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

بَلْسَم
BALSAM

بَلْسَم: مؤشر نضج تقنيات الذكاء الاصطناعي للغة العربية

التقرير الثاني

فبراير- ٢٠٢٦

HUMAIN

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

جامعة الملك سعود
King Saud University

جامعة الملك عبدالعزيز

جامعة قطر
QATAR UNIVERSITY

جامعة محمد بن زايد للذكاء الاصطناعي
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

QEERI
معهد قطر لبحوث البيئة والطاقة
Qatar Environment & Energy Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

جامعة بيشة
University of Bisha

جامعة نيويورك أبوظبي
NYU ABU DHABI

أكاديمية طويق
Tuwaiq Academy

stc

علم
elm

aiXplain

جامعة الأمير سلطان
PRINCE SULTAN UNIVERSITY

MISRAJ AI

واكب
WAKEB

groq

كلمة الأمين العام لمجمع الملك سلمان العالمي للغة العربية

تتكامـل أعمـال مجمـع الملك سـلمان العالـمي للغة العـربية في مسـارات متعـددة تشـمل: الحوسبة، والتخطيـط، والثقـافة، والتعليم؛ بما يـحقق أهدافه وإستراتيجيته وفقًا لمهامـه التي نص عليها تنظيم المجمع؛ خدمةً للغة العربية، وتعزيزًا لمكانتها العالمية.

ويأتي مؤشر (بَلْسَم) ضمن سياق أعمال المجمع المختلفة في الحوسبة العربية؛ سعيًا إلى توظيف الذكاء الاصطناعي في خدمة العربية، واستكمالًا لأعماله، مثل: (مرصد العربية) الذي يقيس واقع حوسبة العربية مقـارنةً باللغـات الأخرى، ويرصد أهم المبادرات فيه، و(مركز ذكاء العربية) الذي يقدّم خدماتٍ متنوعةً لروّاد صناعة تقنيات اللغة العربية، ويجمع الباحثين والأساتذة من اللغويين والحاسوبيين.

ويسـعد المجمـع بنشـر تقرير أداء النماذج اللغـوية في النصف الثاني من عـام ٢٠٢٥م، الذي يمثّل تطورًا واضحًا لمؤشـر (بَلْسَـم)؛ بتوسيع نطاق التقييم ليشـمل أسئلةً جديدةً وفئاتٍ إضافيةً؛ مما يعزز شـمولية المؤشر ودقته في قياس أداء نماذج الذكاء الاصطناعي للغة العربية. ويأتي التقرير بالشراكة مع عدد من الجهات المحلية والدولية التي أسـهمت في بناء بياناته وأدوات قياسـه، وقدّمت استشـاراتٍ في تطوير منصة متاحة للتقييم، تجمع بين التحكيم البشري والتقييم الآلي عالي الدقة؛ مما أسهم في الوصول إلى مقاييس موثوقة، وتعزيز اعتماده على المستوى النظري والتطبيقي.

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 لمشـروعات المجمع وأعماله، وإلى شـركاء المؤشر الذين أسهموا في تحقيق الأهداف النهائية وبناء البيانات الابتدائية له، ونتطلع في التقارير القادمة إلى مزيـد من التعاون والشـراكة مـع العديد من الجهات الأكاديـمية والتقنية؛ لتوسيع نطـاق العمل البحثي والتقني في المعالجة الآلية للغة العربية؛ بما يعزز من ريادة اللغة العربية في المجال التقني.

الأمين العام

أ.د.عبد الله بن صالح الوشمي.

المحتويات

بَلْسَم مؤشر نضج
تقنيات الذكاء الاصطناعي للغة العربية

هو مبادرة تعاونية يقودها مجمع الملك سلمان العالمي للغة العربية، بالتعاون مع عدد من الشركاء من داخل المملكة العربية السعودية وخارجها، ويهدف إلى قيادة تطوير وتجهيز مجموعات بيانات اختبار متخصصة تُعدُّ ضروريةً لتقييم أداء نماذج اللغة الكبيرة (LLMs) في مجموعة متنوعة من مهام معالجة اللغة الطبيعية (NLP) باللغة العربية. وتسهم هذه المبادرة في سد فجوة معرفية مهمة في مجال تقييم النماذج اللغوية العربية؛ بتوفير أطر ومنهجيات تقييم دقيقة تساعد على قياس أداء هذه النماذج كليًّا أو جزئيًّا. وتهدف إلى تأسيس فهم أعمق لنقاط القوة والضعف في النماذج اللغوية عند تقييم أدائها في مهام متعددة، مثل: التدقيق النحوي، وتوليد النصوص، وفهم المحتوى.

يهدف هذا التقرير إلى تقديم تحليل شامل لأداء النماذج اللغوية الكبيرة (LLMs) على مؤشر (بَلْسَم)؛ إذ يستند إلى تصميم وبناء مجموعات بيانات اختبار متخصصة جرى تحليل نتائجها ومقارنتها مع نتائج التقييم البشري؛ لاختبار مدى دقة النماذج في محاكاة التقييم البشري. ويشمل التقييم تحليل أداء النماذج في مجموعة واسعة من المهام ضمن كل فئة، مع ترتيب النماذج من الأعلى إلى الأدنى، إضافةً إلى استعراض متوسط الأداء العام بجميع الفئات.

وشهدت هذه النسخة توسعًا في تصميم مهام التقييم من حيث تنوّع أنماط الأسئلة؛ إذ تضمنت أسئلةً توليديةً وأسئلةَ اختيارٍ من متعدد؛ بما يتيح تقييمًا أكثر شمولية لقدرات النماذج اللغوية؛ بقياس أدائها في إنتاج الإجابات المفتوحة، ودقة الاختيار والاستدلال ضمن اختيارات محددة؛ وبما يعكس استخداماتٍ واقعيةً ومتنوعةً للنماذج اللغوية في التطبيقات العملية. وقد تم تقييم أسئلة جديدة لم ترها النماذج من قبل.

إحصاءات عامة عن التقييم لمؤشر (بَلْسَم)

عدد الفئات:	عدد المهام:	عدد الأسئلة:	عدد النماذج المقيمة:
(١٤)	(٦٧)	(~١٥٠٠٠)	(٢٠)

عدد الأسئلة التي تم التقييم عليها في هذه النسخة

(+٣٠٠٠)

فئات المهام المشمولة في التقييم:

الفئـــة	تعريف الفئة	المهمة	تعريف المهمة
الكتابـــة الإبداعية	كتـابة نصوص أصلية متنوعة تتطلب الإبداع اللغـــــوي.	إكمال النصوص (كلاهما)	إكمـال جمـلة معطـاة بجملة منـطقية ومتوافقة مع السياق.
		توليد الحوارات (كلاهما)	توليـد ردود حـوارية تمثـل تفـاعلًا بين أطراف الحديث.
		الشـــــــرح (كلاهما)	تقديم تفسيرات أو حقائق داعمة لإجابة؛ لتوضيح المعلومة.
		توليد التعليمات (توليدي)	تحويل أسئلة المستخدم أو استفساراته إلى تعليمات واضحة ومصنفة.
		توليد المقالات الإخبارية (توليدي)	كتابة مقالات إخبارية مستندة إلى وصف محدد للحدث.
		توليد القصائد (توليدي)	توليـد قصائـد مرتبـطة بعنـوان أو نص معين.
		توليد الأسئلة (توليدي)	إنشاء سؤال بناءً على إجابة معطاة.
		تأليف الجمل (توليدي)	إنتـاج افتراض جديـد مستمد من فرضية دون تعارض أو إضافة.
		تأليف القصص (توليدي)	كتـابة قصة منـطقية ومتماسـكة لنص سردي معطى.
		توليد العنوانات (توليدي)	توليد عنوان ملائم لنص يصف مضمونه .

الفئـــة	تعريف الفئة	المهمة	تعريف المهمة
الكتابـة الإبداعية	كتـابة نصوص أصلية متنوعة تتطلب الإبداع اللغـــــوي.	**توليد النصوص (توليدي)**	إنتـاج نـص جديـد ذي معنًـى بنـاءً على مدخل معين.
		توليد التعريفات (توليدي)	إنشـاء تعريفات لمفاهيـم أو عبارات بناءً على سياق المستخدم أو طلبه.
		توليد إجابات غير مناسبة (توليدي)	توليد إجابة خاطئة للسؤال.
		تقييم صحّة تتمةالنص (اختياري)	تقييم مـا إذا كانت جملة أو فقرة معينة تُعدُّ امتدادًا منطقيًّا وسليمًا لنص سابق.
التضمين	تحديـد مـا إذا كـان يمكن استنتاج جملـة معينـة منطقيًّـا من أخـرى؛ بهـدف تقييـم العـلاقة الـدلالية بين النص والاستنتاج.	**الاستلزام النصي (كلاهما)**	تحديد مـا إذا كان يمـكن استنتاج جملة مـعينة من نـص مـا استنتاجًـا منطقيًّا؛ لتقييـم الـعلاقة الـدلالية بين النـص والاستنتاج.
		التشابة الدلالي (كلاهما)	تحديد نوع العلاقة الدلالية بين جملتين.
ملء الفراغ	استكمـال الجمل بكلمـات مناسبة دلاليًّـا ولغويًّـا.	**ملء الفراغ (كلاهما)**	استكمـال الجمل بكلمات منـاسبة دلاليًّا ولغويًّا.
استخـلاص المعلومات	تحـــليل النصـوص لاســتـخـــــراج معلومات محددة.	**استخلاص المعلومات المطلوبة (كلاهما)**	تحديـد نـوع الـعلاقة أو المعلـومة المطلوبة من نص معين.
		تحديد أسماء الأمراض في النصوص (توليدي)	استخراج أسـماء الأمراض المذكورة في النص.

التقرير الثاني لمؤشر بلسم.

الفئـــة	تعريف الفئة	المهمة	تعريف المهمة
استخـلاص المعلومات	تحـــليل النصـوص لاسـتـخـــــراج معلومات محددة.	**استخلاص الكلمات الرئيسة (توليدي)**	استخلاص الكلمات أو العبارات المفتاحية التي تلخص النص.
		التعرف إلى أسماء الكيانات (توليدي)	تحديد الكيانات المسماة، مثل: (الأسماء، والأماكن، والجينات) من النص.
		استخلاص العلاقات (توليدي)	استخـراج الـعلاقة بين كيـانين في نص معين.
		تصنيف العلاقات بين الكيانات (اختياري)	تحديـد وجـود علاقة دلالية بين كيـانين مذكورين في النص من عدمها.
		تصنيف الكيانات (اختياري)	تصنيف كيان معين إلى أقرب حقل دلالي مناسب له.
المنطق	استنتـــاج المنطق من خـلال مـــهام تتطـلـــب فـــهم العلاقـات السببية، وحـــل الألغـــاز، والترتيب، والتنبـؤ.	**تصنيف السبب والنتيجة (كلاهما)**	تحديـد الجمـلة النـاتجة عن سبب معين في نص.
		تحليل الإحالات الضميرية (توليدي)	تحديد الكلمة أو العبارة التي يشير إليها ضمير معين في الجملة.
		التحليل التنبؤي (توليدي)	تحليـل الخيارات؛ لتحديـد الأكثر منطقية بناءً على معطًى معين.
		حل الألغاز (توليدي)	حل الألغاز البسيطة باستخدام المنطق.
		التحقق من المنطق السليم (اختياري)	تحديـد الجمـلة التي تخـالف المنـطق السليم أو المعرفة العامة عن العالم.

التقرير الثاني لمؤشر بلسم.

الفئـــة	تعريف الفئة	المهمة	تعريف المهمة
المنطق	استنتـاج المنطق من خـلال مـهام تتطـلـب فـهم العلاقـات السببية، وحـل الألغـاز، والترتيب، والتنبـؤ.	ترتيب الجمل (اختياري)	تـرتيب خطـوة أو إدراجها في تسلسـل منطقي للخطوات.
		الاستدلال المنطقي (اختياري)	تـطبيق التفكيـر المنـطقي لاستخلاص نتيجة صحيحة من نص معين.
		تصنيف التماسك النصي (اختياري)	تحديد ما إذا كانت جملة معيّنة متماسكةً دلاليًّا مع الجملة أو السياق السابق لها.
تنفيذ البرامج	تنفيـذ المعادلات الرياضية أو المهام البرمجية.	تنفيذ البرامج (كلاهما)	تنفيذ المعـادلات الريـاضية أو المهـام البرمجية.
الإجابة عن الأسئلة	الإجـابـة عـن السؤال المطروح.	الإجابة عن الأسئلة المطروحة (كلاهما)	تقديـم إجابة مباشـرة وصحيـحة لسؤال مطروح.
		تحليل السؤال (توليدي)	تفكيك سؤال معقد إلى سلسـلة أسئلة بسيطة للإجابة النهائية.
فهم المقروء	استخـراج وفهـم المعلومات مـن النصوص المكتوبـة مـن خلال الإجابـة عـن أسئلـة مستندة إلى المحتوى.	التحقق من الاستدلال النصي (كلاهما)	قيـاس قـدرة النماذج علـى التحقق من صحة الاستدلالات اعتمادًا على النص.
		فهم السؤال (كلاهما)	التأكد مما إذا كان التوضيح المقدم يعين علـى تفسير الغرض من السؤال الأصلي، ويزيل الغموض عنه.
		تصنيف إمكانية الإجابة (اختياري)	تحديـد مـا إذا كان السؤال يمـكن أن يُجاب عنه إجـابةً تامةً اعتمادًا علـى السياق أو المعلومات المتاحة.
توسيم السلاسل النصية	تعيين تسميات دقيقة لعناصر ضمن تسلسـل نـصي، مثـل: تحديـد الأخطـاء النحـوية، أو استـخـراج الكلمـات المفتاحية.	اكتشاف الأخطاء النحوية (توليدي)	تحديـد الكلمات التي تحتوي على أخطاء نحوية في الجملة.
		تحديد القسم الكلامي (اختياري)	تحديد القسـم الكلامي لكل كلمة داخل سياق الجملة.

