

بَلْسَم: مؤشر نضج النماذج اللغوية الترجيحية

تاريخ التقرير: ١١ أكتوبر ٢٠٢٦م

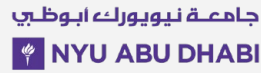




HUMAIN



Entropy



stc



groq

aiXplain

شركاء
المؤشر:

كلمة الأمين العام لمجمع الملك سلمان العالمي للغة العربية

يعمل مجمع الملك سلمان العالمي للغة العربية على خدمة اللغة العربية وترسيخ حضورها العالمي، بمسارات يسند بعضها بعضاً: التخطيط اللغوي، والحوسبة، والتعليم، والثقافة. وقد غدا حضور العربية في تقنيات الذكاء الاصطناعي وجهاً من وجوه مكانتها بين اللغات؛ ولذلك عُني المجمع بتقييم هذه التقنيات وقياس نضجها في تعاملها مع العربية، إيماناً بأن ما لا يُقاس لا يمكن تطويره.

وقد بدأ المجمع عمله في هذا الميدان بإطلاق مؤشر (بَلَسَم) في تقريره الأول، الذي وضع منهجيةً لتقييم النماذج اللغوية التوليدية تجمع بين التحكيم البشري والتقييم الآلي، ثم وسَّع في تقريره الثاني نطاق التقييم بأسئلة جديدة وفئات إضافية. ويأتي هذا التقرير امتداداً لذلك العمل، وخطوةً جديدةً فيه؛ ينتقل بها المؤشر إلى صنف من النماذج اللغوية لم يشمل التقييم من قبل.

ويسعد المجمع بنشر تقرير (مؤشر نضج النماذج اللغوية الترجيحية)؛ وهي النماذج التي لا تولّد نصّاً، بل ترجّح قراراً من بين خيارات محددة. ولا يقف التقرير عند سؤال: هل أصاب النموذج في قراره؟ بل يتجاوزهُ إلى سؤال آخر: هل يمكن الاطمئنان إلى ثقته في ذلك القرار؟ ولذلك يقيس أداء النماذج بأربعة مقاييس، هي: الدقة، والترجيح، والثقة، والترتيب، في (٣٧) مهمة موزعة على (١٢) فئة لغوية ومعرفية، وتخضع نتائجه لتدقيق مستقل يعزز موثوقيتها واعتمادها على المستوى البحثي والتطبيقي.

وإذ تتسع دائرة شركاء المؤشر تقريراً بعد تقرير، فإن المجمع يدعو الجهات الأكاديمية والتقنية، والباحثين والمهتمين، إلى الإسهام في صناعة تقريره القادم؛ ببناء بياناته، وتطوير منهجيته، وترشيح النماذج لتقييمها. ويمكن التواصل للتعاون والعمل معنا على بريد بلسم المنشور في الموقع.

وفي الختام، نتقدّم بالشكر إلى صاحب السمو الأمير/ بدر بن عبدالله بن فرحان آل سعود -حفظه الله- وزير الثقافة، رئيس مجلس أمناء المجمع، على دعمه الدائم لمشروعات المجمع وأعماله، وإلى شركاء المؤشر الذين أسهموا في بناء بياناته وتطوير منهجيته؛ بما يعزز من ريادة اللغة العربية في المجال التقني.

الأمين العام

أ. د. عبد الله بن صالح الوشمي.

٤	مؤشر نضج النماذج اللغوية الترجيحية
٤	إحصاءات عامة عن التقييم بمؤشر نضج النماذج اللغوية الترجيحية
٥	فئات المهام المشمولة في التقييم
٨	منهجية التقييم
٨	طريقة التقييم المعتمدة
٩	مقاييس التقييم
٩	التحقق من النتائج
١٠	النماذج المُقيّمة
١١	النتائج
١٢	المتوسط العام للنماذج
١٢	أولاً: الدقة (Accuracy)
١٣	ثانياً: الترجيح (Brier Score (BS))
١٣	ثالثاً: الثقة (Expected Calibration Error (ECE))
١٤	رابعاً: الترتيب (Area Under the Risk-Coverage Curve (AURC))
١٥	حجم النموذج مقابل الأداء العام
١٥	النتائج حسب الفئات والمهام
٣٨	ملحق (أ): تعريف الفئات والمهام

مؤشر نضج النماذج اللغوية الترجيحية

مؤشر نضج النماذج اللغوية الترجيحية يقيس قدرة النماذج على اختيار قرار من بين خيارات محددة في مهام مكتوبة بالعربية، وذلك بأسئلة اختيار من متعدد تشمل مجالات لغوية ومعرفية متنوعة. يهدف هذا التقرير إلى عرض نتائج التقييمات الرسمية الكاملة كما حُفظت بعد التحقق منها. وتُعرض أربعة مقاييس أساسية بعد توحيدها من (٠) إلى (١٠٠). ويُشتق منها متوسط عام بأوزان متساوية لتقديم ملخص موحد لدرجات العرض، ولا يُعد هذا المتوسط مقياساً علمياً مستقلاً. تتكون كل حالة اختبار من مدخل وسياق حسب المهمة، وتعليمات عند توفرها، ومجموعة خيارات محددة، وإجابة مرجعية واحدة.

إحصاءات عامة عن التقييم لمؤشر (بَلَسَم)

 عدد النماذج المقيمة: (١٣)	 عدد الأسئلة: (٨٥٩)	 عدد المهام: (٣٧)	 عدد الفئات: (١٣)
--	---	---	---

✓ فئات المهام المشمولة في التقييم:

الفئة	الاسم الإنجليزي للمهمة	المهمة	الاسم الإنجليزي للمهمة
الكتابة الإبداعية	Creative Writing	إكمال النصوص	Text Completion
		تقييم صحة تنمة النص	Text Continuation Evaluation
التضمين	Entailment	التشابه الدلالي	Semantic Similarity
		الاستلزام النصي	Textual Entailment
دقة المعلومات	Factuality	التحقق من صحة الإجابة	Answer Verification
		التحقق من صحة الادعاء	Claim Verification
ملء الفراغ	Fill in the Blank	ملء الفراغ	Fill in The Blank
استخلاص المعلومات	Information Extraction	تصنيف الكيانات	Entity Categorization
		تصنيف العلاقات بين الكيانات	Entity Relation Classification
		استخلاص المعلومات المطلوبة	Extracting Required Information

الفئة	الاسم الإنجليزي للفئة	المهمة	الاسم الإنجليزي للمهمة
المنطق	Logic	تصنيف السبب والنتيجة	Cause Effect Classification
		تصنيف التماسك النصي	Coherence Classification
		التحقق من المنطق السليم	Commonsense Validation
		الاستدلال المنطقي	Logical Reasoning
		ترتيب الجمل	Sentence Ordering
تنفيذ البرنامج	Program Execution	تنفيذ البرامج	Program Execution
الإجابة عن الأسئلة	Question Answering	الإجابة عن الأسئلة المطروحة	Answering Given Question
الفهم القرائي	Reading Comprehension	تصنيف إمكان الإجابة	Answerability Classification
		اختيار الحوار المناسب	Dialogue Selection
		اختيار التفسير المناسب	Explanation Selection
		فهم السؤال	Question Understanding
		التحقق من الاستدلال النصي	Textual Inference Validation
		تحديد القسم الكلامي	Part of speech tagging

الفئة	الاسم الإنجليزي للفئة	المهمة	الاسم الإنجليزي للمهمة
تصنيف النصوص	Text Classification	تحديد اللهجة	Dialect Identification
		تعرف أفعال الحوار	Dialogue Act Recognition
		اكتشاف المشاعر	Emotion Detection
		تصنيف السلوكيات الأخلاقية	Ethics Classification
		اكتشاف الإساءة	Offensive Language Detection
		تصنيف الأسئلة	Question Categorization
		اكتشاف السخرية	Sarcasm Detection
		تحليل المشاعر	Sentiment Analysis
		كشف الرسائل المزعجة	Spam Detection
		اكتشاف القوالب النمطية	Stereotype Detection
		التصنيف الفئوي للنصوص	Text Categorization
		تحديد الموضوع	Topic Identification

الفئة	الاسم الإنجليزي للمهمة	المهمة	الاسم الإنجليزي للمهمة
التعديل على النصوص	Text Manipulation	تصحيح القواعد النحوية	Grammar Correction
الترجمة الآلية والنقل الصوتي	Translation/Transliteration	الترجمة الآلية	Machine Translation

منهجية التقييم:

اعتمدت منهجية تقييم موحدة لقياس نضج النماذج اللغوية الترجيحية في مجموعة من المهام العربية، وذلك بأن يُطلب من كل نموذج ترجيح اختيار من بين مجموعة من الخيارات المحددة لكل حالة اختبار. تتضمن كل حالة سياقاً أو مدخلاً، وتعليمات حسب طبيعة المهمة، ومجموعة من الخيارات، وإجابة مرجعية واحدة. وتُعد إجابة النموذج صحيحةً عند اختيار الخيار المطابق للإجابة المرجعية. يُعتمد التقييم الرسمي على بيانات الاختبار في إصدار (بَلْسَم) الثاني (Balsam V2)، ولا تدخل بيانات التطوير ضمن النتائج النهائية. ويضم الإصدار الحالي (٨٥٩) حالة اختبار موزعة على (٣٧) مهمة ضمن (١٢) فئة، وجميع حالات الاختبار في هذا الإصدار قابلة للتقييم.

طريقة التقييم المعتمدة:

تعتمد عملية التقييم على محوّل (Adapter) خاص بكل نموذج، يحوّل مخرجاته إلى صيغة قرار موحّدة قبل التقييم؛ بما يتيح تقييم النماذج بالطريقة نفسها بغض النظر عن طريقة تشغيلها. ويتضمن ناتج الإجابة الخيار الذي حدده النموذج، إضافةً إلى توزيع الاحتمالات على الخيارات عندما يوفره النموذج بصورة أصلية. ولا يُشتق توزيع احتمالي من درجة ثقة منفردة إذا لم يكن النموذج يوفر هذا التوزيع. تُحسب النتائج على ثلاثة مستويات: النتيجة الإجمالية للمؤشر، ونتائج الفئات، ونتائج المهام. وتُحسب النتيجة الإجمالية ونتيجة كل فئة على مستوى حالات الاختبار مباشرةً، وليس بأخذ متوسط بسيط لنتائج المهام؛ بما يحافظ على الوزن الفعلي لعدد الحالات الموجودة في كل مجموعة.

مقاييس التقييم:

تعتمد المنهجية على أربعة مقاييس أساسية لتقييم أداء النماذج ومخرجاتها. ولتسهيل قراءة النتائج والمقارنة بين النماذج، يُجرى تحويل كافة القيم العلمية إلى درجة عرض موحدة تتراوح من ٠ إلى ١٠٠، بحيث تشير الدرجة الأعلى دائمًا إلى أداء أفضل وأعلى جودة.

الدقة (Accuracy)

تقيس نسبة القرارات أو الإجابات الصحيحة التي اتخذها النموذج من إجمالي حالات الاختبار. وتُحسب درجتها المعروضة بضرب القيمة العلمية للمقياس في ١٠٠ مباشرةً، وفق المعادلة: $\text{درجة العرض} = \text{القيمة العلمية} \times 100$

الترجيح (Brier Score - BS)

يقيس جودة التوزيع الاحتمالي الذي يمنحه النموذج للخيارات ومدى قربها من الإجابة الصحيحة، ولا يقتصر على صحة الاختيار النهائي فحسب. ولأن القيمة العلمية للمقياس تُعبّر عن نسبة الخطأ، حيث تكون القيمة الأقل أفضل، تُحوّل إلى درجة العرض وفق المعادلة: $\text{درجة العرض} = (1 - \text{القيمة العلمية}) \times 100$

الثقة (Expected Calibration Error - ECE)

تقيس مدى واقعية وموثوقية مستويات الثقة التي يُسندها النموذج لقراراته مقارنة بدقته الفعلية، لمعرفة مدى توافق ثقة النموذج مع صحة ترجيحه. ولأن القيمة العلمية للمقياس تعبّر عن مقدار خطأ المعايرة، حيث تكون القيمة الأقل أفضل، تُحوّل إلى درجة العرض الموحدة وفق المعادلة: $\text{درجة العرض} = (1 - \text{القيمة العلمية}) \times 100$

الترتيب (Area Under the Risk-Coverage Curve - AURC)

يقيس قدرة درجات الثقة في النموذج على ترتيب القرارات بحسب موثوقيتها، بحيث تُقدّم القرارات الأكثر صحة على غيرها. وتُحسب درجة العرض من القيمة العلمية للمساحة تحت منحنى المخاطرة-التغطية، حيث تكون القيمة العلمية الأقل أفضل، وفق المعادلة: $\text{درجة العرض} = (1 - \text{القيمة العلمية}) \times 100$

التحقق من النتائج:

تخضع النتائج الرسمية لعملية تحقق آلية ضمن نظام التقييم قبل اعتمادها، للتأكد من اكتمال حالات الاختبار وعدم وجود حالات مفقودة أو مكررة، ومطابقة إصدار البيانات وبصمته، والتحقق من صلاحية توزيعات الاحتمالات عند توفرها. كما تشمل عملية التحقق إعادة حساب المقاييس من سجلات الحالات ومقارنتها بالقيم المسجلة في النتائج؛ للتأكد من صحتها واتساقها.

لتسهيل قراءة النتائج ومقارنتها؛ وُحدت المقاييس على مقياس من (٠) إلى (١٠٠) وتشير القيمة الأعلى دائمًا إلى أداء أفضل. لا يغيّر هذا التوحيد طريقة حساب المقاييس العلمية الأصلية، وإنما يوحّد اتجاهها وطريقة عرضها لتسهيل المقارنة بين النماذج. جميع الدرجات المعروضة في الجداول والرسوم من (٠) إلى (١٠٠)، والقيمة الأعلى أفضل. أربعة مقاييس أساسية، ويُشتق منها متوسط عام. المتوسط العام هو متوسط درجات العرض للمقاييس الأربعة بأوزان متساوية، إذ يمثل المتوسط العام المتوسط الحسابي المتساوي الأوزان لدرجات الدقة والترجيح والثقة والترتيب بعد توحيدها على مقياس من (٠) إلى (١٠٠). ويُستخدم المتوسط العام لتقديم ملخص رقمي موحد لأداء النموذج بواسطة الأبعاد الأربعة للمؤشر. لا يمثل المتوسط العام مقياسًا علميًا مستقلًا، ولا يغيّر القيم العلمية الأصلية للمقاييس.

النماذج المُقيّمة:

اسم النموذج	إتاحة المصدر	الحجم (عدد المعاملات)	الوصف	المرجع/ الرابط
Bespoke Labs — Bespoke-Nimble-9B	مفتوح المصدر	٩,٠٠٠,٠٠٠,٠٠٠	نموذج متخصص في تقييم الخيارات المرتبطة بالسياق، واتخاذ القرار بينها دون توليد إجابة حرة.	الرابط
Cloudflare — Clef	مفتوح المصدر API + رسمي	٢٧,٠٠٠,٠٠٠,٠٠٠	نموذج قرار مستضاف عبر Cloudflare Workers AI، يختار من بين خيارات محددة ويعيد احتمالاتها.	الرابط
Convai Innovations — Laya Multilingual	مفتوح المصدر	٣٢٢,٠٠٠,٠٠٠	نموذج قرار متعدد اللغات غير توليدي، مخصص للاختيار بين خيارات محددة وإرجاع احتمالاتها، ويدعم العربية ضمن اللغات التي يشملها.	الرابط
Decider-2B	مفتوح المصدر	١,٨٨١,٨٢٥,٠٨٨	نموذج متخصص في اتخاذ القرار من بين خيارات محددة دون توليد نص.	الرابط
Faisal	مفتوح المصدر	٤,٠٠٠,٠٠٠,٠٠٠	نموذج قرار عربي يعتمد على ثقة مُعايرة في مخرجاته.	الرابط
IJyad — Jeb	مفتوح المصدر	١٧٧,٦٠٩,٩٨٥	نموذج قرار عربي غير توليدي، مخصص لمهام التصنيف والاستدلال واتخاذ القرار.	الرابط
Jared Palmer — Kev-0.8B	مفتوح المصدر	٨٠٠,٠٠٠,٠٠٠	نموذج قرار صغير الحجم وغير توليدي، مخصص للاختيار بين خيارات محددة دون توليد نص.	الرابط
TypeSafe Jev — jev-1.13.0	مغلق المصدر	غير معلن	نموذج متخصص في اتخاذ القرارات المهيكلة، ويدعم الاختيار والتقييم بالدرجات والأحكام الشائنية دون توليد نص.	الرابط
Wouze — Laya-Ara	أوزان متاحة للاستخدام البحثي/غير التجاري	٣٢١,٩٠٨,٩٩٨	نموذج قرار عربي غير توليدي مخصص لفهم اللغة العربية والتصنيف والاستدلال.	الرابط

الرابط	نموذج قرار محلي مفتوح الأوزان من Liquid AI، يختار من بين خيارات محددة ويعيد احتمالاتها في تعريفة واحدة.	٣,١٢٣,٤٨٣,٨٨٨	مفتوح المصدر	Liquid AI — d1-3B
الرابط	نموذج قرار غير توليدي من Microsoft، مبني (post-trained) على Qwen3.5-9B، يختار من بين خيارات محددة ويعيد احتمالاً معيّزاً لكل خيار.	٩,٠٠٠,٠٠٠,٠٠٠	غير معلن (مفتاح عبر API)	Microsoft — Microsoft-Decision-1
الرابط	نموذج قرار من OpenAI يُقَيَّم عبر Decisions API للاختيار بين خيارات محددة مع إرجاع احتمالاتها.	غير معلن	مغلق المصدر	OpenAI — GPT-6 Luna (Decisions)

ملاحظة: أُجري تقييم Clef عبر Cloudflare Workers AI API، والذي يوفّر نسخة مستضافة من نموذج Clef نفسه المتاح بأوزان مفتوحة. وتشير النتائج الواردة في التقرير إلى تشغيله عبر API.

