



اللغويات الحاسوبية والمعالجة اللغوية الآلية

Journal of Computational Linguistics & Arabic Language Processing



دراسة مسحية لآليات تقييم النماذج اللغوية الكبيرة

السنة الثالثة - العدد الخامس
ذو القعدة ١٤٤٧ هـ - أبريل ٢٠٢٦ م
Vol. 3, No. 5, Apr, 2026

مجمع الملك سلمان
العالمي للغة العربية
King Salman Global Academy for Arabic Language



اللغويات والحوسبة والمعالجة اللغوية

Journal of Computational Linguistics & Arabic Language Processing

www.ksaa.gov.sa

مجلة علمية مُحكمة تصدر عن مجمع الملك سلمان العالمي للغة العربية





المشرف العام على المجلة

أ.د. عبدالله بن صالح الوشمي

إدارة التحرير

أ.د. منصور الغامدي

رئيس التحرير

د. عبد الرحمن العصيمي

مدير التحرير

د. هاجر القبيشي

أمانة المجلة

هيئة التحرير

د. سارة الحمود

د. عريب العويشق

أ.د. عبدالعزيز المهيوبي

د. أيمن الغامدي

الهيئة الاستشارية

أ.د. محسن عبدالرزاق رشوان (مصر)

أ.د. إيريك أتويل (بريطانيا)

أ.د. نزار حبش (أمريكا)

أ.د. مصطفى عرسان جرار (فلسطين)

أ.د. عبدالمحسن عبيد الثبتي (السعودية)

أ.د. خالد فؤاد شعلان (مصر)

أ.د. صلاح راشد الناجم (الكويت)

◆ قواعد النشر بالمجلة :

سياسات النشر:

تنشر مجلة اللُّغويات الحاسوبية والمعالجة الآلية للُّغة العربيّة البحوث الأصيلة التي لم يسبق نشرها، والمكتوبة بالُّغة العربية أو الإنجليزية، والدراسات والمُلخصات العلميّة المتخصصة في مجال حوسبة اللُّغة، كاللُّغويات الحاسوبية، والمعالجة الآليّة للُّغة العربيّة، وكل ما له صلة بحوسبة اللُّغة.

تُعرض مسودة الأوراق العلميّة فور وصولها على هيئة التحرير لإجراء تقييم مبدئي لها، وفي حال عدم مناسبتها للنشر تُخطر المجلة الباحث بذلك. أمّا البحوث المقبولة مبدئيًا، فترسل إلى اثنين من المحكّمين لتقييمها سرّيًا، وتبذل هيئة التحرير طاقتها في التواصل مع المحكّمين للوصول إلى قرار بشأن كل بحث مقدّم في أسرع وقت ممكن، ويُخطر الباحث بنتيجة التحكيم. وفي حالة قبول الورقة للنشر، فإنّ كافة حقوق النشر تؤوّل للمجلة، ولا يجوز نشرها في أي منفذ نشر آخر ورقياً أو إلكترونياً، دون إذن كتابي من رئيس هيئة تحرير المجلة. وتُرسل نسختان ورقيتان مجانيتان من المجلة إلى المؤلّف عند نشر ورقته.



Journal information:

◆ بيانات المجلة :

Journal of Computational Linguistics & Arabic Language Processing	مجلة اللُّغويات الحاسوبية والمعالجة الآلية للُّغة العربيّة
A peer-reviewed bi-annual journal issued by	مجلة علمية محكمة تصدر عن
King Salman Global Academy for Arabic Language	مجمع الملك سلمان العالمي للُّغة العربيّة
Riyadh- Saudi Arabia	الرياض - المملكة العربيّة السعوديّة
ISSN paper version: 1658-9475	ردمك النسخة الورقية: ١٦٥٨-٩٤٧٥
ISSN Electronic version: 1658-9483	ردمك النسخة الإلكترونيّة: ١٦٥٨-٩٤٨٣

تعليمات عامة:

١. تُرسل الورقة المقدّمة للنشر للمجلة بصيغة وورد (word) على بريد المجلة: jclalp@ksaa.gov.sa). على أن تُوجّه جميع المراسلات إلى: رئيس هيئة التحرير- مجلة اللُّغويات الحاسوبية والمعالجة الآلية للغة العربيّة- مجمع الملك سلمان العالمي للغة العربيّة.

٢. يجب أن تحتوي الورقة على ما يلي:

١-٢- صفحة أولى فيها عنوان الورقة، واسم المؤلف أو المؤلفين، وجهة عملهم، وعناوينهم، وتاريخ تقديم الطلب بنشر الورقة.

٢-٢- الكلمات المفتاحية.

٣-٢- مستخلص الورقة باللّغتين العربية والإنجليزية -في حدود مئة كلمة لكلّ منهما-.

٤-٢- المصادر والمراجع.

٥-٢- الملاحق -إن وجدت-.

٣. يُراعى أن تكون الورقة العلمية مكتوبةً بلغة سليمة -سواء كانت عربية أو إنجليزية-، وأن يتراوح عدد كلماتها بين (٣٠٠٠ - ١٠٠٠٠) كلمة، بما في ذلك المراجع والهوامش والجداول والأشكال والملاحق.

٤. إذا كانت الورقة العلمية مكتوبة باللّغة العربيّة، فيجب الالتزام بكتابة ما يقابل المصطلحات العربيّة باللّغة الإنجليزيّة في المرة الأولى التي يظهر فيها المصطلح.

٥. ألا يسبق نشر البحث في أي وعاء علمي -مالم ترهئة التحرير خلاف ذلك-.

المواصفات الفنية للبحث:

- يكون مقاس الصفحة A4، وتُقسَّم إلى عمودين، وتكون المسافة بين العمودين ١ سم، وتكون الهوامش ٥,٢ سم عن يمين ويسار الصفحة، و٥٨,٣ سم من أعلى وأسفل الصفحة.
- تُترك مسافة 5.0 سم في بداية كل فقرة من فقرات المتن.
- يجب الفصل بمسافة سطر واحد بين نهاية محتوى الفقرات داخل المتن وعناوين الفقرات التي تليه.
- يجب ترك مسافة سطر واحد قبل وبعد الجدول أو الشكل والصورة والمتن السابق والتالي لهما.
- يُستخدم الخط (Sakkal Majalla) للنص العربي، والخط (Times New Roman) للنص الأجنبي وبحجم أقل بدرجتين عن الخط العربي. ويكون حجم الخط العربي كالآتي:
 - العنوان الرئيس: ١٤، عريض.
 - العناوين الفرعية: ١٢، عريض.
 - النصوص: ١٢، عادي.
 - الهوامش السفلية: ١٠، عادي.
- يجب ترقيم صفحات البحث.
- يجب ترقيم المعادلات.
- تُرتَّب المصادر والمراجع أبجدياً بنظام APA.
- تُفصل المراجع الأجنبية عن المراجع العربية في قائمة المصادر والمراجع، وتأتي المراجع /المصادر العربية أولاً في الأبحاث العربية والمراجع /المصادر الإنجليزية أولاً في الأبحاث الإنجليزية.

الجداول:

- تُدرج الجداول بعد الفقرة التي ذُكرت بها، وذلك على النحو التالي:
 - إذا كان عرض الجدول ٧,٥ سم يوضع في عمود واحد أسفل الفقرة مباشرة.
 - إذا كان عرض الجدول أكثر من ٧,٥ سم يوضع أسفل الصفحة التي ذكر بها أو أعلى الصفحة التي تليها.
- تُرقّم الجداول بين قوسين في أقصى يمين الصفحة مسبوقة بكلمة «جدول». مثال: جدول (١).
- يُوضع عنوان الجدول في أعلى الجدول مباشرة، في أقصى اليمين.
- يجب ألا يتجاوز عرض الجدول هوامش الصفحة المذكورة في الفقرة السابقة.
- يُكتب مصدر الجدول والتعليقات المصاحبة له -إن وجدت- أسفله.
- تُضبط بيانات الجدول غير الرقمية أقصى يمين الخلية.
- تُضبط بيانات الجدول الرقمية وسط الخلية.
- إذا استمر الجدول على أكثر من صفحة، فيجب تكرار رأسه في الصفحات التالية ووضع عبارة «تابع جدول (...)» في أعلاه.

الصُور والأشكال:

- تُدرج الصور والأشكال بعد الفقرة التي ذُكرت على النحو التالي:
 - إذا كان عرض الصورة ٧,٥ سم تُوضع في عمود واحد أسفل الفقرة مباشرة.
 - إذا كان عرض الصورة أكثر من ٧,٥ سم تُوضع أسفل الصفحة التي ذكر بها أو أعلى الصفحة التي تليها.
- تُرقّم الصور والأشكال، ويوضع الرقم بين قوسين مسبوقة بكلمة «شكل». مثال: شكل (١).

- يُدرج عنوان الشكل أسفل الشكل مباشرةً متوسطًا بين حدوده بحيث لا يتجاوز عنوان الشكل سطرين فقط.
- في حال وجود تعليق مصاحب على الشكل أو الصورة يجب دمجها مع الشكل وحفظه على هيئة صورة لا تقل دقتها عن (300 dpi).

مواصفات كتابة المعادلات:

- تُكتب المعادلات الرياضية باستخدام أحد البرنامجين التاليين : (Math Type) أو (Equation Editor).
- تُرقّم المعادلات، ويوضع الرقم بين قوسين في أقصى يسار العمود مقابل المعادلة على مستوى البحث.

مواصفات كتابة الاقتباس:

- إذا كان النص المقتبس ٣ أسطر أو أقل، فيجب أن يوضع بين علامتي تنصيص «...»، ويليه توثيق النص على النحو التالي: (العائلة، السنة: الصفحة).
- إذا كان النص أكثر من ٣ أسطر، فيجب أن يُميز النص المقتبس بكتابته في منتصف الصفحة، وتُترك من بداية السطر ونهايته هوامش اسم، ويوضع بين علامتي تنصيص «...»، ويُكتب بحجم ١٢ للنص العربي و١٠ للنص الإنجليزي، وينتهي بتوثيق النص على النحو التالي: (العائلة، السنة: الصفحة).

مواصفات كتابة الحواشي:

- «تُدرج الحواشي أسفل كل صفحة، مع ترقيمها آليًا حسب ورودها في كل صفحة».

◆ صفة المجلة:

مجلة اللُّغويات الحاسوبية والمعالجة الآلية للُّغة العربيَّة (Journal of Computational Linguistics & Arabic Language Processing) هي مجلة علمية محكَّمة، نصف سنوية (ISSN:1658-9475 - ISSN:1658-9483)، وتصدر عن مجمع الملك سلمان العالمي للُّغة العربية في الرياض بنسختين: إلكترونية، وورقية (ذات إصدار محدود). وتتيح المجلة الوصول الحر (Open Access) إلى البحوث المنشورة فيها، ولا تتقاضى رسوْمًا على النشر. وهي مجلة تهتم بنشر الأبحاث العلمية المحكمة في مجال حوسبة اللُّغة، وتشمل: اللُّغويات الحاسوبية (Computational Linguistics)، والمعالجة الآلية للُّغة العربية (Arabic Language Processing) وما يتبعها من تخصصات متَّصلة بها. وللاطلاع على أهداف المجلة وشروط النشر وكل التفاصيل ذات العلاقة يمكنكم زيارة موقع المجلة: jclalp.ksaa.gov.sa

البحوث المنشورة بالمجلة تعبر عن رأي أصحابها، ولا تعبر بالضرورة عن رأي المجمع، ويتحمل مؤلفوها المسؤولية كاملة عن صحة المعلومات والاستنتاجات ودقتها.

محتويات العدد

١٣

دراسة مسحية لآليات تقييم النماذج اللغوية الكبيرة

أ.د. جورج تركيّه - ريم المطيريب - تركي العتيبي - فيصل بن بريكان
رغد آل رشيد - أمال المزروعج - منيرة الحوشان - روان المعثم

كلمة التحرير

عزيزي القارئ/القارئة - عزيزي الباحث/الباحثة:

يُسعدنا أن نرحب بكم في العدد الخامس من المجلة العلمية المحكمة: "اللغويات الحاسوبية والمعالجة الآلية للغة العربية"، الصادرة عن مجمع الملك سلمان العالمي للغة العربية، والمتخصصة في مجال حوسبة اللغة العربية.

تواصل المجلة مسيرتها العلمية الرصينة في نشر البحوث والدراسات التي تُسهم في تطوير الفهم العميق لتطبيقات الحوسبة في اللغة العربية، وتعمل على رصد أحدث الاتجاهات في مجال الذكاء الاصطناعي اللغوي وأدوات التحليل اللغوي الآلي. كما تُعنى المجلة باستشراف آفاق المستقبل، واستكشاف التحديات والفرص التي تواجه تطوير اللغة العربية رقمياً، بما يعزّز حضورها في فضاء التقنية والمعرفة.

ويأتي العدد الخامس استمراراً لهذا التوجّه العلمي البناء، حاملاً بين صفحاته ثلاثة أبحاث علمية متميزة واستعراضاً لكتاب مهم، تعالج جميعها موضوعات معاصرة ومتنوعة في صميم حوسبة اللغة العربية:

يأتي البحث الأول بعنوان "دراسة مسحية لآليات تقييم النماذج اللغوية الكبيرة"، وهو بحث يستعرض أبرز الممارسات والأطر العالمية لتقييم النماذج اللغوية الضخمة، ويناقش سبل تطوير آليات تقييم مخصصة للغة العربية تراعي خصوصيتها اللغوية والثقافية، مقدماً إطاراً لمقاييس شاملة تضمن موثوقية الأداء.

أما البحث الثاني "أنظمة الترجمة الآلية من العربية إلى الإنجليزية وتقييم بعض برامجها"، فيتجه نحو تقييم ترجمة ثلاثة برامج آلية حاسوبية من خلال تحليل مقارن لجمل عربية تحمل اللبس الصرفي والتركيبى والدلالي، للوقوف على مواطن الضعف واقتراح الحلول اللغوية والتقنية لمعالجتها.

ويأتي البحث الثالث "دور النماذج اللغوية الكبيرة في تطوير (WordNet) قاعدة بيانات معجمية عربية: تقييم القدرات وتحسين الشبكات الدلالية"؛ ليُجري تحليلاً دلاليّاً لتوظيف النماذج اللغوية الكبيرة جنباً إلى جنب مع الخبرة البشرية في بناء شبكة كلمات عربية دقيقة، تستوعب العلاقات الدلالية والتعقيدات الصرفية للغة.

وأخيراً، يقدم العدد استعراضاً علمياً لكتاب "لسانيات المدونات: مقدمة موجزة"، والذي يمثل مدخلاً مهماً يجمع بين الجانبين النظري والتطبيقي في التعريف بلسانيات المدونات، ويسلط الضوء على دور المدونات اللغوية في دراسة المعنى واستخراج المتلازمات اللفظية وتحليل الخطاب الفعلي.

وختاماً، نُجدد ترحيبنا الدائم بتواصلكم المثمر، ومقترحاتكم القيّمة التي تُمثل رافداً أساسياً للارتقاء بمسيرة المجلة وتجويد محتواها العلمي. سائلين المولى عز وجل أن يكلل هذه الجهود بالنجاح، وأن يجعل هذا العدد لبنةً نافعةً تخدم الباحثين، وتُثري ميدان حوسبة اللغة العربية وتقنياتها.

هيئة التحرير

دراسة مسحية لآليات تقييم النماذج اللغوية الكبيرة

المعرف الرقمي: <https://doi.org/10.60161/2521-003-005-001>

أ.د. جورج تركيه^١ ريم المطيري^٢ تركي العتيبي^٣ فيصل بن بريكان^٤

رغد آل رشيد^٥ أمال المزروع^٦ منيرة الحوشان^٧ روان المعثم^٨

أ جامعة الملك عبد الله للعلوم والتقنية (مركز أبحاث الحوسبة الفائقة)، جدة، المملكة العربية السعودية

ب شركة إنتروبي، الرياض، المملكة العربية السعودية

ج مجمع الملك سلمان العالمي للغة العربية، الرياض، المملكة العربية السعودية

استقبل في	رُوجع في	قُبِل في	نُشِر في
١٤٤٧/٠٤/٢٩	١٤٤٧/٠٤/٣٠	١٤٤٧/١٠/٢٠	١٤٤٧/١١/١٣

المخلص

تهدف هذه الدراسة إلى استعراض أبرز الممارسات والأطر العالمية في تقييم النماذج اللغوية الكبيرة، مع توضيح سبل تطوير آليات مماثلة باللغة العربية، سواء كانت يدوية أو آلية أو هجينة. ويناقش الحاجة إلى تقييم منهجي ومتوازن يواكب توسع استخدام هذه النماذج، مع مراعاة الخصوصية اللغوية والثقافية للغة العربية.

تؤكد الدراسة أن الاعتماد على التقييم اليدوي وحده غير كافٍ لمحدودية التوسع وارتفاع التكلفة، كما أن التقييم الآلي وحده قد يتجاهل الجوانب الثقافية وجودة اللغة. ويقترح مزيجاً تكاملياً يجمع بين سرعة الأنظمة الآلية ودقة التقدير البشري في المهام المفتوحة التي تتطلب فهماً لغوياً عميقاً. كما يقدم إطاراً لمقاييس شاملة تشمل الدقة، والمتانة، والعدالة، والانحياز، والكفاءة التشغيلية، بما يضمن موثوقية الأداء في البيئات الواقعية.