الفئـــة	تعريف الفئة	المهمة	تعريف المهمة
التلخيص	مـوجز للنـص يُبـرز أهـم المعلـومات والمفاهيم.	**توليد العنوانات** (توليدي)	اقتـراح عنوان أو موضوع مناسب لنص أو بريد شبكي.
		تلخيص النصوص (توليدي)	تلخيـص نـص طويـل إلى جمـلة أو عدة جمل تختصر المعنى.
تصنيف النصوص	تـصنيف المحتـوى النـصي إلـى فئات محددة اعتـــمادًا على معـــانيه أو خصائصه.	**تحديد اللهجة** (كلاهما)	تحديـد اللـهجة المستخـدمة في النص العربي.
		التصنيف الفئوي للنصوص (كلاهما)	تصنيف النص إلى فئة معينة.
		فهم التعليمات (توليدي)	تحديد التعليمات المطلوبة من النص.
		تحديد المقصود (توليدي)	تصنيف نية المستخدم من استفسـار أو طلب.
		تحديد الإشكالية (توليدي)	تحديـد نـوع الإشـكالية أو الفئة التي وجدت في النص.
		اكتشاف المشاعر (اختياري)	تحديد المشاعر التي يعبّر عنها النص.
		اكتشاف الرسائل المزعجة (اختياري)	تحديـد مـا إذا كانت الرسـالة مزعجةً أو معتادةً.
		اكتشاف الإساءة (اختياري)	اكتشـاف ما إذا كان النـص يحتوي على كراهية أو إساءة.
		اكتشاف السخرية (اختياري)	اكتشـاف مـا إذا كانت الجمـلة تتضـمّن سخريةً.
		تحليل المشاعر (اختياري)	تحليل مشـاعر شخـصية بناءً على موقف أو جملة.

التقرير الثاني لمؤشر بلسم.

الفئـــة	تعريف الفئة	المهمة	تعريف المهمة
تصنيف النصوص	تـصنيف المحتـوى النـصي إلـى فئات محددة اعتـــمادًا على معـــانيه أو خصائصه.	**تعرُّف أفعال الحوار** (اختياري)	تحديـد نـوع الجمـلة أو الغـرض التواصـلي منهـا، مثل: السؤال، أو الأمـر، أو التصريح؛ اعتمادًا على معناها وسياقها.
		تصنيف السلوكيات الأخلاقية (اختياري)	تقييم الأفعال أو السلوكيات وتصنيفها استنـادًا إلى القيم والمعاييـر الأخلاقية العامة.
		تصنيف الأسئلة (اختياري)	تحديـد الفئة المناسبة للسؤال بناءً على معناه ومحتواه.
		تحديد الموضوع (اختياري)	تصنيف محتوى النص بناءً على موضوعه.
		اكتشاف القوالب النمطية (اختياري)	تحديد ما إذا كان النص يتضمّن قولبةً نمطيةً أو يعرض نمطًا مضادًّا لها؛ اعتمادًا على تحليل المعنى والسياق.
تعديـــل النصوص	إجـــراء تغييـــرات لغويـة أو دلاليـــة علـــى النصـوص؛ لتحسينهـــــا أو تكييفهـــا مـــع سياقـــات محددة.	**تصحيح القواعد النحوية** (كلاهما)	تصحيح الأخطاء النحـوية والإملائية في النص.
		تعديل الهوية الجنسية للنص (توليدي)	تعديـل النـص ليتـوافق مـع الهـوية الجنسية (من الذكر للأنثى) والعكس.
		إعادة الصياغة (توليدي)	إعـادة صيـاغة الجمـلة بأسـلوب نحـوي مختلف، مع الحفاظ على المعنى.
		تعديل السؤال (توليدي)	إعادة صياغة سؤال غامض ليصبح واضحًا ومحددًا.
		تبسيط النص (توليدي)	تبسيـط النـص المعقـد؛ ليصبـح أيسـر فهمًا.
الترجمة الآلية والنقل الصوتي	تحويل النصوص آليًّا من لغة إلـى أخرى، أو بيـن لهجـــات مختلفـــــــة.	**الترجمة الآلية** (كلاهما)	ترجمة النصوص من لغة إلى أخرى.
		ترجمة اللهجات (توليدي)	ترجمة لـهجة عربية محلية إلـى العربية الفصحى.

التقرير الثاني لمؤشر بلسم.

الفئـــة	تعريف الفئة	المهمة	تعريف المهمة
الترجمة الآلية والنقل الصوتي	تحويـل النصوص آليًّا من لغة إلى أخرى، أو بيـــن لهجات مختلفة.	النقل الصوتي (توليدي)	تحويـل النـص العـربي إلـى الكتـابة الصوتية.
الدقـــــة المعلوماتية	إجـراء تغيـيـرات لغويـــة أو دلالية عـــلى النصـوص؛ لتحسينـــهـــا أو تكييفـــهـــا مـع سياقات محددة.	التحقق من صحة الادعاء (اختياري)	التـحقق من أنّ مـا ورد في النـص من ادّعـاء مدعـوم بأدلة واردة في النـص نفسه.
		التحقق من صحة الإجابة (اختياري)	التـحقق من أنّ الإجابة المقدّمة صحيحة وتتماشـى مـع المعلومات الـواردة في نص أو محتوى معرفي معيّن.

منهجية التقييم:

يعتمـد هذا التقريـر على منـهجية تقييم متعددة المراحل، تهدف إلى قيـاس أداء النماذج اللغوية بدقة وتوازن؛ وذلك من خلال الجمع بين التقييم البشـري والتقييم الآلي، واستخدام أنماط مختلفة من الأسئلة تعكـس تنـوّع الاستخدامات الواقـعية للنماذج اللغوية. فيما يـلي نستعرض أنماط الأسئـلة وآلية قياس النتائج وطريقة التقييم المعتمدة وتوصيات التقرير الثاني.

أولًا: أنماط الأسئلة

تم تصميم مجموعات التقييم لتشمل نوعين رئيسين من الأسئلة، هما:

الأسئلة التوليدية (إجابات مفتوحة)

تُستخـدم لتقييـم قـدرة النموذج على توليـد محتوًى لغوي متماسـك ودقيق، يشـمل الفهم العميق للسؤال، وصياغة الإجابة، وجودة التعبير، ومدى توافق المخرجات مع الإجابة المرجعية.

الأسئلة الاختيارية (اختيار من متعدد)

تُستخـدم لقياس دقة الفهم، والاستـدلال، والقدرة على اختيار الإجابة الصحيحة من بين بدائل محددة؛ بما يعكس كفاءة النموذج في المهام المقيدة والمعيارية.

تـم اعتماد منهجية تقييـم متعددة المراحل هدفت إلى تحديد الطريقة المنـاسبة لقياس أداء النماذج بـدقة وموثـوقية، شـملت المرحلة الأولـى تجربة عـدد من المنهجيـات الآلية، مثل: مقيـاس ن-وحدة (n-gram)للأسئلة التوليدية، والدقة.

(Accuracy) للأسئلة الاختيارية. تمت مقارنة نتائج هذه المنهجيات بنتائج التقييم البشري؛ لاختيار الأداة الأقرب لمحاكاة الحكم البشري في تقييم جودة الإجابات.

ثانيًا قياس النتائج

يعتمـد قيـاس النتائج في هذا التقرير على آليات تقييم متعـددة، تهدف إلى تحليل أداء النماذج اللغوية من زوايا مختلفة؛ بالجمع بين التقييم الكمي، والتقييم الآلي، والتقييم البشري. ويُسهم هذا التكامل في توفيـر قيـاس أكثر دقة وموثوقية لأداء النماذج، والحد من الانحياز المرتبط بالاعتماد على أسـلوب تقييم واحد فقط.

أ) التقييم الكمي

تـم استخدام مقاييـس كمية معيارية لقياس أداء النمـاذج اللغوية بموضوعية؛ بمقـارنة مخرجات النماذج بالإجابات المرجعية المعتمدة. وقد طُبِّقت هذه المقاييس بما يتناسب مع طبيعة كل نوع من الأسئلة.

- الأسئـلة التوليـدية: استُخـدمت المقاييس المعتمدة علـى ن-وحـدة (n-gram)، التي تقيس درجة التشابه بين النص المُولَّد والنص المرجعي من حيث تطابق تسلسلات الكلمات.
- الأسئـلة الاختيـارية: تم الاعتماد على مقياس الـدقّة (Accuracy)؛ لقياس مدى صحة اختيار النموذج للإجابة الصحيحة من بين البدائل المتاحة.

ب) التقييم الآلي

استخـدام مقيّـم آلي يُقدّر جودة مخرجات النماذج اللغوية؛ بمقـارنة إجابة النموذج بالإجابة المرجعية، وذلك لكل من الأسئلة الاختيارية والتوليدية.

- **درجة (٠): الإجابة غير صحيحة تمامًا.**
- **درجة (١): إجابة ضعيفة تحتوي على أخطاء كبيرة أو غير مكتملة بدرجة كبيرة، وتكاد لا تشبه الإجابة المرجعية**
- **درجة (٢): إجابة مقبولة، ولكنها تختلف بوضوح عن الإجابة المرجعية**
- **درجة (٣): إجابة صحيحة تمامًا أو شبه كاملة.**

وبعد احتساب درجات التقييم الآلي لكل من الأسئلة الاختيارية والتوليدية بطريقة مستقلة، تُدمَج النتائج النهائية وفقًا لأوزان نسبية تعكس أهمية كل نمط من أنماط الأسئلة، وذلك على النحو الآتي:

- **(٣٠٪) من الدرجة النهائية مخصصة لنتائج الأسئلة الاختيارية.**
- **(٧٠٪) من الدرجة النهائية مخصصة لنتائج الأسئلة التوليدية.**

ويهدف هذا التوزيع إلى إعطاء أولوية أكبر للقدرات التوليدية للنماذج اللغوية؛ نظرًا إلى دورها الأساس في التطبيقات الواقعية التي تتطلب فهمًا عميقًا وإنتاجًا لغويًّا عالي الجودة.

ج) التقييم البشري:

أُجـرِي تقييم بشـري علـى عينة مكونة من (٢٠٠) سؤالٍ مـوزعة على المهام المختلفة، وقـد تولى ثلاثة محكمين تقييم كل إجابة وفقًا لأربعة مقاييس رئيسة، واحتُسب المتوسط النهائي لدرجاتهم لكل سؤال، ثم لكل نموذج.

معايير التقييم البشري:

- **الـدقـة (٠ إلى ٣):** تشير إلى مـدى صحـة الإجـابـة وتطـابقهـا مع المطـلـوب.
- **الشمولية (٠ إلى ٣):** تقيس مدى اكتمال الإجابة وتغطيتها للجوانب المهمة في السؤال.
- **الهلوسـة (٠ أو ١):** تشيـر إلى وجـود معلومـات مُـختَلَـقة أو غيـر صحيحـة.
- **الإطنـاب (٠ أو ١):** تقيـس مـدى الإطـالـة أو الخـروج عـن المـوضـوع.

ثالثًا: طريقة التقييم المعتمدة

بنـاءً علـى المقـارنة بين المنهجيات المختلفة؛ تم اعتمـاد التقييم الآلي كونه المنـهجية الأقرب لمحاكاة التقييـم البشـري؛ إذ أظهـر علاقة طـردية قوية بـلغت (٠.٩) مع نتـائج المحكـمين. في المقابل، أظهرت المقاييـس الكـمية التقليدية علاقةً عكسيةً مع التقييم البشـري؛ مما يشير إلـى محدودية قدرتها على عكس الحكم البشـري بـدقة في تقييم جودة الإجابات.والجدير بالذكر أن منـهجية التقييم قابلة للتحديث مـستقبلًا؛ بمـا يتناسب مع طبيعة المهام المختـلفة والتقارير القادمة، مع الاستمـرار في مقارنة أدوات التقييم الآلي بالأحكام البشرية؛ لضمان دقة النتائج وموثوقيتها.

رابعًا: توصيات التقرير الثاني

- **أشكال المهام:** تتفاوت المهام بين مهام مفتوحة (توليدية إبداعية: مثل التلخيص، واستكمال القصص، ونظـم القصائـد، وغيرهـا)، وبين مهام مغـلقة (لها إجابة واحـدة صحيحة: مثل الإعـراب، وتحديد اللهجة، وغيرهـا). ويـوصي فريق بلسـم بأهمية النظـر في آلية مبتكـرة لتقييم المهام الإبداعية سـواء كان ذلك بتقسيمهـا إلـى مهام حتمية (مثل اكتشـاف البيت المكسـور من أجـل تقييم نظم القصائـد)، أو كان عبر المقارنة بين النماذج (بأن تعرض نتائج نموذجين، ويتم المقارنة بينهما).
- **صعـوبة المهـام:** من الملاحظ أن عددًا من المهام التي قيمت النماذج اللغوية عليها لم تعد صعبة على النمـاذج الحديثة، ممـا يستدعي النظر أكثر في مهام أكثر صعوبة. من المهام التي فشـلت فيها النماذج بوضوح توليد القصائد، لذا فهي تستدعي مزيدًا من البحث.
- **علاقة طـردية:** لاحظنـا تقاربًـا وارتباطًا جيدًا بين التقييم البشـري وتقييم النمـوذج المحكِّم (الذي يقوم بالمقـارنة بين الإجـابة النموذجية وإجـابة النموذج اللغوي)، وهو ما يزيد من موثـوقية التقييم الآلي. هذا التقارب أفضل في المهام المغلقة، ولكنه أقل في المهام المفتوحة.

