA)	6.932	0.002	رفض الفرضية الصفرية
الفرضية الرابعة: يؤثر المجال المعرفي للنص في أداء أدوات الكشف بصورة دالة إحصائيًا.	تحليل التباين الأحادي (One-Way ANOVA)	5.184	0.006	رفض الفرضية الصفرية
الفرضية الخامسة: تعاني أدوات الكشف من ارتفاع في معدل الإيجابيات الكاذبة عند اختبار النصوص البشرية العربية.	اختبار مربع كاي (Chi-Square Test)	12.476	0.014	رفض الفرضية الصفرية

* **حجم الأثر:** أظهرت نتائج التحليل قيم حجم الأثر (η^2 الجزئية) البالغة 0.129 لأثر النموذج اللغوي و0.032 لأثر المجال المعرفي، وتُفسَّر هذه القيم بالتكامل مع النتائج الوصفية ومصفوفة الالتباس ومؤشرات أداء الكاشفات، بما يوضح الأهمية العملية للفروق إلى جانب دلالتها الإحصائية.

3-8 مناقشة النتائج في ضوء الدراسات السابقة

تُظهر النتائج أن أداء الأدوات ليس متجانساً في العربية. فقد حقق AraBERT المُعدَّل أعلى دقة إجمالية (92.5%) وحساسية (95%) ونوعية (90%)، يليه Turnitin AI بدقة 87.5%، في حين بلغت دقة GPTZero و ZeroGPT 70% لكل منهما. وتنسجم الأفضلية النسبية للنماذج العربية المتخصصة مع الأدبيات التي تؤكد أثر المعالجة اللغوية والتخصيص العربي، لكنها لا تعني أن أي كاشف عربي يمكن اعتباره دليلاً قطعياً؛ فحتى أفضل أداة في العينة سجلت أخطاء تصنيف.

وتظهر مصفوفة الالتباس أن مشكلة الأدوات العامة ليست في الدقة وحدها، بل في توزيع الخطأ: بلغ FPR نحو 40% لدى GPTZero و35% لدى ZeroGPT، مقابل 10% لدى AraBERT. وهذه النتيجة أكثر أهمية من متوسط الدقة عند اتخاذ قرار أكاديمي؛ لأن الإيجابية الكاذبة قد تنقل نصاً بشرياً سليماً إلى دائرة الاشتباه. وتدعم هذه النتيجة الاتجاه البحثي الذي يحذر من تحيز الكاشفات تجاه أنماط الكتابة غير المألوفة في بيانات التدريب.

ويظهر أثر المجال المعرفي بوضوح؛ إذ كان المجال الإبداعي الأقل قابلية للكشف، حيث انخفضت دقة GPTZero إلى 60% وAraBERT إلى 86%، بينما ارتفعت الدقة في المجالات الأكاديمية والتقنية. ويشير ذلك إلى أن قابلية التنبؤ والانتظام الأسلوبي قد تساعد الكاشف في بعض الأنواع، في حين تزيد المساحة الأسلوبية في النص الإبداعي من صعوبة التمييز.

كما ظهر تباين جغرافي داخل النصوص البشرية العربية؛ فقد بلغ FPR في GPTZero 45% لنصوص المغرب العربي، مقابل 37% في الخليج و38% في المشرق. ويجب تفسير ذلك باعتباره مؤشراً على اختلاف الأداء داخل العربية الفصيحة، لا دليلاً على وجود خاصية لغوية جوهرية في منطقة بعينها؛ إذ قد تتداخل العوامل الأسلوبية والتعليمية والترجمية ومصدر النص.

وعلى مستوى النموذج المولد، أظهرت البيانات تفاوتاً في دقة الكشف؛ وكان من أبرز الحالات انخفاض دقة GPTZero إلى 65% مع نصوص Gemini. ويتفق هذا الاتجاه مع الأدبيات الحديثة التي تبين أن تحسن النماذج التوليدية وتغير أساليبها يضغط باستمرار على قدرة الكاشفات على التعميم.

أمثلة تمثيلية للأخطاء: في حالة الإيجابي الكاذب، يكون النص الفعلي بشرياً لكن الأداة تمنحه تصنيفاً آلياً؛ وفي حالة السلبي الكاذب يكون النص مولداً آلياً لكن الأداة تصنفه بشرياً. ولعدم وجود نصوص خام مرفقة في النسخة الأصلية من الورقة، أُدرجت هنا أمثلة توضيحية لنمط الخطأ وليست مشاهدات إضافية ضمن العينة: (أ) نص أكاديمي بشري منظم يستخدم جُملاً تقريرية قصيرة قد يُلتبس بالنص الآلي؛ (ب) نص آلي في المجال الإبداعي يتضمن تنوعاً في طول الجمل والاستعارات قد يُلتبس بالنص البشري. أما الحالات الفعلية فتُراجع من ملف النصوص المرمزة TXT-001 إلى TXT-480 عند الإتاحة.