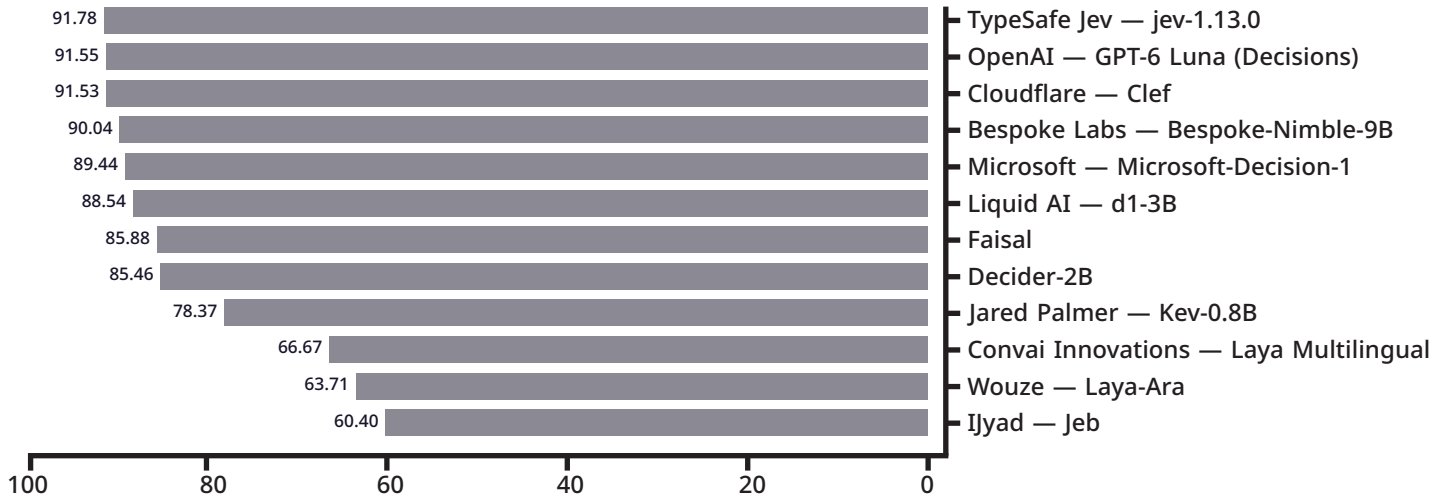
النتائج:

النموذج	الدقة	الترجيح	الثقة	الترتيب	المتوسط العام
TypeSafe Jev — jev-1.13.0	٨٤,٦٨	٩١,٠٨	٩٦,٣٦	٩٥,٤٠	٩١,٧٨
OpenAI — GPT-6 Luna (Decisions)	٨٤,٤٠	٩١,٥٨	٩٥,٧٣	٩٤,٤٧	٩١,٥٥
Cloudflare — Clef	٨٤,٩٨	٩١,٠٨	٩٥,٧١	٩٤,٣٣	٩١,٥٣
Bespoke Labs — Bespoke-Nimble-9B	٨٣,٣٥	٩٠,٠٨	٩٣,٦٦	٩٣,٤٨	٩٠,٠٤
Microsoft — Microsoft-Decision-1	٨٣,٤٧	٨٩,٩٩	٩١,٤٩	٩٦,٨٠	٨٩,٤٤
Liquid AI — d1-3B	٧٦,٩٥	٨٨,٦٨	٩٧,٤٩	٩١,٤٥	٨٨,٥٤
Faisal	٧٦,٨٣	٨٦,٧٦	٨٩,١٦	٩٠,٨٦	٨٥,٨٨
Decider-2B	٧١,٣٦	٨٦,٠٨	٩٧,٣٦	٨٧,٠٩	٨٥,٤٦
Jared Palmer — Kev-0.8B	٥٩,٩٥	٨١,٩١	٩٧,٦٣	٧٤,٣٩	٧٨,٣٧
Convai Innovations — Laya Multilingual	٥٠,٤١	٧٦,١٣	٨٤,٠٦	٥٦,١٤	٦٦,٦٧
Wouze — Laya-Ara	٤٨,٦٦	٧٥,٠٦	٧٩,٦٤	٥١,٨٩	٦٣,٧١
IJyad — Jeb	٣٨,٨٨	٧٥,٦١	٨٤,٧٥	٤٦,٧٧	٦٠,٤٠

جميع الدرجات من (٠) إلى (١٠٠)، والقيمة الأعلى أفضل. المتوسط العام هو المتوسط الحسابي المتساوي الأوزان لدرجات الدقة والترجيح والثقة والترتيب.

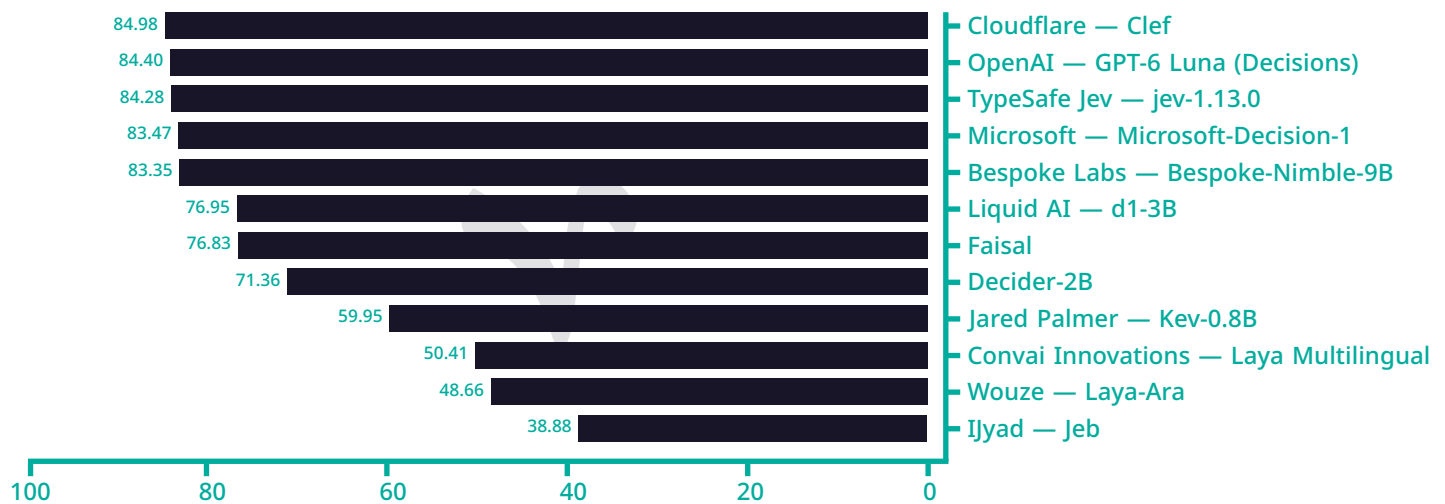
المتوسط العام للنماذج

المتوسط العام للنماذج:



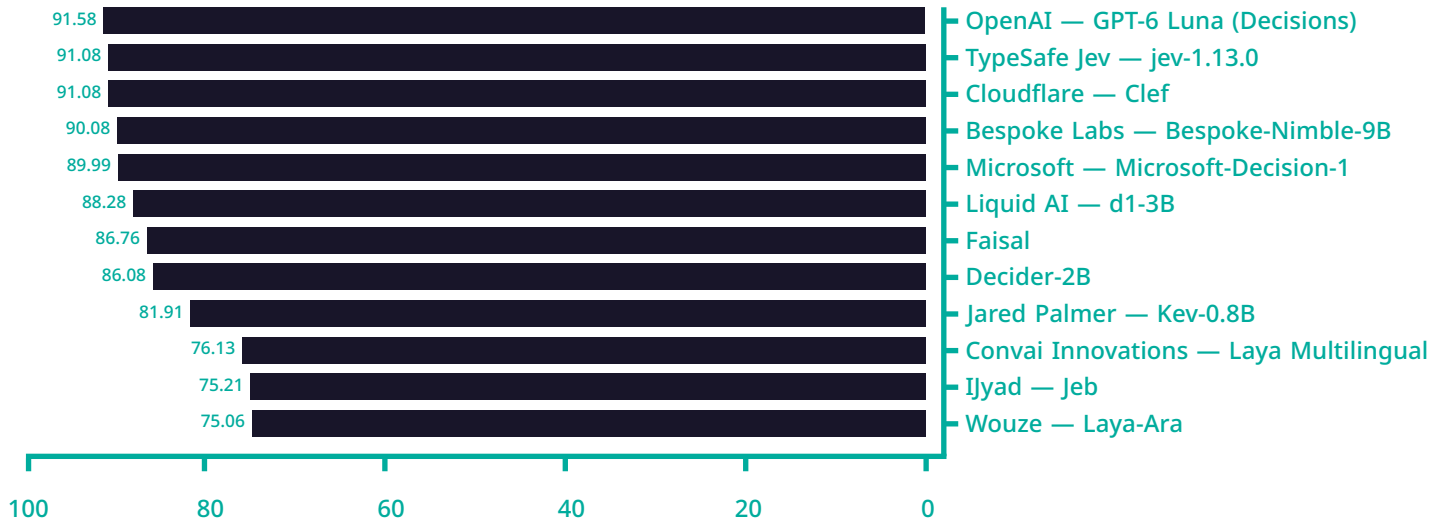
أولاً: الدقة (Accuracy):

الدقة (Accuracy):



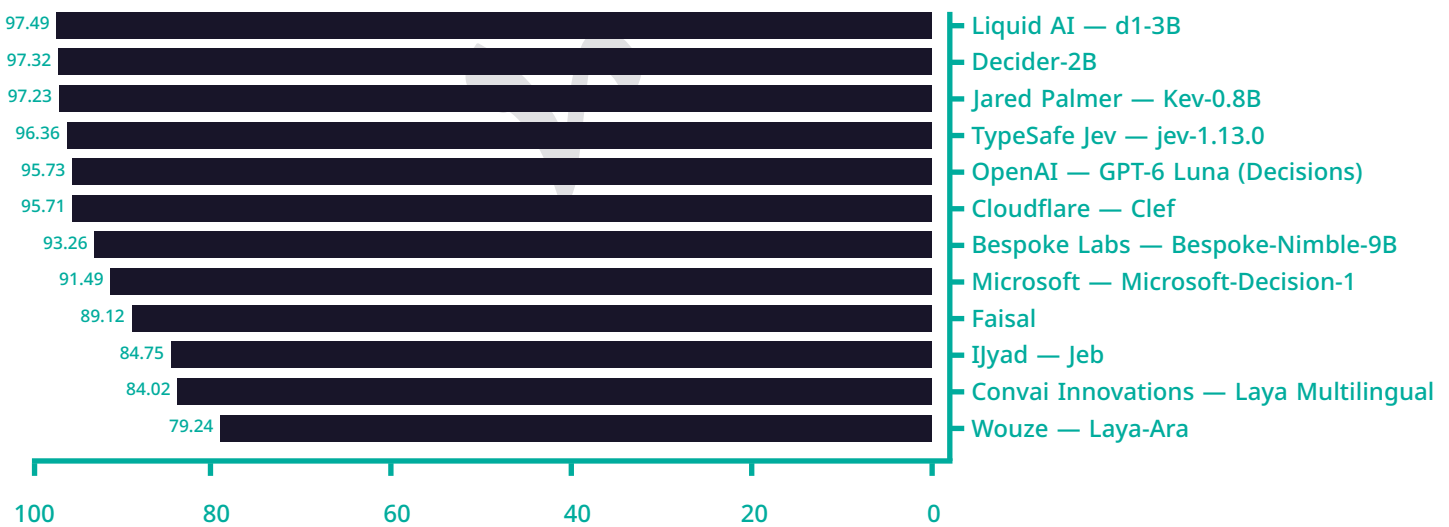
ثانيًا: الترتيب (Brier Score (BS))

ثانيًا: الترتيب (Brier Score (BS))

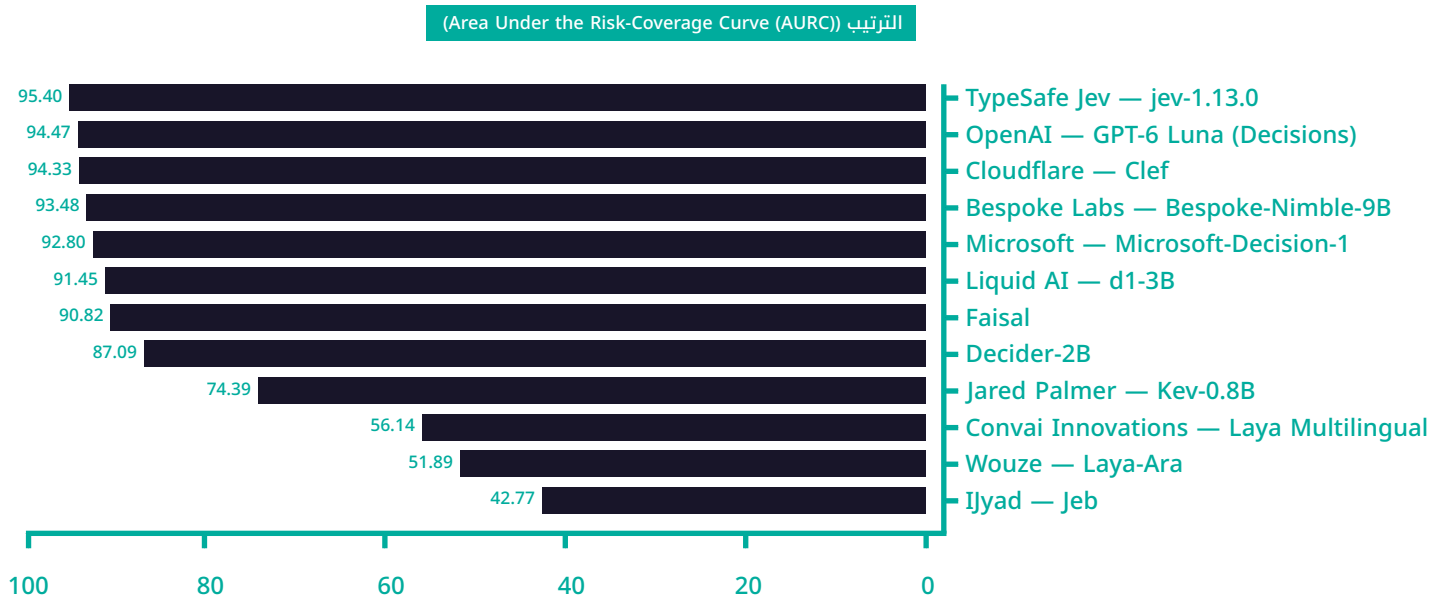


ثالثًا: الثقة (Expected Calibration Error (ECE))

الثقة - (Expected Calibration Error (ECE))



رابعًا: الترتيب (Area Under the Risk-Coverage Curve (AURC))

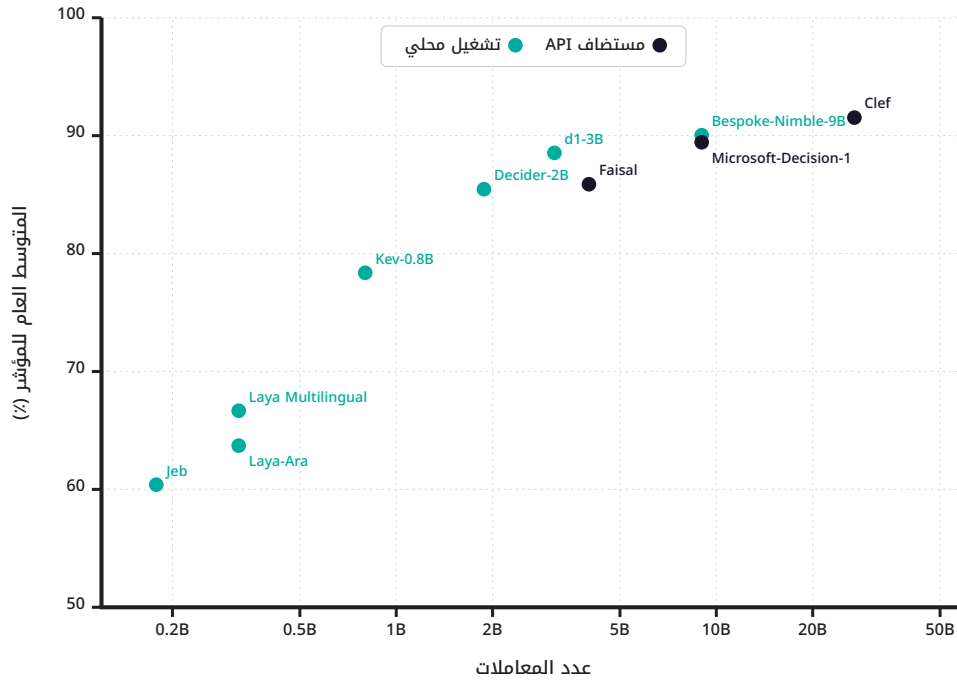


حجم النموذج وعلاقته بالأداء

يعرض الجدول أحجام النماذج المشمولة في المؤشر وفق المعلومات الموثقة والمتاحة، إلى جانب طريقة تقييمها ومتوسط أدائها العام. وتُعرض النماذج التي لا تتوفر معلومات موثقة عن حجمها بعبارة «غير معلن».

النموذج	حجم النموذج (فئة المعاملات)	طريقة التقييم	المتوسط العام
TypeSafe Jev — jev-1.13.0	غير معلن	مستضاف API	91.78
GPT-6 Luna (Decisions)	غير معلن	مستضاف API	91.55
Cloudflare — Clef	27B	مستضاف API	91.53
Bespoke-Nimble-9B	9B	تشغيل محلي	90.04
Microsoft — Microsoft-Decision-1	9B	مستضاف API	89.44
Liquid AI — d1-3B	3.12B	تشغيل محلي	88.54
Faisal	4B	مستضاف API	85.88
Decider-2B	1.88B	تشغيل محلي	85.46
Kev-0.8B	0.8B	تشغيل محلي	78.37
Laya Multilingual	0.322B	تشغيل محلي	66.67
Laya-Ara	0.322B	تشغيل محلي	63.71
Ijyad — Jeb	0.178B	تشغيل محلي	60.40

ملاحظة: تمثل قيم الحجم فئة حجم النموذج أو عدد المعاملات الموثق بحسب المعلومات المتاحة لكل نموذج؛ وفي النماذج المبنية بالتخصيص مثل Faisal، تشير القيمة إلى حجم النموذج الأساسي. ويقتصر الرسم التالي على النماذج التي تتوفر لها معلومات موثقة عن الحجم.

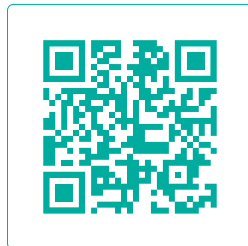


ملاحظة: يقتصر الرسم على النماذج ذات عدد المعاملات المعلن.

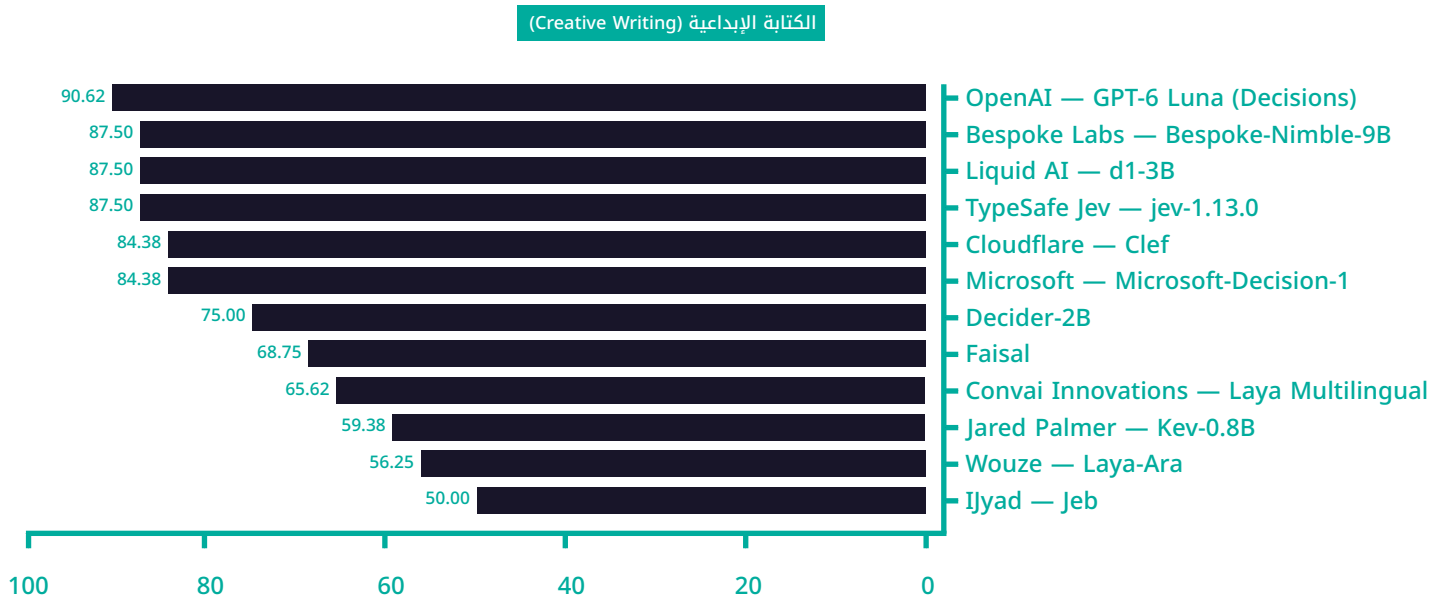
يوضح الشكل العلاقة بين حجم النموذج، ممثلًا بعدد المعاملات، ومتوسط أدائه العام في المؤشر. وتظهر النتائج اتجاهًا عامًا نحو ارتفاع الأداء مع زيادة حجم النموذج، إلا أن هذا الاتجاه لا ينطبق بصورة منتظمة على جميع النماذج؛ إذ حققت بعض النماذج الأصغر أداءً أعلى من نماذج أكبر منها. ويشير ذلك إلى أن حجم النموذج وحده لا يفسر مستوى الأداء، الذي قد يتأثر بعوامل أخرى مثل بنية النموذج ومنهجية تدريبه وتخصيصه للمهمة. كما يوضح الشكل طريقة تشغيل النماذج أثناء التقييم، سواءً محليًا أو عبر واجهة برمجية مستضافة (Hosted API). ويقتصر الرسم على النماذج ذات عدد المعاملات المعلن أو حجم النموذج الأساسي الموثق.