- **النمـوذج المحكـم:** علـى الرغم من أن النموذج المحكِّم لا يجب أن يطلع علـى السؤال، وإنما على الإجابة النمـوذجية وإجابـات النماذج المختلفة، إلا أنه فاق جميع النماذج الأخرى عند تقييمه، مما استلزم استبعاده من القائـمة لأجـل العدالة بين النماذج. ويوصي فريق بلسـم باعتماد نموذج مفتوح المصدر ليكون النموذج المحكم (خاصة بعد النظر في نتائج هذا التقرير).
- **المستويات اللغوية:** ركز التقرير الأول والثاني من مؤشـر بلسـم على المهام النصية فقط في النماذج اللغوية. ولا شـك أن المهام الصوتية والبصرية هي جزء أسـاسي من لغتنا العربية (مثل القدرة على نطق القصائـد بالشـكل الصحيـح، وصناعة الصور بطابـع ثقافي عربي، وغيرهـا). ونأمل في التقاريـر القادمة أن تتضمن النتائج تقييمًا لهذه المهام المختلفة.
- **البحث والتطوير في التخصصات المختلفة:** ما زالت آليات التقييم تتطلب مزيدًا من الأبحاث العلمية خاصة فيمـا يتعـلق بالمهام العربية، مثل مجال اللهجات والتفاوت فيمـا بينها، وقياس قدرات النماذج في فهم الثقافة العربية، ومجال البلاغة والشعر والأدب. ولم يعد مجال التقييم حكرًا على أصحاب التخصصات التقنية. وندعو الباحثين من التخصصات الإنسـانية والشـرعية وغيرها، إلى المشاركة في مؤشر بلسم بإضافة مهام نوعية عربية.

النماذج المُقيّمة في التقرير الثاني

اعتمـد التقريـر في اختيار النماذج المُقيّمة على منهجية واضحة، تهدف إلى ضمان الاستمرارية والاتسـاق في التقييم. وقد شـمل الاختيار أغلب النماذج التي قُيِّمَت في التقرير السـابق، مع اعتماد أحدث إصداراتها المتاحة، إلى جانب إدراج نماذج جديدة طُرحت خلال المدة اللاحقة، وقد رُوعي في عملية الاختيار تمثيل تنوّع النماذج من حيث أحجامها وبناها المعمارية، مع تضمين كل من النماذج المفتوحة المصدر والمغلقة المصدر؛ بما يضمن توافقها مع متطلبات التقرير ودعمها للغة العربية، أو عملها ضمن إطار متعدد اللغات.

اسم النموذج	المصدر	الحجم (عدد المعاملات)	الوصف	المرجع/ الرابط
AceGPT-v8-8B-Chat	مفتوح المصدر	8B	نمـوذج محـادثـة باللـغـة العربية، مكـون من (٨) مليارات معامل، مبني عـلـى (LLaMA٢)، ومخـصـص للتفـاعـلات النصية	المرجـع / الرابـط
c4ai-aya-expanse-32b	مفتوح المصدر	32B	نموذج متعدد اللغات، مكون من (٣٢) مليـار معمـل، ويدعم (٢٣) لغةً من بينها العربية.	المرجـع / الرابـط
Cohere (Command-A 03- 2025)	مفتوح المصدر	111B	نموذج لغـوي كبير، يحتوي على حـوالي (١١١) مليـار معمـل، ويدعم (٢٣) لغةً.	المرجـع / الرابـط

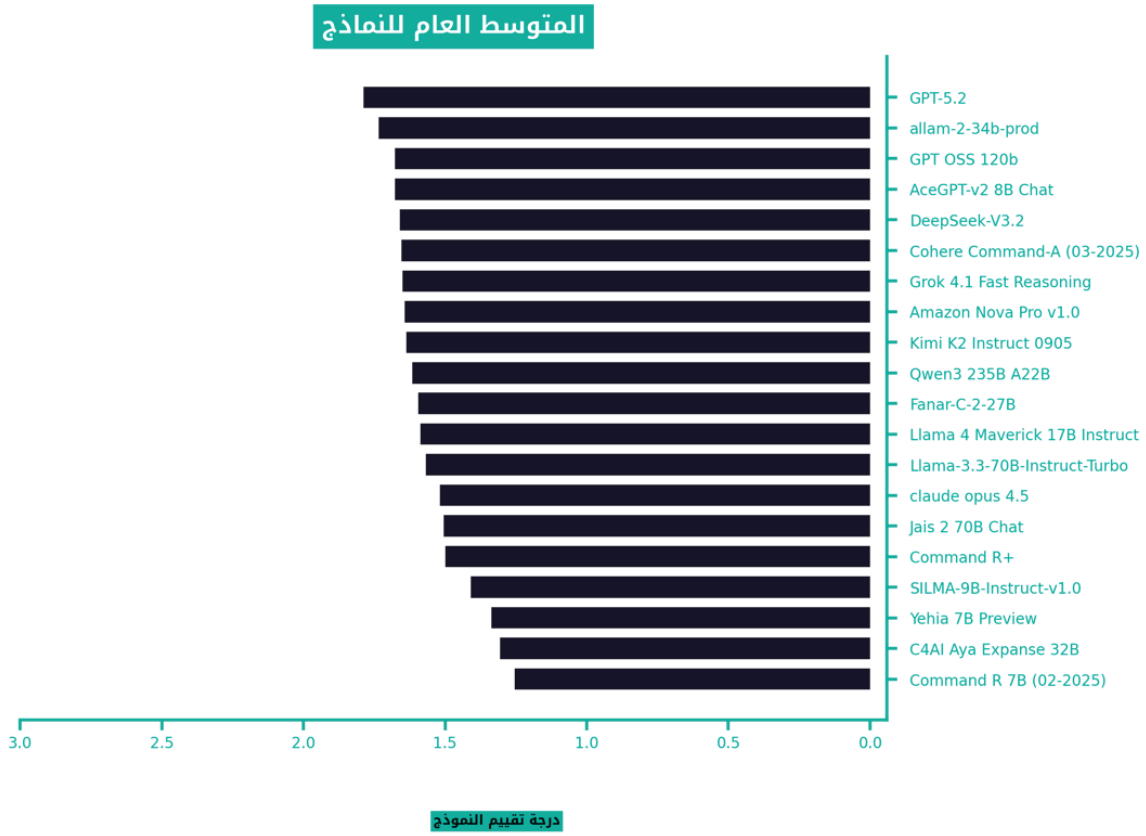
Command R+	مفتوح المصدر	104B	نموذج متعدد اللغات، مكون من (١٠٤) مليارات معامل، ومخصص للتوليد المعزز بالاسترجاع، وتنفيذ التعليمات.	الرابـط
Command-r7b 2024-12	مفتوح المصدر	7B	نمـوذج لغـوي متعـدد اللغـات، يتكون من ٧ مليارات معامل.	الرابـط
DeepSeek V3	مفتوح المصدر	685B MoE	نموذج متعدد الخبراء، مكون من (٦٨٥) مليار معمـل يدعم مهام الفهـــم اللـــغـوي، والبرمجة، والتفكير المنطقي.	المرجـع الرابـط
Jais-70-2B-Chat	مفتوح المصدر	70B	نمـوذج ثنـائي الـلغة (عـربي-إنجليزي)، مكـون من (٧٠) مليار معامل.	المرجـع الرابـط
Kimi K2-Instruct-0905	مفتوح المصدر	1T MOE	نمـوذج لغوي كبيـر قائم على تعدد الخبـراء، ومُـدرَّب علـى اتبـاع التعليمات، ومهيأ لمهام الاستدلال والبرمجة بسياق طويل.	المرجـع الرابـط
Llama-70-3.3B-Instruct-Turbo	مفتوح المصدر	70B	نمـوذج لغـوي كبيـر صمّمتـه (Meta)، ويحتوي على (٧٠) مليار معامل، ويدعم عدة لغات.	المرجـع الرابـط
Llama 4 Maverick 17B Instruct	مفتوح المصدر	17B	نمـوذج لغـوي كبيـر صمّمتـه (Meta)، ويحتوي على (١٧) مليار معمل، ويدعم عدة لغات.	المرجـع الرابـط
Qwen235 3B A22B	مفتوح المصدر	235B	نمـوذج لغوي مكـون من (٢٣٥) مليار معامل.	المرجـع الرابـط
SILMA9-B Instruct-v1.0	مفتوح المصدر	9B	نمـوذج لغـوي بالـلغة العـربية مكـون من (٩) مليـارات معامل، ومبني على بنية (Gemma).	المرجـع الرابـط
Yehia7-B preview	مفتوح المصدر	7B	نمـوذج ثنـائي الـلغة (عـربي-إنجليزي) مكـون من (٧) مليارات معامل.	المرجـع الرابـط

التقرير الثاني لمؤشر بلسم.

اسم النموذج	المصدر	الحجم (عدد المعاملات)	الوصف	المرجع/ الرابط
Fanar-C-27-2B	مفتوح المصدر	27B	نموذج ثنائي اللغة (عربي-إنجليزي) مدرب على ما يقرب (٢٧) مليار معامل.	الرابط
gpt-oss-120b	مفتوح المصدر	120B	نموذج متعدد اللغات، مكون من (١٢٠) مليار معمل، وطوّرته (OpenAI).	الرابط
Amazon Nova Pro	مغلق المصدر	غير معلن	نموذج من (Amazon Bedrock).	الرابط
Grok-4.1 Fas Reasoning	مغلق المصدر	غير معلن	نموذج متعدد اللغات، ويدعم الاستدلال المتقدم، وطوّرته (xAI).	الرابط
Claude Opus 4.5Instruct	مغلق المصدر	>130B	نموذج متعدد اللغات من (Anthropic).	الرابط
GPT 5.2	مغلق المصدر	غير معلن	نموذج من (OpenAI)، يدعم مدخلات متعددة الوسائط واللغات.	الرابط
ALLaM34 – 2B	مغلق المصدر	غير معلن	نموذج ثنائي اللغة (عربي-إنجليزي) مدرب على ما يقرب (٢٧) مليار معمل.	الرابط

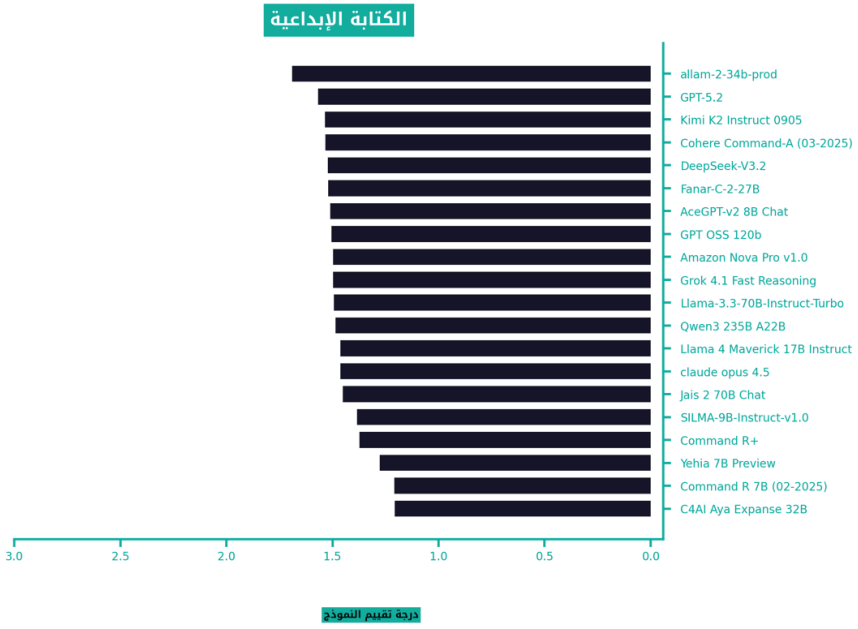
النتائج

أولًا: متوسط الأداء العام للنماذج:



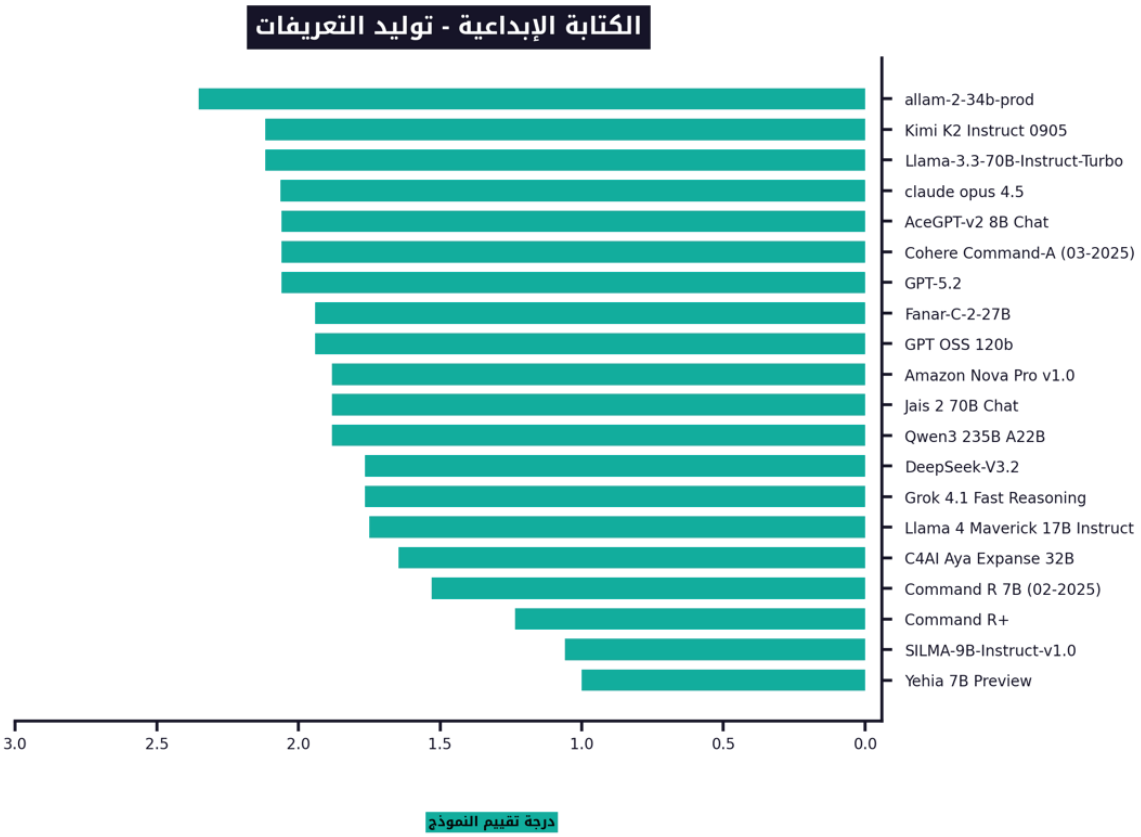
ثانيًا: تقييم النماذج حسب الفئات، ومهام كل فئة:

١. الكتابة الإبداعية (Creative Writing).