1. النتيجة الأولى: يتفوق AraBERT المُعدَّل في الدقة والحساسية والنوعية داخل العينة، ما يدعم قيمة التخصيص اللغوي العربي.
2. النتيجة الثانية: تمثل الإيجابيات الكاذبة مصدر الخطر الأكبر في الأدوات العامة؛ لذلك لا ينبغي تحويل نسبة "AI" الصادرة عن الكاشف إلى حكم اتهامي مستقل.

3. النتيجة الثالثة: يتأثر الأداء بالمجال المعرفي؛ ويظهر المجال الإبداعي باعتباره الحالة الأكثر صعوبة في العينة.
4. النتيجة الرابعة: يتغير الأداء باختلاف النموذج المولد، بما يفرض تحديث اختبارات الكاشفات دورياً بدلاً من التعامل مع نتائجها باعتبارها ثابتة زمنياً.

3-9 توصيات الدراسة

في ضوء النتائج الإحصائية ومؤشرات الخطأ، تُصاغ التوصيات على النحو الآتي:

1. عدم استخدام درجة الكاشف وحدها لاتخاذ قرار تأديبي أو اتهامي؛ لأن FPR بلـغ 40% في GPTZero و 35% في ZeroGPT و10% حتى في AraBERT.
2. إعطاء أولوية للأنظمة التي تراعي خصائص العربية، مع إعادة اختبارها دورياً عـبر مجـالات ونمـاذج مولـدة جديدة، وعدم تفسير التفوق الحالي باعتباره ضماناً دائماً.
3. اعتماد بروتوكول تحقق متعدد الأدلة يجمع نتيجة الكاشف مع فحص النسخ والمسودات والمناقشة الشفهية أو الأدلة السياقية الأخرى، خصوصاً في الحالات الحدودية.
4. إنشاء قاعدة بيانات عربية معيارية مستمرة التحديث، تتضمن تنوع المجالات والنماذج والمناطق العربية، مع حفظ إصدارات النماذج والموجهات لضمان قابلية التكرار.
5. إدراج متغيرات الخطـأ والحساسـية والنوعيـة وحجم الأثـر في تقـارير المؤسسـات، وعـدم الاكتفـاء بالدقـة الإجمالية أو النسبة المئوية التي يعرضها المنتج التجاري.

الإعلانات والإفصاحات

تضارب المصالح: يصرح المؤلفون بعدم وجود أي تضارب في المصالح.

التمويل: لم يتلقَّ البحث تمويلًا خارجيًا.

الموافقة الأخلاقية (عند الاقتضاء): [[تم الالـتزام بالمعـايير الأخلاقيـة وحمايـة خصوصـية المشـاركين، والحصـول على الموافقات الرسمية والموافقة المستنيرة [رقم الموافقة: .....].]]

إتاحة البيانات (عند الاقتضاء): البيانات متاحة عند الطلب.

الإفصاح عن استخدام الذكاء الاصطناعي (عند الاقتضاء): يفصح المؤلفون عن أي استخدام لأدوات الذكاء الاصطناعي التوليدي في تنفيذ البحث أو إعداد المخطوط، مع بيان الغرض من استخدامها بإيجاز.

4. المراجع

Alshammari, H., & El-Sayed, A. (2023). AIRABIC: Arabic dataset for performance evaluation of AI detectors. In Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA), Jacksonville, FL, USA, December 15–17. IEEE. https://doi.org/10.1109/ICMLA58977.2023.00127

Alshammari, H., & Elleithy, K. (2024). Toward robust Arabic AI-generated text detection: Tackling diacritics challenges. Information, 15(7), 419. https://doi.org/10.3390/info15070419

Alshammari, H., El-Sayed, A., & Elleithy, K. (2024). AI-generated text detector for Arabic language using encoder-based transformer architecture. Big Data and Cognitive Computing, 8(3), 32. https://doi.org/10.3390/bdcc8030032

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the Fourth Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 9–15). European Language Resource Association. https://doi.org/10.48550/arXiv.2003.00104

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan,

T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1-25.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Eaton, S. E. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. International Journal for Educational Integrity, 19(23). https://doi.org/10.1007/s40979-023-00144-1

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 111–116). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3019

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597. https://arxiv.org/abs/2301.07597

Habash, N. Y. (2010). Introduction to Arabic natural language processing. Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201008HLT010

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1808–1822). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.164

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305. https://arxiv.org/abs/2301.11305

Odri, G.-A., & Yoon, Y. (2023). Détection de l'intelligence artificielle générative dans les articles scientifiques: techniques d'évasion et implications pour l'intégrité scientifique. Revue de Chirurgie Orthopédique et Traumatologique, 109(8), ١-6. https://doi.org/10.1016/j.otsr.2023.103706

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

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Schaaff, K., Schlippe, T., & Mindner, L. (2023). Classification of human- and AI-generated texts for English, French, German, and Spanish. In Proceedings of the 6th International

Conference on Natural Language and Speech Processing (ICNLSP 2023) (pp. 1–10). Association for Computational Linguistics https://arxiv.org/abs/2312.04882

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 1–15. https://doi.org/10.48550/arXiv.1706.03762