النتائج حسب الفئات والمهام:

رسوم المهام التالية تعرض درجة عرض الدقة من (٠) إلى (١٠٠) لكل مهمة، والقيمة الأعلى أفضل وتظهر في أعلى الرسم. أما نتيجة الفئة فتؤخذ من ملخصها المخزن على عناصرها مباشرة، لا من متوسط درجات المهام. وللإطلاع على نتائج المقاييس الأخرى لكل فئة؛ يمكن زيارة [الواجهة التفاعلية للتقرير](#).

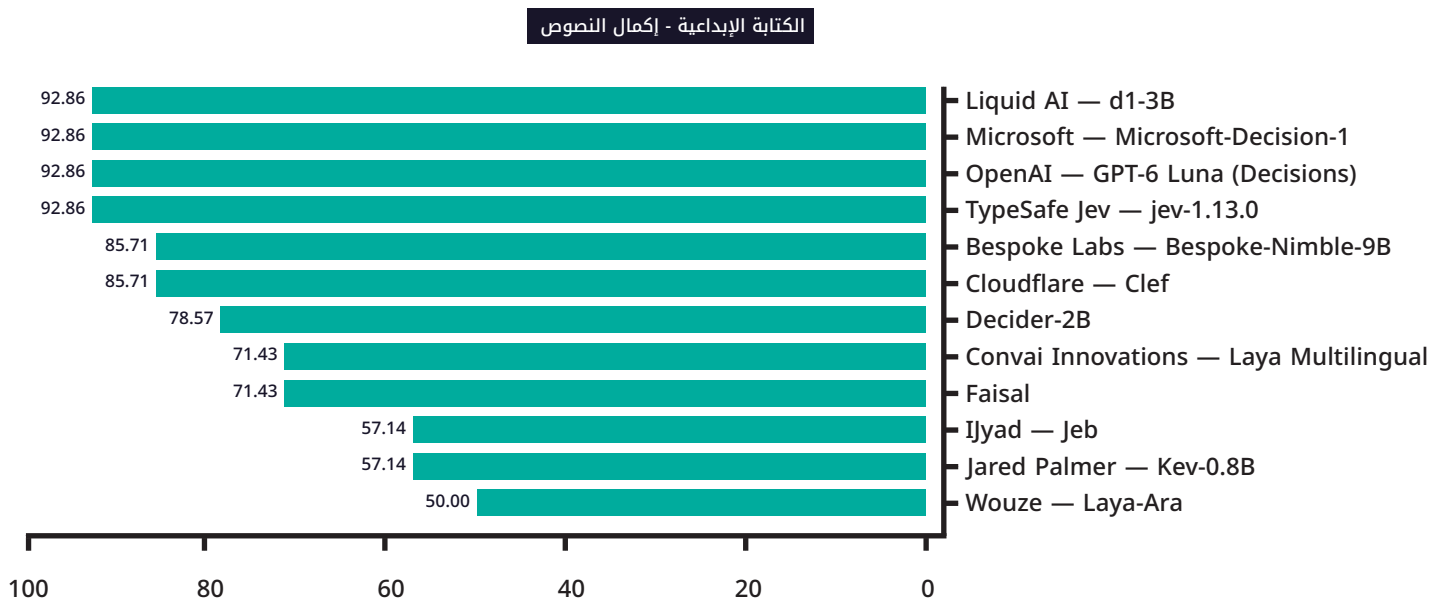


١. الكتابة الإبداعية (Creative Writing).



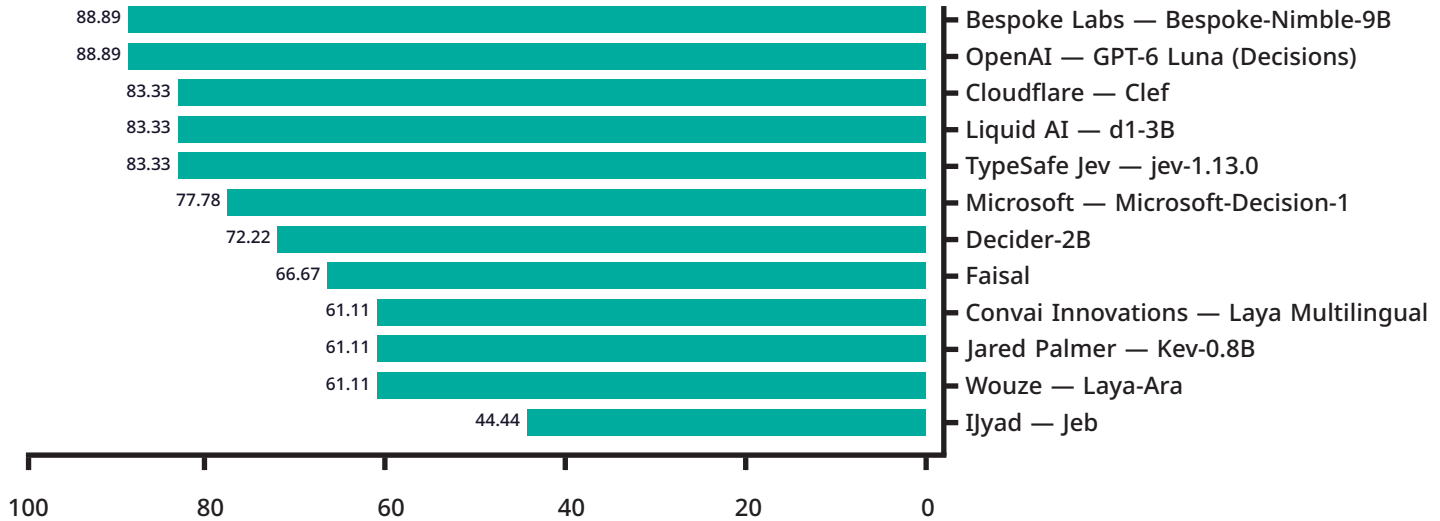
المهام الفرعية:

١.١ إكمال النصوص (Text Completion)



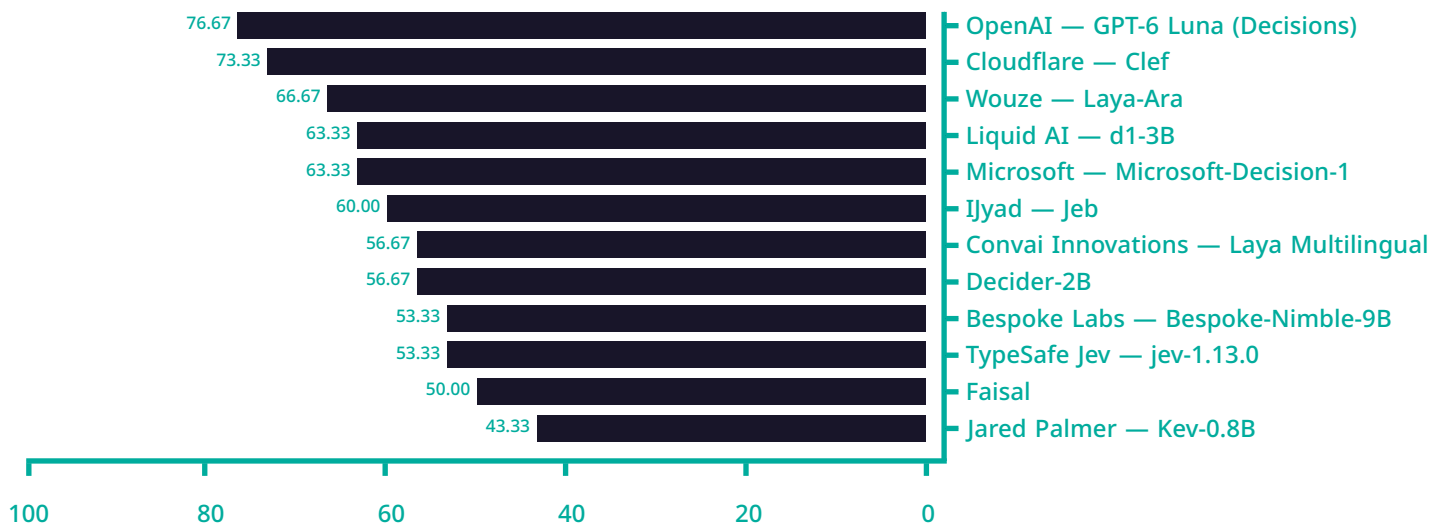
١.٢ تقييم صحة تتمة النص (Text Continuation Evaluation)

الكتابة الإبداعية - تقييم صحة تتمة النص



٢. التضمين (Entailment)

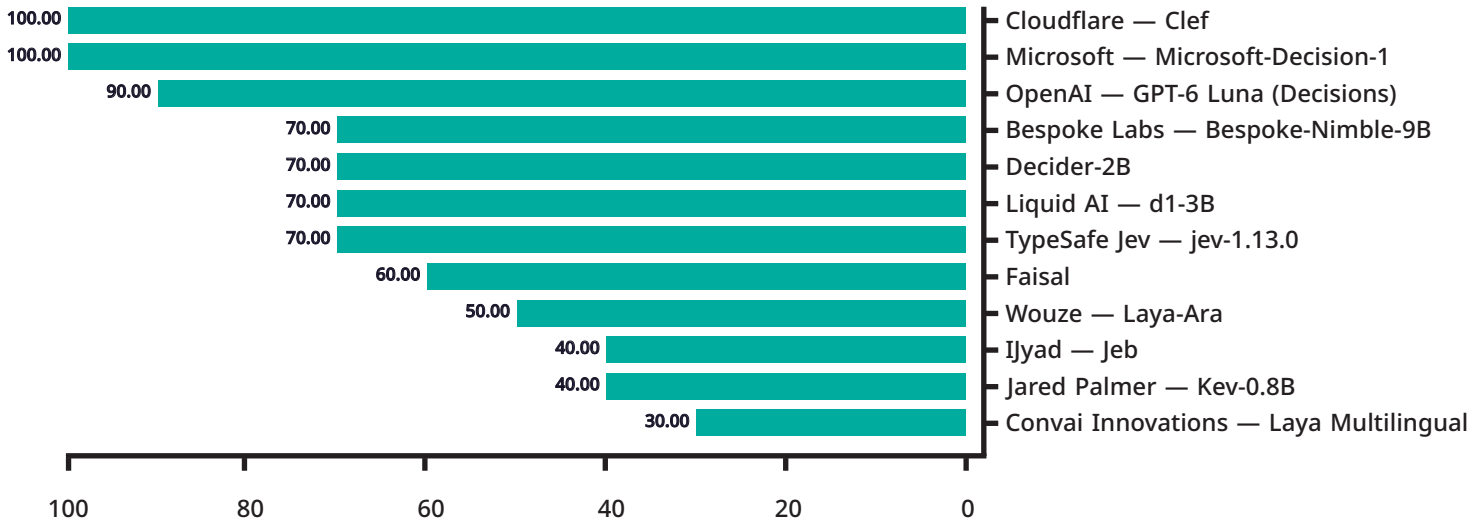
التضمين



المهام الفرعية:

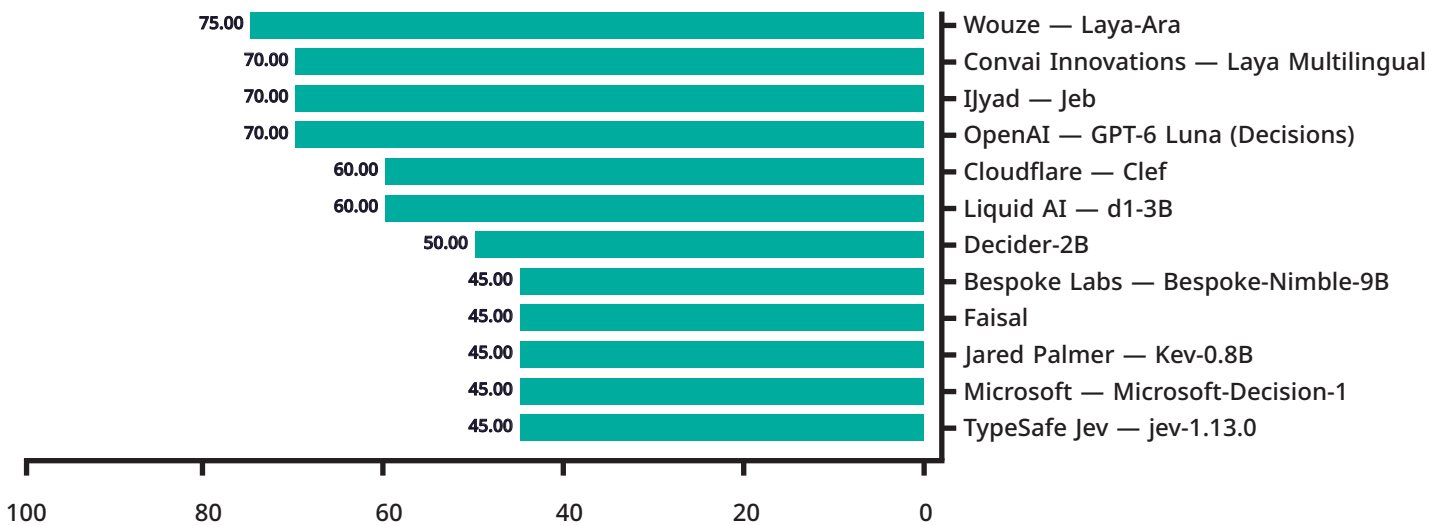
٢.١ التشابه الدلالي (Semantic Similarity)

التضمين - التشابه الدلالي

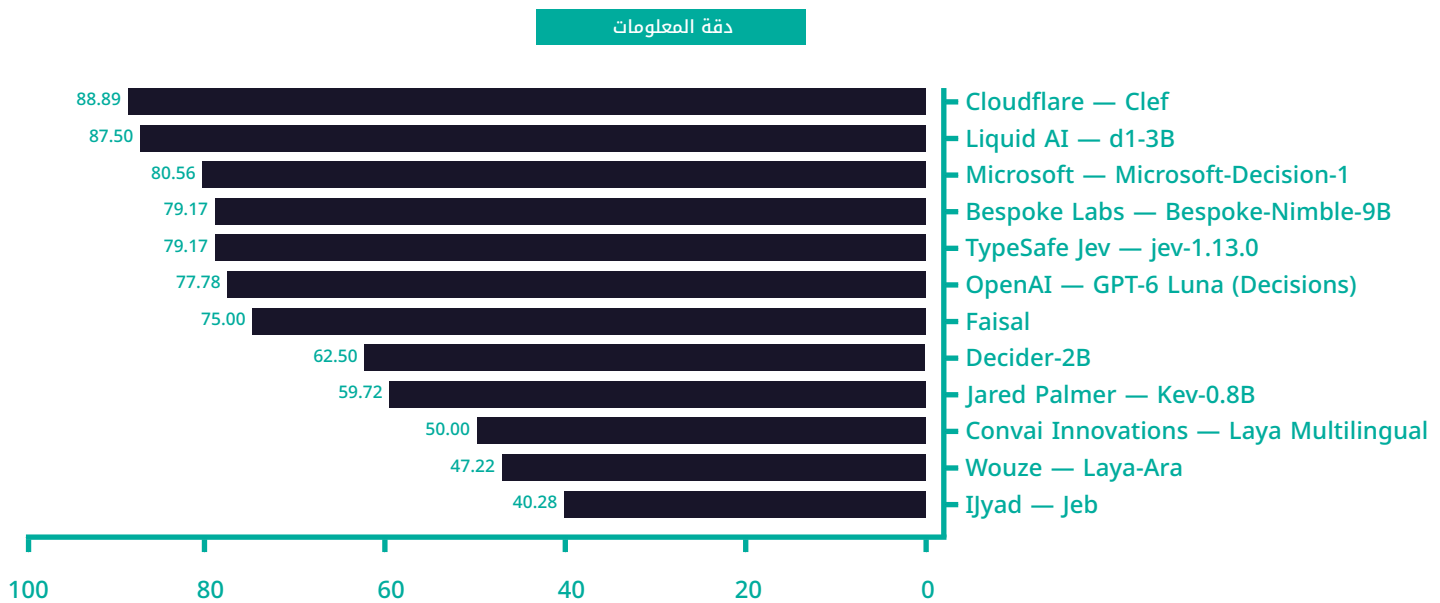


٢.٢ الاستلزام النصي (Textual Entailment)

التضمين - الاستلزام النصي

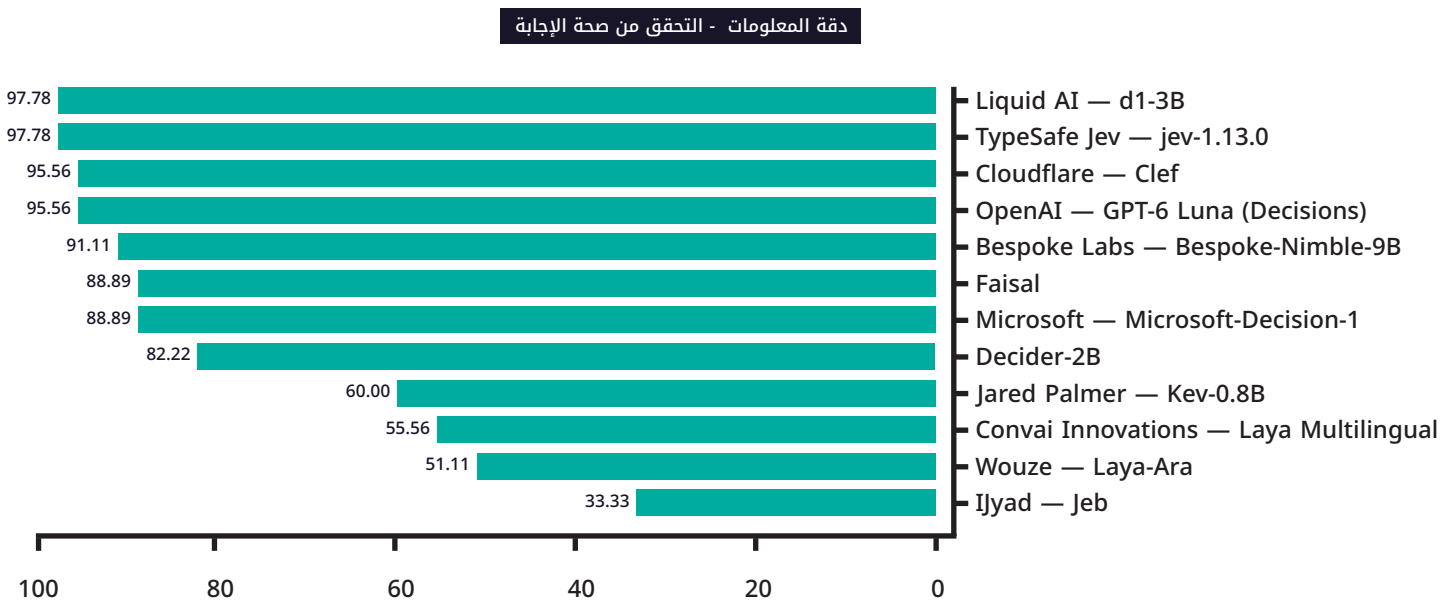


٣. دقة المعلومات (Factuality)