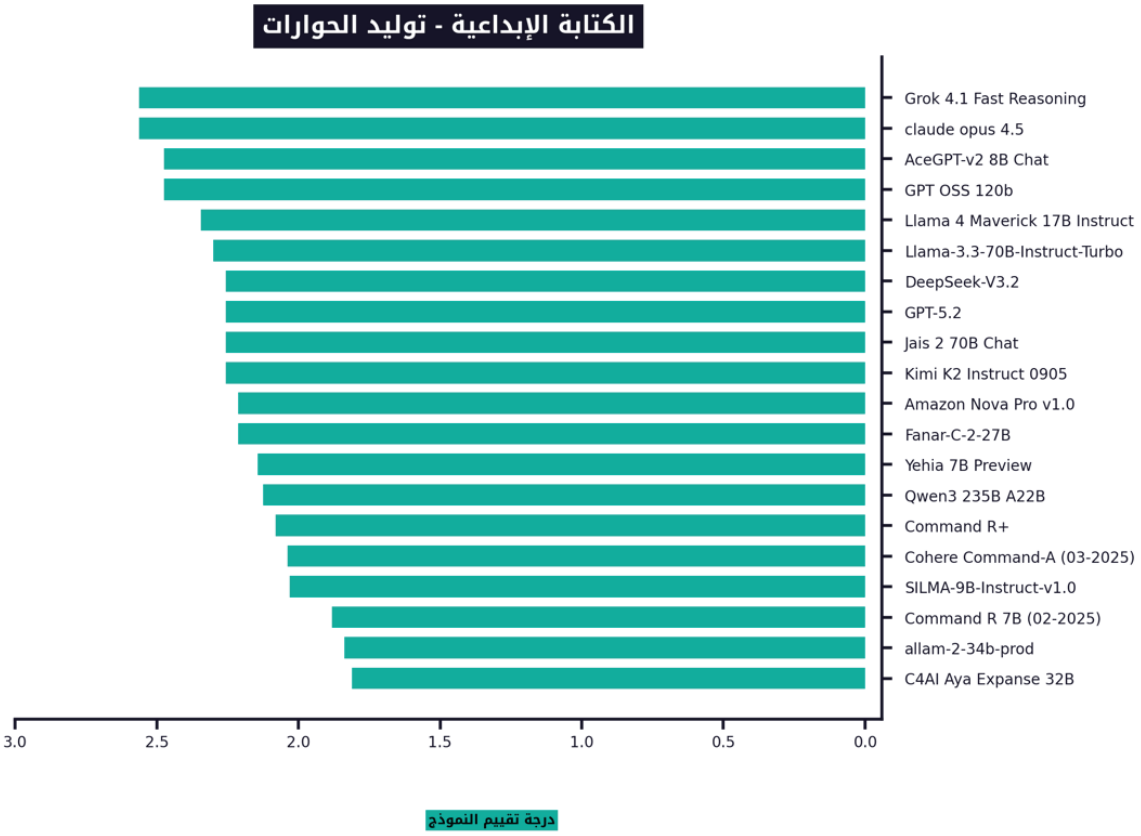


المهام الفرعية:

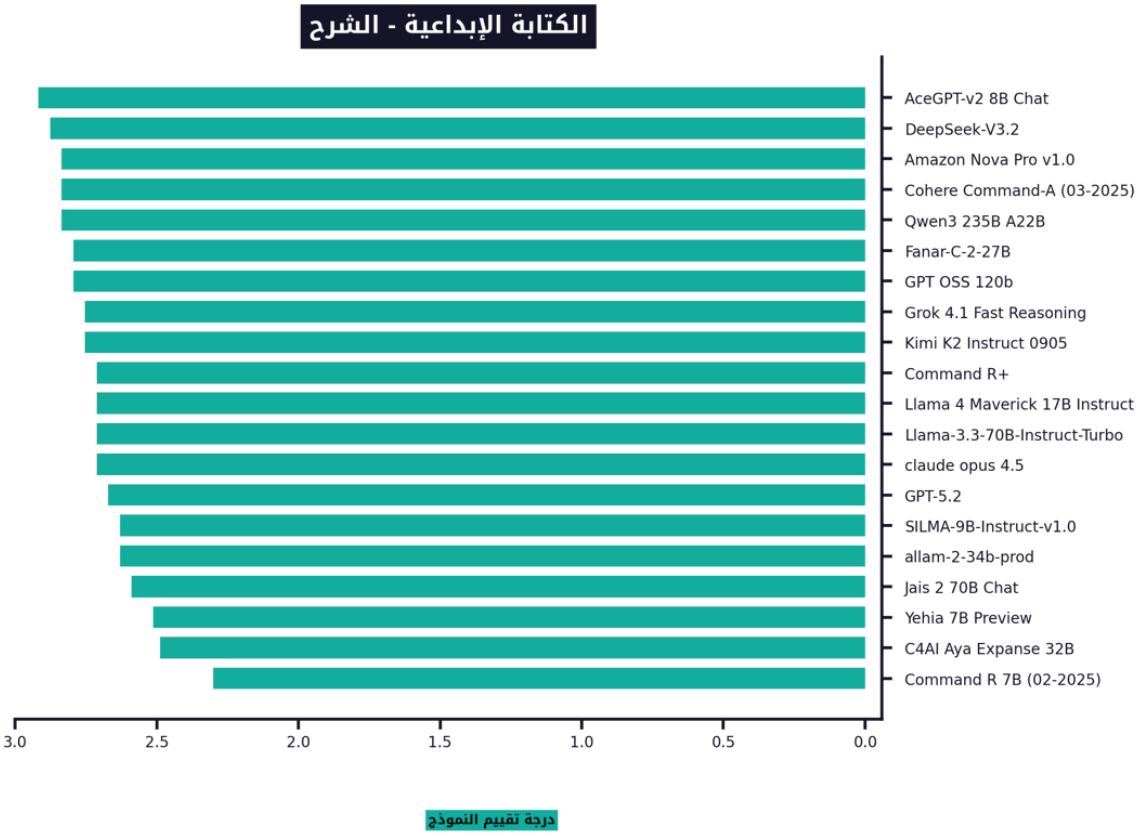
١.١ توليد التعريفات (Definition Generation).



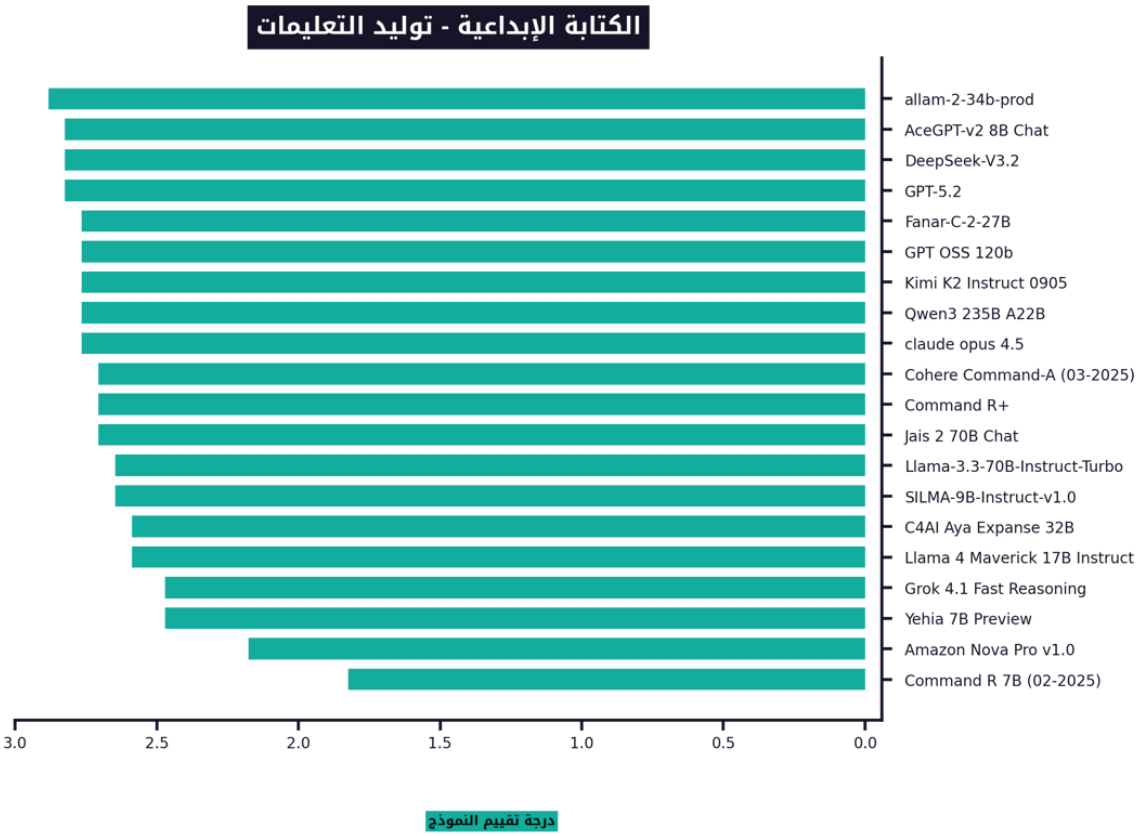
١.٢ توليد الحوارات (Dialogue Generation).



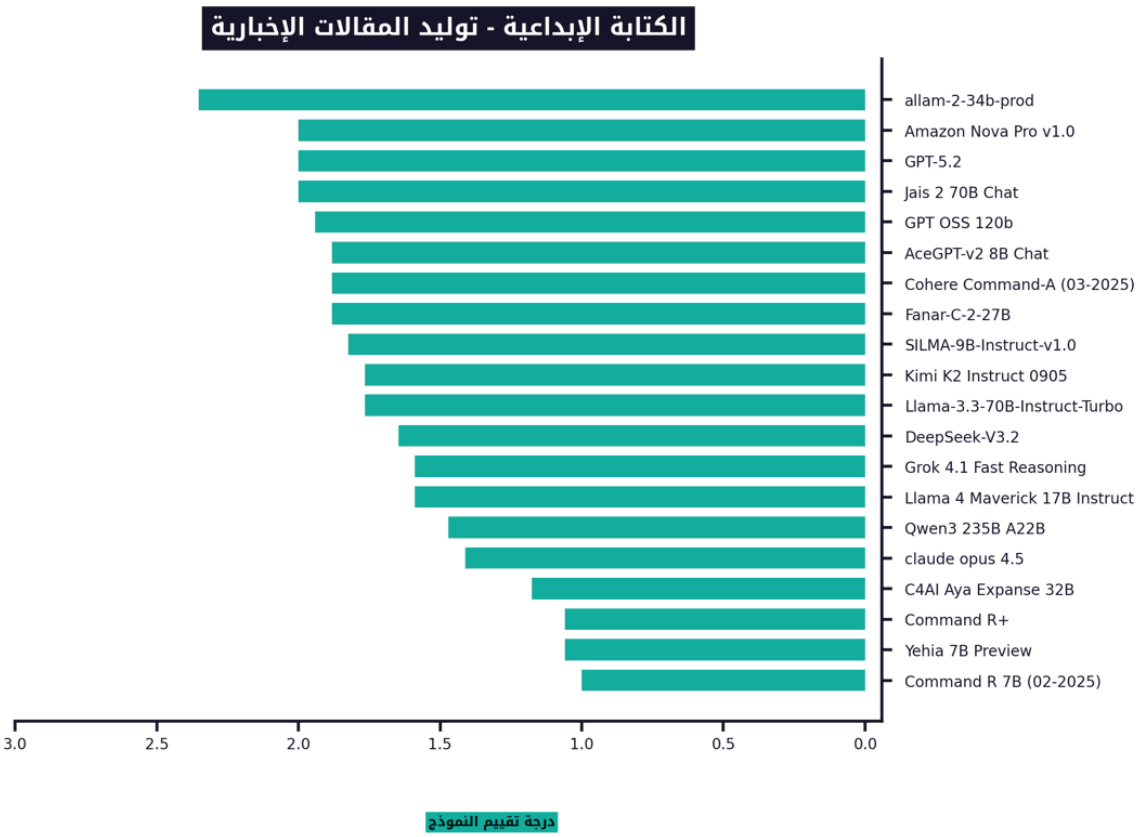
١.٣ الشرح (Explanation).