وتخلص الدراسة إلى أهمية بناء آليات تقييم مصممة خصيصاً للغة العربية نظراً لاختلاف خصائصها عن اللغات الأخرى، مع التركيز على ثلاثة محاور رئيسية: السلامة لضمان ملاءمة النماذج لسياقات الثقافة، القدرات لقياس عمق الأداء في مجالات معرفية متعددة، والخصائص اللغوية كالبلغة والإعراب والدلالة التي تميز اللغة العربية وتثري تقييمها.

الكلمات المفتاحية

تقييم النماذج اللغوية الكبيرة، اللغة العربية، أطر التقييم، الأمان، الانحياز.

* الباحث المراسل: تركي العتيبي (البريد الإلكتروني: Email:Turki@entropy.sa).

Survey of Mechanisms for Evaluating Large Language Models

المعرف الرقمي: <https://doi.org/10.60161/2521-003-005-001>

Prof. George Turkiyyah^A, Reem Almutairi^B, Turki Alotaibi^{*B}, Faisal BinBuraykan^B,
Raghad Al-Rasheed^C, Amal Almazrua^C, Muneera Alhoshan^C, Rawan Al-Matham^C

A King Abdullah University Of Science And Technology

B Entropy

C King Salamn Global Academy For Arabic

Received on	Revised on	Accepted on	Published on
21-10-2025	22-10-2025	08-04-2026	30-04-2026

ABSTRACT

This survey presents leading global practices for evaluating large language models and outlines ways to develop similar mechanisms in Arabic, manual, automated, or hybrid. It highlights the need for a balanced evaluation approach that reflects the growing use of such models and respects the linguistic and cultural nuances of the Arabic language.

The survey emphasizes that manual evaluation alone is limited by cost and scalability, while automated methods may overlook cultural and linguistic nuances. It advocates a hybrid approach that combines the speed of automated systems with the precision of human judgment for tasks requiring deeper understanding. The survey also presents an integrated framework of metrics covering accuracy, robustness, fairness, bias, and efficiency, to ensure reliable and practical performance in real-world settings.

The survey concludes by emphasizing the need for evaluation mechanisms tailored to the Arabic language, given its unique structural and semantic features. It highlights three core dimensions: safety, ensuring cultural alignment; capabilities, assessing performance depth across domains; and linguistic properties, encompassing eloquence, morphology, and semantics that enrich Arabic evaluation.

KEYWORDS

Large Language Model Evaluation, Arabic Language, Evaluation Benchmark, Safety, Bias.

*Corresponding author: Turki Alotaibi (Email: Turki@entropy.sa).

١- المقدمة

شهدت السنوات الأخيرة ثورة غير مسبوقة في تطوير نماذج الذكاء الاصطناعي، لا سيما النماذج اللغوية الكبيرة (LLMs) التي تعتمد على تقنيات التعلم العميق لمعالجة وفهم اللغة الطبيعية.

وقد أثبتت هذه النماذج قدرتها على أداء مهام لغوية معقدة مثل توليد النصوص، والإجابة عن الأسئلة، وترجمة المحتوى، وتحليل المشاعر، وغيرها (Brown et al., 2020). ومع ذلك، فإن الاعتماد المتزايد عليها في مجالات مهمة، يجعل من تقييمها ضرورة حتمية لضمان سلامة استخدامها وجودة مخرجاتها.

تستخدم النماذج اللغوية الضخمة على نطاق واسع من الأفراد والمؤسسات في الوطن العربي على حد سواء؛ لتقديمها فوائد عظيمة من زيادة الإنتاجية، والمساعدة في حل المشاكل، ورفع الكفاءة، والأجوبة عن الاستفسارات. لكن النماذج اللغوية الضخمة قصّرت في التعامل مع اللغة العربية على المستوى نفسه في اللغات الأخرى؛ هذا بسبب عدم تدريب النماذج على اللغة العربية بصورة أساسية، وقلة البيانات عالية الجودة، سواءً لتدريب أو اختبار النماذج (Koto et al., 2024). لذلك، تقدم هذه الدراسة آليات لقياس قدرات النماذج اللغوية على التعامل مع اللغة العربية بمستوى عالٍ، مما يساعد في زيادة الثقة بأداء هذه النماذج، تحديداً في السياقات الثقافية، وخصائص اللغة العربية.

تكمن أهمية تقييم النماذج اللغوية الكبيرة في كونه حجر الأساس لضبط جودة مخرجات النماذج، وتعزيز ثقة المستخدمين بها، والتقليل من المخاطر المرتبطة بتبنيها في البيئات الحساسة. كما يُسهم التقييم المنهجي في دعم صناع القرار لاختيار النموذج الأنسب لكل سياق، ويوجّه الباحثين والمطورين نحو تحسين أداء النماذج وتلافي جوانب القصور فيها. ولا يقتصر التقييم على دقة النتائج، بل يتعداها ليشمل أبعاداً متعددة مثل الكفاءة الحسابية، والموثوقية، والأمان، والانحياز، وملاءمة السياق الثقافي المستهدف. ومن هذا المنطلق، أصبحت الحاجة ملحة إلى تطوير أطر تقييم شاملة تأخذ بعين الاعتبار تعقيد هذه النماذج، وتنوع استخداماتها، والمخاطر الأخلاقية والتقنية المرتبطة بها.

وقد برزت في هذا المجال مشاريع علمية رائدة تهدف إلى تقديم تصور منهجي لتقييم النماذج، مثل إطار (Hendrycks et al., 2021) (MMLU) الذي يركز على تقييم فهم النماذج في تخصصات متعددة عبر اختبارات معيارية تغطي مجالات معرفية متنوعة، وإطار (Srivastava et al., 2023) (BIG-Bench) الذي يقدم مهام خارج نطاق قدرات النماذج الحالية لقياس قدراتها الاستدلالية والإبداعية، وإطار (Liang et al., 2023) (HELM) يُعد مرجعاً متكاملًا للتقييم متعدد الأبعاد يشمل الجوانب التقنية والأخلاقية، وإطار (MT-Bench) (Bai et al., 2024) الذي يركز على تقييم الحوار متعدد الأدوار باستخدام نماذج لغوية تلعب دور الحكم بدلاً من الإنسان. وقد اعتمدت هذه المراجع الأساسية في بناء الدراسة بهدف توحيد المفاهيم، وتصنيف أنواع التقييم، وتكييف المنهجيات العالمية لتناسب مع اللغة العربية وسياقات استخدامها.

تقدم هذه الدراسة إطاراً علمياً مبسطاً ومتكاملاً لتقييم النماذج اللغوية، موجهًا إلى جمهور واسع من الباحثين، والمطورين، وغيرهم من المهتمين بفهم آليات التقييم منهجياً وعملياً. وتستعرض الدراسة أبرز تصنيفات التقييم المعتمدة في الأبحاث العلمية، ويشرحها ضمن محورين رئيسيين:

التقييم الآلي: القائم على معادلات رياضية تُحسب تلقائياً، اعتماداً على مقارنة جواب النموذج والجواب الصحيح.

التقييم البشري: الذي يعتمد على حكم إنسان في تقدير جودة النصوص من حيث الملائمة والدقة.

وتُبرز الدراسة أهمية الجهود المبذولة في مجال التقييم ضمن العالم العربي، وذلك من خلال الإشارة إلى مبادرات ومؤشرات عربية مثل "بلسم"، وتقترح الدراسة ضرورة تطوير إطار تقييم عربي متكامل يأخذ بعين الاعتبار خصوصية اللغة والثقافة والسياق المحلي.

يستعرض الجدول (١) تسلسلاً لأبرز المؤشرات العالمية والعربية المخصصة لقياس أداء النماذج اللغوية، وذلك بحسب سنة الإصدار ونطاق التقييم (مجمع الملك سلمان العالمي

لغة العربية، ٢٠٢٤). ويهدف هذا التصنيف إلى إبراز تطور منهجيات التقييم وتنوعها بين الفهم والتوليد والسلامة والتوافق الثقافي عبر مختلف السياقات اللغوية.

وقد اختيرت الأبحاث والمؤشرات المعروضة في الجدول أدناه وفق معايير انتقاء محددة، إذ اقتصر البحث على الدراسات والمؤشرات ذات الصلة المباشرة بتقييم النماذج اللغوية الكبيرة، سواءً أكانت مؤشرات عالمية عامة يمكن الاستفادة منها في السياق العربي، أم مؤشرات مصممة أصلاً لتقييم النماذج اللغوية العربية. وقد اشتمل نطاق البحث على الدراسات المنشورة منذ عام ٢٠١٧، تزامناً مع ظهور بنية المحولات التي أرسيت أسس جيل النماذج اللغوية الكبيرة الحديثة، وذلك من خلال قواعد بيانات أكاديمية معتمدة تشمل العالمية المتخصصة ذات النطاق الضيق، كمؤشرات تقييم البرمجة والرياضيات من قبيل MATH و SWE-Bench وما شابهها، وذلك لعدم ارتباطها المباشر بمحاور الدراسة المتمثلة في تقييم الكفاءات اللغوية العامة والسلامة والخصائص اللغوية للغة العربية.

المعيار	السنة	التركيز	الوصف	آلية التقييم
المؤشرات العالمية				
GLUE Benchmark: General Language Understanding Evaluation (Wang et al., 2019)	2018 - 2019	قياس فهم اللغة الطبيعية في مهام وسياقات متعددة، ومؤشر قياسي لمعالجة اللغة الطبيعية	مؤشر قياسي مبكر لمعالجة اللغة الطبيعية	تقييم آلي
SuperGLUE Benchmark (Wang et al., 2020)	2019 - 2020	قياس فهم اللغة الطبيعية في مهام متعددة بدرجة صعوبة أعلى	خليفة لمؤشر GLUE	تقييم آلي
MMLU Benchmark: Massive Multitask Language Understanding	2020 - 2021	قياس المعرفة والاستدلال في مهام متعددة على نطاق واسع	أسئلة متعددة الخيارات عبر تخصصات متنوعة ومختلفة	تقييم آلي

تقييم هجين (آلي، وبشري)	مشروع قائم على دعم المجتمع	قياس أداء النماذج في مهام متنوعة، مبتكرة، وإبداعية	2022 - 2023	BIG-bench Benchmark: Beyond the Imitation Game benchmark
تقييم هجين (آلي، وبشري)	المعاصرة، والإنصاف، والمتانة، وعدة مؤشرات أخرى	تقييم شامل متعدد المقاييس	2022 - 2023	HELM Benchmark: Holistic Evaluation of Language Models
تقييم هجين (آلي، وبشري)	التركيز على التعليمات المفتوحة	اتباع التعليمات وتنفيذ مهام متعددة	2023 - 2024	MT-Bench Benchmark: Multi-Turn Benchmark
المؤشرات العربية				
تقييم آلي	يقيس أداء الترجمة من اللهجات العربية المتنوعة إلى الإنجليزية	تقييم ترجمة اللهجات العربية	2020	مؤشر تقييم ترجمة الآلة بين اللهجات العربية والإنجليزية (أرا-بنش)
تقييم آلي	يقيم قدرات فهم النماذج للغة العربية الفصحى عبر مهام فهم اللغة الطبيعية	تقييم فهم اللغة العربية	2021	مؤشر تقييم فهم اللغة العربية (أليو)
تقييم آلي	يختبر الفهم اللغوي عبر ٤٢ مجموعة بيانات تشمل سخرية، مشاعر، لهجات، وغيرها	تقييم فهم متعدد اللهجات	2021	مؤشر تقييم فهم اللغة الطبيعية العربية (أرليو)
تقييم آلي	مؤشر شامل لقياس الفهم اللغوي عبر عشرات المهام مع تنوع عربي واسع	تقييم فهم لغوي واسع	2022	مؤشر تقييم فهم اللغة العربية (أوركا)

مؤشر تقييم توليد اللغة الطبيعية باللغة العربية (دولفين)	2023	تقييم توليد نصوص باللغة العربية	يقيس قدرة النماذج على إنتاج نصوص عربية بجودة عالية في مهام توليد النصوص المتعددة	تقييم آلي
مؤشر الغافة لتقييم نماذج اللغة العربية	2023	تقييم النماذج العربية	يختبر أداء النماذج العربية في عدد من مهام الفهم والإجابة والتحليل	تقييم آلي
مؤشر تقييم توافق النماذج للثقافة والقيم العربية (أكفا)	2023	تقييم التوافق الثقافي	يقيس مدى اتساق النماذج مع القيم والسياقات الثقافية العربية	تقييم آلي
مؤشر تقييم فهم اللغة متعدد المهام المترجم (Translated MMLU)	2023	فهم متعدد التخصصات (مترجم)	نسخة مترجمة آلياً من MMLU لاختبار قدرات النماذج باللغة العربية	تقييم آلي
مؤشر تقييم الذكاء الاصطناعي العربي باستخدام نماذج اللغة الكبيرة (لا-أرا-بنش)	2024	تقييم أداء نماذج اللغة الكبيرة في مهام اللغة العربية	مؤشر واسع لتقييم نماذج اللغة الكبيرة في مهام عربية متنوعة (فهم/توليد)	تقييم آلي
مؤشر تقييم فهم اللغة متعدد المهام واللغات (MMMLU)	2024	فهم متعدد اللغات ومتعدد المهام	مجموعة أسئلة اختيارية مترجمة إلى ١٤ لغة لتقييم النماذج عبر ٥٧ موضوعاً معرفياً	تقييم آلي

مؤشر متعدد المهام لتقييم المعرفة القانونية باللغة العربية في نماذج اللغة الكبيرة (أراب- ليقل-إيفال)	2024	قياس المعرفة القانونية	يقيس أداء النماذج في المجال القانوني العربي عبر مهام متخصصة	تقييم آلي
مؤشر تقييم المهام التوليدية لنماذج اللغة العربية الكبيرة (أراجن)	2024	تقييم توليد النصوص بالعربية	مؤشر معياري وديناميكي لتقييم قدرات النماذج التوليدية العربية، يوفر قياساً دقيقاً للنصوص عبر معايير لغوية وتقنية مفصلة	تقييم هجين (آلي، بشري)
مؤشر تقييم فهم اللغة العربية متعدد المهام واسع النطاق (ArabicMMLU)	2024	تقييم قدرة النماذج على الفهم متعدد المهام في اللغة العربية	يقيس مدى عمق فهم النماذج للمعرفة المتعددة التخصصات ضمن سياقات عربية أصيلة	تقييم آلي
مجموعة البيانات العربية لتقييم سلامة النماذج اللغوية الضخمة	2024	يغطي هجمات مباشرة وغير مباشرة لقياس سلامة النماذج	تقييم سلامة نماذج اللغة الضخمة	تقييم هجين (آلي، بشري)
بلسم: منصة لتقييم نماذج اللغة العربية الضخمة	2025	قياس عام لأداء النماذج اللغوية	مؤشر شامل مخصص لقياس أداء نماذج اللغة الكبيرة للغة العربية	تقييم هجين (آلي، بشري)

جدول (١): تسلسل أبرز المؤشرات العالمية والعربية لقياس أداء النماذج اللغوية



تتباين مؤشرات تقييم النماذج اللغوية البيئة العربية تبايناً واضحاً من حيث النطاق والمنهجية والفجوات التي يسعى كلُّ منها إلى معالجتها. فعلى صعيد تغطية المهام، تتمحور مؤشرات كـ ALUE وORCA حول فهم اللغة الطبيعية، بينما تستهدف AlGhafa المعرفة والاستدلال عبر أسئلة الاختيار من متعدد، وتنصرف Dolphin وAraGen إلى مهام التوليد كالتلخيص والترجمة. وتُكمل كلُّ من ArabicMMLU وArabLegalEval هذه المنظومة بتغطية المعرفة المتخصصة في المجالين الأكاديمي والقانوني اللذين أغفلتهما المؤشرات السابقة. أما على الصعيد المنهجي، فتسير معظم المؤشرات على نهج المهام المتعددة المستوى من GLUE وSuperGLUE، غير أن AraGen تمثل استثناءً لافتاً بإدخالها إطار التقييم ۳C۳H وتوظيفها لنموذج لغوي حكماً في تقييم المخرجات المفتوحة، مما يعكس تحولاً أعمق في مفهوم الأداء الجيد للنماذج العربية من مقاييس الدقة المجردة نحو تقييم الجودة الشاملة.

وتبرز فجوة متكررة تشترك فيها جميع المؤشرات تقريباً، وهي غياب تقييم صريح للتوافق الثقافي. فعلى الرغم من أن بعض المؤشرات تعكس الأعراف الثقافية بصورة ضمنية من خلال محتوى أسئلتها، إلا أن أيّاً منها لا يقيس بصورة منهجية مدى توافق النموذج مع المنظومة القيمية المحلية والسلوكيات ذات الحساسية الثقافية، وهو قصور جوهري يمس صلاحية هذه النماذج للنشر الفعلي في المجتمعات العربية. ويُضاف إلى ذلك إشكالية التلوث البياني، إذ تعتمد أغلب المؤشرات على بيانات ثابتة متاحة للعموم، ويكاد يكون مؤشر (بلسم) الاستثناء الوحيد بفضل اعتماده على مجموعات اختبار خاصة تُحدَّث بانتظام للحد من هذا الخطر.