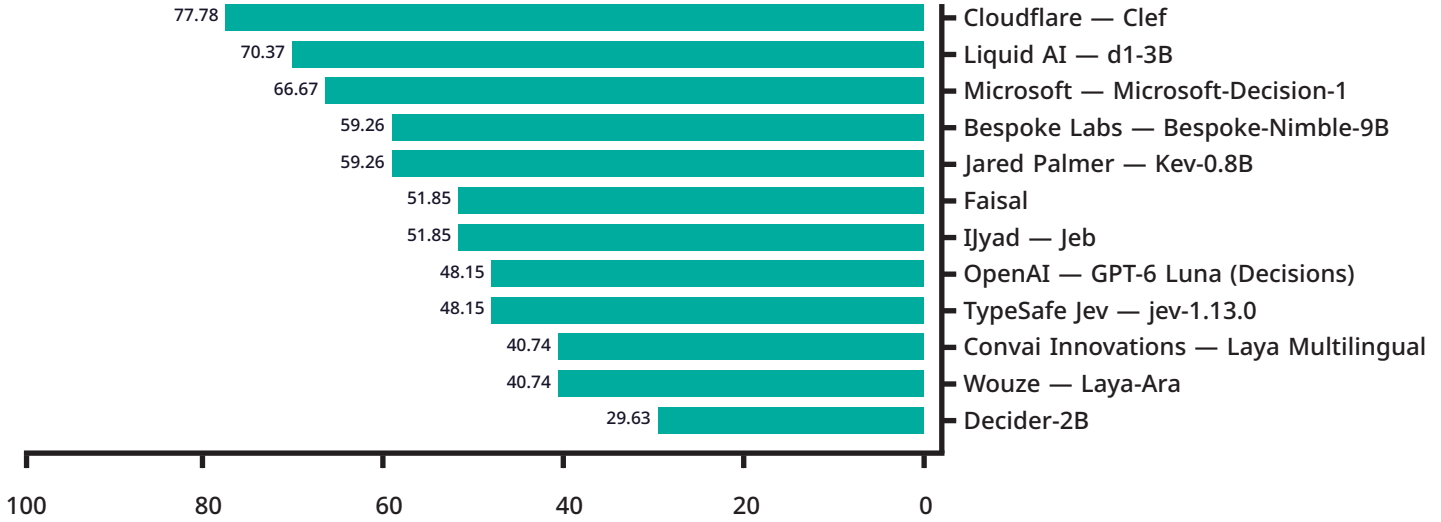
المهام الفرعية:

٣.١ التحقق من صحة الإجابة (Answer Verification)



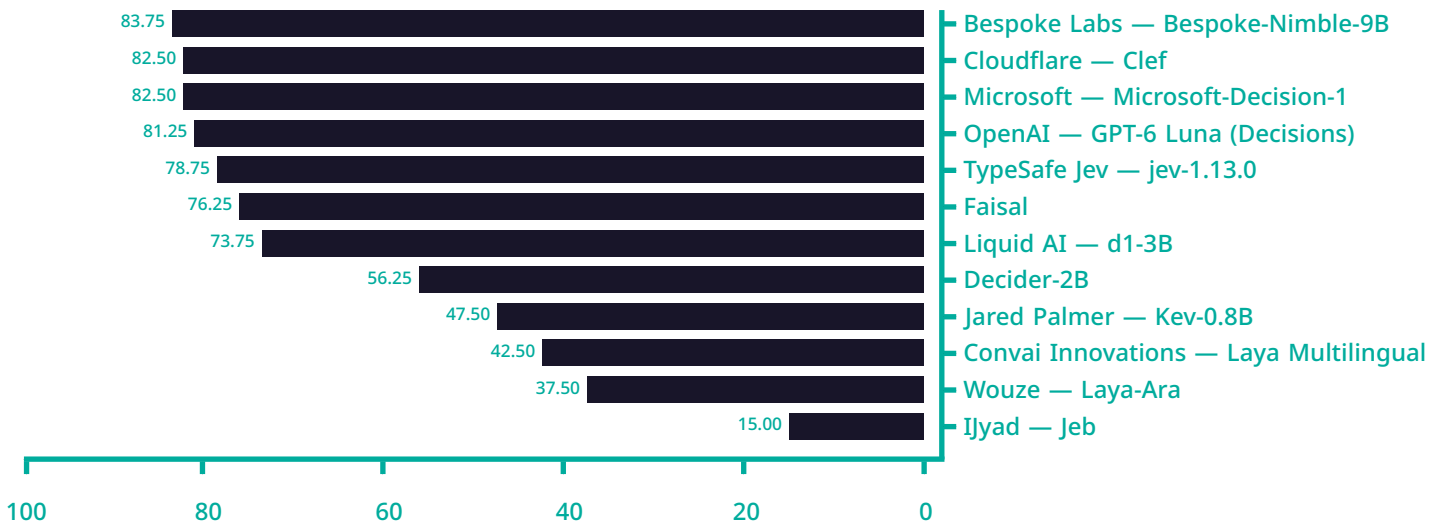
٣.٢ التحقق من صحة الادعاء (Claim Verification)

دقة المعلومات - التحقق من صحة الادعاء

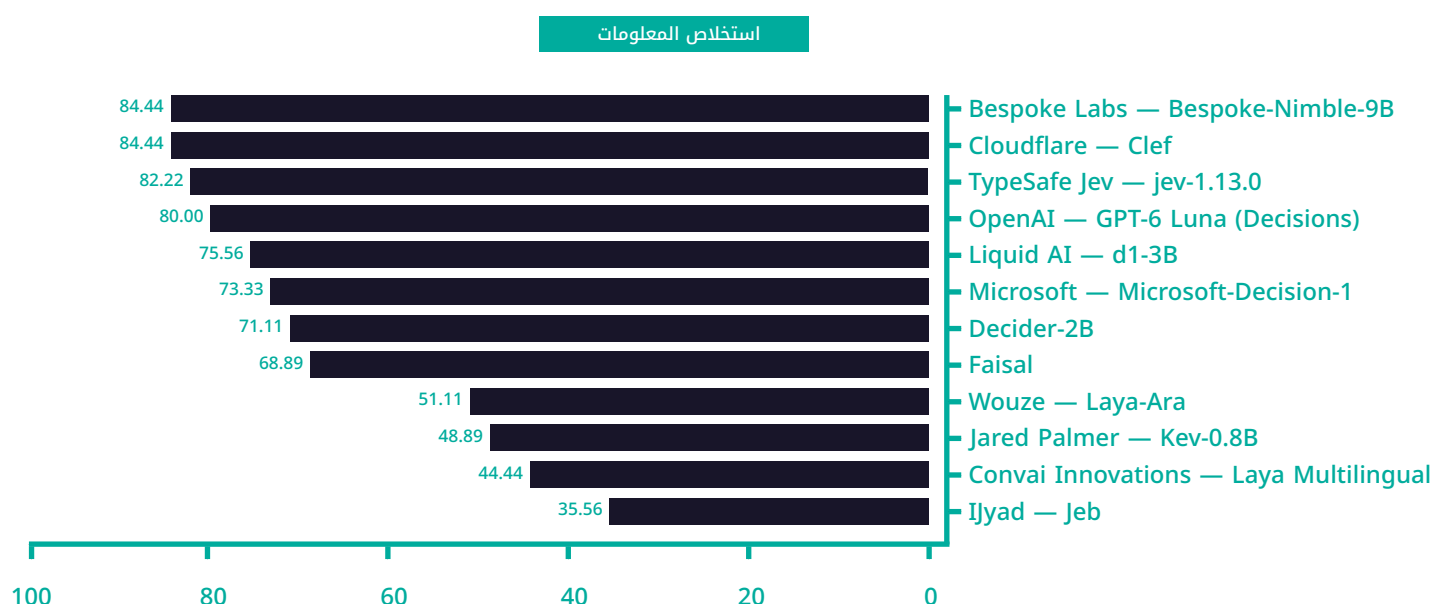


٤. ملء الفراغ (Fill in the Blank)

ملء الفراغ

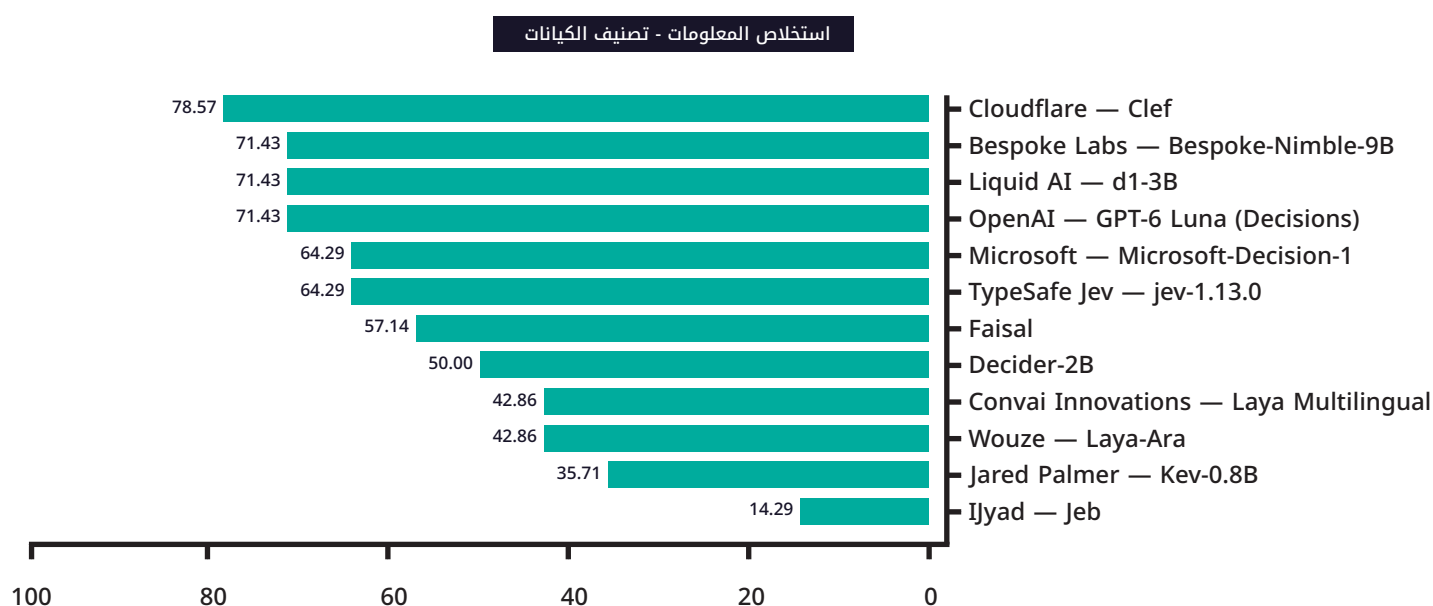


0. استخلاص المعلومات (Information Extraction)



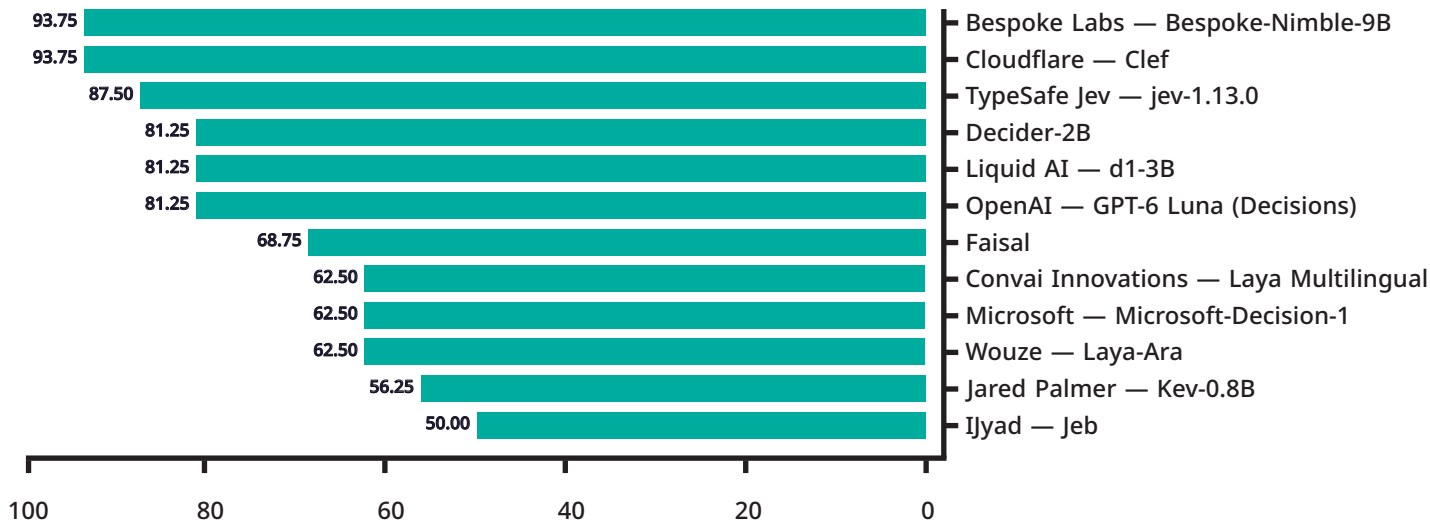
المهام الفرعية:

0.1 تصنيف الكيانات (Entity Categorization)



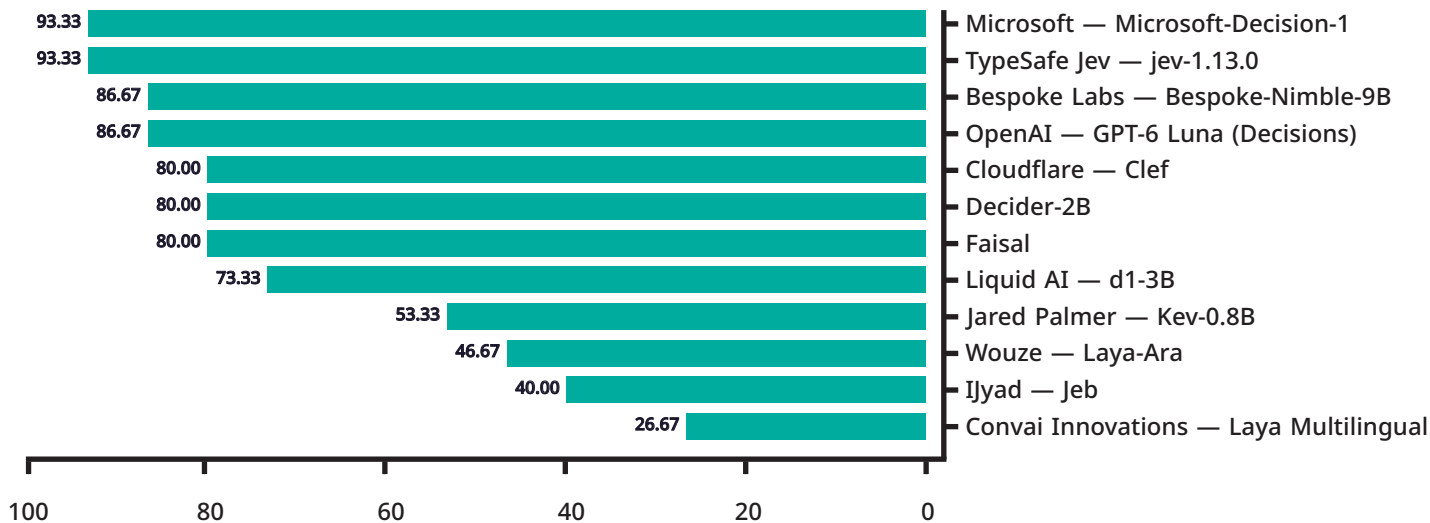
0.٢ تصنيف العلاقات بين الكيانات (Entity Relation Classification)

استخلاص المعلومات - تصنيف العلاقات بين الكيانات

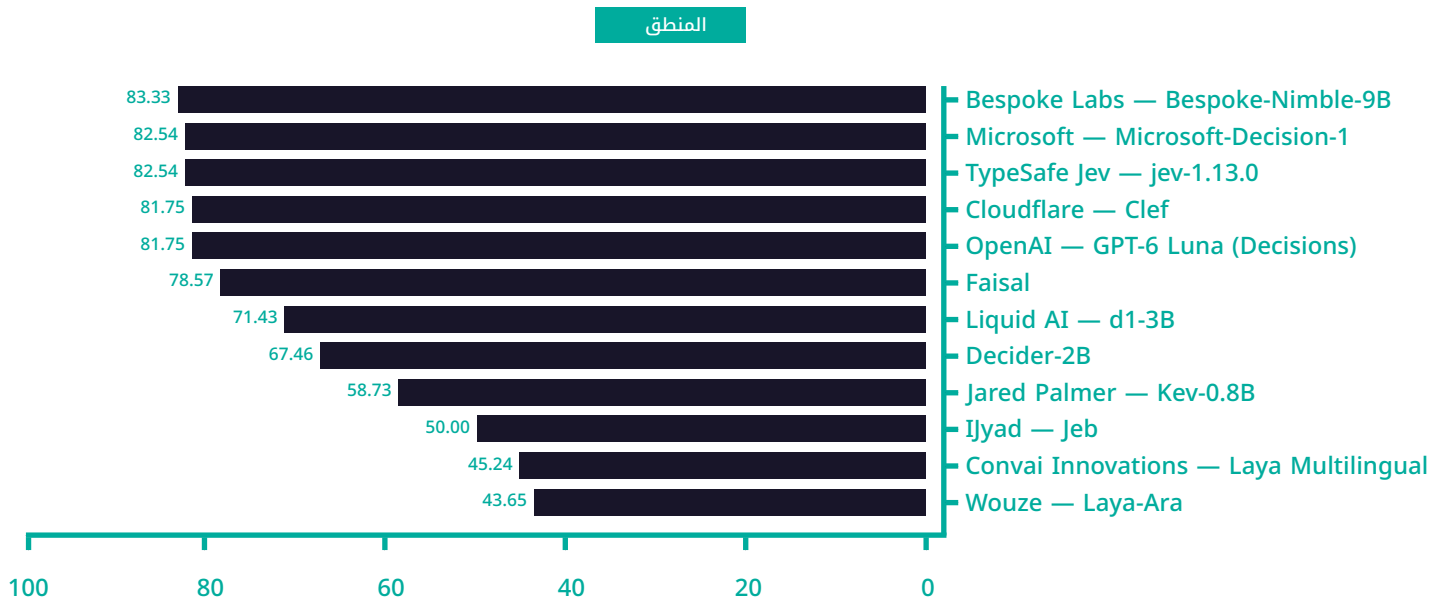


0.٣ استخلاص المعلومات المطلوبة (Extracting Required Information)

استخلاص المعلومات - استخلاص المعلومات المطلوبة

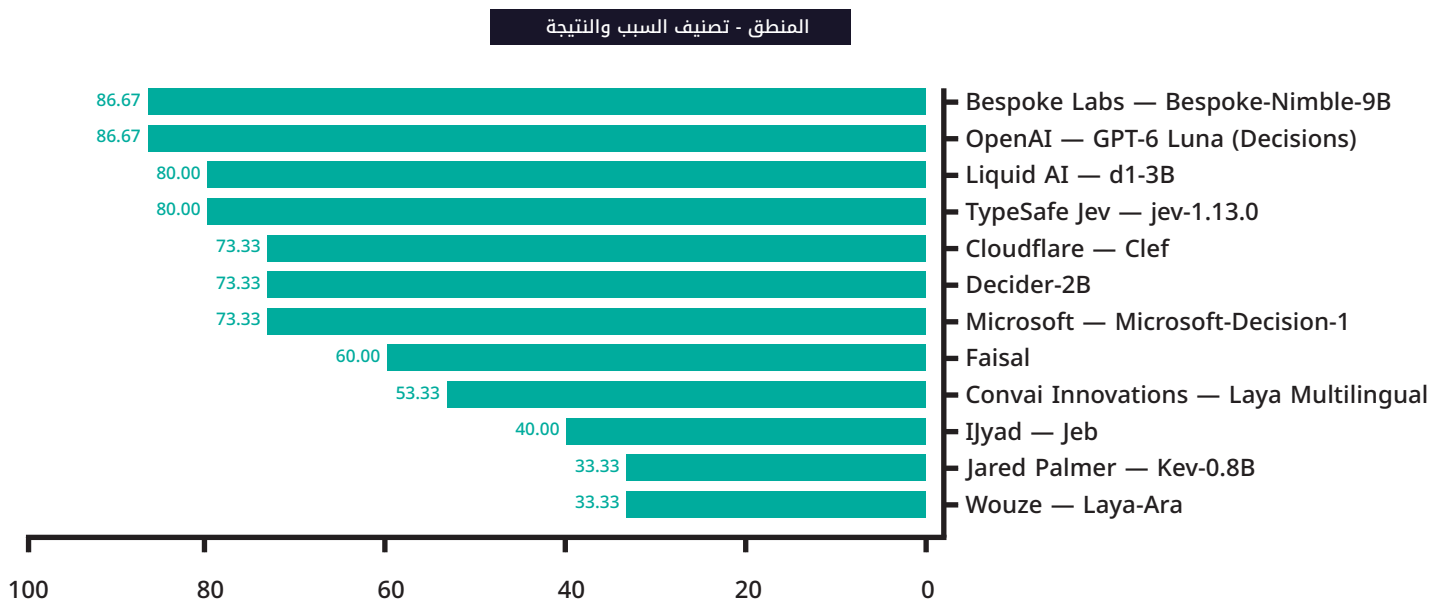


٦. المنطق (Logic)