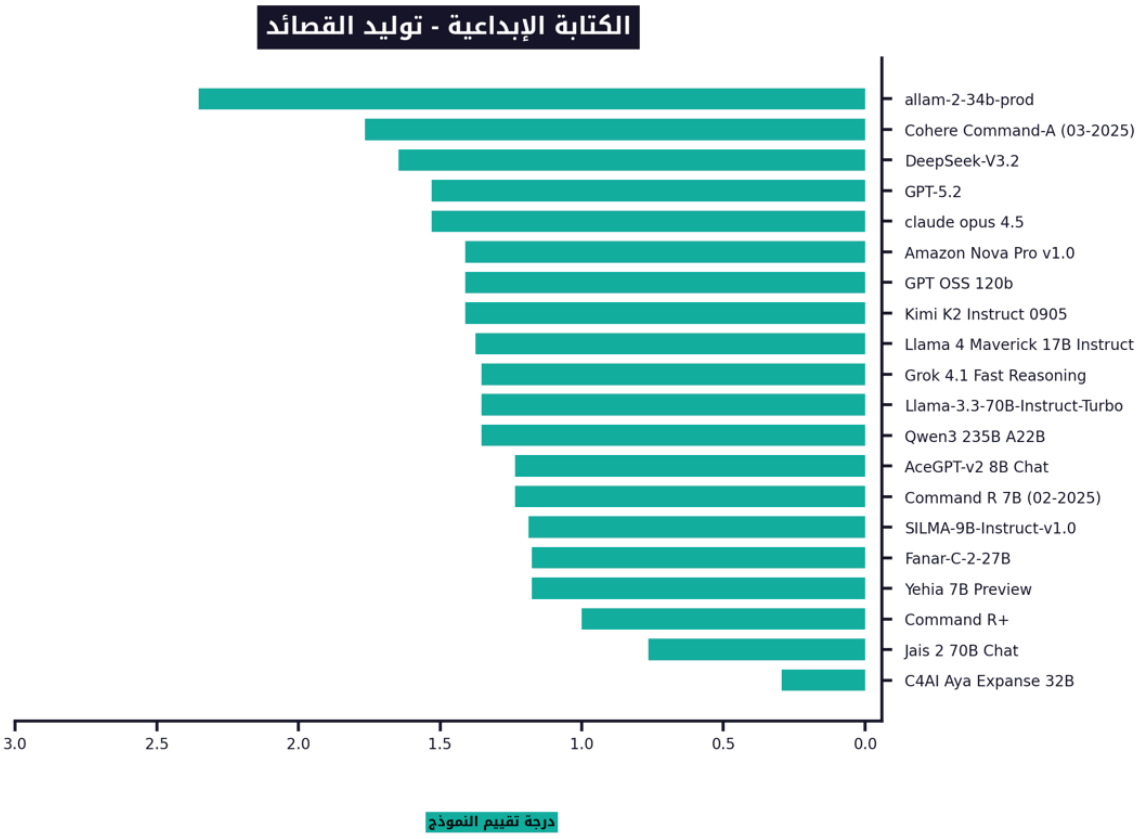
١.٤ توليد التعليمات (Instruction Generation).



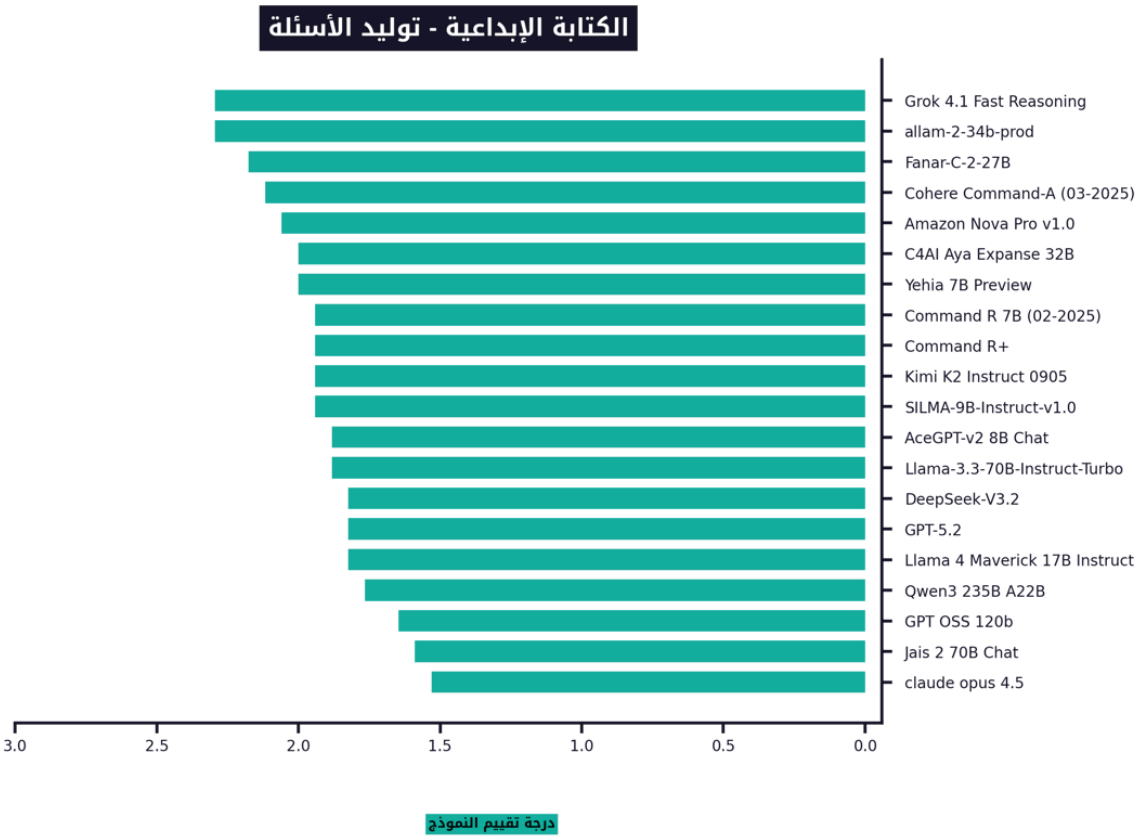
١.٥ توليد المقالات الإخبارية (News Article Generation).



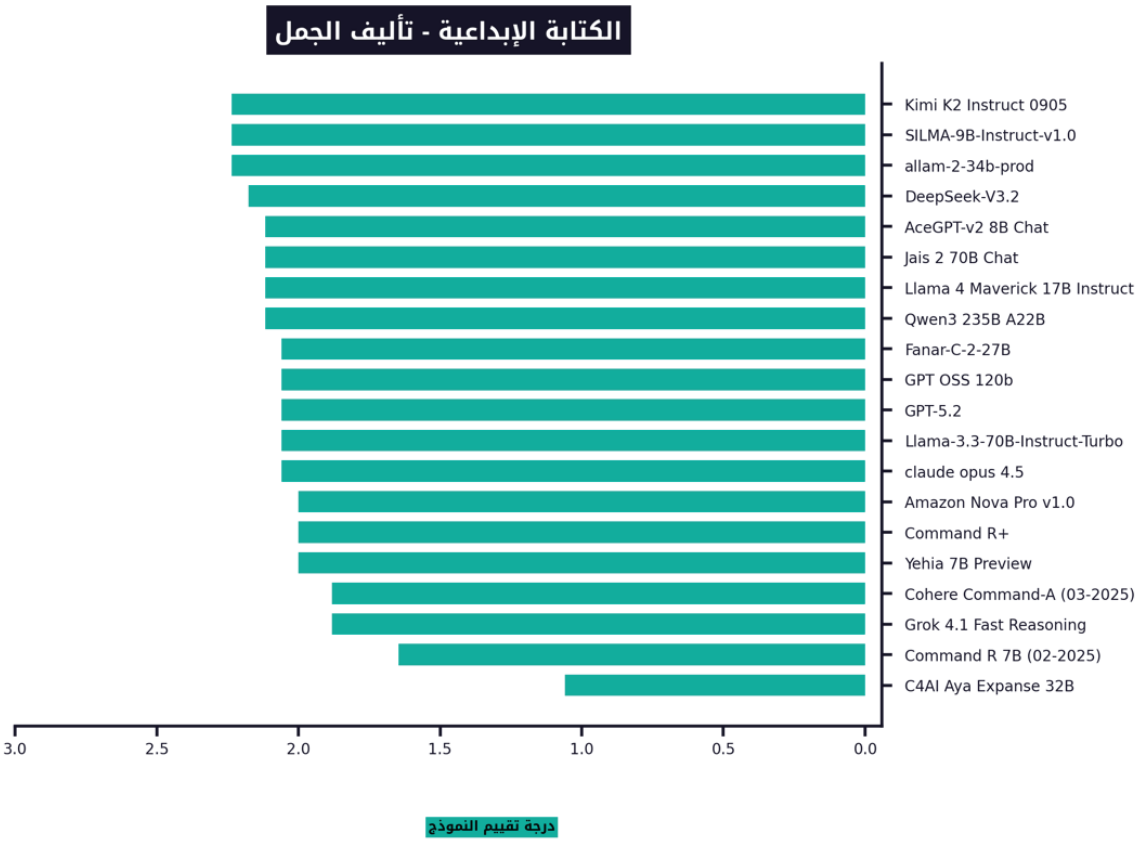
١.٦ توليد القصائد (Poem Generation).



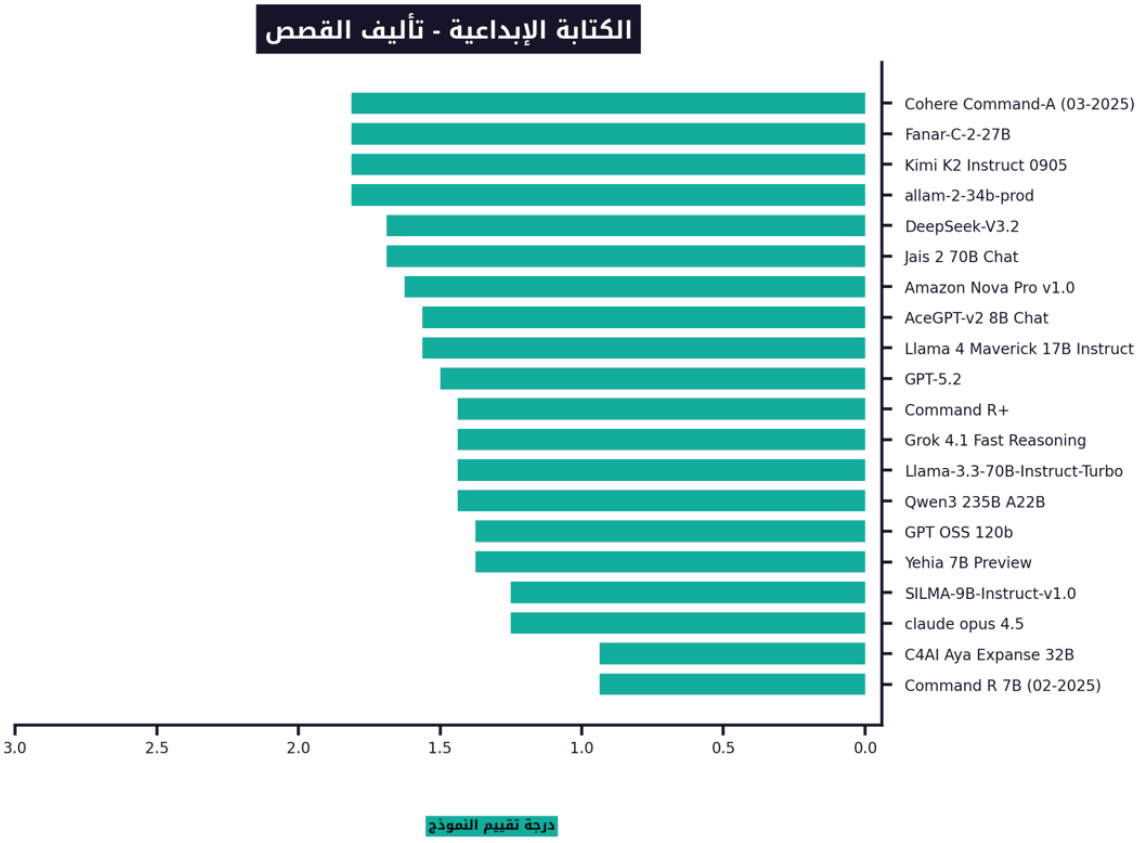
١.٧ توليد الأسئلة (Question Generation).



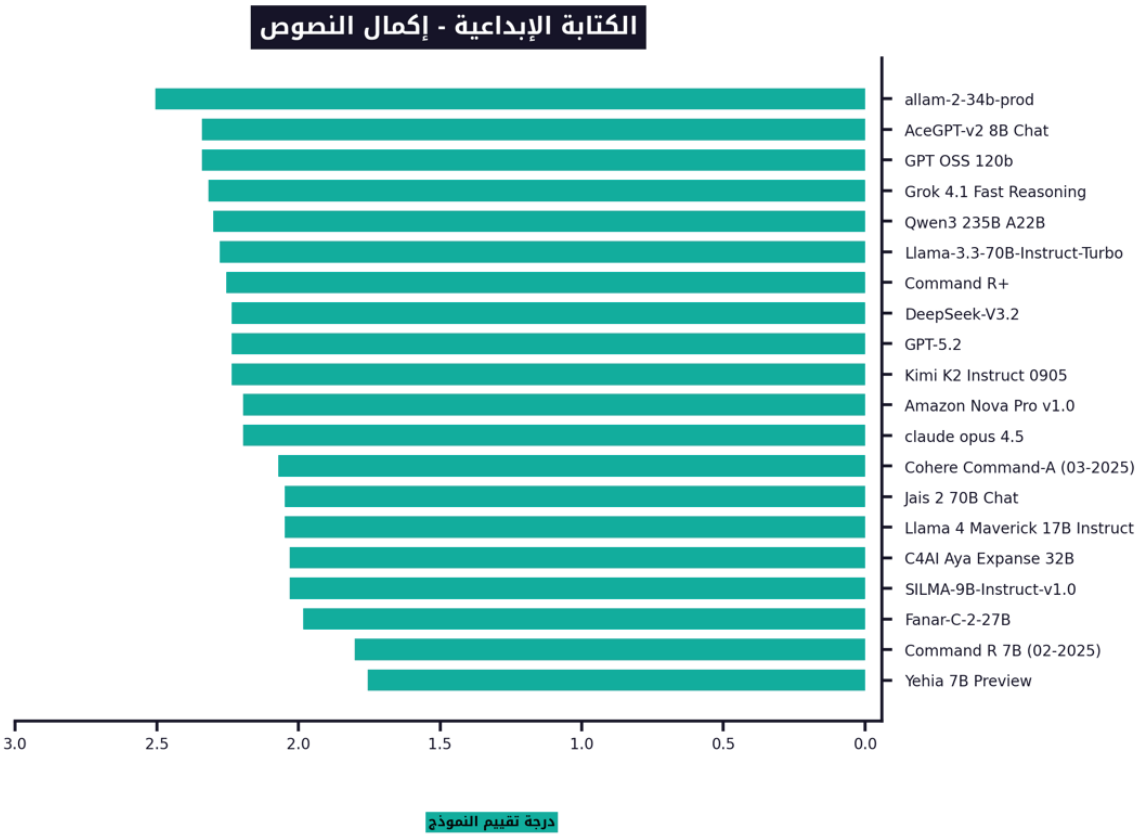
١.٨ تأليف الجمل (Sentence Composition).