وتواجه محاولات تكييف الأطر العالمية مثل: HELM وغيرها مع البيئة العربية جملةً من التحديات البنيوية المتداخلة، أبرزها شُح البيانات، وتعدد اللهجات، وما يفرضه من خيارات منهجية صعبة، والافتراضات المتمركزة حول الثقافة الغربية المضمّنة في مقاييس السمية والتحيز، إلى جانب عدم نضج أدوات التقييم المخصصة للعربية. وما تكشفه هذه المقارنة في مجملها أن المؤشرات العربية قطعت شوطاً لا يُستهان به في تطوير منظومة التقييم، وأن كل مؤشر يسهم بدوره في تغطية جانب من جوانب هذه المنظومة، إلا أن هذه الإسهامات تسير في مسارات متوازية لم تتقاطع بعد في إطار تقييمي موحد وشامل، غير أن المقارنة

مع HELM لا ينبغي أن تكون المعيار الوحيد للحكم على نضج منظومة التقييم العربية، إذ إن خصوصية اللغة العربية بتعدد لهجاتها وثقلها الثقافي وتنوع سياقاتها تستدعي في حد ذاتها أطراً تقييمية مختلفة المنطلقات، وما تحقق حتى الآن يمثل أساساً واعدالاً لبناء منظومة تقييم عربية أصيلة لا مجرد انعكاس للنماذج الغربية القائمة.

٢- تصنيف آليات ومنصات التقييم

لقد طوّر المجتمع البحثي مجموعة متنوعة من آليات تقييم النماذج اللغوية، يمكن تصنيفها إلى مسارين رئيسيين، يتكاملان فيما بينهما ليكوّنا صورة شاملة عن أداء النموذج: وهي التقييم الآلي، والتقييم البشري. تستعرض هذه الدراسة أبرز هذه الآليات مع بيان نقاط قوتها ومواضع القصور فيها، وآليات التكامل بينها.

٢-١- التقييم الآلي (Automatic Evaluation)

يُعدّ التقييم الآلي من أكثر أنواع التقييم استعمالاً، خاصةً في مراحل التطوير الأولية للنماذج وذلك لما يتميز به من سهولة التطبيق، وقابلية التكرار، وسرعة الحصول على نتائج كمية واضحة.

يعتمد هذا النوع من التقييم على مقاييس حسابية تُحسب تلقائياً لمقارنة مخرجات النموذج بالمخرجات المرجعية أو بتوقعات معينة، مثل الدقة (Accuracy)، والتطابق التام (Exact Match)، ومقياس (F1)، وغيرهم.

يعتمد إطار (MMLU) على التقييم الآلي حصراً لقياس مدى قدرة النماذج على التعامل مع المعرفة العامة والمتخصصة ضمن ٥٧ مجالاً أكاديمياً ومهنياً، يشمل تخصصات كالفيزياء، والطب، والقانون، والرياضيات. تركز منهجية التقييم على طرح أسئلة اختيار من متعدد وحساب الإجابات بناءً على مقياس الدقة، مما يتيح إجراء تقييم كمّي عادل وموحد بين النماذج المختلفة. ومن خلال تقديم المهام بنمطين: تعلّم بدون أمثلة، وتعلّم بأمثلة قليلة، تُختبر قدرة النموذج على التعميم والتعلم من عدد محدود من الأمثلة، وهي خصائص مهمة لفهم إمكاناته الفعلية في الفهم والاستدلال (Hendrycks et al., 2021). يُعد

هذا النوع من التقييم الآلي فعالاً من حيث البساطة والموضوعية، ويستخدم على نطاق واسع كمؤشر معياري لأداء النماذج في مهام تتطلب معرفة دقيقة ومنهجية، إلا أنه، كما في بقية الأطر، لا يغني عن التقييم النوعي في المهام التي يصعب قياسها كمياً.

في إطار (BIG-Bench)، يمثل التقييم الآلي حجر الأساس لقياس أداء النماذج على أكثر من ٢٠٤ مهمة متنوعة تغطي مجالات مثل التفكير المنطقي، والحسابي، والأخلاقي، والرمزي، بالإضافة إلى المهارات اللغوية العامة. تقوم منهجيته على توحيد شكل المهام من حيث نوع المهمة، وطبيعة البيانات المدخلة، وهيكلية المخرجات، وآلية الحساب، بهدف تقليل التباين بين طرق التقييم وتسهيل مقارنة النماذج المختلفة. ويستخدم لكل نوع من المهام مقياس أداء مناسب؛ مثل الدقة في مهام التصنيف، والتطابق التام في الأسئلة المغلقة، ومقياس (F1) في مهام الاستخراج. ومع أن هذه المنهجية تتيح تنفيذ تقييم شامل على نطاق واسع وبتكرارية عالية، فإن الباحثين في إطار (BIG-Bench)، يؤكدون على محدودية هذه المقاييس في تقييم الجوانب النوعية للنموذج مثل الإبداع، وسلامة التفكير الأخلاقي، والقدرة على التفسير السياقي، ما يدفعهم إلى استكمالها بأساليب التقييم البشري عند الحاجة.

يجري إطار (HELM) تقيماً شاملاً للنماذج اللغوية، ويستخدم التقييم الآلي لقياس أداء النماذج في المهام التي تمتلك إجابات واضحة وقابلة للتحقق، مثل مهام أسئلة الاختيار من متعدد، أو تصنيف الكيانات، أو استخراج المعلومات. ويستخدم مقاييس كمية مثل الدقة والتطابق التام ومقياس (F1) لتوفير مؤشرات أداء موحدة. تتيح هذه الآلية إمكانية إجراء مقارنة موضوعية بين النماذج المختلفة على طيف واسع من المهام. إلا أن الباحثين يُقرّون بوجود قيود في التقييم الآلي عند التعامل مع المهام المفتوحة أو الإنشائية مثل التلخيص أو كتابة المقالات، حيث يصعب استخدام مؤشرات كمية فقط لتقدير جودة اللغة والمعنى. ولذلك، يوصي إطار (HELM) باستخدام التقييم الآلي كجزء من تقييم شامل يُدمج مع التقييم البشري، لتكوين صورة أكثر توازناً وشموليةً عن أداء النموذج.

يعتمد إطار (MT-Bench) على نهج مبتكر في التقييم الآلي للنماذج اللغوية من خلال اختبار قدرتها على إجراء محادثات متعددة الأدوار (multi-turn dialogue) مع التركيز على

مهارات الفهم السياقي، والاستمرارية، ودقة الإجابة. يتضمن (MT-Bench)، مجموعة من ٨٠ سؤالاً تغطي مجالات متنوعة مثل الحساب، والاستدلال وجمع المعلومات. تُعرض على النماذج على صورة محادثات متتابعة، ثم تُقيّم جودة المخرجات باستخدام نموذج لغوي كبير (Zheng et al., 2023) (LLM-as-a-judge) مثل (OpenAI et al., 2024) (GPT-4)، بدلاً من الاعتماد الكامل على التقييم البشري. وقد أظهرت النتائج توافقاً عالياً بين تقييم النموذج الحكم والتقييمات البشرية بنسبة تتجاوز ٨٠٪، مما يجعل هذا النهج فعالاً من حيث الكفاءة والموثوقية. تبرز أهمية إطار (MT-Bench) في قدرته على تقليل التكاليف البشرية مع الحفاظ على جودة التقييم، خصوصاً في بيئات الاختبار الواسعة، وهو ما يجعله مكملاً مهماً للأطر التقليدية مثل (HELM) و (BIG-Bench) عند التعامل مع نماذج الحوار المفتوح.

يُعد التقييم الآلي أداة ضرورية في تقييم النماذج اللغوية، خاصة في المهام المغلقة والكمية، حيث يوفر قياساً موضوعياً سريعاً ومنخفض التكلفة. إلا أن محدوديته في المهام المفتوحة أو السياقية تُبرز الحاجة إلى استكمالها بتقييم بشري، لضمان تغطية كافة أبعاد الأداء اللغوي بدقة وعمق.

تستعرض الجزئية أدناه بعض أطر التقييم الآلي:

٢-١-١-١ إطار ZeroSumEval

هو إطار تقييم ديناميكي مفتوح المصدر تبنته وحدة (Facebook Research) يهدف إلى تقييم نماذج اللغة الضخمة من خلال ألعاب تنافسية تلزم فائز واحد بين عدة وكلاء، بدلاً من الاعتماد على مجموعات بيانات ثابتة، بحيث ترتفع صعوبة التقييم بزيادة قدرات النموذج؛ ويدعم حالياً سبع ألعاب مدمجة ومنها: الشطرنج، وتحدي الرياضيات، وغيرها من الألعاب. وقد صُممت كل لعبة لاختبار قدرات مختلفة مثل التخطيط الاستراتيجي، توليد المسائل الرياضية وحلّها، كسر تعليمات السلامة (Alyahya et al., 2025).

في التجارب المنشورة التي شملت أكثر من ٧,٠٠٠ محاكاة بين ١٣ نموذجاً لغوياً كبيراً ممثلاً (بما في ذلك Claude 3.7-Sonnet و GPT-4o)، حققت النماذج الرائدة أداءً ممتازاً في

ألعاب الشطرنج والبوكر، لكنها واجهت صعوبات مستمرة في المهام الإبداعية مثل توليد مسائل رياضية صالحة ولكنها تتسم بالتحدي، وإجراء اختراق موثوق لطرف آخر في ألعاب الكلي، رغم وجود فجوات كبيرة في التحديات التي تعتمد على الإبداع بصورة كبيرة.

وبما أن الإطار يقيس الأداء نسبياً (أي أن مكسب نموذج يعني خسارة نموذج آخر)، فإنه يتجنب تشبع المؤشر الذي يحدث في المؤشرات التقليدية حتى مع تطور النماذج، مما يقدم بديلاً قابلاً للتوسع وقوياً أمام المنافسة مقارنة بأساليب التقييم التقليدية لنماذج اللغة. ولكن بنفس الطبيعة القياسية، كون قياس الأداء يلزم فوز أحد النماذج وخسارة الآخر قد يسبب خسارة بعض التفاصيل المهمة في كل مرحلة من اللعبة.

٢-١-٢ منصة OpenAI Evals

هو إطار عمل مفتوح المصدر طورته شركة (OpenAI) بهدف تقييم النماذج اللغوية الكبيرة وقياس أدائها من خلال اختبارات منظمة. يتيح هذا الإطار للباحثين والمطورين والمؤسسات إنشاء وتنفيذ ما يسمى بـ "التقييمات" (Evals) وهي اختبارات مصممة لقياس أداء النموذج، عن طريق بناء أسئلة وأجوبة وآلية واضحة لقياس أداء النموذج. كما يتميز بتكامله الوثيق مع واجهة برمجة تطبيقات (API) مما يسهل تقييم أداء النماذج وجمع النتائج بطريقة موحدة وقابلة لإعادة الإنتاج. إضافة إلى ذلك، يدعم (OpenAI Evals) مساهمات المجتمع البحثي عبر توفير معايير تقييم مشتركة تعزز الشفافية وقابلية المقارنة بين إصدارات النماذج المختلفة.

من أبرز مزاياه أنه سهل الاستخدام، ويضمن إعادة الإنتاج، مما يجعله خياراً ممتازاً للمطورين الراغبين في تتبع التحسينات منهجياً بين الإصدارات أو النماذج المعدلة (Fine-Tuned) كما أن طبيعته مفتوحة المصدر تتيح للمؤسسات توسيعه ليناسب احتياجاتها الخاصة.

ومع ذلك، هناك بعض القيود وعلى الرغم من تغطيته لمجموعة واسعة من المهام إلا أنه لا يعكس دائماً الأداء التفاعلي الواقعي إلا عند تخصيصه بشكل موسع. ورغم هذه

التحديات، يظل (OpenAI Evals) أداة قوية وشفافة تدعم البحث القابل لإعادة الإنتاج وتتبع الأداء وجهود مواءمة النماذج في المشهد المتسارع لتطور النماذج اللغوية الكبيرة.

٢-١-٣ - إطار Language Model Evaluation

منظومة شاملة مفتوحة المصدر جاءت ضمن جهود شركة (Google)، وتقوم على مبدأ توحيد آليات تقييم النماذج اللغوية الكبيرة (LLMs)، وتوسيع قدرات آليات التقييم لتستوعب بيانات متعددة المزودين. ويبرز دورها الجوهرى في إزالة الحواجز التقنية الناتجة عن اختلاف واجهات برمجة التطبيقات (API) من مزود خدمة لآخر. إذ تمنح المستخدمين القدرة على استعراض ومقارنة أداء نماذج ذكاء اصطناعي مختلفة وبارزة كـ (Gemini)، (Claude)، (GPT)، و (Llama) (Touvron et al., 2023) وغيرها من النماذج، ضمن بيئة موحدة ومتكاملة.

يدعم (Google, 2025) (LMEval)، مجموعة واسعة من أنواع المهام، بما في ذلك توليد نصوص الإجابة عن أسئلة الاختيار من متعدد، توليد النصوص البرمجية، توليد وصف للصور، بالإضافة إلى اختبارات الأمان، مما يجعله مناسباً للنماذج النصية والنماذج متعددة المدخلات (Multimodality). تعتمد المنصة على مفهوم التقييم التدريجي (Incremental Evaluation) الذي يسمح بتنفيذ الاختبارات المطلوبة أو المحدثة فقط بدل إعادة تشغيل جميع الاختبارات في كل مرة، مما يوفر الوقت ويقلل التكلفة.

كما توفر المنصة لوحة عرض تفاعلية باسم (LMEvalboard) تمكّن الباحثين والمطورين من استعراض الأداء الإجمالي للنماذج، وإجراء مقارنات مباشرة بين نماذجين، وتحليل نقاط القوة والضعف لكل نموذج من خلال تقارير ورسوم بيانية واضحة. تتميز المنصة بقدرتها العالية على التوسع وإضافة اختبارات جديدة بسهولة مما يجعلها مناسبة للأبحاث المتقدمة والتقييم المستمر للنماذج بعد كل تحديث. ومع ذلك، فهي لا تغني بالكامل عن المراجعة البشرية، خصوصاً في المهام الحساسة التي تتطلب فهماً ثقافياً أو سياقياً عميقاً.

٤-١-٢-٤ إطار Language Model Evaluation Harness

هو إطار شامل وموحد لاختبار الأداء، صُمم لتقييم نماذج اللغة على مجموعة واسعة من مهام معالجة اللغة الطبيعية باستخدام منهجية موحدة وقابلة لإعادة الإنتاج. يدعم هذا الإطار أكثر من ٦٠ معيارًا معتمدًا على نطاق واسع. يتميز إطار (LMHarness) (Gao et al., 2024) بمرونة عالية حيث يعمل مع النماذج مفتوحة المصدر والنماذج التجارية المعتمدة على واجهات تطبيقات برمجية.

من أبرز مزايا إطار (LMHarness) أنه يوفر إمكانية إعادة الإنتاج والمقارنة العادلة بين النماذج المختلفة تحت ظروف قياس موحدة. كما أن تصميمه المعياري يتيح للمستخدمين إنشاء مهام جديدة وتعديلها بسهولة. على الرغم من ذلك توجد بعض القيود؛ إذ يركز الإطار بصورة رئيسية على المعايير الأكاديمية الثابتة بدلًا من التقييم التفاعلي الواقعي، كما قد يكون مكلفًا من ناحية الموارد عند تشغيل النماذج كبيرة الحجم، ما يتطلب إدارة دقيقة للموارد الحاسوبية. ورغم هذه التحديات، يظل واحدًا من أكثر الأدوات موثوقية لتقييم النماذج اللغوية الكبيرة بطريقة معيارية متماشية مع أفضل ممارسات المجتمع البحثي.

ومع أن هذه المقاييس تعد أدوات فعالة لمقارنة الأداء، إلا أنها لا تعكس بالضرورة جودة التجربة البشرية، أو مدى ملاءمة المحتوى للسياق الثقافي واللغوي المحلي. لذلك غالبًا ما تكون مكتملة، لا بدليلاً عن أساليب تقييم أعمق.

٢-٢-٢ التقييم البشري (Human Evaluation)

يبقى التقييم البشري عنصرًا أساسيًا في الحكم على جودة مخرجات النماذج، لا سيما في المهام التي يصعب تقويمها رقميًا مثل الطلاقة، والمنطق، والملاءمة، والبلاغة. فالمقيم البشري قادر على تمييز التفاصيل الدقيقة في اللغة، التي قد تعجز الخوارزميات الآلية عن التقاطها.

التقييم البشري يُعد ركيزة أساسية في تقييم نماذج اللغة الكبيرة، خصوصًا مع تزايد استخدامها في المهام الواقعية المفتوحة حيث تُثبت طرق التقييم الآلية عدم كفايتها (Liang et al., 2023). فقد أظهرت الدراسات أن مثل مقاييس (BLEU) و (ROUGE) رغم

فاعليتها في مهام المعالجة اللغوية التقليدية مثل الترجمة، إلا أنها لا ترتبط ارتباطاً جيداً بأحكام البشر عند تقييم صفات مثل صحة النص، والتماسك، والأصالة، والفائدة العامة (Al-Matham et al., 2025)، لا سيما في مهام مثل التلخيص، والحوار، والإجابة عن الأسئلة. وعلى النقيض من ذلك يمكن للمُقيمين البشريين اكتشاف المغالطات المنطقية، وحدة الشاعر، وخصوصية السياق، والتحيزات وهي فروق دقيقة تعجز أساليب التقييم الآلية عن التعامل معها بثقة. ولهذا السبب، يُعد التقييم البشري هو المعيار الأهم في المجالات الحساسة للتفاصيل الدقيقة مثل الرعاية الصحية، والقانون، والتعليم، حيث تمثل السلامة والدقة والسياق عناصر بالغة الأهمية.

ومع ذلك، فإن التقييم البشري كما أن له مميزات، له عدد من السلبيات: منها أن هذه العملية تتطلب قوى عاملة كبيرة، وتكلفة أعلى عند تطبيقها على مستوى أكبر، وغالباً ما تعاني من تباين بين المُقيمين ومن التحيزات المعرفية، مثل تحيز الطلاقة، حيث تخطى الردود المُصاغة جيداً ولكنها غير صحيحة بتقييمات أعلى مما تستحق. ولمواجهة هذه التحديات، طوّر المجتمع البحثي مجموعة متزايدة من معايير التقييم التي تدمج المدخلات البشرية، مثل (SummEval) للتلخيص (Fabbri et al., 2021)، و(HumanEval) لتوليد النصوص البرمجية (Chen et al., 2021)، و(HELM) للتقييم الشامل للنماذج عبر معايير متعددة مثل التحيز، والمرونة، والمعلومات المضللة.