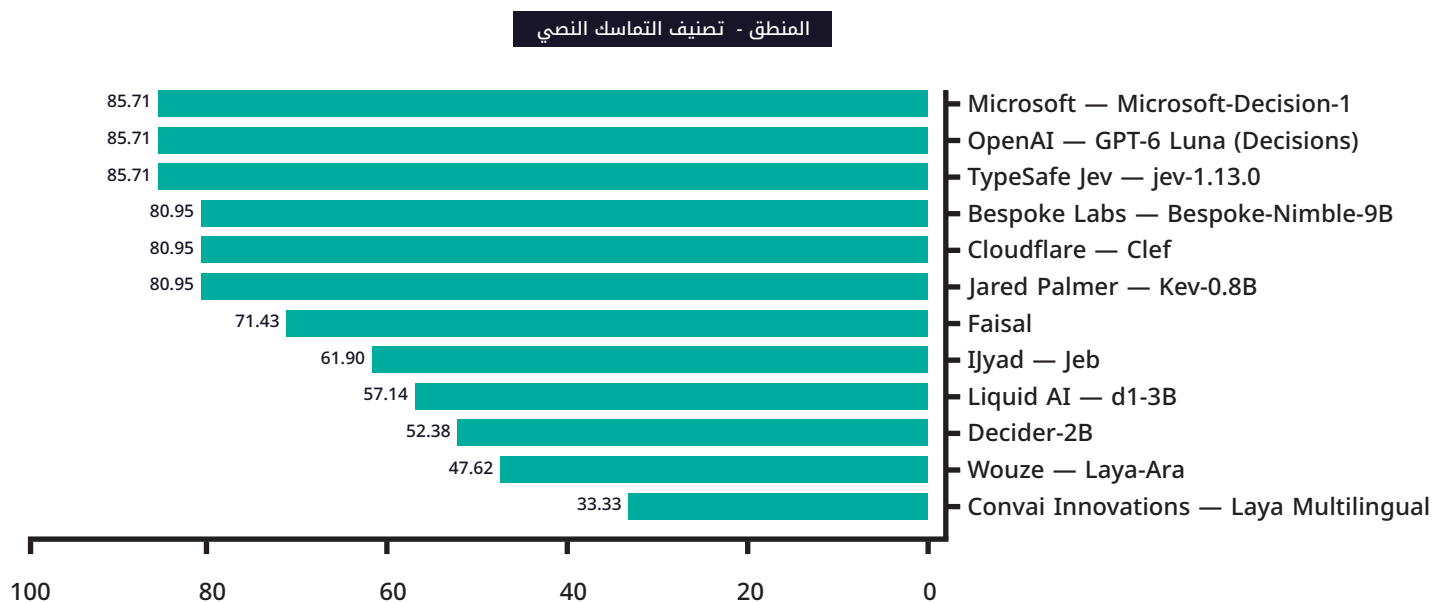


المهام الفرعية:

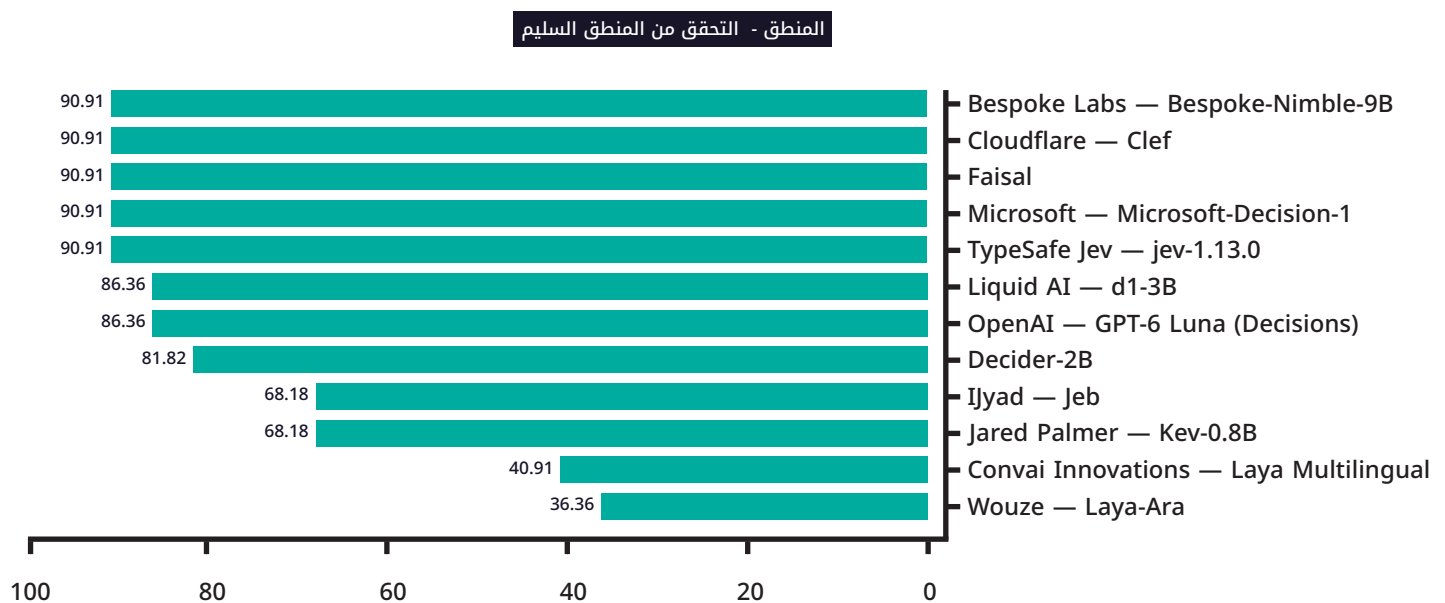
٦.١ تصنيف السبب والنتيجة (Cause Effect Classification)



٦.٢ تصنيف التماسك النصي (Coherence Classification)

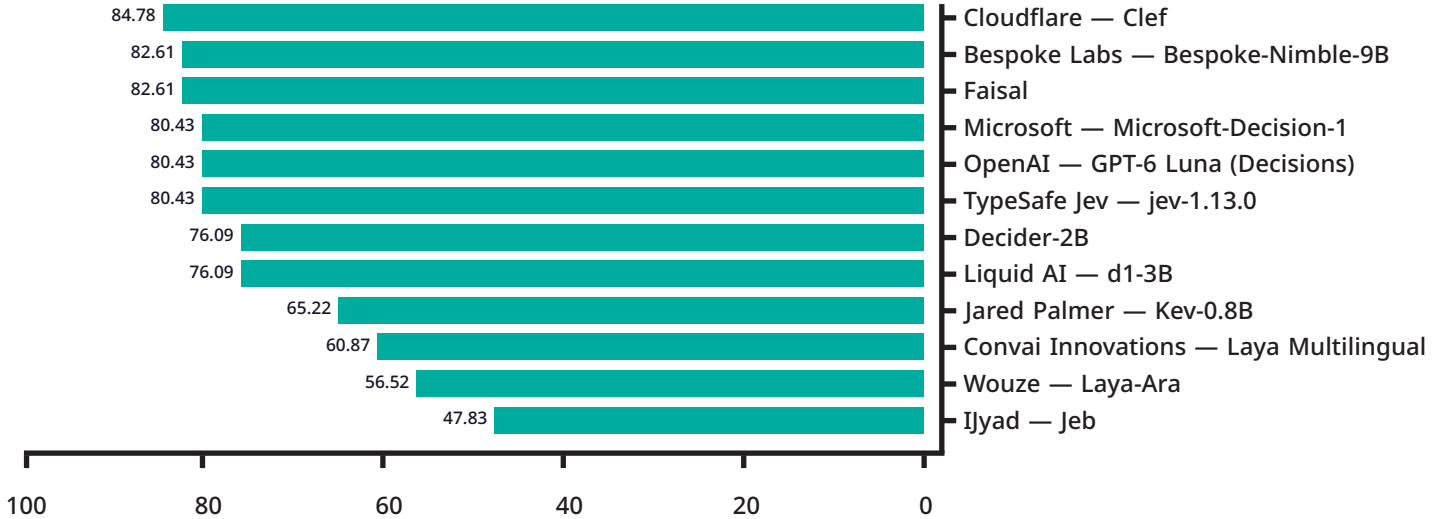


٦.٣ التحقق من المنطق السليم (Commonsense Validation)



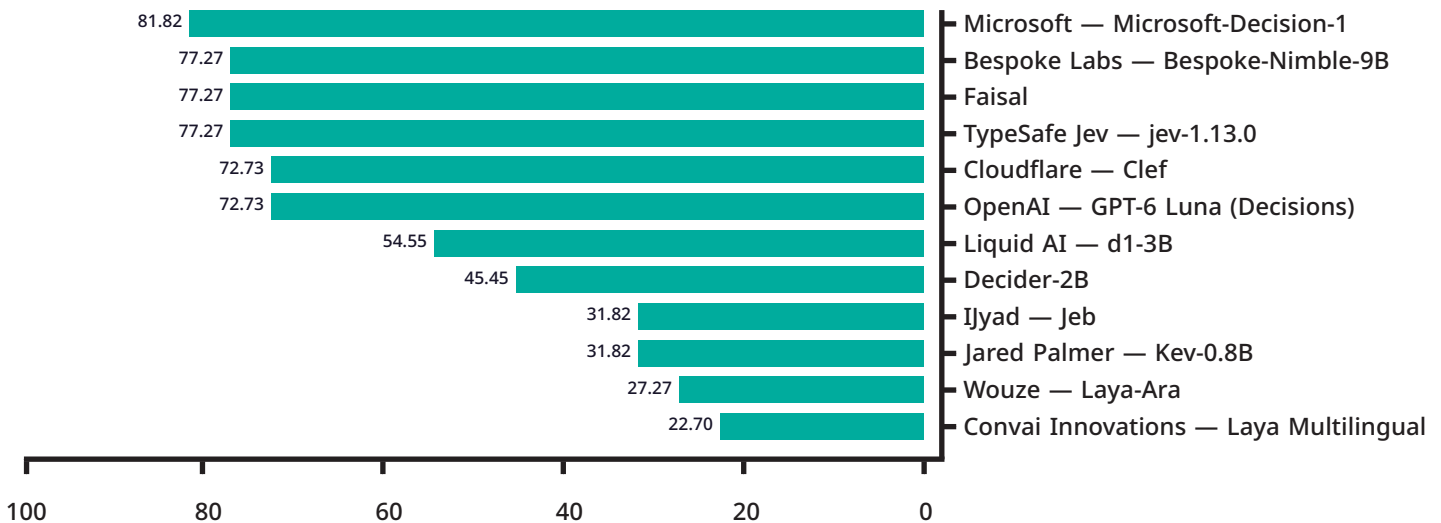
٦.٤ الاستدلال المنطقي (Logical Reasoning)

المنطق - الاستدلال المنطقي

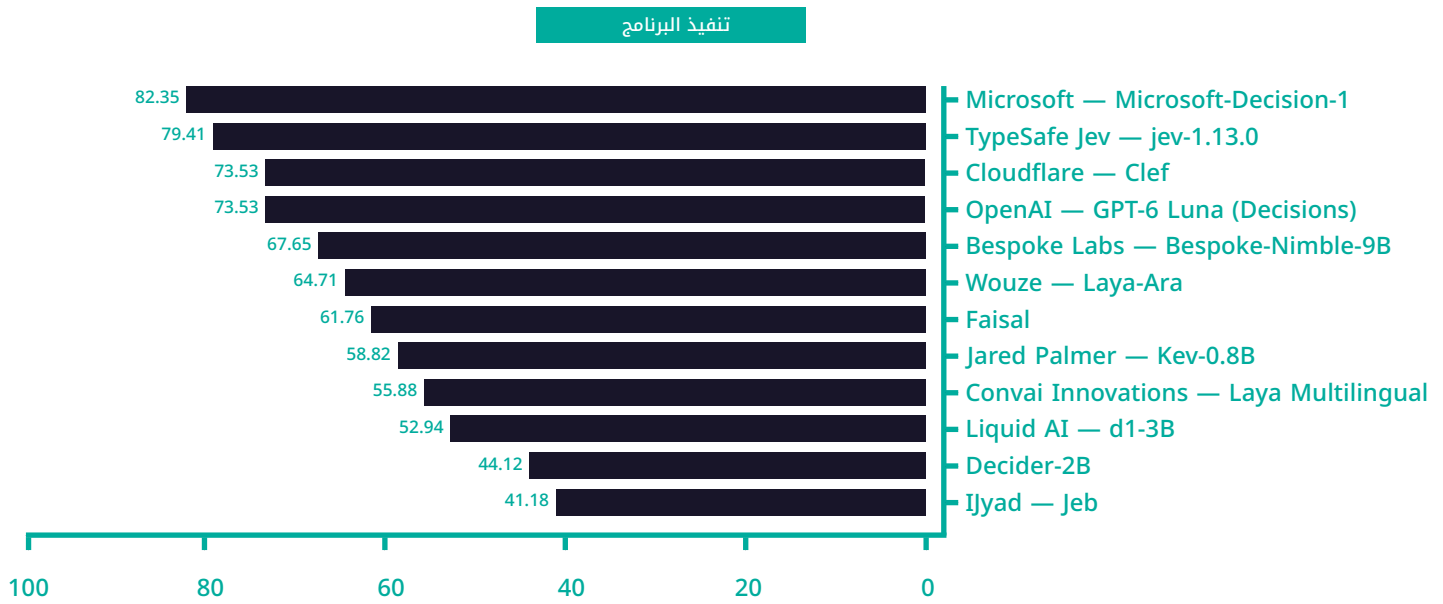


٦.٥ ترتيب الجمل (Sentence Ordering)

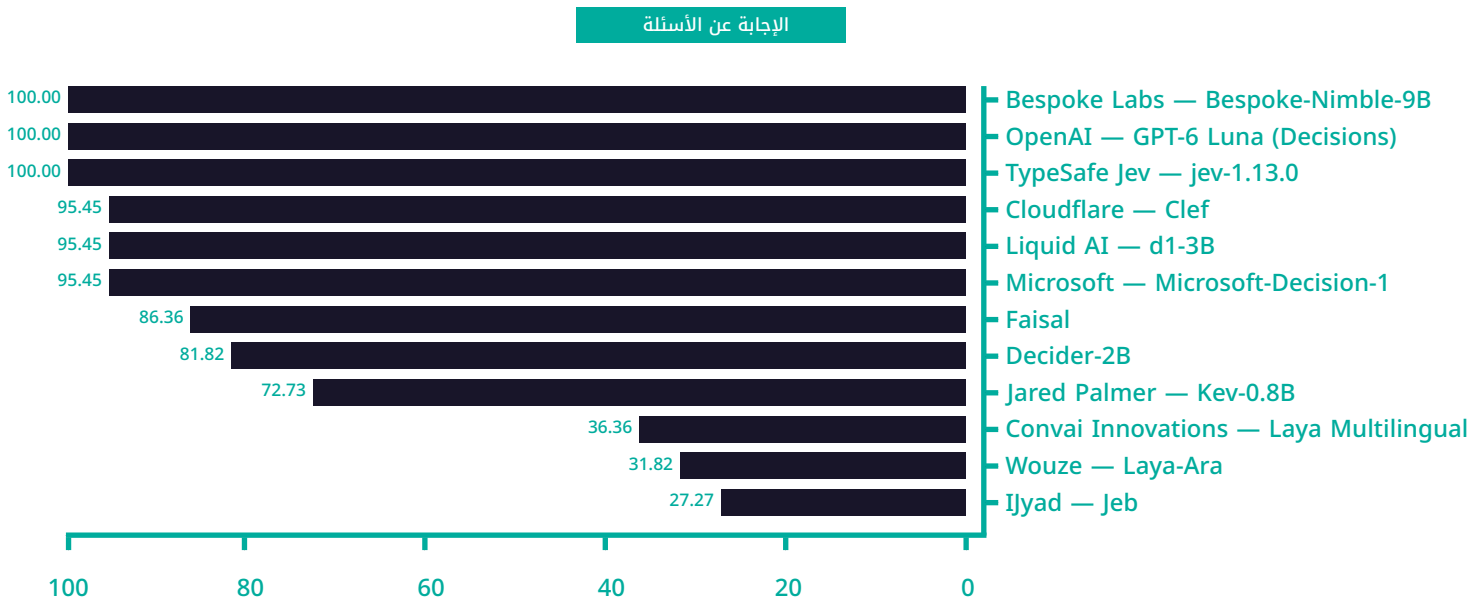
المنطق - ترتيب الجمل



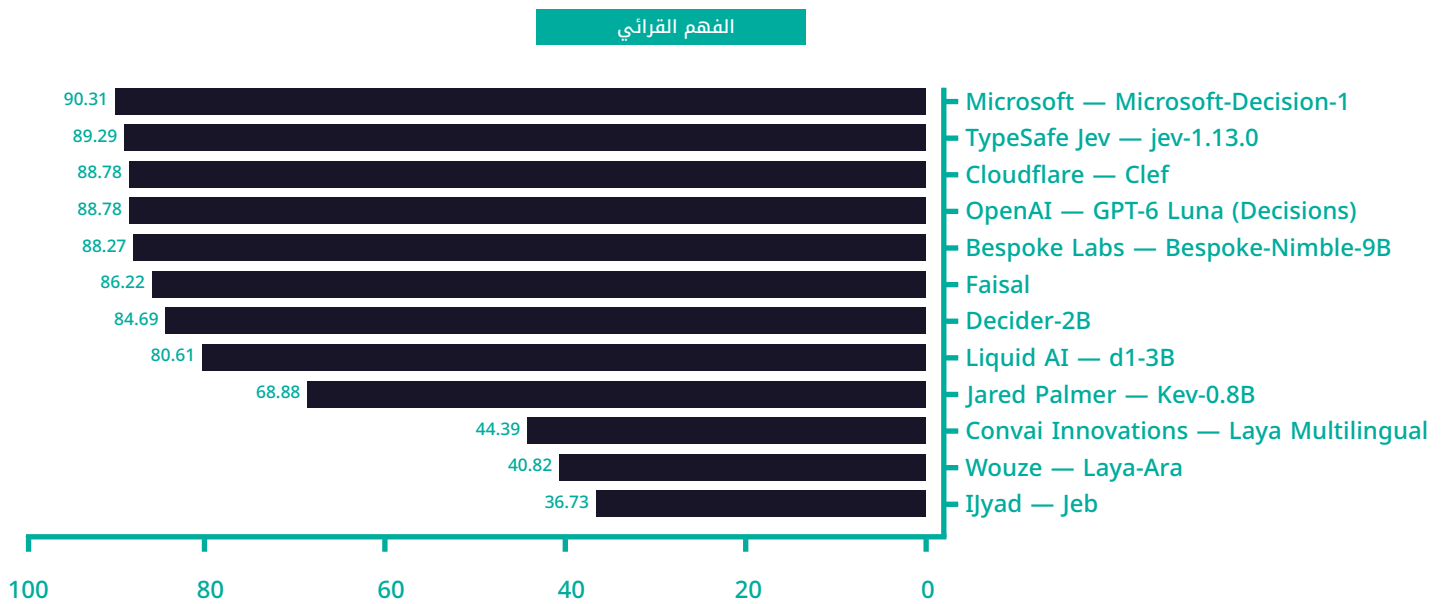
٧. تنفيذ البرنامج (Program Execution)