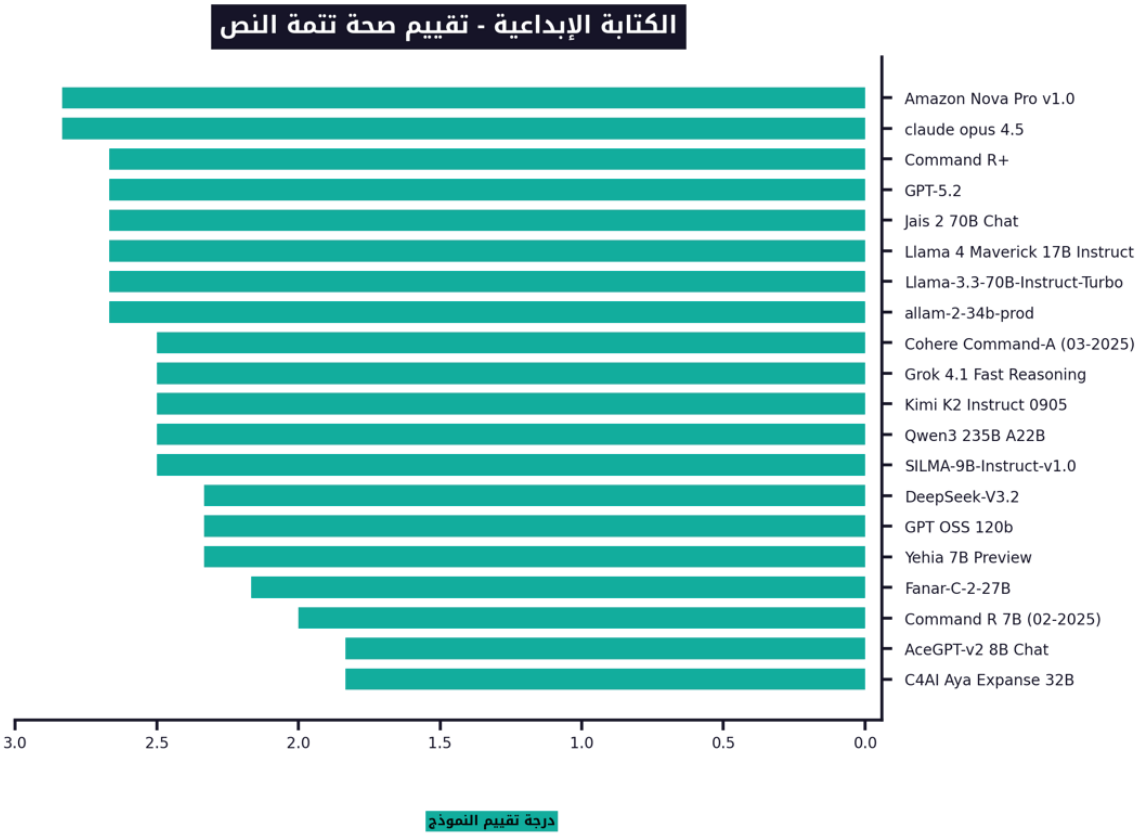
١.٩ تأليف القصص (Story Composition).



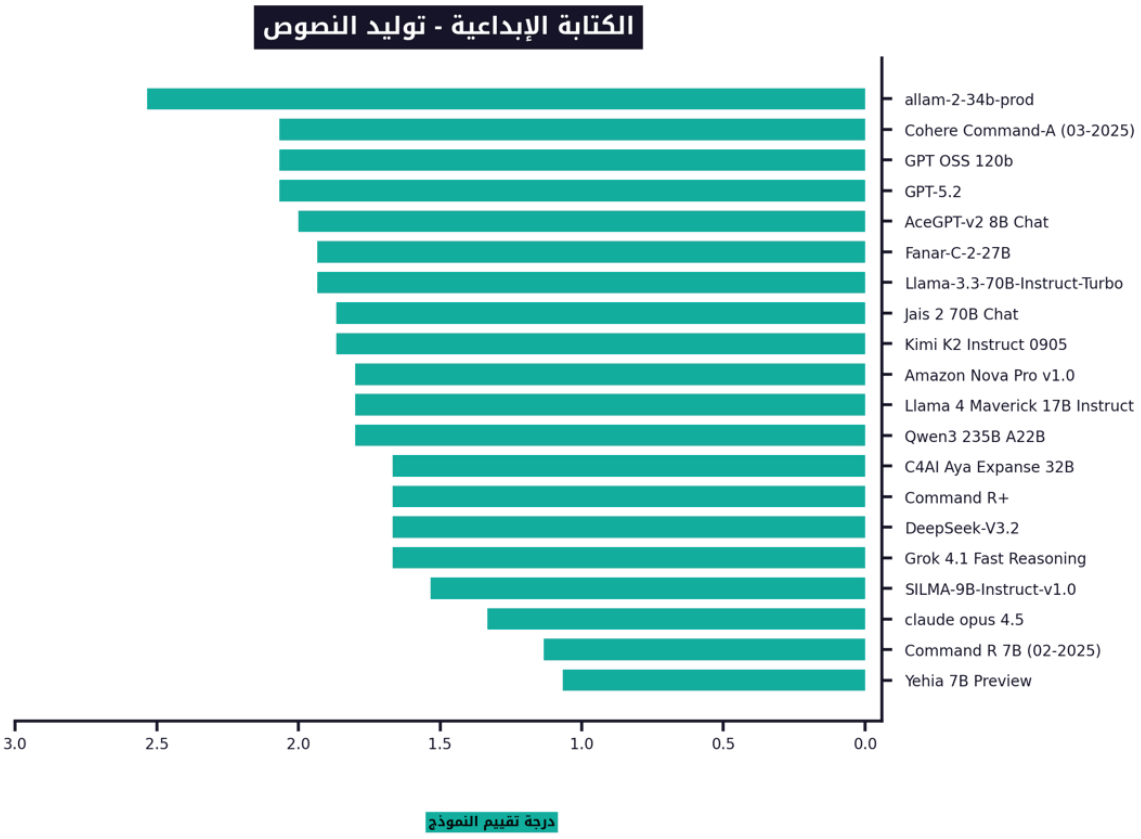
١.١٠ إكمال النصوص (Text Completion).



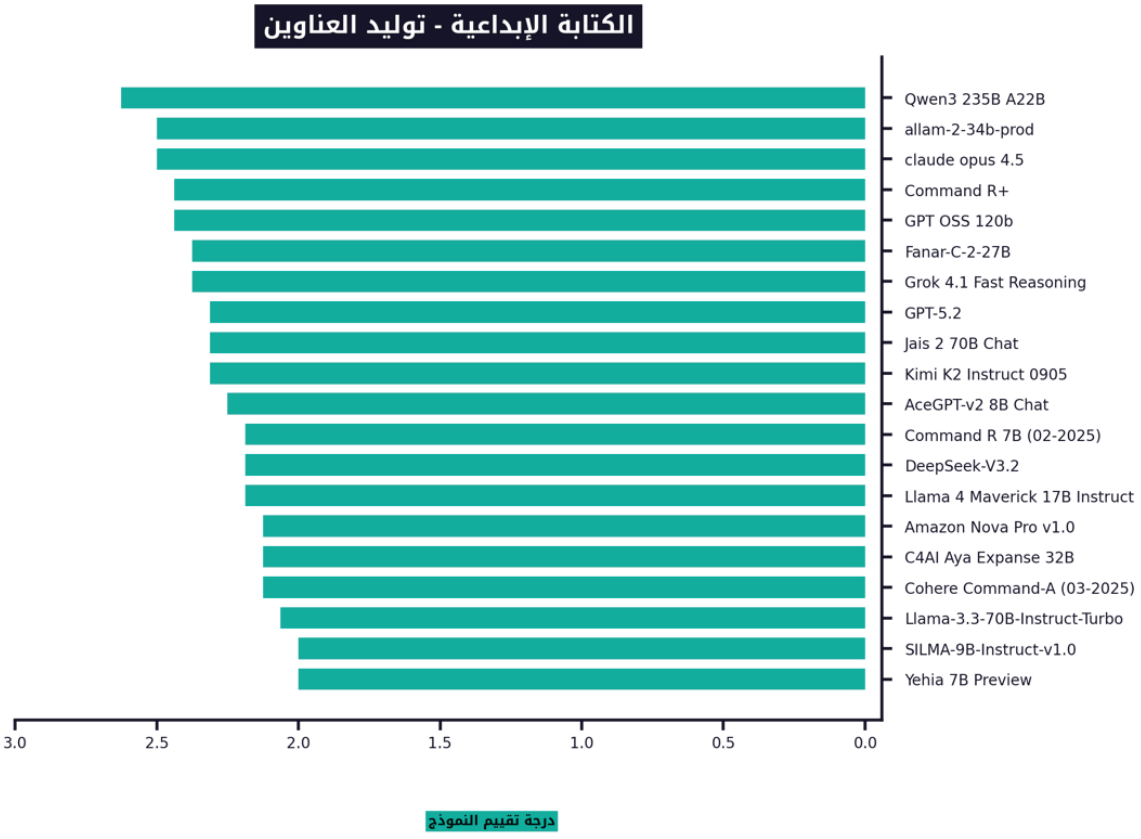
١.١١ تقييم صحة تتمة النص (Text Continuation Evaluation).



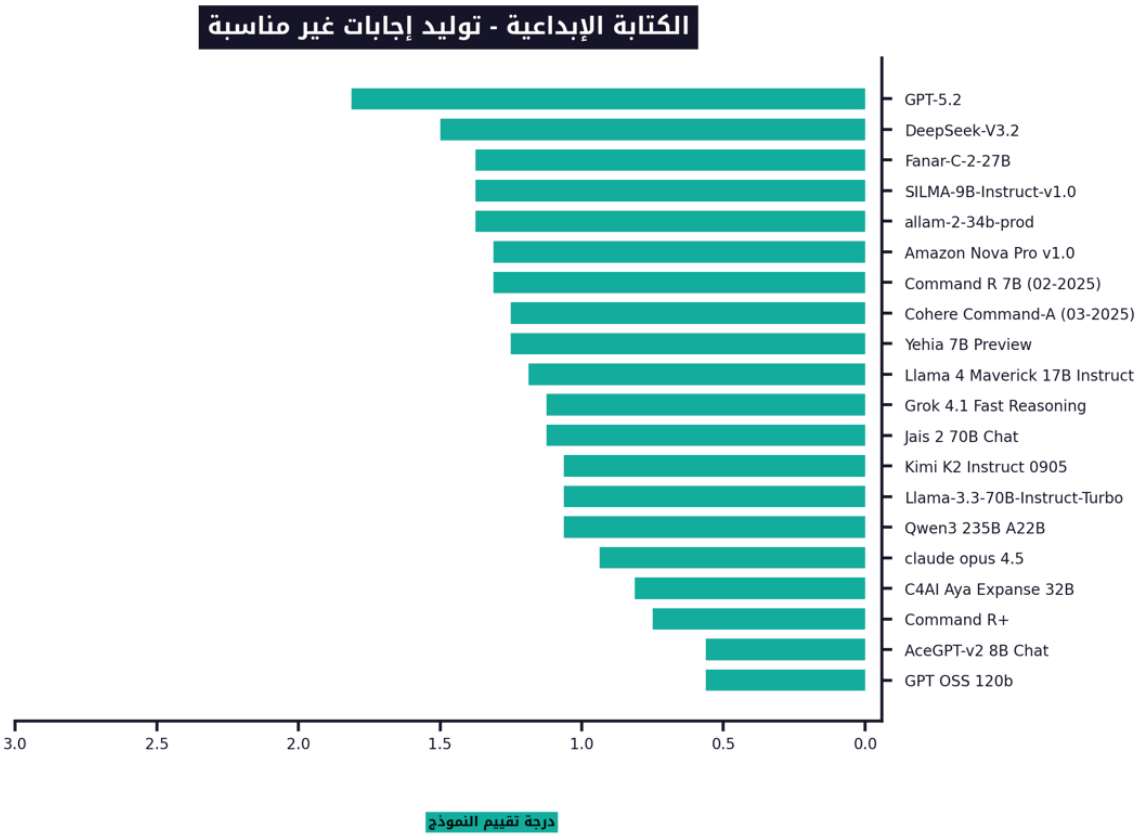
١.١٢ توليد النصوص (Text Generation).