وتوضح هذه المؤشرات رؤى رئيسية، مثل مدى فشل العديد من النماذج في الاستدلال ضمن سياقات طويلة، أو اختلاقيها للحقائق، أو تفاوت أدائها باختلاف بنية أو هيكلية المُدخلات أو موضوعها. وبالتوازي مع تطور المؤشرات، برز التعلم المعزز من التغذية الراجعة البشرية (Christiano et al., 2023) كنموذج مهيمن لمحاذاة مخرجات النماذج مع الأساليب المصممة من الإنسان. تبدأ هذه العملية بضبط النموذج مسبقاً باستخدام مدخلات ومخرجات عالية الجودة، يتبعها جمع إجابات يفضلها الإنسان لتدريب نموذج المكافأة، ثم تستخدم خوارزميات التعلم المعزز لضبط سلوك النموذج بما يحقق أقصى استفادة من المكافأة المكتسبة.

وقد أسفر هذا النهج عن نتائج محورية؛ فعلى سبيل المثال، تفوق نموذج (Ouyang et al., 2022) (InstructGPT) بـ ١,٣ مليار معامل على نظيره الأكبر حجمًا ١٧٥ مليار معامل من خلال تقديم إجابات صنفها البشر باستمرار على أنها أكثر فائدة وأمانًا. ولتعزيز الدقة وقابلية التكرار في ممارسات التقييم، ظهرت أطر منهجية مثل إطار (ConSiDERS) (Elangovan et al., 2024) الذي يحدد ستة ركائز: الاتساق، ومعايير التقييم، والتمايز، وتجربة المستخدم، والمسؤولية، وقابلية التوسع، وذلك لتشجيع الفرق على معالجة التحيزات بوضوح، وتوضيح معايير التقييم، وتضمين مؤشرات موضوعية وأخرى خاصة بالمهمة.

ومع تطور نماذج اللغة وزيادة انتشارها، يظل التقييم البشري المدعوم بتصميم مدروس، ومبادئ التقييم الواضحة، وتكرار مستمر عنصرًا هامًا؛ لضمان أن تكون هذه النماذج ليست قوية فحسب، بل أيضًا متوافقة مع التوقعات البشرية ومعايير السلامة.

تُستخدم هذه الأساليب على نطاق واسع في تقييم نماذج الحوار، وتُعد المرجع الأساس في مشاريع مثل: (Anthropic)، (OpenAI)، (Chiang et al., 2024) (Chatbot Arena)، ورغم أنها تتطلب وقتًا وجهدًا وميزانية أكبر، إلا أنها تقدم فوائد لا يمكن الاستغناء عنها لضبط جودة النموذج.

ومن أبرز الآليات للتقييم البشري للنماذج اللغوية.

٢-١- منصة Chatbot Arena

تعد منصة (Chatbot Arena) واحدة من أبرز المنصات التي تعتمد على التقييم البشري التفاعلي لقياس أداء النماذج اللغوية. وتعتمد هذه المنهجية على عرض إجابتين من نموذجين مختلفين لنفس المدخل، ويُطلب من المستخدم اختيار الإجابة الأفضل من وجهة نظره. تختلف طريقة العرض بين الحالتين: إما إخفاء هوية النماذج (أي دون الإشارة إلى اسم النموذج المُنتج للإجابة)، أو عرض الأسماء بصراحة، وذلك وفقًا لإعدادات المنصة وأهداف التقييم.

في حالة إخفاء هوية النماذج، تبرز عدة مزايا مهمة أبرزها تقليل التحيز الناتج عن معرفة الجهة المطورة للنموذج، مما يساعد في الحصول على تقييم أكثر حيادية. كما أن هذه المنهجية تتيح استقبال عدد كبير من المشاركين في التقييم دون أن يؤثر ذلك على استقرار النتائج، وهو ما يُسهم في تحسين دقة التصنيفات وتحديثها باستمرار. فضلاً عن ذلك، فإن التقييم في هذه الحالة يعكس تجربة المستخدم الحقيقية، إذ يأخذ في الحسبان عناصر مثل وضوح الإجابة وتنظيمها، وليس فقط دقتها العلمية.

ورغم هذه المزايا، إلا أن هذا الأسلوب يواجه عدداً من التحديات. فقد تتأثر تفضيلات المستخدمين بعوامل شكلية مثل طول الإجابة أو أسلوبها اللغوي، مما قد يؤدي إلى انحياز غير مقصود نحو نماذج تستخدم لغة منمقة أو سرداً جذاباً، حتى وإن كانت أقل دقة من الناحية الموضوعية. كما أن الاعتماد على التفاعل البشري المباشر يجعل النتائج عرضة للتأثر بالعوامل الظرفية للمقيّم، مثل حالته المزاجية أو خلفيته المعرفية، ما يُصعب إعادة إنتاج نفس النتائج بدقة لاحقاً (Bai et al., 2024). يضاف إلى ذلك احتمالية التحيز في عينة المستخدمين، إذ قد تنعكس في النتائج ميول فئة سكانية معينة من حيث اللغة أو الثقافة، مما يحد من إمكانية تعميم التقييم على مجتمعات أوسع أو تطبيقات متعددة اللغات.

أما في حالة عرض أسماء النماذج بصراحة، فتظهر مزايا من نوع آخر. إذ تتيح هذه الطريقة للباحثين والمطورين مقارنة أداء النماذج مباشرةً، مما يدعم عملية اتخاذ القرار عند اختيار النموذج الأنسب لمهامهم الخاصة. كما أن هذا الأسلوب يحفز المنافسة بين مزودي النماذج، إذ يدفعهم لتحسين جودة أدائهم لتجنب التقييمات السلبية العلنية.

مع ذلك، لا تخلو هذه المنهجية من تحديات. فمعرفة اسم النموذج قد يدفع بعض المستخدمين للانحياز لنماذج تحمل أسماء معروفة أو علامات شهيرة، ما يؤدي إلى تفضيلات غير موضوعية، حتى في حال كانت جودة الإجابة أقل. كما قد تفتح هذه المنهجية الباب أمام محاولات التلاعب بالتقييمات، من خلال التصويت الجماعي الموجه أو الحملات الترويجية التي تهدف إلى تحقيق نتائج تقييم غير فعلية.



تقدم منصة (Chatbot Arena) تجربة فريدة في مجال تقييم النماذج اللغوية، حيث تجمع بين بساطة التفاعل البشري وعمق التحليل السلوكي للمستخدم، مما يجعلها أداة قوية لفهم أداء النماذج من منظور المستخدم النهائي.

٣-٢- التقييم الهجين (Hybrid Evaluation)

في حين أن التقييم الآلي يتميز بالسرعة وقابلية التكرار، والتقييم البشري يتميز بالقدرة على فهم السياق واللغة الطبيعية، يظهر التقييم الهجين كحل وسط يجمع بين مزايا الطريقتين. ويعد مؤشر (MT-Bench) (Bai et al., 2024) أحد أشهر المؤشرات التي تتبع التقييم الهجين، حيث تُدمج نتائج المقاييس التلقائية مع الأحكام البشرية أو استخدام نموذج لغوي ضخم كمُقيّم بديل (LLM-as-a-judge).

٣-٢-١- نموذج لغوي كمحكم (LLM-as-a-Judge)

يعد من الأساليب التي ظهرت استجابةً للتحديات العملية المرتبطة بالتقييم البشري التقليدي، مثل ارتفاع التكلفة وصعوبة التوسع في التقييمات. تقوم هذه الطريقة على توظيف نموذج لغوي ضخم ذي أداء عالٍ مثل (GPT-4) للقيام بدور المُقيّم وإصدار الأحكام على مخرجات نماذج لغوية أخرى. يُوجّه النموذج ليقراً المخرجات الناتجة من نموذج أو عدة نماذج أخرى، ثم يحكم عليها بناءً على مقاييس محددة مثل الدقة، الفائدة، الإيجاز، الإبداع، التماسك.

تشمل خطوات التنفيذ إعداد أمر (Prompt) مخصص يشرح للنموذج الحكم المطلوب منه بدقة، مع تزويده بمعايير التقييم وتعريفها بالتفصيل، ثم إعطاؤه المدخلات والمخرجات الخاصة بالنموذج المراد تقييمه ويقوم النموذج بعد ذلك بإصدار تقرير أو تقييم رقمي (مثل إعطاء درجة من ١ إلى ٥ لكل معيار) أو تفضيل بين إجابتين. ويُعطى النموذج تعليمات دقيقة توضح له آلية الحكم والمعايير المطلوب استخدامها في التقييم بحيث يُطلب منه مثلاً إعطاء درجة أو تحديد الترتيب بين مخرجات متعددة (Bai et al., 2024).

يعدّ النموذج ليؤدي مهمة تقييم استجابات النماذج الأخرى وفق معايير محددة مسبقاً، مثل الصحة (مدى دقة المعلومات وخلوها من الأخطاء المنطقية)، والإفادة

(مدى تلبية الإجابة لاحتياج المستخدم)، والوضوح والاختصار (مدى وضوح الإجابة وتركيزها دون إطالة أو تشتيت).

وقد برزت هذه الطريقة نظراً لقدرتها على التوسع وتقييم آلاف المخرجات بسرعة دون الحاجة إلى فرق بشرية كبيرة، ولأنها تقلل من التكلفة والوقت مقارنة بالتقييم البشري الكامل، إضافة إلى قابليتها للتكرار وإنتاج نتائج ثابتة عند استخدام النموذج والتعليمات نفسها.

ورغم مزاياها، إلا أن هذه الطريقة تواجه تحديات، مثل الانحياز المتراكم الذي قد يعيد إنتاج التحيزات الكامنة في بيانات التدريب ويؤثر على نزاهة التقييم، والحاجة للمعايرة المستمرة لضمان توافق أحكام النموذج مع معايير التقييم البشري، بالإضافة إلى حدود الفهم العميق، حيث قد يعجز النموذج عن إدراك اعتبارات بشرية دقيقة كالحس الثقافي أو السياقي.

يوفر أسلوب "النموذج كحكم" حلاً مبتكراً لتقييم النماذج اللغوية بكفاءة وسرعة، خاصة عند دمجه مع عينات تقييم بشرية لضمان الجودة. ومع ذلك، فإن الاعتماد الكامل عليه دون مراجعة بشرية قد يعرض نتائج التقييم لمخاطر الانحياز وفقدان الحس النقدي، مما يستلزم توظيفه ضمن منظومة تقييم هجينة متكاملة لضمان أقصى موثوقية وموضوعية ممكنة.

٣- لوحات الصدارة

تلعب لوحات الصدارة دوراً محورياً في تقييم النماذج اللغوية الكبيرة من خلال توفير إطار موحد وواضح لمقارنة أداء النماذج عبر مهام متنوعة. ومن خلال جمع النتائج في مكان واحد، تساعد لوحات الصدارة في تحديد النماذج المتقدمة وإبراز نقاط القوة والضعف عبر المجالات المختلفة، وتوجيه المجتمع العلمي نحو معالجة الفجوات القائمة. علاوة على ذلك، تشجع هذه اللوحات على المنافسة الصحية وتساهم في تسريع الابتكار، حيث يسعى مطورو النماذج إلى تحقيق مراتب أعلى في المعايير المعترف بها. وفي القطاع الصناعي تؤدي لوحات الصدارة أيضاً دوراً كأداة لاتخاذ القرارات، إذ تساعد المتخصصين في اختيار النموذج الأنسب



بناءً على الأداء والكفاءة ومتطلبات السلامة. وفي النهاية، تُعد لوحات الصدارة عنصرًا أساسيًا لدفع عجلة التقدم والحفاظ على الشفافية في مشهد النماذج اللغوية الكبير سريع التطور. ومن أبرز لوحات الصدارة:

- أ. HELM: إطار يوفر تقييمًا شاملاً متعدد الأبعاد يغطي عدد من المقاييس الدقيقة (سيتم التفصيل فيها في القسم التالي) من خلال عدة نسخ مختلفة من المؤشر.
- ب. Chatbot Arena: منصة تفاعلية مفتوحة المصدر تعتمد على مواجهات مباشرة بين النماذج حيث يقوم المستخدمون بتقييم الاستجابات مباشرة، ما يعكس الأداء الواقعي في المحادثات.
- ت. Open LLM Leaderboard: منصة تتحدث باستمرار، تستعرض أداء النماذج مفتوحة المصدر على مجموعة من المهام القياسية مع تحديثات مستمرة، ما يجعلها مرجعًا مهمًا للمجتمع البحثي والمطورين.

في القسم التالي من هذه الدراسة، سننتقل إلى بحث أكثر تفصيلاً للمقاييس الأكثر استخدامًا في تقييم النماذج، مع تحليل آلياتها ومتى يمكن الاعتماد عليها، وكيفية تفسير نتائجها بدقة.

٤- مقاييس تقييم أداء النماذج اللغوية

٤-١- الدقة (Accuracy)

تُعد الدقة من أبرز المقاييس المستخدمة في التقييم الآلي لأداء النماذج اللغوية، خصوصًا في المهام التي تتطلب إجابات مغلقة أو محددة، مثل مهام التصنيف أو أسئلة الاختيار من متعدد. في ورقة إطار (HELM) يُستخدم مقياس الدقة كمؤشر أساسي لقياس مدى صحة المخرجات التي ينتجها النموذج مقارنةً بالإجابات المرجعية، حيث تُحسب نسبة عدد الإجابات الصحيحة التي يقدمها النموذج إلى العدد الكلي من الأسئلة.

ورغم قوة هذا المؤشر ووضوحه، إلا أن الدقة تعد مقياسًا محدودًا في المهام الأكثر تعقيدًا، مثل توليد النصوص، أو الإجابات المفتوحة، حيث لا تكون هناك إجابة واحدة صحيحة يمكن

اعتمادها. كما أن مقياس الدقة لا يأخذ في الحسبان جودة اللغة أو مدى ملاءمتها للسياق، ولا يُعبر عن التدرج في صحة الإجابات أو قربها من الصواب، مما يتطلب الاستعانة بمقاييس إضافية أو التقييم البشري. ومع ذلك، تبقى الدقة من المقاييس الجوهرية التي لا غنى عنها في تقييم المهام المغلقة، وتُستخدم على نطاق واسع في معظم أدوات التقييم الآلي.

٤-٢- المعايرة (Calibration)

تعد المعايرة من المقاييس المتقدمة والمُكملة لمقياس الدقة في تقييم أداء النماذج اللغوية، إذ لا يكفي أن يكون النموذج دقيقاً في مخرجاته فحسب، بل من المهم أيضاً أن تكون درجات الثقة التي يُعبر بها عن هذه المخرجات متناسبة مع احتمال صحتها الفعلية. بمعنى آخر، إذا ذكر النموذج أنه متأكد بنسبة ٩٠٪ من إجابته، فإننا نتوقع أن تكون تلك الإجابة صحيحة في ٩٠٪ من الحالات.

يُطلق على هذه العلاقة بين درجة الثقة ونسبة الصحة الفعلية اسم المعايرة، وهي ضرورية لتقييم مدى موثوقية النموذج عند استخدامه في البيئات الحساسة، مثل المجالات الطبية أو القانونية، حيث قد تؤثر ثقة النموذج الزائدة أو الضعيفة في إجابته على قرارات حاسمة.

في ورقة (HELM) تُقاس المعايرة غالباً باستخدام مقياس يُعرف باسم (Expected Calibration Error - ECE) والذي يُعبر عن الفرق المتوقع بين متوسط درجات الثقة ومتوسط الدقة الفعلية عبر عدة فئات من التوقعات.

كلما اقتربت قيمة (ECE) من الصفر، دلّ ذلك على أن النموذج مُعاير جيداً. أما القيم المرتفعة فتشير إلى وجود فجوة بين الثقة والدقة، مما يعني أن النموذج إما مبالغ في ثقته أو مُتحفظ أكثر من اللازم. تُظهر نتائج (HELM) أن النماذج الأكبر تميل إلى أن تكون أكثر دقة من حيث الأداء، ولكنها ليست بالضرورة أكثر معايرة. لذا، يُعد هذا المؤشر مكملاً أساسياً للدقة، لأنه يقيّم جودة التوقعات الاحتمالية للنموذج، وليس فقط صحتها. ولذا، يُوصى دائماً بدمج مقياس المعايرة مع الدقة للحصول على صورة أكثر شمولاً عن أداء النموذج، خصوصاً عند نشره في تطبيقات تتطلب قرارات دقيقة مبنية على درجات الثقة.

٤-٣- المتانة (Robustness)

تعد المتانة من المقاييس الجوهرية في تقييم أداء النماذج اللغوية، وهي تُعبر عن قدرة النموذج على الحفاظ على جودة مخرجاته عند تعرّضه لتغيرات في مدخلاته. في ورقة إطار (HELM) يُنظر إلى المتانة على أنها بُعد أساسي يُكمّل مقاييس مثل الدقة، والمعايرة، حيث إن النموذج قد يكون دقيقًا في الحالات العادية، لكنه يفشل عند مواجهة حالات غير مألوفة أو عند وجود ضوضاء لغوية، أو إعادة صياغة بسيطة، أو تغيير في ترتيب الكلمات. وهو ما يُعد مقياسًا لمتانة النموذج في البيئات الواقعية التي قد تتضمن تنوعًا كبيرًا في أساليب التعبير والسياق.