٨. الإجابة عن الأسئلة (Question Answering)

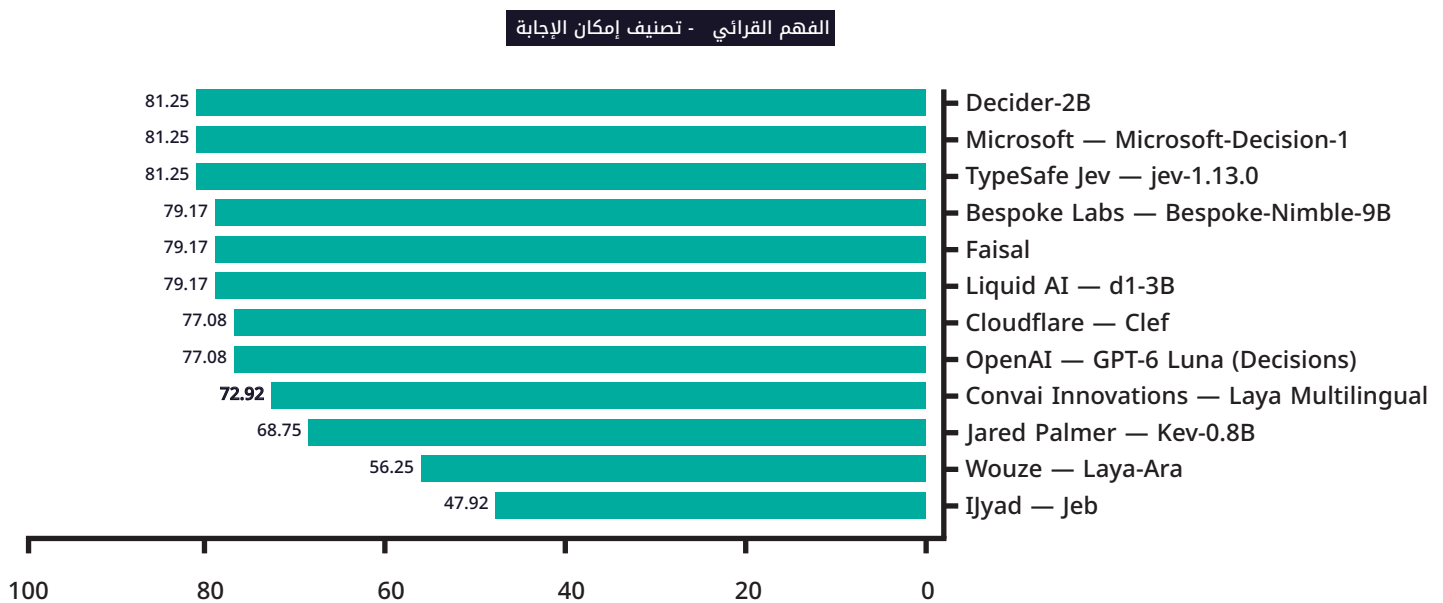


٩. الفهم القرائي (Reading Comprehension)



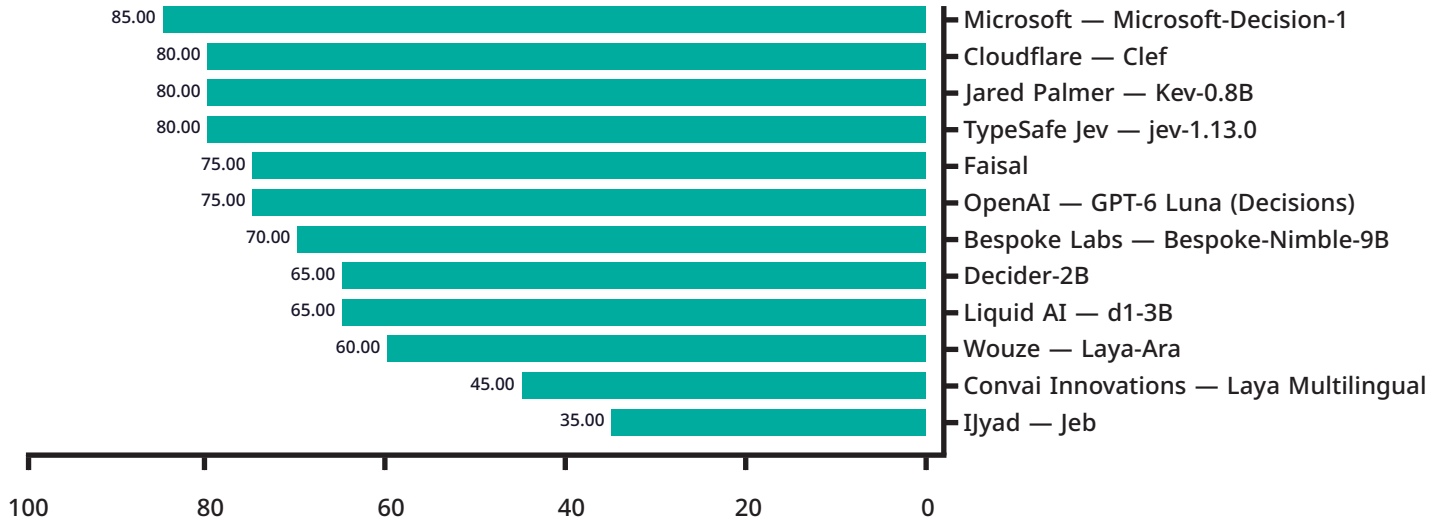
المهام الفرعية:

٩.١ تصنيف إمكان الإجابة (Answerability Classification)



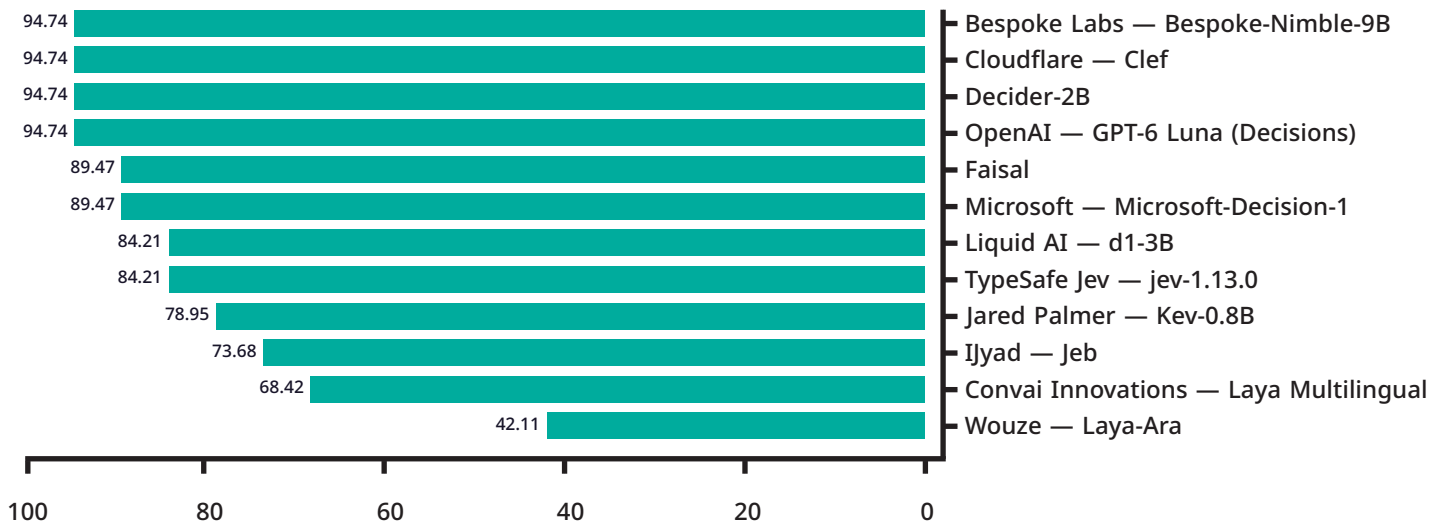
٩.٢ فهم السؤال (Question Understanding)

الفهم القرآني - فهم السؤال



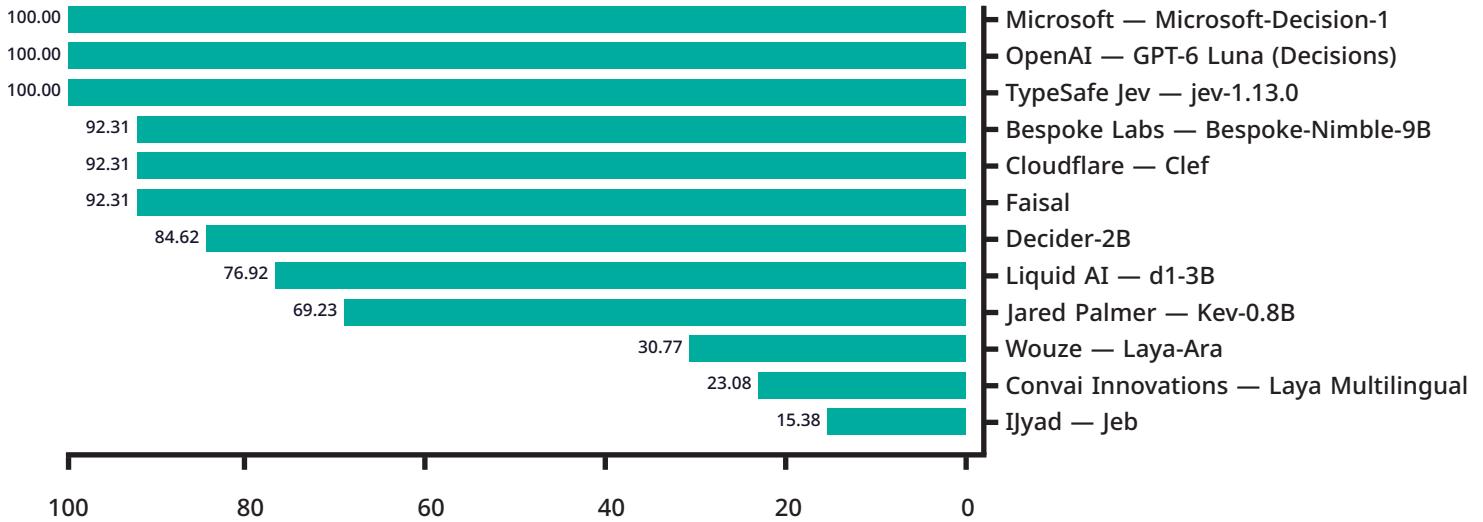
٩.٣ التحقق من الاستدلال النصي (Textual Inference Validation)

الفهم القرآني - التحقق من الاستدلال النصي



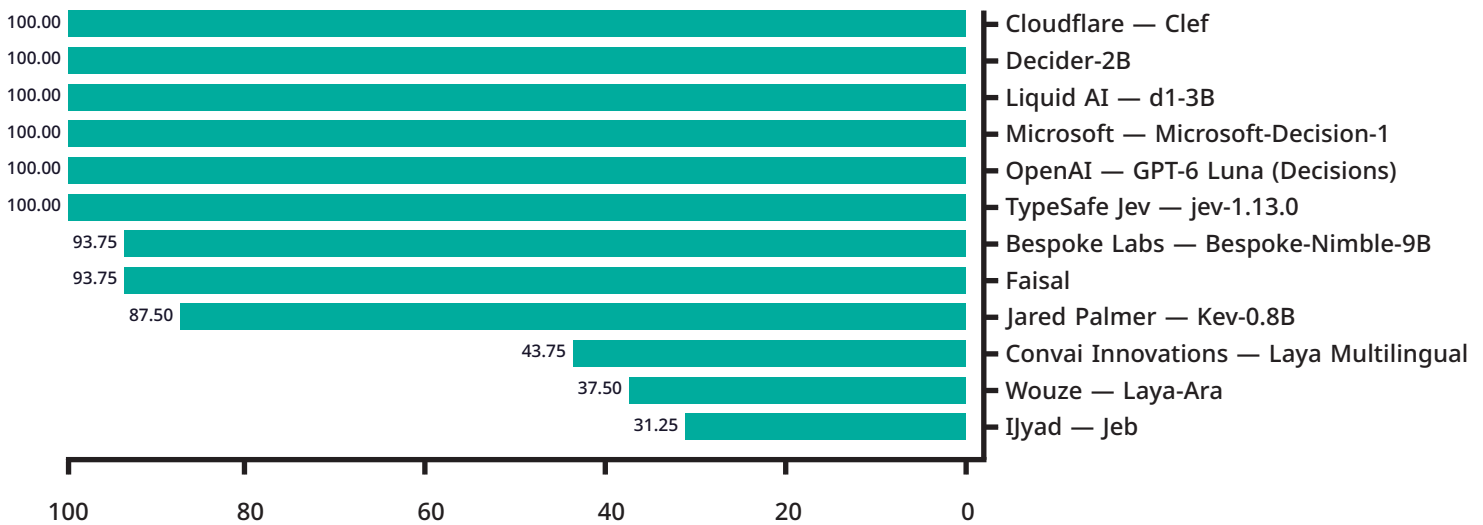
٩.٤ اختيار الحوار المناسب (Dialogue Selection)

الفهم القرائي - اختيار الحوار



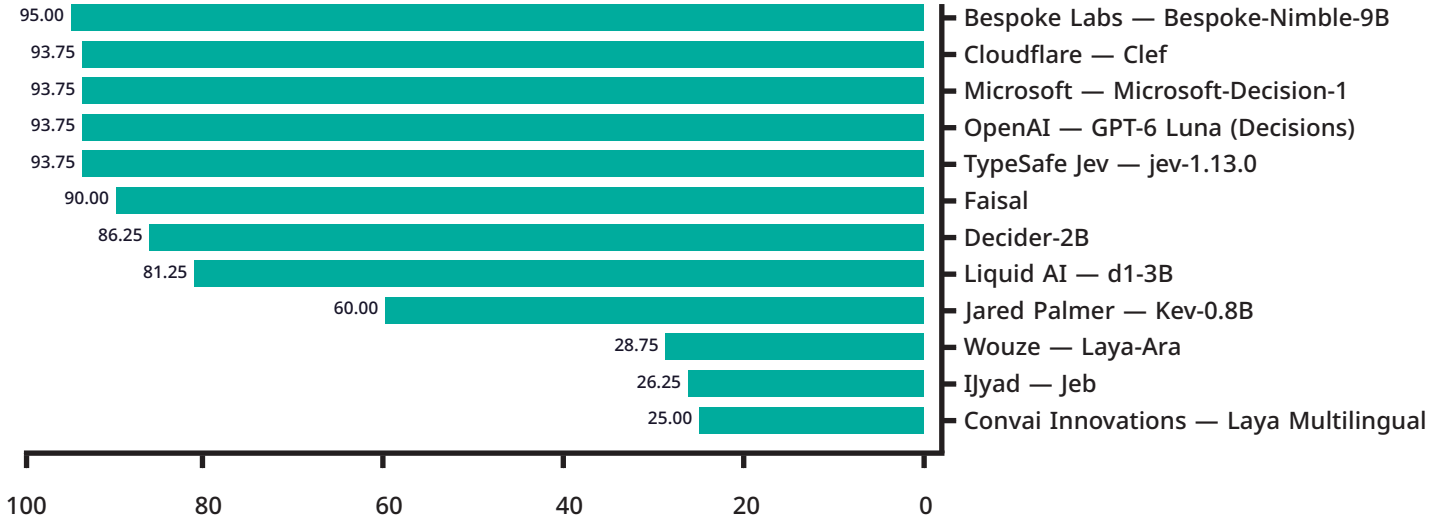
٩.٥ اختيار التفسير المناسب (Explanation Selection)

الفهم القرائي - اختيار التفسير



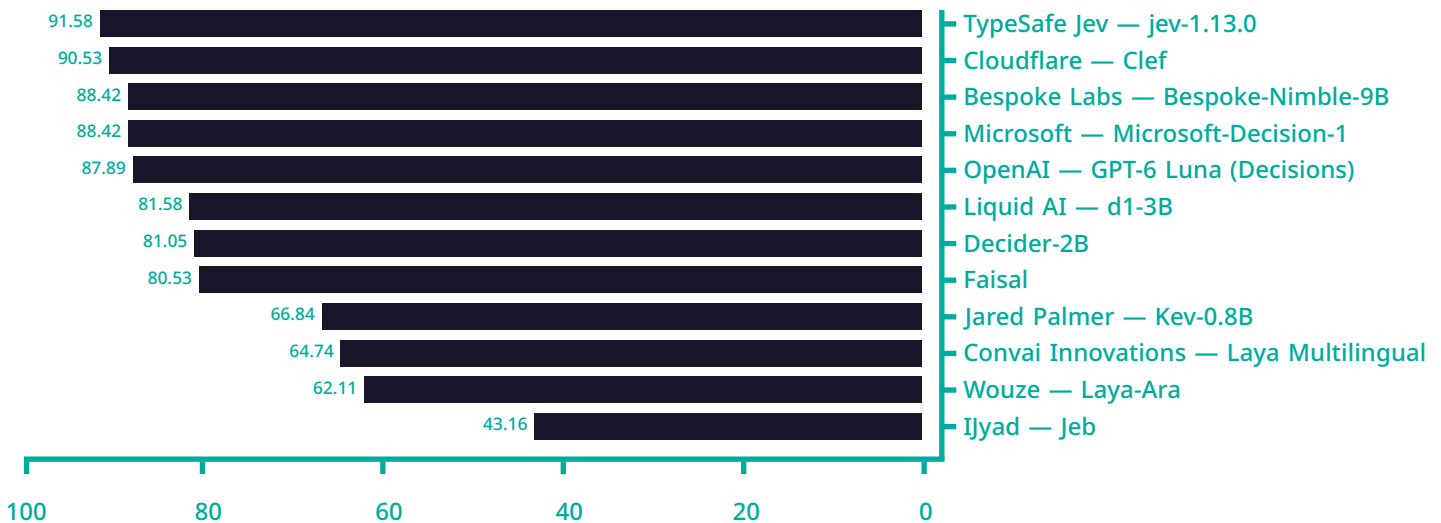
٩.٦ تحديد القسم الكلامي (Part of speech tagging)

الفهم القرائي - تحديد القسم الكلامي



١٠. تصنيف النصوص (Text Classification)