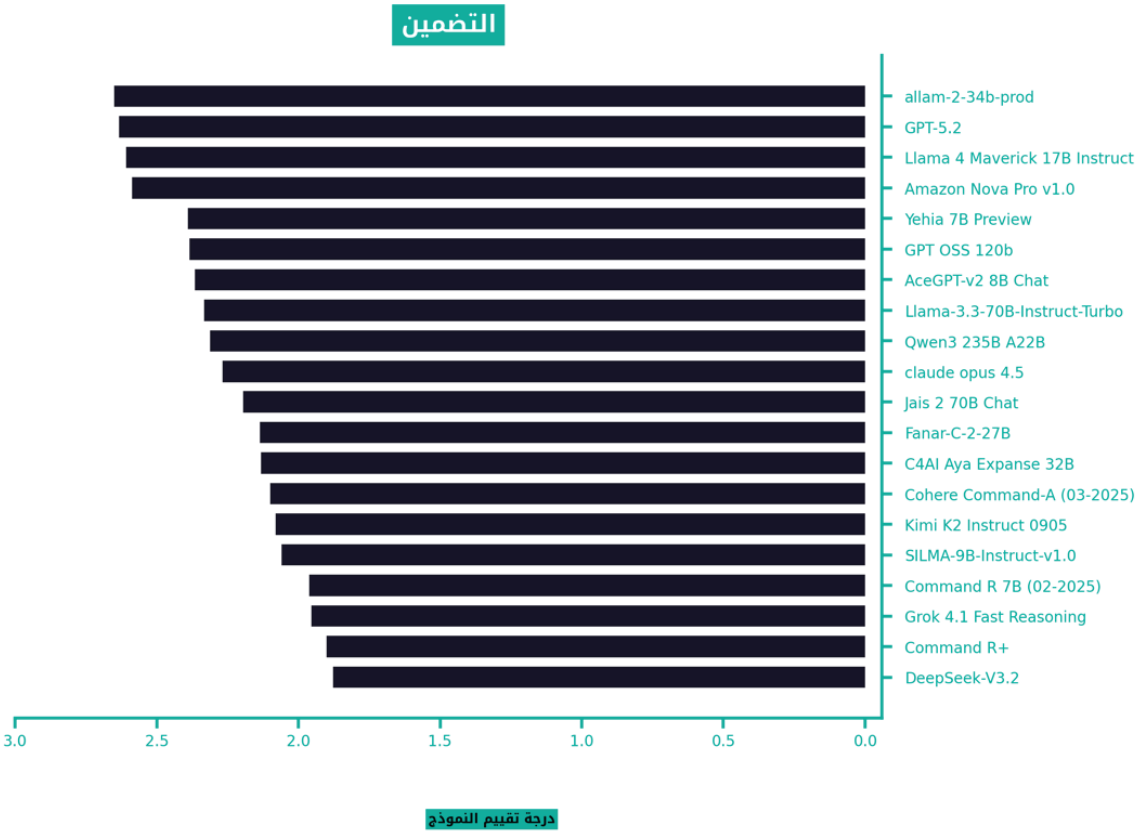
١.١٣ توليد العناوين (Title Generation).



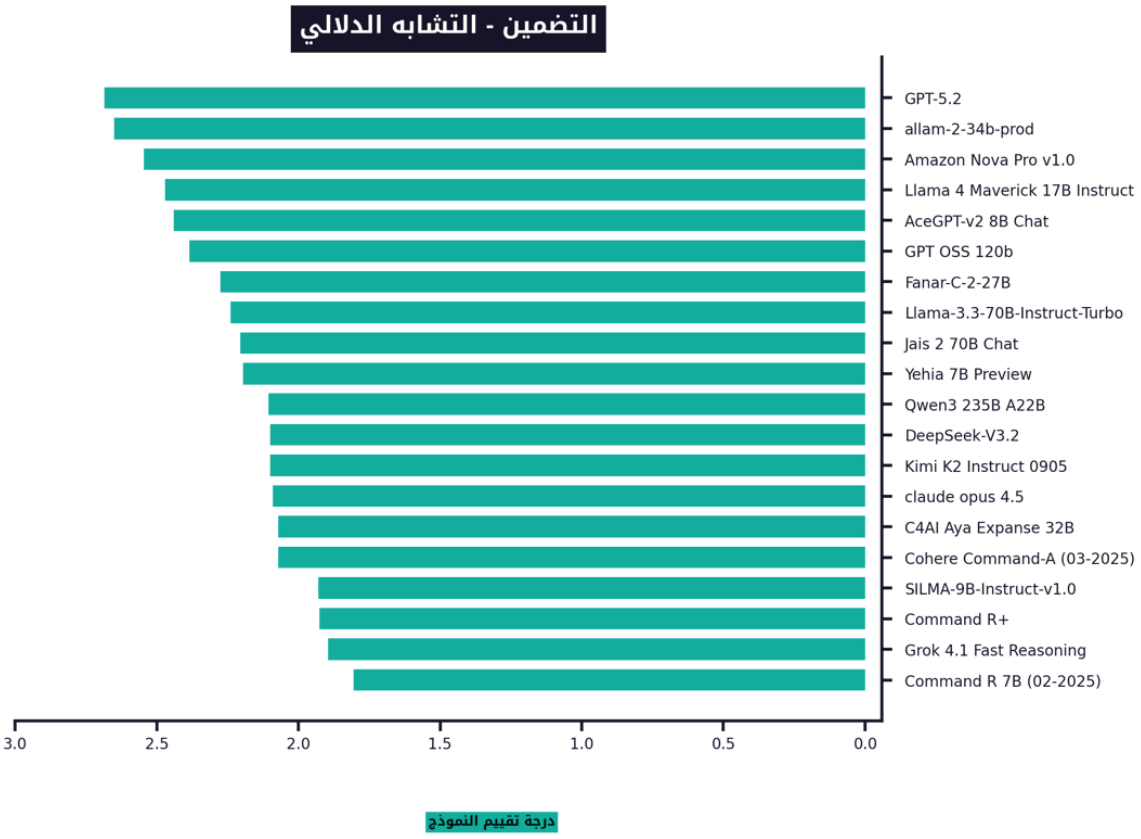
١.١٤ توليد إجابات غير مناسبة (Wrong Candidate Generation).



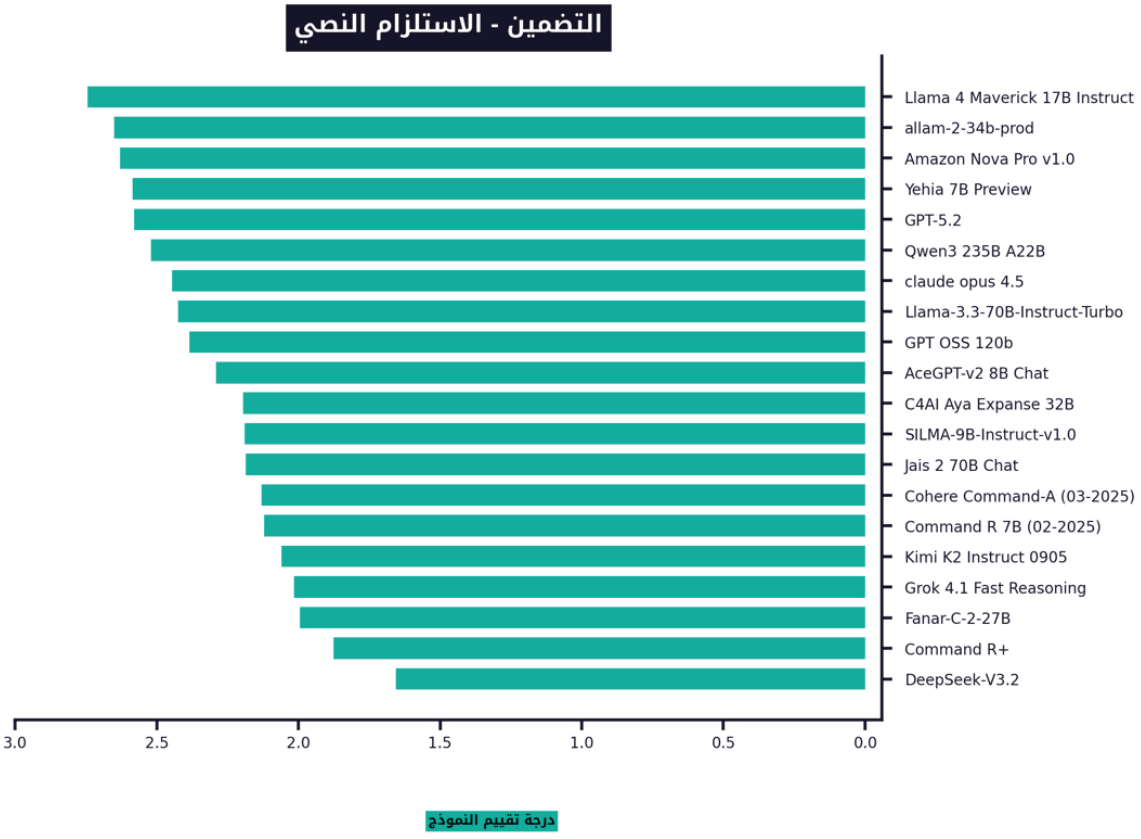
٢. التضمين (Entailment).



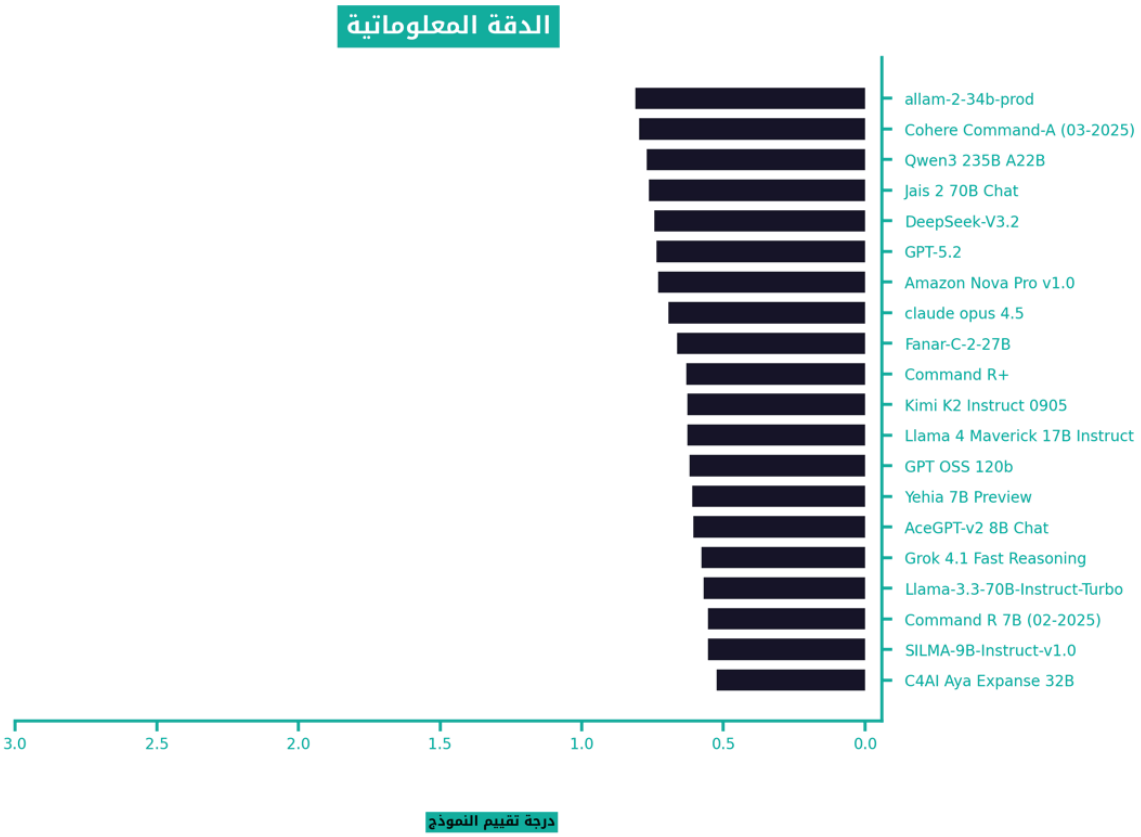
٢.١ التشابه الدلالي (Semantic Similarity).



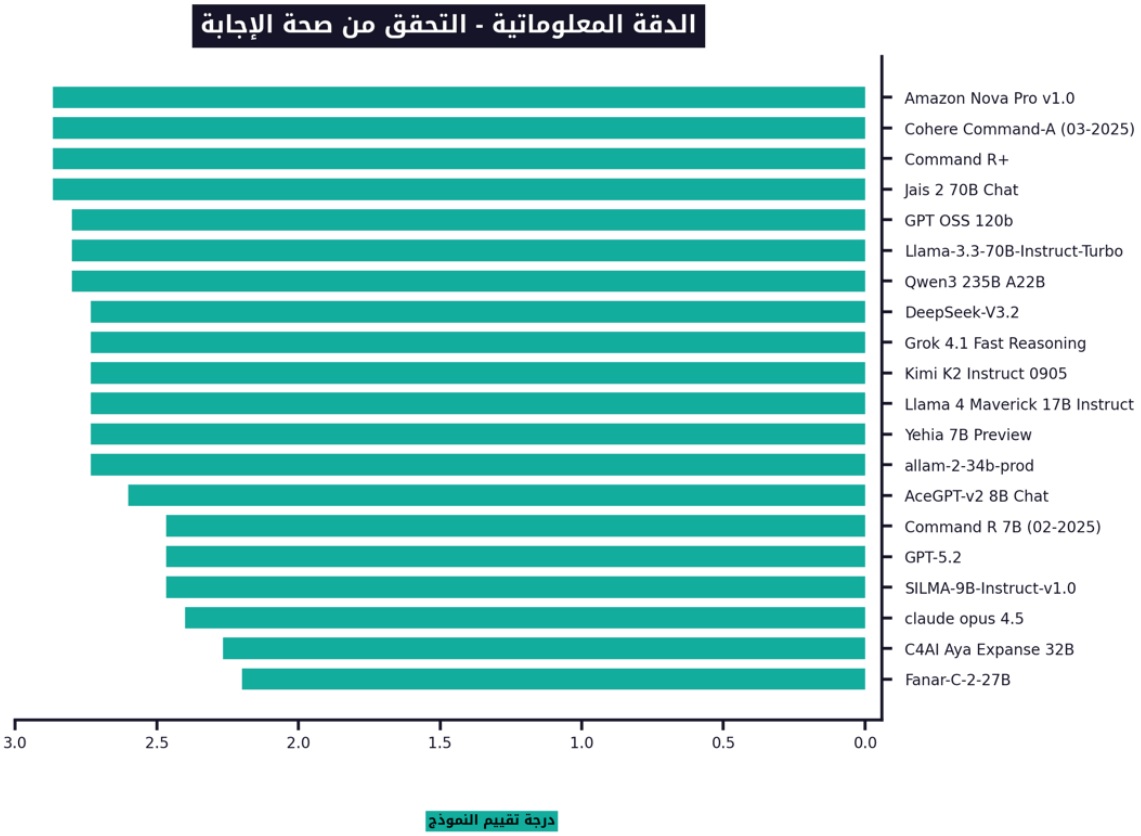
٢.٢ الاستلزام النصي (Textual Entailment).