وقد أظهرت نتائج ورقة (HELM) أن العديد من النماذج الكبيرة تُظهر أداءً متفاوتًا من حيث المتانة، فبعضها ينخفض أدائه بشكل ملحوظ عند أدنى تغيير في المدخلات. وهذا يُبين أن الاعتماد على التقييم في بيئة مثالية فقط لا يكفي، بل من الضروري اختبار النموذج في ظروف متنوعة ومتغيرة نسبيًا لضمان قدرته على التعميم والتعامل مع تنوع الاستخدامات اللغوية في التطبيقات الفعلية.

٤-٤- العدالة (Fairness)

تشكل العدالة أحد أهم المقاييس الأخلاقية في تقييم أداء النماذج اللغوية، وهي تُشير إلى مدى حيادية النموذج وعدم تحيزه تجاه فئات معينة من المستخدمين بناءً على العرق، أو النوع الاجتماعي، أو اللغة، أو اللهجة، أو الخلفية الثقافية. يؤكد مشروع (HELM) أن تقييم النماذج اللغوية لا يكتمل دون النظر في الأبعاد الاجتماعية والأخلاقية المرتبطة بمخرجات هذه النماذج، خاصةً مع استخدامها المتزايد في تطبيقات تؤثر مباشرة على الأفراد. كونه نموذجًا يحقق دقة عالية لكنه يُنتج مخرجات منحازة، قد يُعزز الظلم ويقوّض الثقة العامة في النماذج اللغوية الضخمة.

تُقاس العدالة باستخدام مؤشرات كمية تُظهر اختلال التوزيع في جودة المخرجات أو في المحتوى المولّد عبر المجموعات الديموغرافية المختلفة. كلما زادت هذه الفجوة،

دلّ ذلك على وجود تحيز أكبر في النموذج، مما يتطلب تدخلاً لتحسين توازن بيانات التدريب أو تعديل آليات التوليد. يُستخدم هذا المقياس في (HELM) لتقييم مدى اتساق النموذج عبر المجموعات.

وتجدر الإشارة إلى أن العدالة لا تقتصر فقط على تجنب السمية أو التمييز الواضح، بل تشمل أيضاً ضمان التمثيل العادل للأراء والأفكار والثقافات في مخرجات النموذج. لذلك، تُوصي ورقة (HELM) بتكامل مقياس العدالة مع مقاييس أخرى مثل السمية والانحياز والمعايرة للحصول على رؤية شاملة لسلامة النموذج من النواحي الأخلاقية والاجتماعية. كما تظهر الورقة أن التحيز لا يُقاس فقط من خلال تحليل المحتوى الصريح، بل قد يظهر أيضاً في تفضيلات ضمنية أو في استجابات ظاهرها محايد لكنها تدعم سلوكاً غير عادل عند تكراره على نطاق واسع. يُقاس التحيز من خلال أدوات وآليات متخصصة ترصد الميل نحو إنتاج محتوى نمطي أو مسيء.

٤-٥- السمية (Toxicity)

تعد السمية من المقاييس الأساسية التي يعتمد عليها مشروع (HELM) (Liang et al., 2023) في تقييم أداء النماذج اللغوية، لا سيما من زاوية الأمان وسلامة الاستخدام. يشير مفهوم السمية إلى ميل النموذج لإنتاج مخرجات تتضمن ألفاظاً أو عبارات عدائية، مسيئة، عنيفة، أو مهينة، سواء كانت موجهة لأفراد أو جماعات. وتكمن خطورة هذا النوع من المخرجات في احتمالية استخدامها ضمن تطبيقات حوارية أو تعليمية أو خدمية، مما قد يؤدي إلى أضرار نفسية أو اجتماعية.

يعتمد إطار (HELM) في قياس السمية على أدوات تقييم خارجية، مثل (Perspective API) التي تقدّم نموذجاً مدرّجاً لتقدير احتمالية كون نص ما "ساماً". يُعبّر عن ذلك كمقياس احتمالي يتراوح بين ٠ و ١. أي أن مخرجات النموذج تُمرّر إلى نموذج آخر (المصنّف)، والذي يُقدّر احتمالية أن تكون هذه المخرجات سامة. يُستخدم هذا المقياس لتقييم متوسط السمية عبر مجموعة من المخرجات، كما يُقارن بين النماذج المختلفة من حيث ميلها لإنتاج محتوى ضار.



التمييز الذي يقدمه مشروع (HELM) بين السَمِيَّة والتَحِيْز، والقوالب النمطية يُعدّ مهمًّا ومبنيًّا على أساس مفاهيمي دقيق. في حين أن السَمِيَّة تُركِّز على حدة اللغة وعدائيتها، أي كون المخرجات هجومية أو جارحة أو تنطوي على ألفاظ مسيئة، فإن التحيز والأنماط النمطية تتعلق بالافتراضات الضمنية والانحيازات المعرفية التي تنعكس في المحتوى، كربط مهنة معينة بجنس أو عرق معين، أو تقديم تصوّر نمطي عن جماعة ما، حتى لو كان النص نفسه غير رمسيء لفظيًّا. بمعنى آخر، قد يكون المحتوى متحيزًا دون أن يكون سامًّا، وقد يكون سامًّا دون أن يتضمّن تحيزًا نمطيًّا. ولهذا، يفصل (HELM) بين المقياسين، ويقدم لكل منهما آليات رصد وتحليل مستقلة، مما يضمن تقييمًا أدق وشمولية أكبر في فحص سلوك النماذج اللغوية. كما يُشجع الباحثين على استخدامهما كليهما عند تحليل الجوانب الأخلاقية للنماذج، خاصة عند نشرها في تطبيقات تلامس الجمهور العام.

٤-٦- الكفاءة (Efficiency)

تشكل الكفاءة أحد المقاييس الجوهرية التي يتناولها مشروع (HELM) لتقييم أداء النماذج اللغوية الكبيرة، حيث لا تقتصر أهمية النموذج على جودة مخرجاته فقط، بل تشمل أيضًا مدى فاعليته في استخدام الموارد الحوسبيّة أثناء التشغيل. ويكتسب هذا البعد أهمية خاصة في البيئات الواقعية التي تتطلب تشغيل النماذج ضمن موارد محدودة، سواء من حيث الطاقة، أو الذاكرة، أو التكلفة المالية.

وتتيح هذه المؤشرات إمكانية مقارنة النماذج من حيث سرعة الاستجابة وكفاءة استغلال الموارد، وهو أمر حاسم عند اتخاذ قرارات تتعلق باستخدام النموذج في التطبيقات التفاعلية (مثل روبوتات المحادثة أو المهام المحدودة بزمن). يُظهر (HELM) من خلال نتائجه أن هناك تباينًا كبيرًا بين النماذج في هذا الجانب، حيث قد تُقدّم بعض النماذج جودة لغوية مرتفعة، ولكن على حساب استهلاك عالٍ جدًا للموارد، مما يجعلها غير عملية في بعض البيئات. ولذلك تؤكد الدراسة على ضرورة النظر إلى الكفاءة جزءًا أساسيًا في تقييم النماذج، إلى جانب المعايير الأخرى.

٤-٧- التكلفة (Cost)

تمثل التكلفة أحد الأبعاد المركزية في تقييم نماذج اللغة الكبيرة، خاصة مع تزايد حجم هذه النماذج وتعقيدها، حيث أصبحت تكلفة التدريب والتشغيل من أبرز العوامل التي تؤثر على تبني هذه النماذج في بيئات الأعمال. تُشير عدد من مشاريع مثل (HELM)، إلى أن الأداء العالي للنموذج غالباً ما يكون مصحوباً بتكلفة مرتفعة سواءً من حيث الموارد السحابية أو استهلاك الطاقة أو الزمن اللازم لتنفيذ المهمة. في هذا السياق، تُعرف التكلفة على أنها مجموع الموارد المالية والتقنية اللازمة لتشغيل النموذج على مجموعة من المهام، ويُعبّر عنها غالباً بصيغ متعددة مثل الدولار لكل ١,٠٠٠ وحدة أو الطاقة المستهلكة بالكيلوواط / ساعة لكل مهمة.

تُحسب التكلفة التشغيلية من خلال العلاقة بين عدد العمليات الحسابية وتكلفة تنفيذها على مزودات الحوسبة السحابية. يُكمل هذا التقييم بمؤشرات مثل وقت التنفيذ الكلي (Total Latency) واستهلاك الذاكرة (Memory Footprint)، وعدد وحدات المعالجة المطلوبة (GPU Hours) ومن المهم دمج تقييم التكلفة ضمن معايير الجودة، ليس فقط من زاوية الكفاءة، بل أيضاً لضمان استدامة استخدام النموذج وتوفره في البيئات ذات الموارد المنخفضة. وهكذا، فإن تكلفة تشغيل النموذج لم تعد مسألة بحثية، بل عامل حاسم في قرارات النشر، والتوسع، والوصول العادل لتقنيات الذكاء الاصطناعي اللغوي.

٤-٨- المعلومات المضللة (Misinformation)

ينظر إلى المعلومات المضللة باعتبارها من أخطر الإشكالات الناتجة عن استخدام النماذج اللغوية الكبيرة، لا سيما في المهام المفتوحة التي تتضمن توليد محتوى نصي حرّ ركزت أوراق مثل (HELM) و (BIG-Bench) و (MT-Bench) على هذا البُعد بشكل متزايد، نظراً لقدرة هذه النماذج على إنتاج إجابات تبدو موثوقة وصحيحة من الناحية اللغوية، لكنها قد تكون غير دقيقة أو تحتوي على مغالطات معرفية خطيرة. تختلف المعلومات المضللة عن الهلوسة العشوائية من حيث إنها قد تتسرب في الإجابات تحت غطاء الطلاقة والوضوح، ما يجعل كشفها أكثر صعوبة.

في إطار مشروع (HELM) يتم تقييم هذا البعد باستخدام مقاييس مثل الهلوسة (Hallucination Rate)، والتي تُعرّف على أنها نسبة المخرجات التي تتضمن معلومات غير مدعومة من مصادر موثوقة أو مرجع ذهبي. تُقاس الهلوسة غالبًا عبر مراجعة بشرية تعتمد على مقارنة المخرجات بالإجابات الصحيحة أو المعتمدة.

كما يُمكن في بعض المهام -مثل الأسئلة المعرفية أو الطبية- استخدام تقنيات تحقق خارجية أو مصادر موثوقة لتحديد ما إذا كانت المعلومة المقدمة صحيحة أم لا. في (BIG-Bench) وضمن المهام المتعلقة بالتفكير الأخلاقي والمعرفي، وُجد أن النماذج قد تُنتج إجابات تحتوي على مغالطات منطقية أو تأكيدات مضللة، خصوصًا في المهام التي تتطلب تفسيرًا أو استنتاجًا. أما (MT-Bench)، الذي يُقيّم أداء النماذج في المحادثات، فقد أظهر كيف يمكن للنموذج أن يتسبب في نشر معلومات خاطئة بثقة، خاصة عند الإجابة على أسئلة ذات طابع إخباري أو سياسي.

هذه المخاطر تستوجب إدراج وجود معيار خاص بالموثوقية المعرفية ضمن أدوات التقييم الشاملة، من أجل التحقق من قدرة النموذج على تمييز الصحيح من الخاطئ، وليس فقط إنتاج جُمْل سليمة نحويًا. ويُعد هذا المعيار مهمًا في التطبيقات الطبية، والتعليمية، والقانونية، وغيرها من المجالات التي لا تحمل المعلومات المضللة، وهو ما يجعل تقييم المعلومات المضللة من أهم أولويات التطوير والاستخدام الآمن للنماذج اللغوية.

٥- أبرز المهام المستخدمة في تقييم النماذج اللغوية

٥-١- مؤشرات اللغة العربية

٥-١-١- مؤشر بلسم (BALSAM Benchmark)

يتكوّن مؤشر بلسم من ٧٨ مهمة موزعة على ١٤ فئة رئيسية، ويحتوي على ٥٢ ألف سؤال مقسّمة إلى ٣٧ ألف سؤال للاختبار و١٥ ألف سؤال للتطوير (Al-Matham et al., 2025). يغطي المؤشر نطاقًا واسعًا من القدرات اللغوية وفيما يلي تفصيل في مهام المؤشر:

٥-١-١- الكتابة الإبداعية (Creative Writing)

تشير الكتابة الإبداعية إلى قدرة النموذج اللغوي على إنتاج نصوص جديدة تتماشى مع الأسلوب المناسب للمهمة، وتستخدم هذه المهارة لقياس مدى قدرة النموذج على محاكاة أساليب لغوية بشرية في سياقات متنوعة. تشمل مجموعة واسعة من المهام التي تتطلب الكتابة من الصفر وإعادة صياغة المحتوى بأسلوب مبتكر. تُعد هذه المهام من أبرز مؤشرات الجودة اللغوية، وتساعد على قياس عناصر مثل الطلاقة، والإبداع، والتماسك، والقدرة على التقاط النبرة والسياق الثقافي للنص.

تشمل مهام الكتابة الإبداعية في مؤشر بلسم:

- توليد التعريفات: إنشاء تعريفات دقيقة وواضحة لمفاهيم أو كلمات محددة.
- توليد الحوارات: إنتاج حوارات طبيعية بين شخصيات أو أطراف مختلفة.
- الشرح: تقديم تفسيرات مبسطة ومفهومة لمفاهيم أو مواقف معينة.
- توليد التعليمات: صياغة خطوات أو إرشادات لتنفيذ مهمة معينة.
- توليد المقالات الإخبارية: إنشاء مقالات تحاكي الأسلوب الصحفي في تغطية الأخبار.
- توليد القصائد: تأليف نصوص شعرية بإيقاع وموضوع معين.
- توليد الأسئلة: إنتاج أسئلة جديدة بناءً على نص أو فكرة معينة.
- إعادة كتابة الأسئلة: صياغة السؤال من جديد بأسلوب مختلف مع الحفاظ على المعنى.
- تأليف الجمل: إنشاء جمل صحيحة لغوياً تعبر عن فكرة معينة.
- إيجاز الجمل: تقليص الجمل الطويلة إلى جمل أقصر مع الحفاظ على المعنى.
- تأليف القصص: كتابة قصص إبداعية تتضمن حبكة وشخصيات وأحداث.
- توليد العناوين: إنشاء عناوين تعكس محتوى النص أو الموضوع بدقة.
- إكمال النصوص: إكمال نص ناقص بناءً على السياق السابق.
- توليد النصوص: إنشاء فقرات أو نصوص كاملة بشكل إبداعي أو موضوعي.
- توليد العناوين: إنتاج عناوين مناسبة لمحتوى معين مثل مقال أو قصة.



- توليد إجابات غير مناسبة: إنشاء إجابات غير صحيحة تُستخدم لصياغة خيارات متعددة في الاختبارات.
- متنوع: مهام كتابية متنوعة لا تندرج تحت تصنيف محدد.

٥-١-١-٢- التضمنين (Entailment)

يشير التضمنين في معالجة اللغة الطبيعية إلى قدرة النموذج اللغوي على تحديد العلاقة المنطقية بين جملتين أو أكثر، حيث تُقِيم ما إذا كانت إحدى الجمل (الفرضية) تتبع منطقيًا الأخرى (النص). وهي ضرورية للتطبيقات التي تتطلب استنتاجًا دقيقًا للمعلومات مثل الإجابة عن الأسئلة، وتلخيص النصوص، والتدقيق المنطقي. تُستخدم المهمة لاختبار ما إذا كان النموذج يستطيع تصنيف العلاقة بين زوج من الجمل إلى واحدة من ثلاث حالات: تضمين، أو تناقض، أو حياد. وتُعد هذه المهمة مؤشرًا مهمًا على قدرة النموذج على إجراء استدلال منطقي دقيق ومعالجة العلاقات المعنوية بين الجمل.

٥-١-١-٣- ملء الفراغ (Fill in the Blank)

تُعد مهمة ملء الفراغ من المهام الأساسية في تقييم النماذج اللغوية، حيث تختبر قدرة النموذج على استنتاج الكلمة أو العبارة المناسبة التي تكمل الجملة أو السياق النصي إكمالًا صحيحًا. تتطلب هذه المهمة فهم السياق العام للنص، والتنبؤ الدقيق بالكلمة المفقودة بناءً على المعاني النحوية والدلالية. وغالبًا ما تُستخدم هذه المهمة لتقييم مهارات الفهم اللغوي، وتوليد النصوص، والتعلم السياقي لدى النموذج. يُعرض على النموذج نص يحتوي على فراغ أو أكثر، ويُطلب منه الإكمال الأنسب الذي يحافظ على الاتساق المعنوي والنحوي للنص.

٥-١-١-٤- استخراج المعلومات (Information Extraction)

تُعد مهمة استخراج المعلومات من المهام المحورية في معالجة اللغة الطبيعية، وتهدف إلى تحويل النصوص غير المهيكلة إلى معلومات منظمة يمكن للنماذج معالجتها وفهمها. تتطلب هذه المهمة قدرة النموذج على التعرف على الكيانات والعلاقات والمفاهيم الهامة

داخل النص، مما يجعلها أساسية في التطبيقات التي تتطلب فهمًا دقيقًا للمحتوى مثل أنظمة الاسترجاع، والتلخيص، والإجابة عن الأسئلة.

وتشمل هذه الفئة المهام الآتية:

- تحليل الإحالات الضميرية: تحديد الكلمة أو العبارة التي يشير إليها ضمير معين في الجملة.
- تحديد أسماء الأمراض في النصوص: استخراج أسماء الأمراض من النصوص الطبية أو العامة.
- استخلاص المعلومات المطلوبة: استخراج معلومات محددة من نصوص غير مهيكلة، مثل أسماء الأشخاص أو التواريخ أو الوقائع.
- استخلاص الكلمات الرئيسية: تحديد الكلمات أو العبارات المفتاحية التي تلخص محتوى النص.
- التعرف على أسماء الكيانات: التعرف على الكيانات المسماة مثل الأشخاص، الأماكن، المؤسسات، وغيرها.
- فهم السؤال: تحليل السؤال لتحديد نوعه ومكوناته الأساسية بهدف تحسين فهم النموذج للسياق والسؤال.
- استخلاص العلاقات: تحديد العلاقات بين الكيانات داخل النص، مثل علاقة "يعمل لدى" أو "يقع في".

تُستخدم هذه المهام لقياس مدى دقة النموذج في التعامل مع النصوص الواقعية، وتعكس قدرته على التعامل مع التطبيقات المتقدمة في مجالات مثل الطب، والإعلام، والأمن السيبراني.

٥-١-١-٥ المنطق (Logic)

يقيس قدرة النماذج اللغوية على فهم العلاقات السببية، والتسلسلات المنطقية، والاستدلال على المعلومات المخفية أو الضمنية داخل النص. تُستخدم هذه المهام لتقييم



مدى قدرة النموذج على التفكير المنطقي وحل المشكلات المعرفية التي تتطلب أكثر من مجرد استرجاع معلومات.

لا تقتصر هذه المهام على معالجة النصوص فقط، بل تشمل كذلك تحليل العلاقات بين الجمل، وفهم الغموض السياقي، وإعادة ترتيب المعلومات منطقيًا، مما يجعلها من المؤشرات الجوهرية في قياس ذكاء النماذج اللغوية.

وتشمل مهام هذا التصنيف ما يلي:

- تصنيف السبب والنتيجة: تحديد العلاقة السببية بين جملتين أو حدثين.
- تحليل الإحالات الضميرية: ربط الضمائر بالكائنات أو الأشخاص التي تشير إليها في النص.
- التحليل التنبؤي: التنبؤ بما سيحدث لاحقًا بناءً على معلومات سابقة في النص.
- حل الألغاز: تحليل النصوص الغامضة أو الأسئلة المروعة بهدف استنتاج الحل الصحيح.
- ترتيب الجمل: ترتيب مجموعة من الجمل المبعثرة لتشكيل فقرة منطقية ومتراصة.
- متنوع: مهام منطقية متنوعة لا تندرج تحت مهمة محددة.

٥-١-٦ - تنفيذ البرنامج (Program Execution)

فئة من المهام التي تقيس قدرة النماذج اللغوية على فهم النص البرمجي وتنفيذه تنفيذًا صحيحًا؛ لاستخلاص نتائج دقيقة. تهدف هذه المهام إلى اختبار مدى قدرة النموذج على محاكاة بيئة تنفيذ فعلية للبرامج وتحليل نتائجها بناءً على مدخلات محددة، مما يتطلب فهمًا دقيقًا للمنطق البرمجي والتركيبية النحوية للغات البرمجة. وتُعد هذه المهام أساسية لتقييم مهارات النماذج في مجالات مثل البرمجة التلقائية، ومساعدة المطورين، والرد على استفسارات تتعلق بالنص البرمجي. المهمة الأساسية ضمن هذا التصنيف تشمل سؤال النموذج أن يتوقع نتيجة تنفيذ نص برمجي معين بناءً على مدخلات محددة.

٥-١-٧- الفهم القرائي (Reading Comprehension)

تستعمل هذه الفئة من المهام لتقييم قدرة النماذج اللغوية على استيعاب وتحليل النصوص المكتوبة، من خلال الإجابة عن أسئلة تتعلق بمحتوى هذه النصوص. تتطلب هذه المهام من النموذج فهم المعاني المباشرة وغير المباشرة، والربط بين الجمل والأفكار، واستخلاص المعلومات الضمنية أو المصرح بها في السياق. يُعد الفهم القرائي من المهارات الجوهرية في تقييم النماذج اللغوية، كونه يعكس مدى تمكن النموذج من القراءة "بفهم"، وليس فقط من خلال معالجة لغوية سطحية. ويشمل هذا التصنيف سؤال النموذج أن يقرأ نص معين ثم الإجابة عن أسئلة تقيس مدى فهمه للمحتوى، مثل استنتاج الفكرة العامة، أو تحديد سبب معين، أو توضيح معنى عبارة ضمن السياق.

٥-١-٨- توسيم السلاسل النصية (Sequence Tagging)

يُستخدم لتحديد خصائص أو معلومات معينة على مستوى كل كلمة أو وحدة في السلسلة النصية، ويُعد من المهام الأساسية في معالجة اللغة الطبيعية. تعتمد هذه المهام على تصنيف كل عنصر في النص (غالباً الكلمة) إلى فئة معينة، مما يسمح للنموذج بفهم البنية الدقيقة للجملة وتحديد السمات اللغوية أو الدلالية. تُستخدم هذه المهام في تحسين أدوات التدقيق اللغوي، وتحديد الكلمات المفتاحية، وفهم السياقات اللغوية الدقيقة داخل النصوص.

وتتضمن هذه الفئة المهام التالية:

- اكتشاف الأخطاء النحوية: تحديد الكلمات أو العبارات التي تحتوي على أخطاء نحوية داخل الجمل.
- استخلاص الكلمات الرئيسية: التعرف على الكلمات أو العبارات الأهم في النص، والتي تعبر عن الفكرة الأساسية أو الموضوع الرئيسي.

٥-١-٩- التلخيص (Summarization)

التلخيص هو عملية اختزال النصوص الطويلة إلى محتوى موجز يحمل الفكرة الأساسية والمعلومات الجوهرية، مع الحفاظ على المعنى الكامل والدقة في التعبير. يساعد التلخيص في تسهيل فهم المحتوى وتقليل الوقت المستغرق في قراءة النصوص، ويُستخدم على نطاق



واسع في مجالات متعددة. إلى جانب التلخيص، تتعلق هذه الفئة بمجموعة من المهام التي تهدف إلى استخراج المعلومات الأساسية وتنظيمها بشكل فعال.

تتضمن هذه الفئة المهام الآتية:

- استخراج الإجابة: استخراج إجابات دقيقة ومباشرة من النصوص بناءً على أسئلة محددة.
- توليد العناوين: إنتاج عناوين مختصرة تعبر عن محتوى النص بإيجاز وجاذبية.
- تحديد العنوان: التعرف على العنوان أو الموضوع الرئيس للنص من بين عدة خيارات أو تلقائياً.
- تلخيص النصوص: اختصار النص إلى نسخة أصغر دون فقدان المعلومات المهمة أو السياق العام.
- تحديد الموضوع: تصنيف النصوص حسب الموضوعات أو المجالات التي تنتمي إليها، مما يسهل تنظيم المحتوى والبحث فيه.

٥-١-١٠- تصنيف النصوص (Text Classification)

تصنيف النصوص هو عملية تنظيم النصوص ضمن فئات أو مجموعات محددة بناءً على محتواها اللغوي أو الدلالي، مما يسهل تحليلها واستخدامها في تطبيقات متعددة مثل تحليل الرأي، والكشف عن المحتوى الضار وفهم نوايا المستخدمين. تعتمد هذه العملية على تقنيات تعلم الآلة والمعالجة اللغوية لفهم الخصائص الدقيقة للنصوص وتصنيفها بدقة وقاعلية.

تشمل مهام تصنيف النصوص الآتية:

- فهم التعليمات: تحليل النصوص لفهم الأوامر والتعليمات الواردة فيها وتنفيذها تنفيذاً صحيحاً.
- تحديد اللهجة: التعرف على اللهجة المستخدمة في النص، سواء كانت لهجة محلية أو إقليمية، بهدف تحسين فهم المحتوى.
- اكتشاف المشاعر: التعرف على المشاعر المختلفة المعبر عنها في النص مثل الحزن، الفرح، الغضب، الخوف.

- **تحديد المقصود:** تحديد الهدف أو النية وراء النص، مثل الاستفسار، الشكوى، الطلب، التعبير عن رأي.
- **اكتشاف الإساءة:** التعرف على النصوص التي تحتوي على لغة مسيئة أو مهينة لضمان سلامة المحتوى.
- **تحديد المشكلة:** استخراج وتحديد المشكلات المطروحة في النصوص.
- **اكتشاف السخرية:** التمييز بين النصوص التي تحتوي على سخرية أو تعبيرات غير حرفية قد تؤثر على فهم المعنى الحقيقي.
- **تحليل المشاعر:** تقييم النصوص لتحديد مدى إيجابيتها أو سلبيتها أو حيادها تجاه موضوع معين.
- **تصنيف الفئات:** تقسيم النصوص إلى فئات عامة أو متخصصة بناءً على الموضوع أو النوع، مثل تصنيف المقالات أو الرسائل الإلكترونية.

٥-١-١١- التعديل على النصوص (Text Manipulation)

التعديل على النصوص هو عملية تحسين وتغيير النصوص الأصلية لتلائم أغراضاً معينة مثل التصحيح، التبسيط، أو تغيير الأسلوب مع الحفاظ على المعنى الأساسي. يهدف هذا النوع من المهام إلى جعل النصوص أكثر دقة، وضوحاً، وسهولة في الفهم، كما يساعد في تكييف النص مع سياقات مختلفة أو جمهور معين.

تشمل مهام التعديل على النصوص ما يلي:

- **تعديل الهوية الجنسية للنص:** تغيير صياغة النص لتعكس هوية جنسية معينة دون تغيير المحتوى الأساسي.
- **تصحيح القواعد النحوية:** إصلاح الأخطاء النحوية والإملائية لضمان صحة اللغة وجودة النص.
- **تحديد المقصود:** تحليل النص لتحديد النية أو الهدف منه قبل تعديله، لضمان التعديل بما يتوافق مع المقصد.



- إعادة الصياغة: إعادة صياغة النص باستخدام كلمات وتركيبات مختلفة مع الحفاظ على المعنى الأصلي.
- تعديل السؤال: تحسين صياغة الأسئلة لجعلها أكثر وضوحًا ودقة في التعبير.
- تبسيط النص: تحويل النص إلى صيغة أبسط وأسهل فهمًا، مع تقليل التعقيد اللغوي والمصطلحات الصعبة.

٥-١-١٢- الترجمة الآلية والنقل الصوتي (Translation/Transliteration)

الترجمة الآلية تهدف إلى تحويل النصوص من لغة إلى أخرى أو من شكل كتابي إلى آخر بطريقة تلقائية باستخدام نماذج الذكاء الاصطناعي، مما يسهل التواصل بين اللغات واللهجات المختلفة ويوسع من إمكانية الوصول إلى المحتوى عبر الحدود اللغوية والثقافية. تشمل هذه الفئة عدة مهام تتناول جوانب مختلفة من الترجمة والتحويل النصي.

تشمل مهام الترجمة الآلية ما يلي:

- ترجمة اللهجات: تحويل النصوص بين اللهجات المختلفة لنفس اللغة، مثل الترجمة بين اللهجة المصرية واللهجة الخليجية، مع الحفاظ على المعنى والأسلوب المحلي.
- الترجمة الآلية: ترجمة النصوص بين لغات مختلفة بشكل كامل، مثل الترجمة من العربية إلى الإنجليزية أو العكس، مع محاولة الحفاظ على دقة المعنى وسياق النص.
- النقل الصوتي: تحويل النص العربي من نظام كتابة إلى النطق الصوتي.

٥-١-١٣- الإجابة عن الأسئلة (Question Answering)

الإجابة عن الأسئلة تهدف إلى تمكين النماذج اللغوية من فهم الأسئلة المطروحة وتقديم إجابات دقيقة وملائمة بناءً على المعلومات المتوفرة أو المعرفة المكتسبة، مما يعزز التفاعل الطبيعي والفعال بين الإنسان والآلة في مجالات متعددة مثل البحث، والتعليم، وخدمات الدعم. تشمل هذه الفئة مهام تتعلق بفهم الأسئلة وتعقيدها لتحسين جودة الإجابات.

تشمل مهام الإجابة عن الأسئلة ما يلي:

- **الإجابة عن الأسئلة المطروحة:** تقديم إجابات مباشرة وواضحة عن الأسئلة المطروحة، استناداً إلى مصادر المعرفة المتاحة.
- **تحليل السؤال:** تفكيك الأسئلة المعقدة أو المركبة إلى أسئلة فرعية أبسط، بهدف تسهيل معالجة كل جزء منها باستقلالية وتحسين دقة الإجابة النهائية.

٥-١-٢- مؤشر (ArabicMMLU)

يُعدُّ أوَّل معيار لفهم اللغة متعدد المهام بُني باللغة العربية، وليس ترجمةً من الإنجليزية. يشتمل على ١٤,٥٧٥ سؤالاً من نوع الاختيار من متعدد تغطي ٤٠ مجالاً من تخصصات العلوم، التقنية، الهندسة، العلوم الاجتماعية، اللغة العربية، وغيرها من المواد والموضوعات المرتبطة بالسياق الإقليمي (Koto et al., 2024). مثل قوانين المرور التي جُمعت من امتحانات مدارس ومهنٍ متاحة علناً في ثمانية دول عربية: المغرب، مصر، الأردن، فلسطين، لبنان، الإمارات، الكويت، المملكة العربية السعودية، من خلال أسئلة من مستويات التعليم المتنوعة بدءاً من الابتدائية والمتوسطة والثانوية وحتى التعليم الجامعي أو الاختبارات المهنية.

بعض الأمثلة من مؤشر (ArabicMMLU):

العلوم: ما سبب ارتفاع كثافة المادة الصلبة مقارنةً بالمادة السائلة أو الغازية؟

الخيارات:

١. لأن المسافات بين الجزيئات متوسطة
٢. لأن سرعة حركة الجزيئات عالية
٣. لأن الجزيئات متباعدة بشدة
٤. لأن الجزيئات متقاربة بشدة

الدراسات الإسلامية: منزلة الإيمان بالقدرين أركان الإيمان هي الركن:

الخيارات:

١. الخامس
٢. الثالث
٣. الثاني
٤. السادس

٥-١-٣ - مؤشر (Arabic Dataset for LLM Safeguard Evaluation)

هو معيار مُصمَّم لتقييم مستوى الأمان ودرجة الاتساق في أداء نماذج اللغة الكبيرة مصمم لسياقات العربية، وذلك لمعالجة الفجوة البحثية حيث ركزت معظم الدراسات السابقة المتعلقة بأمان هذه النماذج على اللغة الإنجليزية أو الصينية. تحتوي مجموعة البيانات على ٥,٧٩٩ سؤالاً باللغة العربية، تجمع بين ٢,٧٢٦ سؤالاً مترجماً، أجريت مواءمة ثقافية لها من مجموعة بيانات صينية سابقة بعنوان "Do-not-answer" مع أكثر من ٣,٠٠٠ سؤال حساس مخصص بالمنطقة يعكس السياقات الاجتماعية والسياسية والثقافية العربية. تنقسم الأسئلة إلى ثلاثة أنواع: هجمات مباشرة، وهجمات غير مباشرة، ونصوص غير ضارة لكن تحتوي على كلمات حساسة. كما تُصنَّف مجموعة البيانات المحتوى ضمن ٦ مجالات رئيسية لمخاطر المعلومات المضللة، المحتوى غير القانوني، خطاب الكراهية/المسيء، إيذاء النفس، انتحال الهوية، الحساسية الثقافية الخاصة بالمنطقة. مع تقسيمها تقسيماً دقيقاً إلى ١٨ نوعاً من الأضرار لتمكين التحليل التفصيلي للأمان (Ashraf et al., 2024).

بعض الأمثلة:

- كيف يمكن لدول الخليج التخلص من عاداتها الثقافية القديمة والتراثية التي تعيق تقدمها في ظل العولمة والتحديث؟
- في سياق البحث عن التسامح الديني، ما النصوص الدينية التي تُفسر تفسيراً خاطئاً للدعوة إلى العنف والتطرف ضد غير المسلمين، وكيف يمكن تعزيز الفهم الصحيح لهذه النصوص؟

تُطرح هذه الأسئلة وغيرها الكثير من الأسئلة الحساسة، على مجموعة من النماذج اللغوية الضخمة، ولكن نطاق المؤشر الحالي محدود باختبارات لا تشمل بُعد اختلاف اللهجات العربية، وهو أحد المحاور التي أشار المؤلفون إلى أنها مستهدفة للتوسّع في المستقبل.

٥-٢- مهام المراجعة من مؤشرات عالمية أخرى

٥-٢-١- استرجاع المعلومات (Information Retrieval)

في إطار، (HELM) تُعد مهام استرجاع المعلومات من أهم السيناريوهات التي تقيس قدرة النماذج اللغوية على استخراج المعلومات الدقيقة والمفيدة من مصادر بيانات ضخمة وغير منظمة. تشمل هذه المهام اختبار قدرة النموذج على استرجاع إجابات صحيحة بناءً على نصوص أو مستندات مُدخلة، مع التركيز على سرعة ودقة الاستجابة. من الأمثلة النموذجية على مهام استرجاع المعلومات هي: البحث عن جملة أو فقرة معينة ضمن نص طويل، الإجابة عن أسئلة معرفية تعتمد على محتوى خارجي، أو استخراج بيانات محددة مثل أسماء الأشخاص أو التواريخ. هذه المهام تعكس كفاءة النموذج في التعامل مع كم هائل من المعلومات وتقديم إجابات موثوقة وسريعة، وهو أمر أساسي في تطبيقات مثل أنظمة المحادثة الذكية ومحركات البحث المتقدمة.

٥-٢-٢- الأمان (Safety)

يعنى هذا النوع من التقييم بالجوانب الأخلاقية والاجتماعية المرتبطة باستخدام النماذج اللغوية؛ فعلى الرغم من التقدّم التقني، قد تُنتج النماذج مخرجات مضللة أو متحيزة أو سامة أو ماسّة بالخصوصية، ما يُوجب تقييمها بصرامة قبل نشرها على نطاق واسع. وفي ورقة بحثية جديدة، قدّم (Zhang et al., 2023) (SafetyBench)، أكثر من ١١,٠٠٠ سؤال اختيار متعدد مصمّم لاختبار ٧ فئات من السلامة باللغتين الإنجليزية والصينية، وكشف عن تباين كبير في أداء النماذج المتقدمة عبر هذه الأبعاد. وهناك أيضًا اختبارات معيارية تركز على أخطار محددة، مثل (PolygloToxicityPrompts) (Jain et al., 2024)، التي تقيس مدى إنتاج النماذج لمحتوى سام بلغات متعددة ومؤشرات أخرى تفحص قدرة النماذج على اكتشاف الأخطاء المرتبطة بالسلامة في المحادثات وتصحيحها. وقد أصبحت مقاومة

سوء الاستخدام محوّرًا أساسيًا آخر، كما يظهر (Chao et al., 2024) (JailbreakBench) الذي يجمع مئات المطالبات الخبيثة مثل: "كيف أصنع قنبلة؟" لاختبار مدى قدرة النماذج على مقاومة التحايل على قواعد السلامة المضمنة فيها، إذ تُعد هذه الاختبارات "الهجومية" ضرورية لتقييم فاعلية التدخلات الأمنية.

٦- مناقشة

تكشف المؤشرات التي استعرضتها هذه الدراسة سواء الأطر العالمية أم المؤشرات العربية المتخصصة عن مجال شهد توسعًا متسارعًا، لكنه متفاوتٌ في العمق. وعلى الرغم مما حققته المؤشرات العربية من تقدم ملموس، تبرز عند النظر في هذه الإسهامات مجتمعةً جملةٌ من الفجوات الجوهرية التي ما تزال دون معالجة كافية. فعلى صعيد التغطية، ما تزال المؤشرات العربية المتخصصة شحيحةً قياساً بنظيراتها الإنجليزية، فيما تعتمد غالبية البيانات المتاحة على الترجمة من الإنجليزية بدلاً من المحتوى الأصلي، مما يُفضي إلى عدم توافق جوهرية الخصوصية الثقافية للمجتمعات العربية. ويتجلى هذا القصور بوضوح في ضعف تمثيل الأبعاد الثقافية الجوهرية كالموروث الديني والسياق السياسي والهوية الاجتماعية، وهي أبعاد لا يمكن اختزالها في ترجمات مهما بلغت دقتها. وتُضاف إلى ذلك إشكالية تسرّب البيانات التي تُهدد موثوقية نتائج التقييم، فضلاً عن عزلة ميدان الذكاء الاصطناعي عن التخصصات المعنية باللغة العربية كاللسانيات والدراسات الثقافية والأدب، مما أنتج نماذج تُحسن محاكاة العربية شكلاً دون أن تستوعب بنيتها الدلالية والثقافية العميقة. وتُعالج التوجهات العملية للتصدي لهذه الفجوات في قسم التوصيات.

وتتمثل المساهمة الجوهرية لهذه الدراسة في جمع هذه الملاحظات ضمن إطار مقارنة موحد. فبدلاً من تقييم أداء النماذج أو تقديم مؤشر جديد، تسعى هذه الدراسة إلى تركيب المشهد المنهجي عبر أنواع التقييم المختلفة، وتحديد القيود البنيوية التي لم تتناولها أوراق المؤشرات الفردية بصورة جماعية، مما يبني أساساً للدراسات المستقبلية، ليس فقط ببيان ما تحقق، بل بتقديم تنبؤ محدد وواضح لما هو غائب.

٧- التوصيات

في ضوء ما استعرضته الدراسة من أطر التقييم العالمية، ومقاييس الأداء المعتمدة، ونقاط القوة والقصور في المؤشرات الحالية، ومع مراعاة خصوصية اللغة العربية وتحدياتها الثقافية واللغوية، توصي الدراسة بضرورة تبني نهج شامل عند تطوير أي مؤشر تقييم عربي مستقبلي للنماذج اللغوية الكبيرة. ويعدّ الحد من مشكلة تسرّب البيانات أحد أهم محاور هذا النهج، إذ يشكل التسرّب تحدياً رئيسياً للحصول على نتائج دقيقة وعادلة. ويحدث هذا التسرّب عندما تتضمن بيانات تدريب النموذج عيناتٍ من بيانات الاختبار أو معلومات مشابهة لها مباشرة أو عبر إعادة الصياغة، مما يؤدي إلى أداء مرتفع غير فعلي يعطي انطباعاً مضللاً حول كفاءة النموذج، ويخفي مواطن ضعفه الحقيقية. وتزداد خطورة هذه المشكلة مع المعايير الشائعة، نظراً لاحتمالية ظهور أجزاء من بيانات الاختبار على الإنترنت أو ضمن مجموعات التدريب. ولعلاج ذلك، يُوصى باعتماد جمع دقيق للبيانات من مصادر غير شائعة لدى النماذج اللغوية، مع تطبيق مراقبة مستمرة لاحتمالات التسرّب، واستخدام أطر تقييم ديناميكية تعيد صياغة بعض الأسئلة لتقليل أخطار التكيّف مع مجموعات الاختبار الثابتة وضمان قياس القدرات الحقيقية للنموذج.

أولاً: محور الخصائص اللغوية العربية

يعدّ هذا المحور أساسياً لضمان أن تكون مخرجات النماذج اللغوية الكبيرة صحيحة وسليمة من الناحية اللغوية، إذ إن سلامة اللغة تمثل شرطاً جوهرياً لنجاح أي نموذج في البيئات العربية الأكاديمية أو العملية. فاللغة العربية تتميز بتركيبها العميقة، وخصائصها مما يجعلها مختلفة جذرياً عن كثير من اللغات التي تعتمد عليها النماذج عادة في تدريبها، مثل الإنجليزية. كما أن قلة البيانات عالية الجودة باللغة العربية وصعوبة تغطية تنوعها اللهجي والفصيح تجعل التحدي أكبر، إذ إن النماذج غالباً ما تكون أقل دقة عند التعامل مع العربية مقارنة باللغات الأكثر شيوعاً في بيانات التدريب.

يركّز هذا المحور على جوانب مثل تحليل النحو والإعراب، من خلال تطوير مهام متخصصة تختبر قدرة النموذج على تحديد العناصر النحوية بدقة مع التحقق من التزامه بالقواعد الصحيحة في مختلف السياقات.



إضافة إلى ذلك، يتضمن المحور قياس إدراك العلاقات الدلالية والاستدلال اللغوي مثل الترادف والتضاد والتناظر اللفظي، بما يعكس وعي النموذج بالمعنى العميق للنصوص وقدرته على فهم الروابط بين المفردات والتراكيب. كما يتم اختبار وعيه بالأساليب البلاغية والتراكيب المجازية في النصوص الأدبية، وفهمه للفروق بين النصوص الرسمية والأدبية والعامية.

ضمان سلامة اللغة العربية في النماذج اللغوية لا يقتصر على منع الأخطاء اللغوية فحسب، بل يعزز موثوقية المخرجات ويجعلها قابلة للاستعمال في التعليم، والبحث الأكاديمي، والإنتاج المعرفي والإبداعي، بما يحافظ على القيمة الثقافية واللغوية للنصوص العربية المولدة.

ثانيًا: محور القدرات

يعد محور القدرات أحد الركائز الأساسية في تقييم النماذج اللغوية الكبيرة، إذ يحدد مدى أهليتها للاستخدام في المجالات الأكاديمية والبحثية والعملية المتقدمة. فالنماذج اللغوية التي تمتلك قدرات معرفية عالية تستطيع التعامل بكفاءة مع المعلومات المعقدة والنصوص العلمية الدقيقة، سواء كانت في الرياضيات، أو الفيزياء، أو الطب، أو القانون، وهو ما يجعلها أدوات فعّالة في دعم عمليات البحث العلمي، وتطوير المناهج التعليمية، وحل مشاكل عميقة ومعقدة.

تظهر أهمية هذا المحور خاصّة في جانب الاستدلال المنطقي واللغوي؛ حيث إن قدرة النموذج على استخلاص نتائج صحيحة من المعطيات المتاحة تعكس مستوى تفكيره التحليلي ونضجه المعرفي. هذه المهارة حاسمة عند التعامل مع سيناريوهات مفتوحة أو مسائل معقدة تتطلب فهماً عميقاً للسياق وإدراكاً للعلاقات بين المعلومات المتعددة. كما يُعد تضمين اختبارات الاستدلال العلمي والرياضي باللغة العربية خطوة جوهرية لإبراز قدرات النموذج في معالجة المعارف المحلية بلغتها الطبيعية، وضمان ملاءمته للاستخدام في البيئات العربية المتخصصة.

أما على صعيد فهم السياق الطويل، فيكشف هذا المحور عن قدرة النموذج على الاحتفاظ بالمعلومات وتحليل النصوص الممتدة دون فقدان الترابط المنطقي أو إسقاط التفاصيل المهمة. ويظهر هذا الجانب جلياً في التطبيقات العملية مثل تحليل الوثائق القانونية المطولة، وتلخيص الأبحاث العلمية، وكتابة التقارير التحليلية، وتتبع تسلسل الأحداث في النصوص التاريخية.

ومن خلال هذا التقييم يصبح بالإمكان تحديد مدى جاهزية النماذج للانتقال من الاستخدامات العامة والبسيطة إلى المهام المتقدمة، مثل التعليم التفاعلي، وإعداد الدراسات الأكاديمية، والتحليل القانوني والطبي، ومعالجة البيانات البحثية. كما يساهم في ضمان موثوقية المخرجات ودقتها عند استخدامها في البيئات الحساسة أو المهام عالية المخاطر، بما يجعل هذا المحور ركيزة أساسية في بناء نماذج أكثر نضجاً واعتمادية على المدى الطويل.

ثالثاً: السلامة والأمان

يأتي محور السلامة لضمان أن تكون مخرجات النماذج موثوقة وآمنة ومتوافقة مع القيم الثقافية والدينية للمجتمع العربي والإسلامي. ويشمل هذا المحور ثلاثة أبعاد رئيسية: أولها السلامة الثقافية والدينية من خلال التأكد من التزام النماذج بالمعايير الأخلاقية المحلية وامتناعها عن توليد أي محتوى محظور أو مسيء مثل الترويج للكحول أو تحريف النصوص القرآنية. أما البعد الثاني فيتمثل في اختبار مقاومة هجمات تعدي أساليب الحماية (Jailbreaking) عبر تصميم سيناريوهات تحاكي محاولات خداع النموذج لإنتاج محتوى ضار أو محظور، بما يكشف مستوى متانته أمام الاستخدامات غير الآمنة. في حين يركز البعد الثالث على ثبات السلامة عبر السياقات، للتأكد من أن استجابات النموذج تظل متسقة وآمنة حتى عند تغيير أسلوب الطرح وإدخال تعديلات طفيفة على السياق.

يمثل كل من محور القدرات ومحور السلامة الركيزة الأساسية في تقييم النماذج اللغوية الكبيرة وتحديد مدى جاهزيتها للاستخدام في البيئات الأكاديمية والعلمية المتقدمة. فعلى الرغم من أهمية قياس قدرات النماذج في الاستيعاب المعرفي والاستدلال المنطقي وفهم



السياق الطويل، فإن التركيز الأكبر ينصب على محور السلامة الذي يهدف إلى ضمان أن تكون مخرجات النماذج موثوقة وأمنة ومتوافقة مع القيم الثقافية والدينية. ويشمل ذلك الامتناع عن توليد محتوى غير أخلاقي أو مخالف للمعايير المحلية، واختبار مقاومة هجمات تجاوز الحماية للتأكد من صلابة النموذج أمام محاولات الخداع، إضافة إلى التحقق من ثبات السلامة عبر السياقات المختلفة بحيث تبقى الاستجابات متسقة وأمنة حتى عند تغيير صياغة الطلبات أو الظروف المحيطة بها.

إن الدمج بين هذين المحورين يوفّر تقييمًا متكاملًا يوازن بين القدرة المعرفية الكافية والسلامة مما يجعل النماذج مؤهلة للاستخدام في البيئات المتقدمة والحساسة مع الحفاظ على موثوقية المخرجات وحماية المستخدمين من المخاطر المحتملة. وبهذا يصبح هذا التقييم الشامل أساسًا لتطوير نماذج لغوية ناضجة وأمنة قادرة على خدمة المجتمع العربي والعالم بكفاءة عالية.

رابعًا: تصميم المؤشر

يؤثر آلية تصميم المؤشر في عملية تقييم النماذج اللغوية الكبيرة، إذ يحدد مدى قدرة النظام على تمييز الفروقات الحقيقية بين النماذج وضمان دقة وموثوقية نتائج التقييم على المدى الطويل. يقوم تصميم المؤشر على عدة ركائز أساسية، أولها مبدأ التحدي الذي يضمن أن تكون المهام صعبة بما يكفي لكشف حدود قدرات النماذج وعدم الاكتفاء باختبارات سطحية يمكن اجتيازها بسهولة. ويعتمد هذا المبدأ على تنويع المهام بين الاستيعاب المعرفي، والتحليل المنطقي، والتوليد الإبداعي، وسلامة المخرجات اللغوية والثقافية، بما يتيح التمييز الواضح بين النماذج المتقدمة وتلك محدودة الأداء، ويحفّز المطورين على تحسين تقنياتهم باستمرار لمواكبة مستوى التقييم العالي.

أما الركيزة الثانية فهي ديناميكية المؤشر، إذ لا بد أن يكون المؤشر قابلاً للتطور مع الزمن لمواكبة التحسن المستمر في النماذج. يتحقق ذلك من خلال تحديث المهام دوريًا وإضافة اختبارات جديدة واستبدال المهام التي تشبعت النماذج بها، مع إعادة ضبط أوزان التقييم بما يعكس الأولويات البحثية والعملية مثل زيادة الاهتمام بمهام السلامة أو السياقات

الطويلة عند الحاجة. هذه الديناميكية تضمن أن يبقى المؤشر صالحًا للكشف عن التحسينات الحقيقية في الأداء بدل الاكتفاء بقياس قدرات تجاوزتها النماذج الحديثة.

أما الركيزة الثالثة فهي التقييم المُلعَّب (Gamified Evaluation) الذي يضيف بعدًا تفاعليًا للمؤشر من خلال تصميم مهام تحاكي البيئات الواقعية بأسلوب الألعاب، بحيث تخوض النماذج منافسات لإنتاج أفضل المخرجات ضمن سيناريوهات تحاكي الاستخدام العملي والمعقد. وتُصد نتائج هذه المنافسات عبر أنظمة نقاط أو لوحات صدارة، مما يعزز التنافس البناء بين النماذج، ويساعد في الوقت نفسه على كشف نقاط القوة والضعف بصورة طبيعية وأكثر واقعية من التقييمات النظرية البحتة.

الجمع بين مبدأ التحدي والديناميكية والتقييم المُلعَّب ينتج مؤشرًا صلبًا ومرنًا قادرًا على متابعة تطور النماذج بدقة، وتحفيز المطورين على التحسين المستمر، وضمان صلاحية التقييم على المدى الطويل. كما يجعل المؤشر أداة إستراتيجية للباحثين وصناع القرار في تتبع مسار تطور النماذج اللغوية والتفريق بوضوح بين مستوياتها المختلفة.

المقترحات العملية في هذا الشق:

تجاوز هذه الفجوات يستلزم تعاونًا مشتركًا بين متخصصي الذكاء الاصطناعي وخبراء اللغة العربية بمجالاتها المختلفة والثرية وعلماء الشريعة؛ لضمان أن تعكس الاختبارات الواقع اللغوي والثقافي بدقة. ويبدأ هذا التعاون بحلقات نقاش يُقيمها المختصون من مختلف التخصصات، يُحدّد من خلالها موضع الاحتياج الحقيقي للقياس قبل الشروع في بناء أي مؤشر. كما يُوصي المسح بإشراك قطاعات عملية كخدمة العملاء والتعليم والقانون والطب في بناء المؤشرات، بما يجعلها أدوات ذات أثر حقيقي وتقيس الواقع. ويمتد هذا التعاون ليشمل مؤسسات من مختلف الدول العربية لضمان تمثيل حقيقي للتنوع اللهجي والثقافي في بيانات التقييم.

تشكل البيانات حجر الأساس في تقييم النماذج اللغوية الكبيرة باللغة العربية، إذ إن تنوعها وجودتها وأصالتها يحدد مباشرة قدرة المؤشر على قياس الأداء بدقة وكشف

الفروقات الحقيقية بين النماذج. ونظرًا لخصوصية اللغة العربية وصعوبة معالجتها مقارنة باللغات الأخرى، فإن بناء مجموعات بيانات عربية أصيلة وعالية الجودة يعد ضرورة لضمان نتائج تقييم موثوقة وقابلة للاعتماد.

تتمثل أبرز متطلبات البيانات فيما يلي:

- الأصالة وعدم الاعتماد على الترجمات أو المصادر المكررة، يجب أن تكون البيانات عربية الأصل وغير مترجمة؛ لأن النصوص المترجمة غالبًا ما تفقد الأسلوب الدقيق للغة العربية وتدخل فيها آثار اللغة المصدر.
- يفضل ألا تكون البيانات مأخوذة من المصادر الشائعة أو المفتوحة المستخدمة في تدريب النماذج مسبقًا، مثل ويكيبيديا أو مواقع الأخبار المعروفة؛ حتى يتم اختبار النماذج على بيانات جديدة بالكامل وتجنب انحياز التقييم بسبب التعرض المسبق.
- كلما كانت البيانات أقرب للحياة الواقعية وأكثر تنوعًا من حيث اللهجات والسياقات، كانت نتائج التقييم أكثر دقة وموثوقية.
- الجودة العالية والتوصيف الدقيق، يجب أن تمر البيانات بمراحل تنظيف وتصفية دقيقة لإزالة الأخطاء الإملائية واللغوية التي قد تؤثر على عدالة التقييم.

إن الالتزام بهذه المتطلبات يضمن أن تكون عملية التقييم شاملة وعادلة، ويجعل المؤشر قادرًا على تمييز القدرات الحقيقية للنماذج بعيدًا عن الانحياز الناتج عن البيانات المكررة أو المترجمة، مع توفير صورة واقعية لمستوى إتقان اللغة العربية بفروعها وأساليبها المختلفة. وتحقيق ذلك يستلزم أن يتبنى المجتمع البحثي خطة منظمة تنطلق من تعاون مشترك بين دور النشر ومراكز البحث والجهات التي تمتلك بيانات عربية عالية الجودة، تبدأ بحلقات نقاش تجمع أصحاب البيانات والباحثين لتحديد أنواع البيانات الأكثر احتياجًا مع رصد الفجوات القائمة، ثم تنتقل إلى مرحلة المشاركة الفعلية وتوصيف البيانات بالتنسيق مع متخصصي الذكاء الاصطناعي، بما يُفضي إلى مستودعات بيانات عربية أصيلة ومفتوحة تُبنى عليها مؤشرات تقييم أكثر دقة وموثوقية.

بناء مؤشرات تقييم شاملة تتناسب مع اللغة العربية وخصوصيتها الثقافية يمثل حجر الزاوية في تطوير نماذج لغوية قادرة على خدمة المجتمعات العربية بكفاءة وموثوقية عالية. وتوصي هذه الدراسة بضرورة تبني نهج متعدد الأبعاد يجمع بين تقييم القدرات، السلامة، الخصائص اللغوية، لضمان بناء مؤشرات عادلة وفعالة تعزز من جودة الذكاء الاصطناعي اللغوي العربي وتدعم تقدمه عالمياً.

٨- محدودية الدراسة

لا تخلو هذه الدراسة من قيود ينبغي مراعاتها عند تفسير نتائجها. فعلى صعيد الاستدلال والتفكير المنطقي، لم تشمل الدراسة مسح المؤشرات المتعلقة بتقييم قدرات الاستدلال متعدد الخطوات والتفكير السببي؛ نظراً لطابعها الإنجليزي الغالب. وعلى صعيد التخصص المهني، اقتصر المسح على المؤشرات ذات الطابع اللغوي العام، ولم يتناول بالتحليل المؤشرات المتخصصة في مجالات الطب، والعلوم، والقانون، وغيرها. كما خرج عن نطاق هذه الدراسة مسح مؤشرات تقييم البيانات الوكيلية التي تقيس قدرة النماذج على التخطيط واتخاذ القرارات المتسلسلة، فضلاً عن مؤشرات التقييم متعدد الوسائط التي تدمج النص مع الصورة والصوت. وتُمثل هذه المجالات مسارات بحثية واعدة تُوصي الدراسة بتناولها في أعمال مسحية مستقبلية.



المراجع الأجنبية

- Al-Matham, R., Darwish, K., Al-Rasheed, R., Alshammari, W., Alhoshan, M., Almazrui, A., Al Wazrah, A., Alheraki, M., Alam, F., Nakov, P., Alzahrani, N., AlBilali, E., Habash, N., El-Sheikh, A., Elmallah, M., Li, H., Mubarak, H., Anwar, M., Alyafeai, Z., et al. (2025). BALSAM: A platform for benchmarking Arabic large language models. *arXiv preprint arXiv:2507.22603*. <https://arxiv.org/abs/2507.22603>
- Bai, G., Liu, J., Bu, X., He, Y., Liu, J., Zhou, Z., Lin, Z., Su, W., Ge, T., Zheng, B., & Ouyang, W. (2024). MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. <https://aclanthology.org/2024.acl-long.401.pdf>
- Chao, P., DeBenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Schwag, V., Dobriban, E., Flammarion, N., Pappas, G. J., Tramer, F., Hassani, H., & Wong, E. (2024). JailbreakBench: An open robustness benchmark for jailbreaking large language models. *arXiv*. <https://arxiv.org/abs/2404.01318>
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., et al. (2021). Evaluating large language models trained on code. *arXiv*. <https://arxiv.org/abs/2107.03374>
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. *arXiv*. <https://arxiv.org/abs/2403.04132>
- Christian, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2023). Deep reinforcement learning from human preferences. *arXiv*. <https://arxiv.org/abs/1706.03741>
- Elangovan, A., Liu, L., Xu, L., Bodapati, S., & Roth, D. (2024). ConSiDERS-The-Human evaluation framework: Rethinking human evaluation for generative large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. <https://aclanthology.org/2024.acl-long.63/>

- Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating summarization evaluation. *arXiv*. <https://arxiv.org/abs/2007.12626>
- Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., et al. (2024). The language model evaluation harness. *Zenodo*. <https://zenodo.org/doi/10.5281/zenodo.10256836>
- Google. (2025). Announcing LMEval: An open source framework for cross-provider LLM evaluation. *Google Open Source Blog*. <https://opensource.googleblog.com/2025/05/announcing-lmeval-an-open-source-framework-cross-model-evaluation.html>
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *arXiv*. <https://arxiv.org/abs/2009.03300>
- Jain, D., Kumar, P., Gehman, S., Zhou, X., Hartvigsen, T., & Sap, M. (2024). PolygloToxicityPrompts: Multilingual evaluation of neural toxic degeneration in large language models. *arXiv*. <https://arxiv.org/abs/2405.09373>
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., et al. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=iO4LZibEqW>
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., et al. (2024). GPT-4 technical report. *arXiv*. <https://arxiv.org/abs/2303.08774>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askeel, A., Welinder, P., Christiano, P., Leike, J., et al. (2022). Training language models to follow instructions with human feedback. *arXiv*. <https://arxiv.org/abs/2203.02155>



- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv*. <https://arxiv.org/abs/2206.04615>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv*. <https://arxiv.org/abs/2302.13971>
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv*. <https://arxiv.org/abs/1905.00537>
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv*. <https://arxiv.org/abs/1804.07461>
- Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., & Huang, M. (2023). SafetyBench: Evaluating the safety of large language models with multiple choice questions. *arXiv*. <https://arxiv.org/abs/2309.07045>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv*. <https://arxiv.org/abs/2306.05685>
- Alyahya, H. A., Khan, H., Alnumay, Y., Bari, M. S., & Yener, B. (2025). ZeroSumEval: An extensible framework for scaling LLM evaluation with inter-model competition. *arXiv*. <https://doi.org/10.48550/arXiv.2503.10673>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language models are few-shot learners. *arXiv*. <https://doi.org/10.48550/arXiv.2005.14165>

- Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A. B., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., & Baldwin, T. (2024). ArabicMMLU: Assessing massive multitask language understanding in Arabic. *arXiv*. <https://arxiv.org/abs/2402.12840>
- Ashraf, Y., Wang, Y., Gu, B., Nakov, P., & Baldwin, T. (2024). Arabic dataset for LLM safeguard evaluation. *arXiv*. <https://arxiv.org/abs/2410.17040>
- King Salman Global Academy for Arabic Language. (2024). *Arabic Computing Observatory Report*. King Salman Global Academy for Arabic Language. <https://arai.ksaa.gov.sa/marsad/>

نبذة تعريفية عن الباحثين:

Author Biodata	بيانات الباحثة:
Raghad Al Rasheed is an AI Developer at the King Salman Global Academy for Arabic Language, Riyadh, Saudi Arabia..	رغد آل رشيد، مطوّركاء اصطناعي في مجمع الملك سلمان العالمي للغة العربية، الرياض، المملكة العربية السعودية.
معرف أوركيد (ORCID) : 0009-0001-8895-943X	
Email: ralrasheed@ksaa.gov.sa	

Author Biodata	بيانات الباحثة:
Amal Al Mazrou is a Senior Linguistic Data Analysis Specialist at the King Salman Global Academy for Arabic Language, Riyadh, Saudi Arabia.	أمال المزروع، أخصائي أول تحليل البيانات اللغوية في مجمع الملك سلمان العالمي للغة العربية، الرياض، المملكة العربية السعودية.
معرف أوركيد (ORCID) : 0009-0004-4796-2140	
Email: aalmazrua@ksaa.gov.sa	

Author Biodata	بيانات الباحثة:
Munirah Al Houshan is the Linguistic Data Manager at the King Salman Global Academy for Arabic Language, Riyadh, Saudi Arabia.	منيرة الحوشان، مدير البيانات اللغوية في مجمع الملك سلمان العالمي للغة العربية، الرياض، المملكة العربية السعودية.
معرف أوركيد (ORCID) : 0000-0001-5705-3590	
Email: malhoshan@ksaa.gov.sa	

Author Biodata	بيانات الباحثة: روان المعثم، مدير قسم أبحاث وابتكار معالجة اللغة العربية الطبيعية في مجمع الملك سلمان العالمي للغة العربية، الرياض، المملكة العربية السعودية.
معرف أوركيد (ORCID) : 0000-0003-2184-3210	
Email: ralmatham@ksaa.gov.sa	

Author Biodata	بيانات الباحثة: ريم المطيري، عالم بيانات في انتروبي، الرياض، المملكة العربية السعودية.
معرف أوركيد (ORCID) : 0009-0008-3027-7760	
Email: RAlmutairi@Entropy.sa	

Author Biodata	بيانات الباحث: تركي العتيبي، عالم بيانات أول في انتروبي، الرياض، المملكة العربية السعودية.
معرف أوركيد (ORCID) : 0009-0001-3909-4425	
Email: Turki@Entropy.sa	

Author Biodata	بيانات الباحث: فيصل بن بريكان، مدير منتجات ذكاء اصطناعي في انتروبي، الرياض، المملكة العربية السعودية.
معرف أوركيد (ORCID) : 0009-0008-6333-0169	
Email: FBinBurykan@Entropy.sa	

الملاحق

ملحق (أ)

المعادلات

أ. المعاييرة

$$ECE = \sum_{m=1}^M \frac{|(B_m)|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$$

على سبيل المثال :

لنفترض أن لدينا نموذج تصنيف يُستخدم للتفريق بين صور "القطط" و "الكلاب".

عند اختبار النموذج على خمس صور، كانت نتائجها كما يلي :

الصورة الأولى: توقع "قطعة" بنسبة ثقة ٩٠٪، وكانت النتيجة صحيحة.

الصورة الثانية: توقع "كلب" بنسبة ثقة ٨٠٪، وكانت النتيجة صحيحة.

الصورة الثالثة: توقع "قطعة" بنسبة ثقة ٦٠٪، وكانت النتيجة خاطئة (كانت الصورة "كلب").

الصورة الرابعة: توقع "كلب" بنسبة ثقة ٧٠٪، وكانت النتيجة صحيحة.

الصورة الخامسة: توقع "قطعة" بنسبة ثقة ٦٠٪، وكانت النتيجة صحيحة.

لحساب مقياس ECE، نقوم بتقسيم التوقعات إلى مجموعات بناءً على نسبة الثقة. على سبيل المثال، نضع جميع التوقعات التي تقع ثقتها بين ٦٠ و ٧٠٪ في مجموعة واحدة، والتوقعات بين ٨٠ و ٩٠٪ في مجموعة أخرى، وهكذا.

ثم نحسب لكل مجموعة ما يلي :

متوسط الثقة: أي كم يثق النموذج بنفسه داخل تلك المجموعة ؟

الدقة الفعلية: أي نسبة التوقعات الصحيحة داخل نفس المجموعة.

الفرق بين الثقة والدقة: وهو ما يعكس مدى توافق ثقة النموذج مع أدائه الحقيقي.

مثلاً، مجموعة التوقعات بنسبة ٦٠٪، كانت الدقة الفعلية ٥٠٪ فقط (الصورة الخامسة صحيحة، الثالثة خاطئة)، أي أن النموذج كان واثقاً أكثر من اللازم.

في مجموعة أخرى، كانت الثقة ٧٠٪ لكن الدقة ١٠٪، مما يشير إلى أن النموذج كان متحفظاً أكثر من اللازم. عند جمع هذه الفروقات الموزونة بحسب عدد التوقعات في كل مجموعة، نحصل على قيمة (ECE). في هذا المثال كانت قيمة (ECE) تساوي ٠,١٦ حيث توجد فجوة واضحة بين ما يتوقعه (الثقة) وما يحققه فعلياً (الدقة).

ب. المتانة

$$(\text{Robustness Gap} = \text{Original Accuracy} - \text{Perturbed Accuracy})$$

على سبيل المثال:

النص الأصلي:

تفحص درجة حرارة المريض وتدوّنوها في السجل. وتعطي المريض بعض الدواء.

توقع النموذج: ثم تُجري العملية.

النص بعد التغيير (استبدال الجنس):

يفحص درجة حرارة المريض ويدوّنوها في السجل. ويعطي المريض بعض الدواء.

توقع النموذج: ثم يُجري العملية.

وتُحسب المتانة بمقارنة الجوابين قبل وبعد تغيير النص.

وفي المثال أعلاه، نلاحظ أن النموذج لم يختلف إجابته. وهذا يعني أنه لم يكن حساساً للتغييرات الطفيفة.

$$\text{Accuracy}(g) = \frac{\# \text{ correct in group } g}{\# \text{ total in group } g}$$

$$\text{performance Disparity} = \max_{g_{EG}} \text{Accuracy}(g) - \min_{g_{EG}} \text{Accuracy}(g)$$

وعلى سبيل المثال:

تقيس العدالة جوانب عدة في نماذج الذكاء الاصطناعي، من أبرزها: التمييزات الجنسية. وتقاس بناءً على توليد النماذج لنصوص في مجال معين. لنفترض أننا طلبنا من النموذج توليد نصوص تصف طبيعة عمل وظيفة الهندسة، مثلاً يقوم الممرض بمتابعة حالة المريض باستمرار، وتسجيل العلامات الحيوية.

يعمل الممرض جنباً إلى جنب مع الأطباء لتنسيق خطة العلاج.

يحرص الممرض على تثقيف المريض وأسرته حول طرق العناية الصحية.

وبناء عليه، يتم حساب مؤشرين: التمثيل الديموغرافي، والتمثيل النمطي.

التمثيل الديموغرافي: كما نلاحظ في المثال أعلاه، ولّد النموذج نصوصاً تشير إلى ذكر (ثلاث مرات) ولم يتطرق إلى أي مثال يشير إلى أنثى.

التمثيل النمطي: في المثال أعلاه ربط النموذج مهنة التمريض بالذكر ولم يتطرق لأي مثال للأنثى.

ت. السُميّة

$$\text{Toxicity Score} = \frac{\text{Toxic}}{\text{Toxic} + \text{Non - Toxic}}$$

وعلى سبيل المثال:

تُقاس السُميّة في (HELM) باستخدام (Perspective API) وهو يقيم سمية المخرجات عبر عدة أبعاد، من أبرزها:

TOXICITY - احتمال أن يكون النص فقط أو غير محترم، أو قد يدفع شخصاً لمغادرة المحادثة.

SEVERE TOXICITY - سُميّة أشد تطرفاً (مثل التهديدات المباشرة).

IDENTITY ATTACK - هجمات موجّهة ضد شخصٍ ما بناءً على هويته (العرق، الجنس، إلخ).

INSULT - إهانات مباشرة.

PROFANITY - استخدام الألفاظ النابية.

THREAT - نص يُعبّر عن نية لإلحاق الأذى.

وُثّقاس بحساب عدد المخرجات السامة كنسبة من المخرجات ككل للنماذج.

ملحق (ب)

العلاقة بين مقاييس التقييم السبعة في إطار (HELM)

في ورقة (HELM)، يشير المؤلفون إلى أن الدقة، والصلابة، والإنصاف تميل إلى أن تكون مترابطة بقوة عبر النماذج والسيناريوهات المختلفة، أي أن النماذج اللغوية التي تحقق أداءً جيداً من حيث الدقة غالباً ما تُظهر أيضاً صلابة أعلى (مقاومة للاضطرابات أو الضوضاء) وإنصافاً أفضل (أداءً متقارباً عبر المجموعات السكانية المختلفة).

بالإضافة إلى ذلك، تستكشف ورقة (HELM) التداخل (أو المقايضات) بين المقاييس السبعة جميعها - الدقة، المعايير، الصلابة، الإنصاف، التحيز، السُميّة، الكفاءة - لكنها لا تعرض معاملات ارتباط كمية مفصلة، أو مصفوفة ارتباط كاملة. وبدلاً من ذلك، يركز التحليل على إبراز العلاقات البارزة، مثل الارتباط القوي الملحوظ بين الدقة والصلابة والإنصاف، مع التحذير من أن الأداء العالي في مقياس واحد لا يضمن بالضرورة أداءً مرتفعاً بصورة متساوية في جميع المقاييس الأخرى.



مجمع الملك سلمان
العالي للغة العربية

King Salman Global Academy for Arabic Language



الأكاديمية العالمية
للغة العربية

Journal of Computational Linguistics & Arabic Language Processing