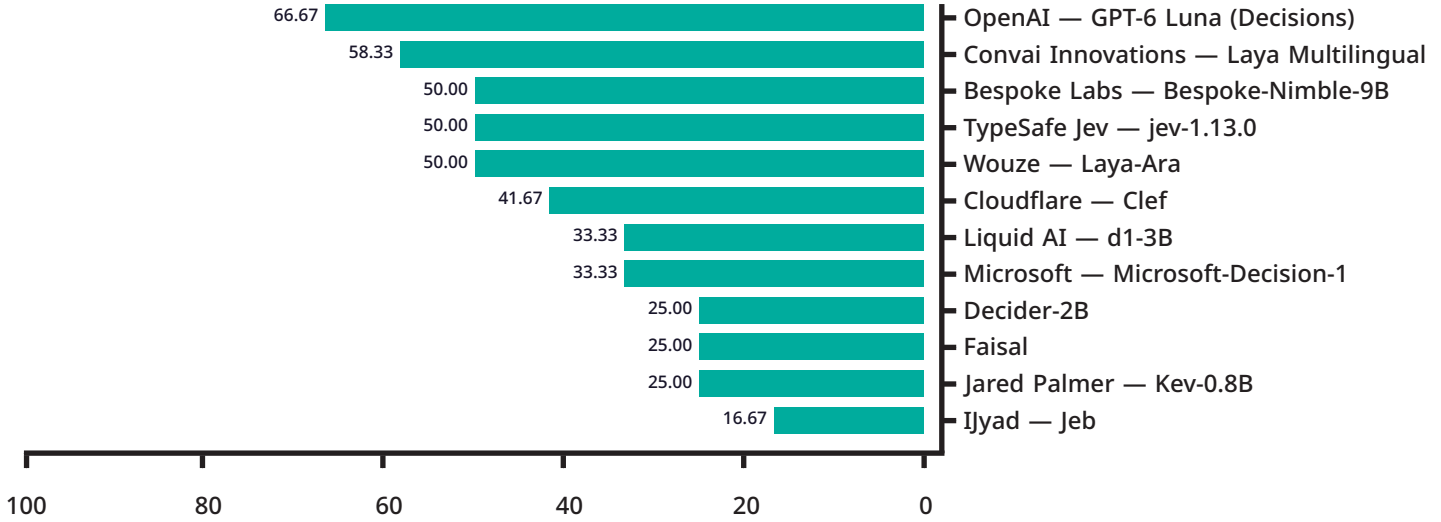
تصنيف النصوص



المهام الفرعية:

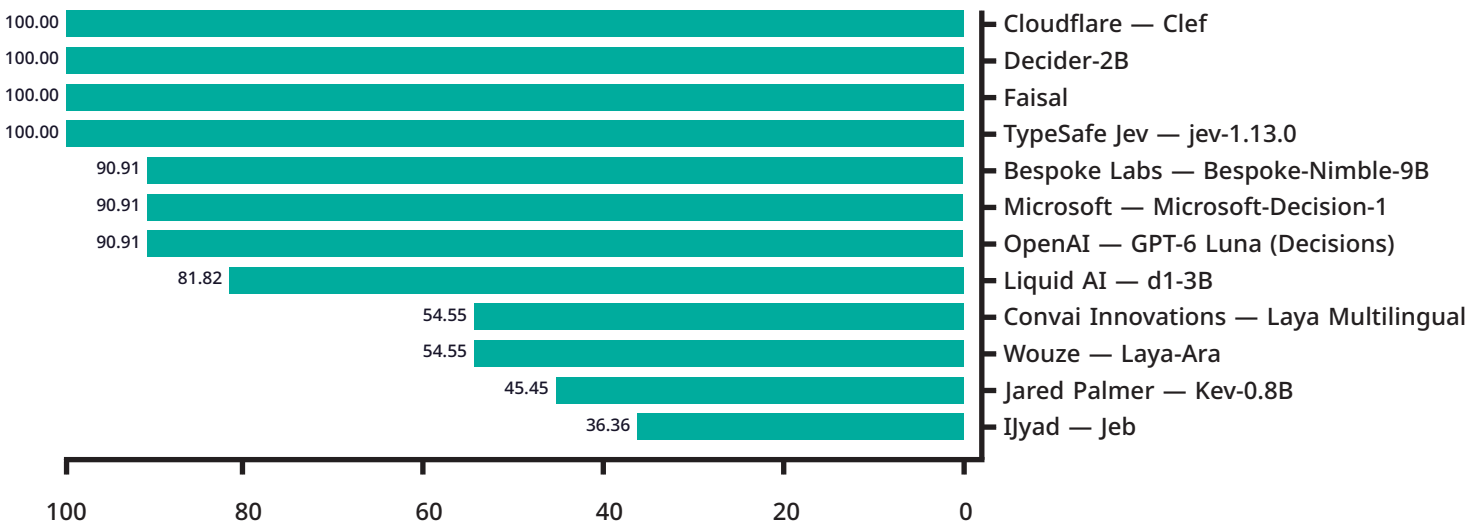
١٠.١ تحديد اللهجة (Dialect Identification)

تصنيف النصوص - تحديد اللهجة



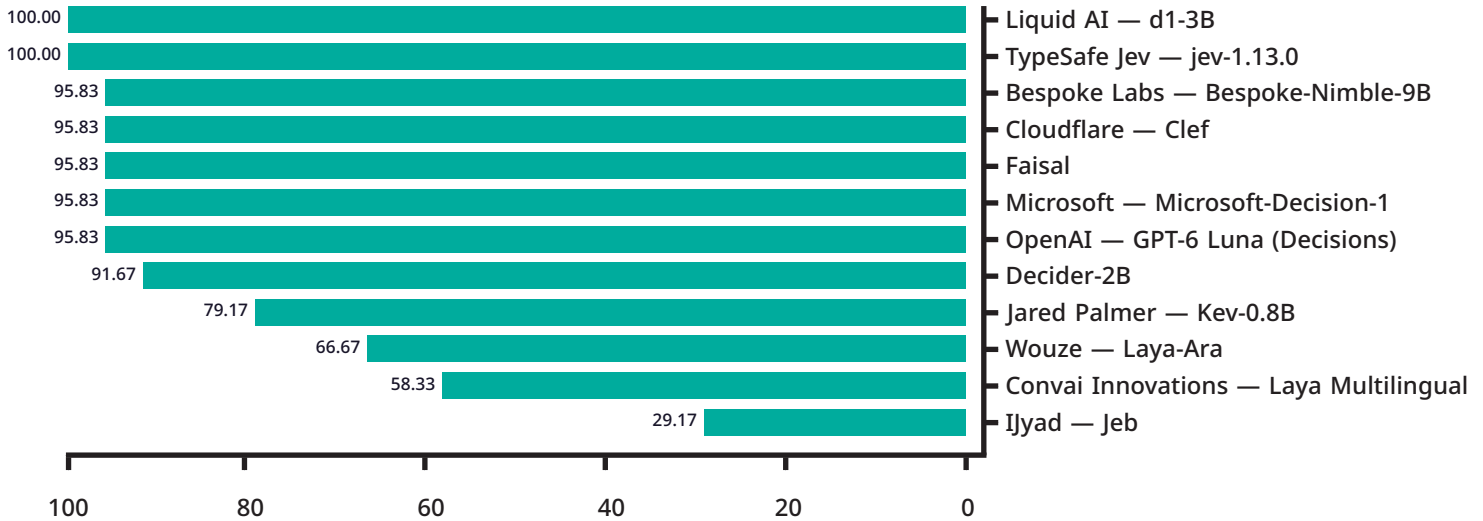
١٠.٢ تعرّف أفعال الحوار (Dialogue Act Recognition)

تصنيف النصوص - تعرّف أفعال الحوار



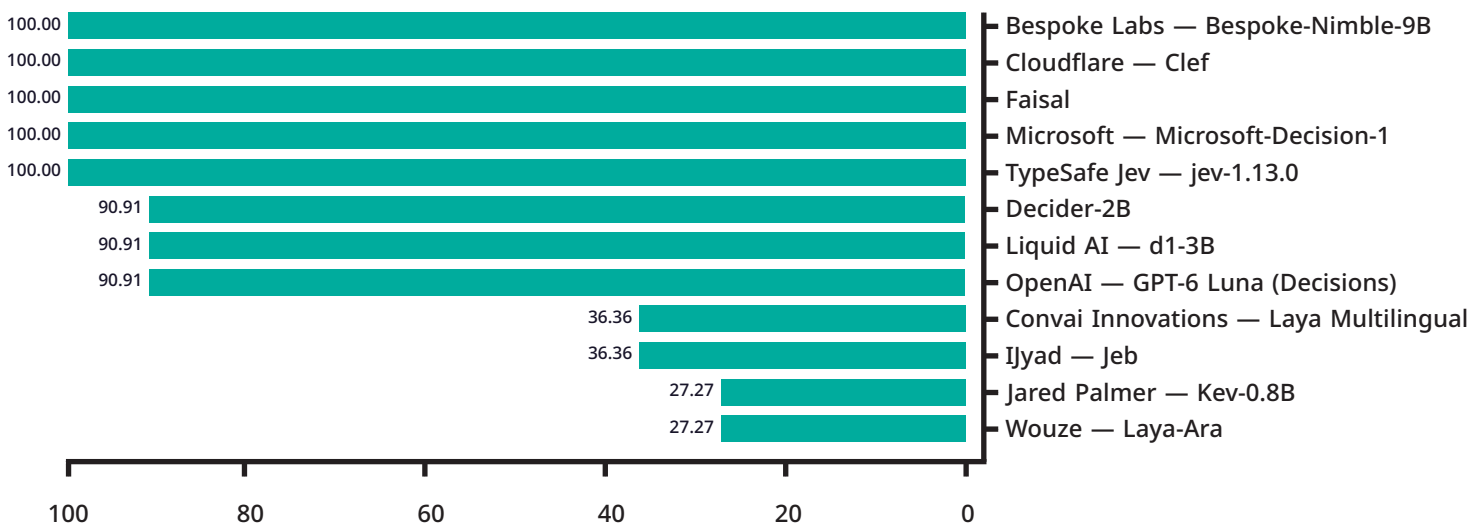
١٠.٣ اكتشاف المشاعر (Emotion Detection)

تصنيف النصوص - اكتشاف المشاعر



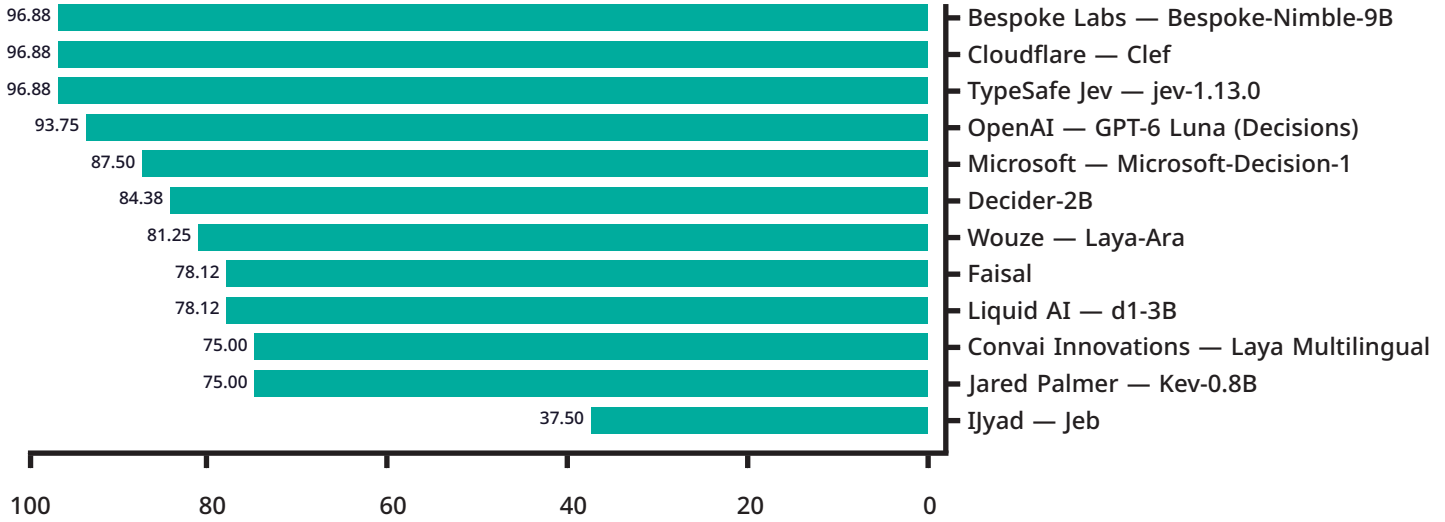
١٠.٤ تصنيف السلوكيات الأخلاقية (Ethics Classification)

تصنيف النصوص - تصنيف السلوكيات الأخلاقية



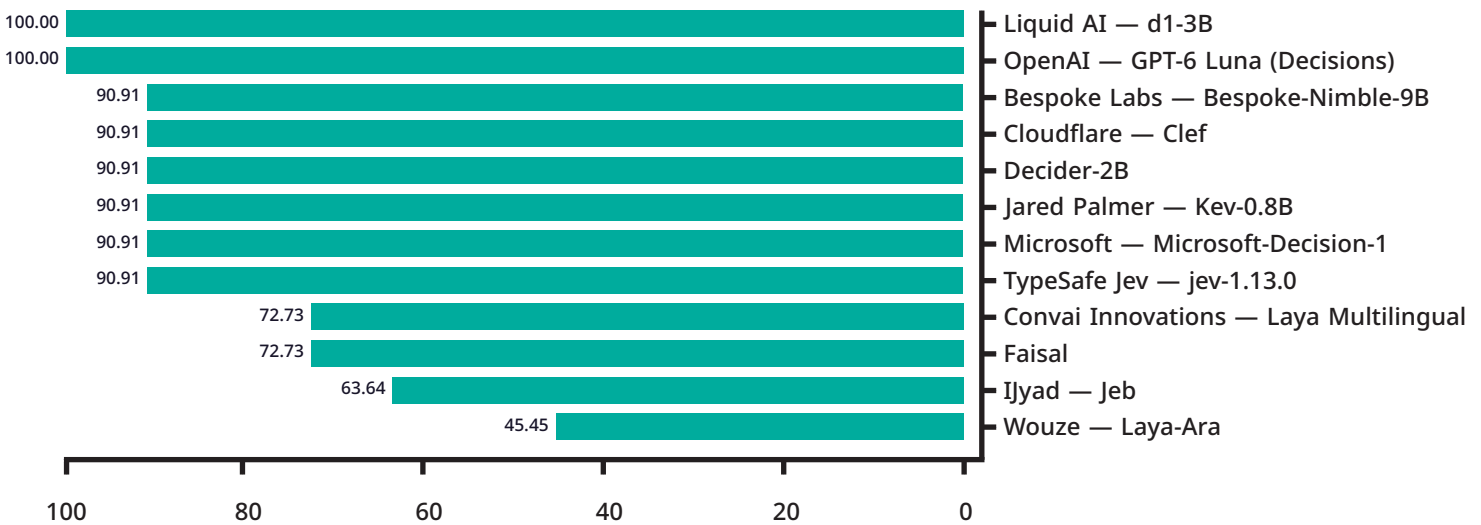
١٠.٥ اكتشاف الإساءة (Offensive Language Detection)

تصنيف النصوص - اكتشاف الإساءة



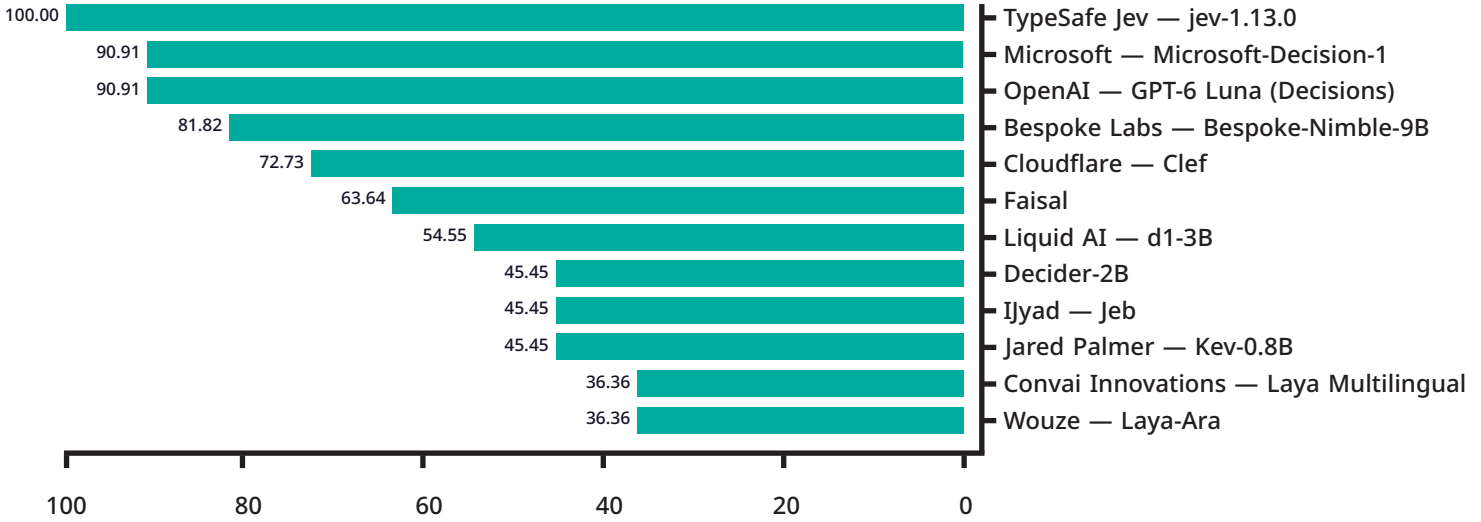
١٠.٦ تصنيف الأسئلة (Question Categorization)

تصنيف النصوص تصنيف الأسئلة



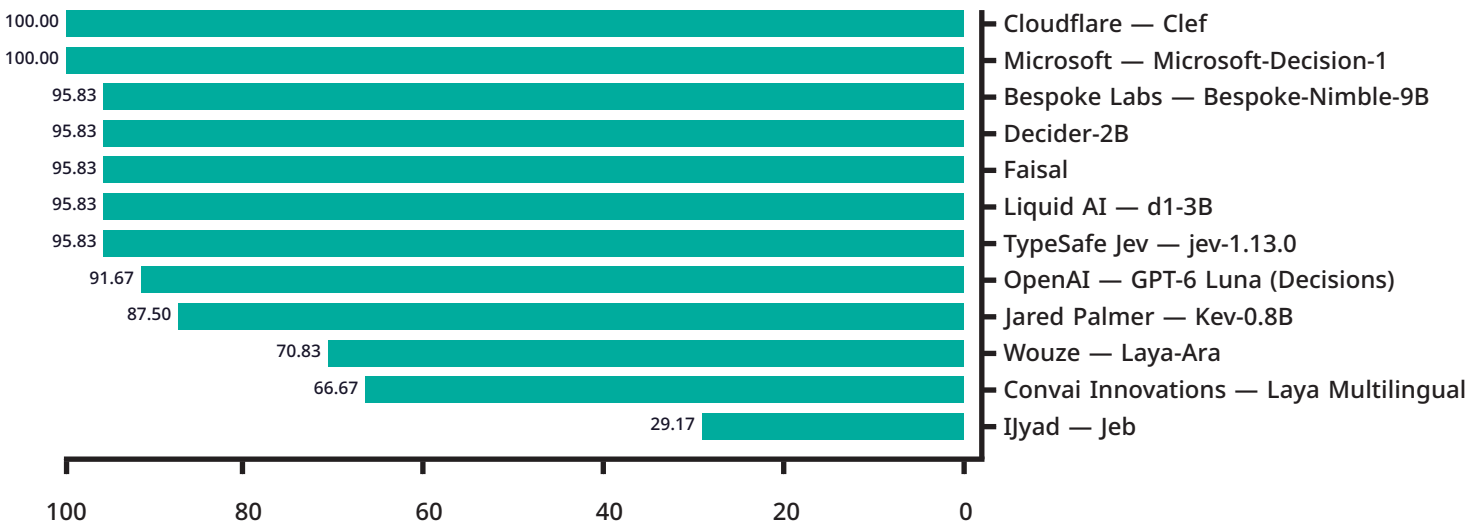
١٠.٧ اكتشاف السخرية (Sarcasm Detection)

تصنيف النصوص - اكتشاف السخرية



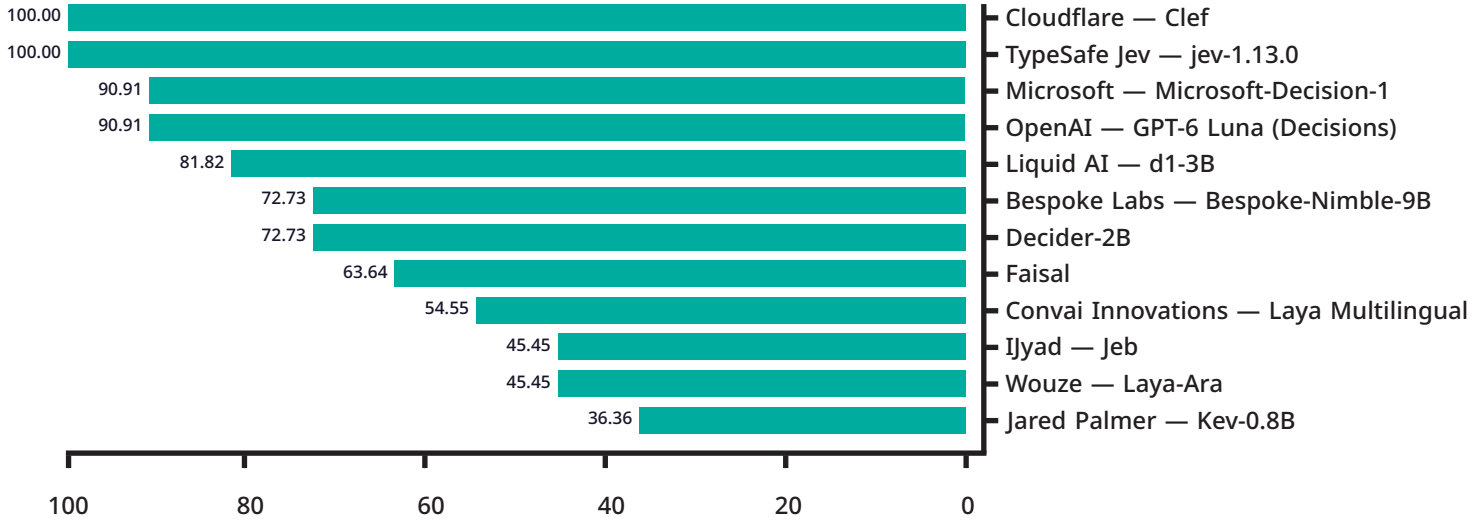
١٠.٨ تحليل المشاعر (Sentiment Analysis)

تصنيف النصوص - تحليل المشاعر



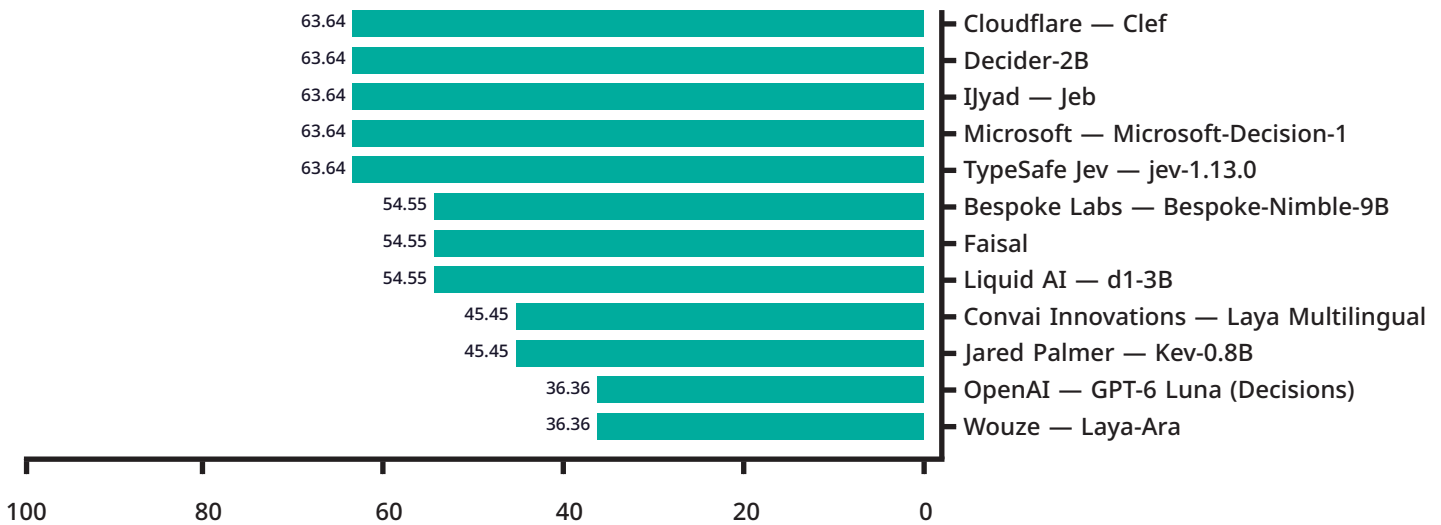
١٠.٩ كشف الرسائل المزعجة (Spam Detection)

تصنيف النصوص - كشف الرسائل المزعجة



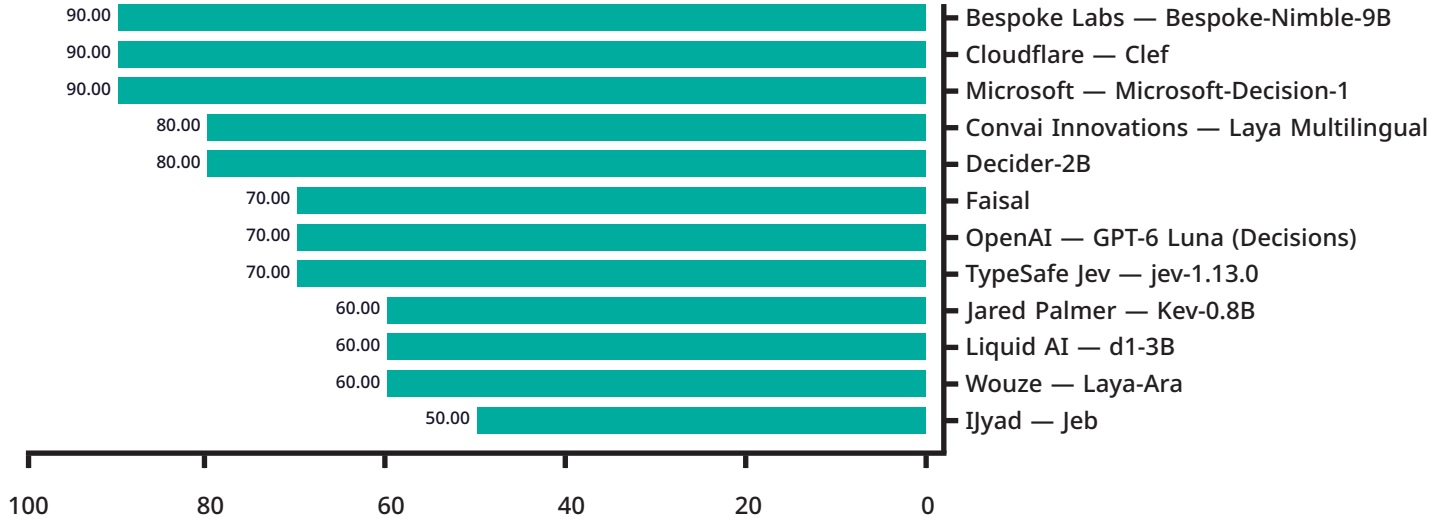
١٠.١٠ اكتشاف القوالب النمطية (Stereotype Detection)

تصنيف النصوص - اكتشاف القوالب النمطية



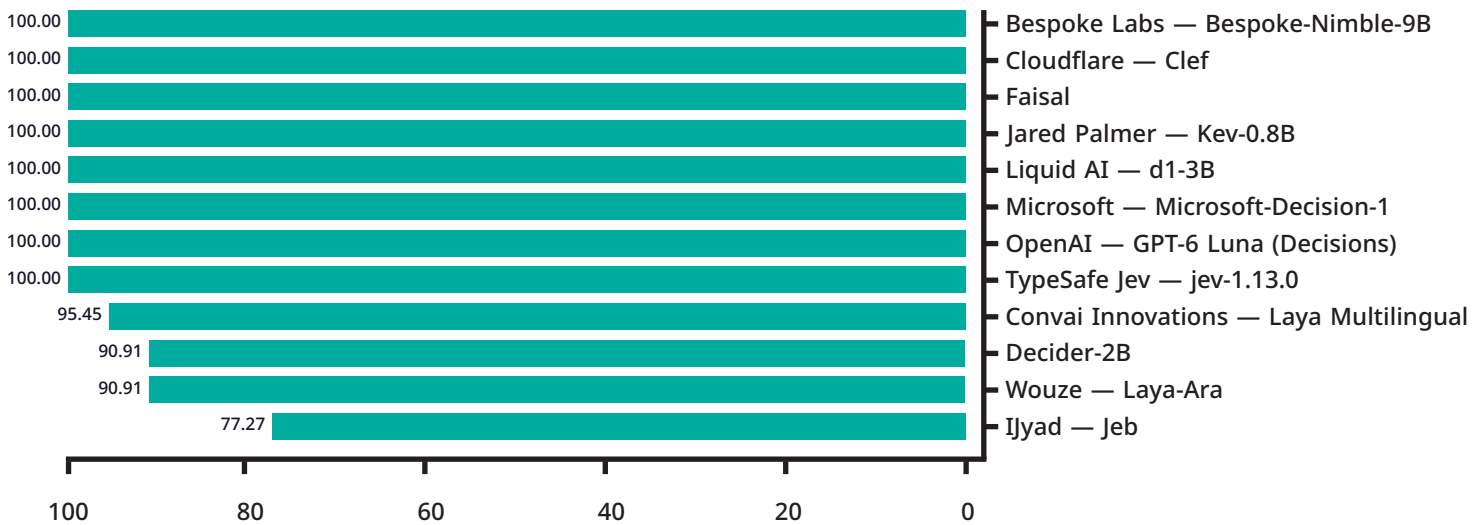
١٠.١١ التصنيف الفئوي للنصوص (Text Categorization)

تصنيف النصوص - التصنيف الفئوي للنصوص



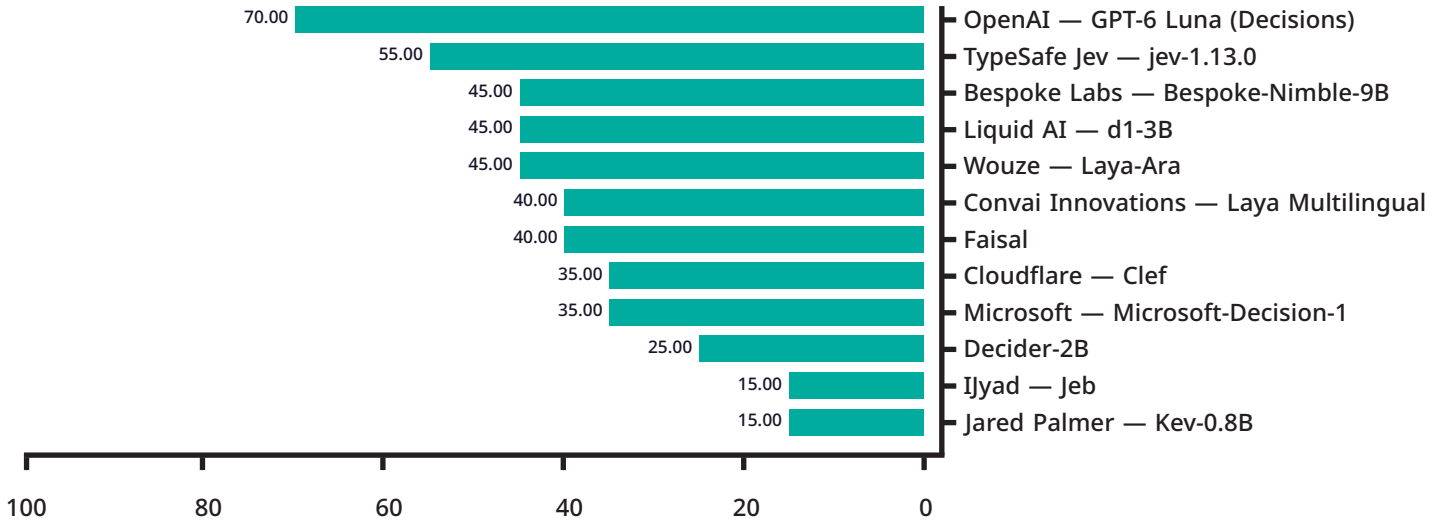
١٠.١٢ تحديد الموضوع (Topic Identification)

تصنيف النصوص - تحديد الموضوع



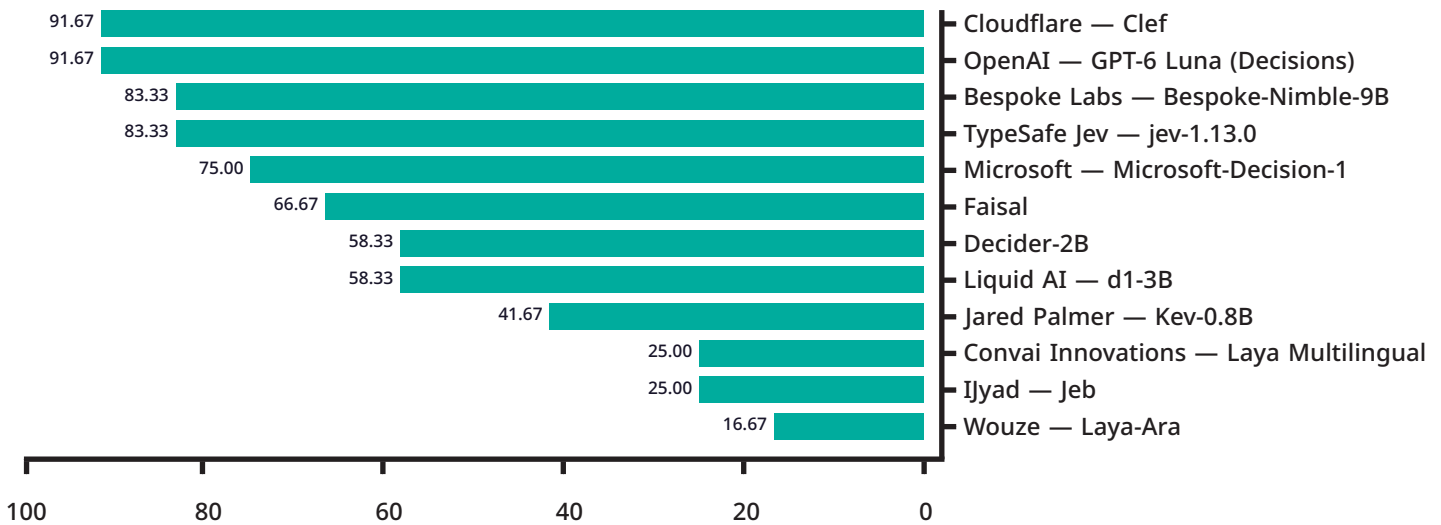
١١. تصحيح القواعد النحوية (Grammar Correction)

تصنيف النصوص - تصحيح القواعد النحوية



١٢. الترجمة الآلية (Translation/Transliteration)

تصنيف النصوص - الترجمة الآلية



ملحق (أ): تعريف الفئات والمهام

الفئة	تعريف الفئة	المهمة	تعريف المهمة
الكتابة الإبداعية	مهام تتعلق بإنتاج النص الإبداعي وتقييم مدى ملاءمة نتيجته.	إكمال النصوص	اختيار التكملة الأنسب لنص غير مكتمل.
التضمين	قياس العلاقة الدلالية بين نصين: التشابه في المعنى أو استلزام أحدهما للآخر.	تقييم صحة تنمة النص	الحكم على مدى ملاءمة تنمة مقترحة للنص.
دقة المعلومات	التحقق من صحة إجابة أو ادعاء بالاستناد إلى معلومات مرجعية.	التشابه الدلالي	تقدير درجة تقارب المعنى بين نصين.
ملء الفراغ	اختيار الكلمة أو العبارة المناسبة لإكمال فراغ داخل النص.	الاستلزام النصي	تحديد ما إذا كان نص يستلزم آخر أو يناقضه أو محايد عنه.
استخلاص المعلومات	تحديد الكيانات والعلاقات والمعلومات المطلوبة داخل النص.	التحقق من صحة الإجابة	الحكم على صحة إجابة مقابل سؤالها أو مرجعها.
		التحقق من صحة الادعاء	الحكم على صحة ادعاء بالاستناد إلى معلومات.
		ملء الفراغ	اختيار ما يكمل الفراغ في الجملة بصورة صحيحة.
		تصنيف الكيانات	تحديد نوع أو فئة كيان مذكور في النص.
		تصنيف العلاقات بين الكيانات	تحديد نوع العلاقة الرابطة بين كيانين.
		استخلاص المعلومات المطلوبة	انتقاء معلومة محددة مطلوبة من النص.

الفئة	تعريف الفئة	المهمة	تعريف المهمة
المنطق	الاستدلال المنطقي وتماسك النص والعلاقات السببية وترتيب الجُمْل.	تصنيف السبب والنتيجة	تمييز علاقة السبب من النتيجة بين حدثين.
		تصنيف التماسك النصي	الحكم على مدى ترابط النص وتماسكه.
		التحقق من المنطق السليم	تمييز العبارة المنطقية من المخالفة للبدئية.
		الاستدلال المنطقي	استنتاج النتيجة الصحيحة من مقدمات معطاة.
		ترتيب الجمل	اختيار الترتيب الصحيح لجُمْل مبعثرة.
تنفيذ البرنامج	استنتاج ناتج تنفيذ خطوات أو تعليمات برمجية.	تنفيذ البرامج	استنتاج مخرجات تنفيذ تعليمات برمجية.
الإجابة عن الأسئلة	اختيار الإجابة الصحيحة لسؤال مطروح.	الإجابة عن الأسئلة المطروحة	اختيار الإجابة الصحيحة لسؤال معطى.
الفهم القرائي	فهم النص: قابلية الإجابة وفهم السؤال والاستدلال وتحليل مكوّناته.	تصنيف إمكان الإجابة	تحديد ما إذا كان النص يتيح الإجابة عن السؤال.
		اختيار الحوار المناسب	اختيار الرد أو الدور الحوارى الأنسب.
		اختيار التفسير المناسب	اختيار التفسير الأنسب لظاهرة أو نص.
		فهم السؤال	إدراك مقصد السؤال وتحديد تفسيره الصحيح.
		التحقق من الاستدلال النصي	الحكم على صحة استدلال مستخرج من النص.
		تحديد القسم الكلامي	تحديد النوع الصرفي للكلمة (اسم/فعل/حرف...).

الفئة	تعريف الفئة	المهمة	تعريف المهمة
تصنيف النصوص	تصنيف النص وفق فئة دلالية	تحديد اللهجة	تمييز اللهجة العربية للنص.
		تعرفّ أفعال الحوار	تصنيف وظيفة العبارة في الحوار (سؤال، طلب...).
		اكتشاف المشاعر	تحديد الشعور المعبر عنه في النص.
		تصنيف السلوكيات الأخلاقية	الحكم على مقبولية سلوك أو عبارة أخلاقياً.
		اكتشاف الإساءة	تحديد وجود لغة مسيئة في النص.
		تصنيف الأسئلة	تصنيف السؤال وفق نوعه أو مجاله.
		اكتشاف السخرية	تمييز النص الساخر من غيره.
		تحليل المشاعر	تحديد الاتجاه الوجداني (إيجابي/سلبي/محايد).
		كشف الرسائل المزعجة	تمييز الرسالة المزعجة من السليمة.
		اكتشاف القوالب النمطية	تحديد وجود قالب نمطي في النص.
		التصنيف الفئوي للنصوص	إسناد النص إلى فئة موضوعية.
		تحديد الموضوع	تحديد موضوع النص الرئيس.

الفئة	تعريف الفئة	المهمة	تعريف المهمة
التعديل على النصوص	اكتشاف الأخطاء النحوية في النص وتصويبها.	تصحيح القواعد النحوية	اكتشاف الخطأ النحوي واختيار التصويب الصحيح.
الترجمة الآلية	ترجمة النص بين اللغات.	الترجمة الآلية	اختيار الترجمة الصحيحة للنص بين اللغتين.



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



برنامج تنمية
القدرات البشرية
Human Capability
Development Program



رؤية VISION
2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA