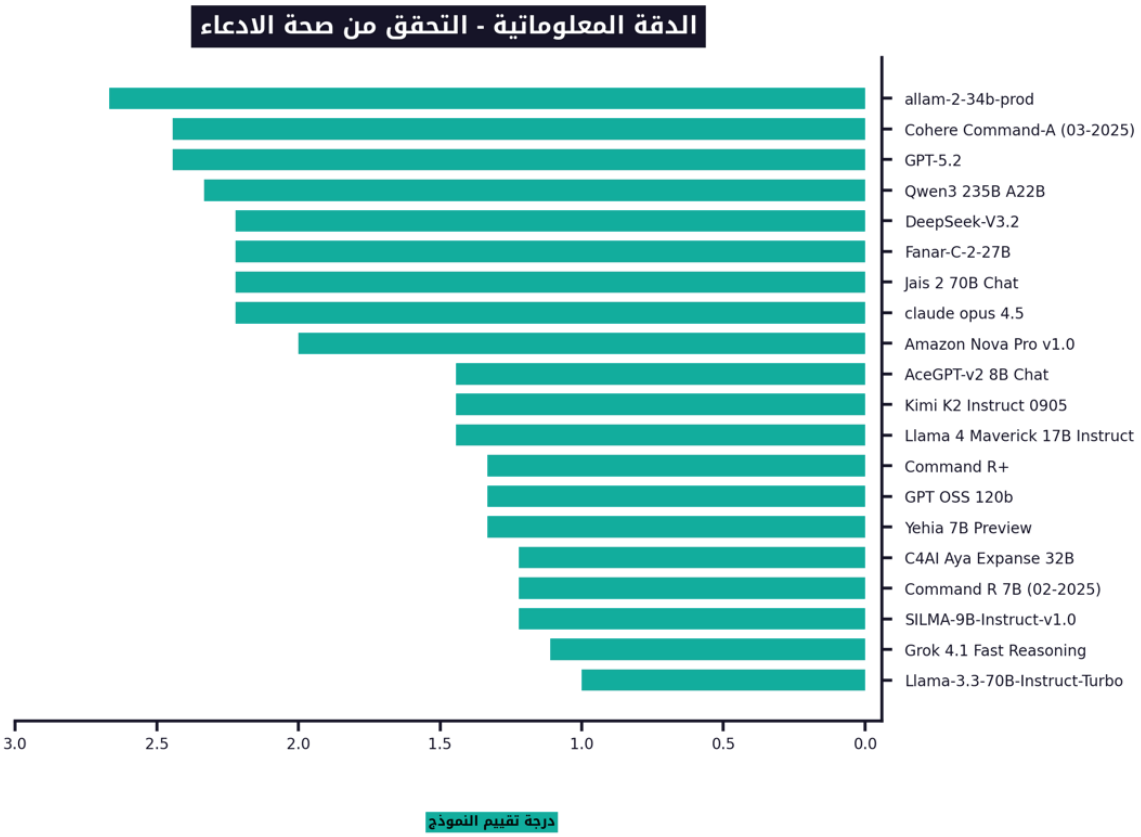
٣. الدقة المعلوماتية (Factuality).



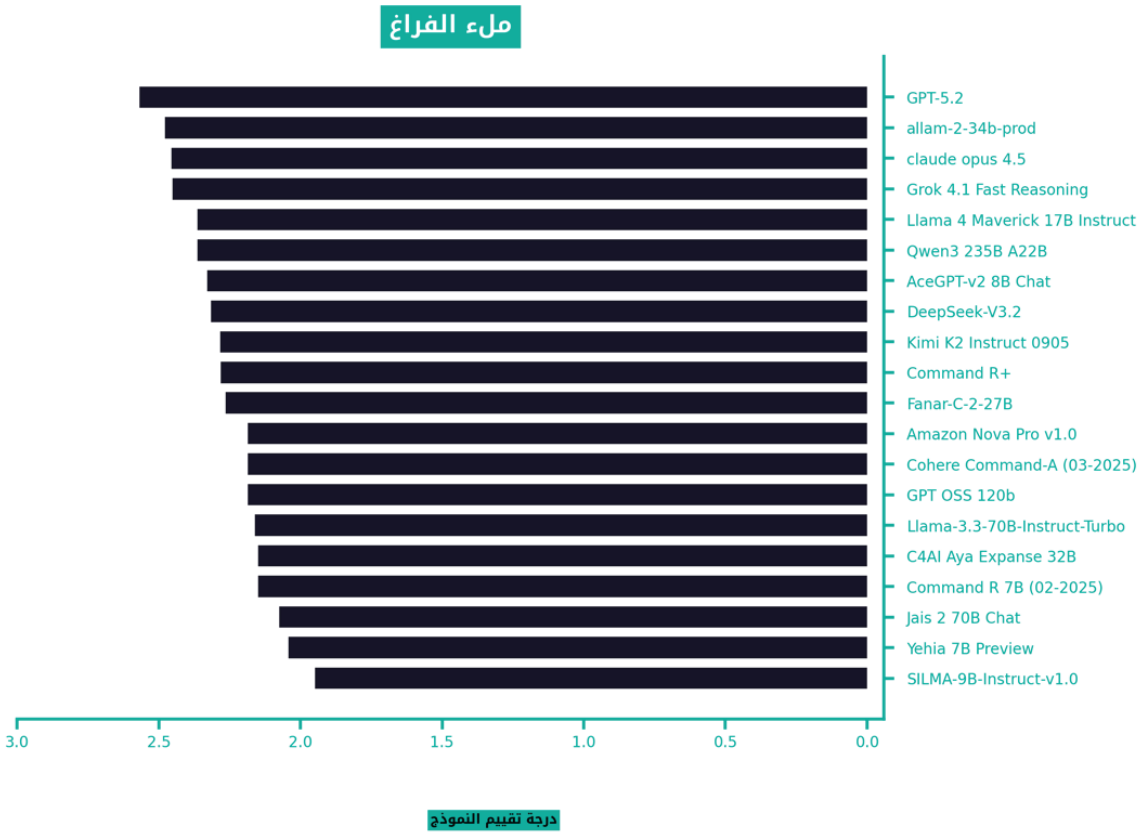
٣.١ التحقق من صحة الإجابة (Answer Verification).



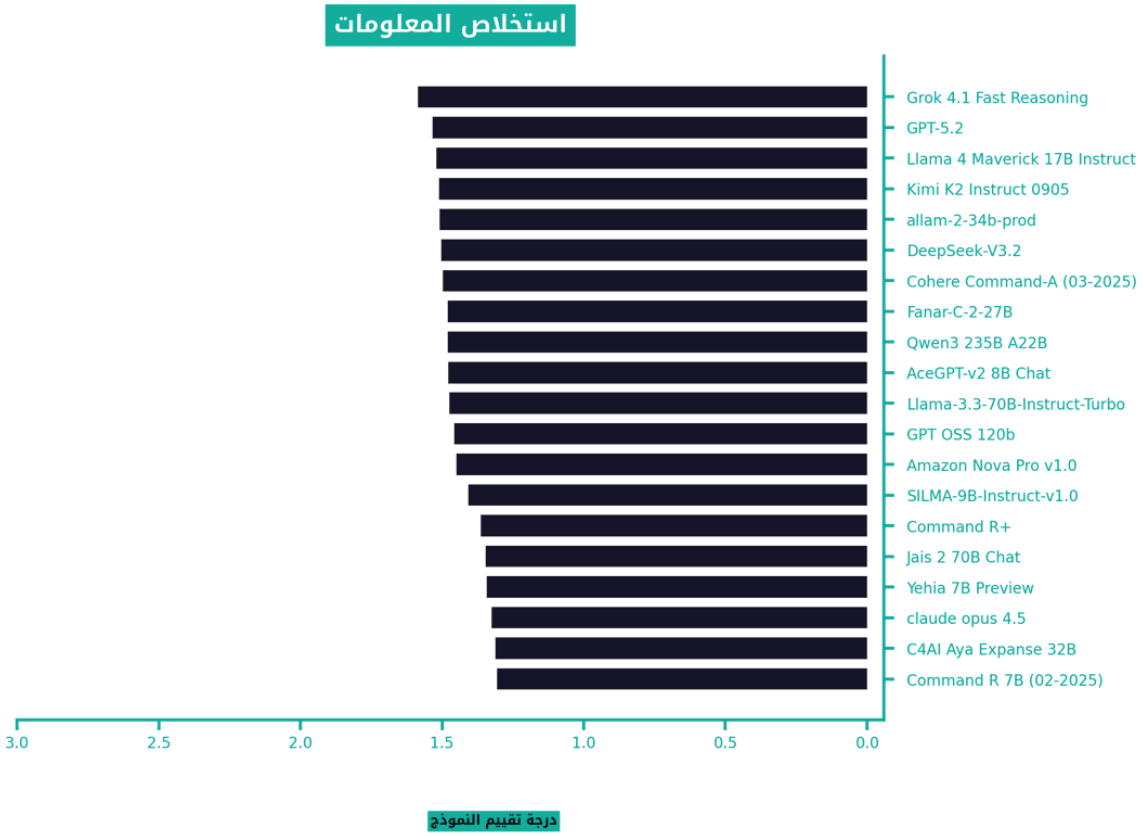
٣.٢ التحقق من صحة الادعاء (Claim Verification).



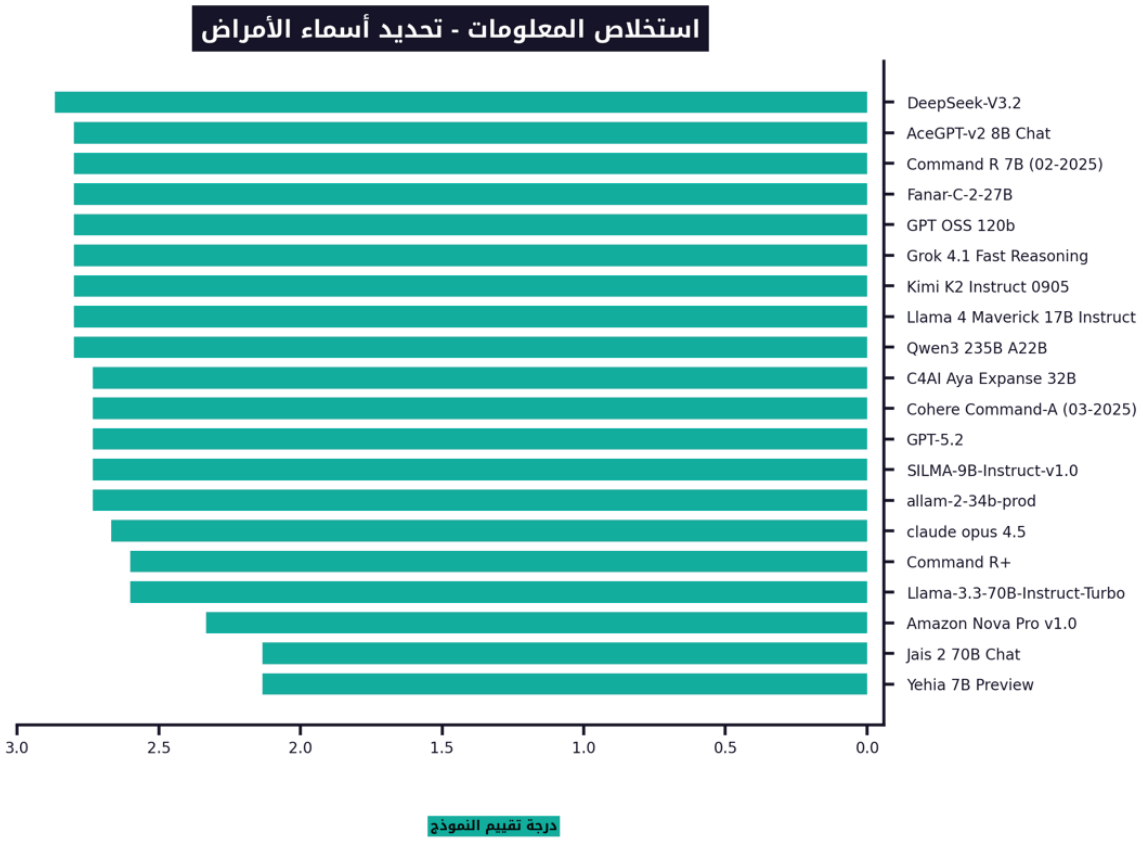
٤. ملء الفراغ (Fill in the Blank).



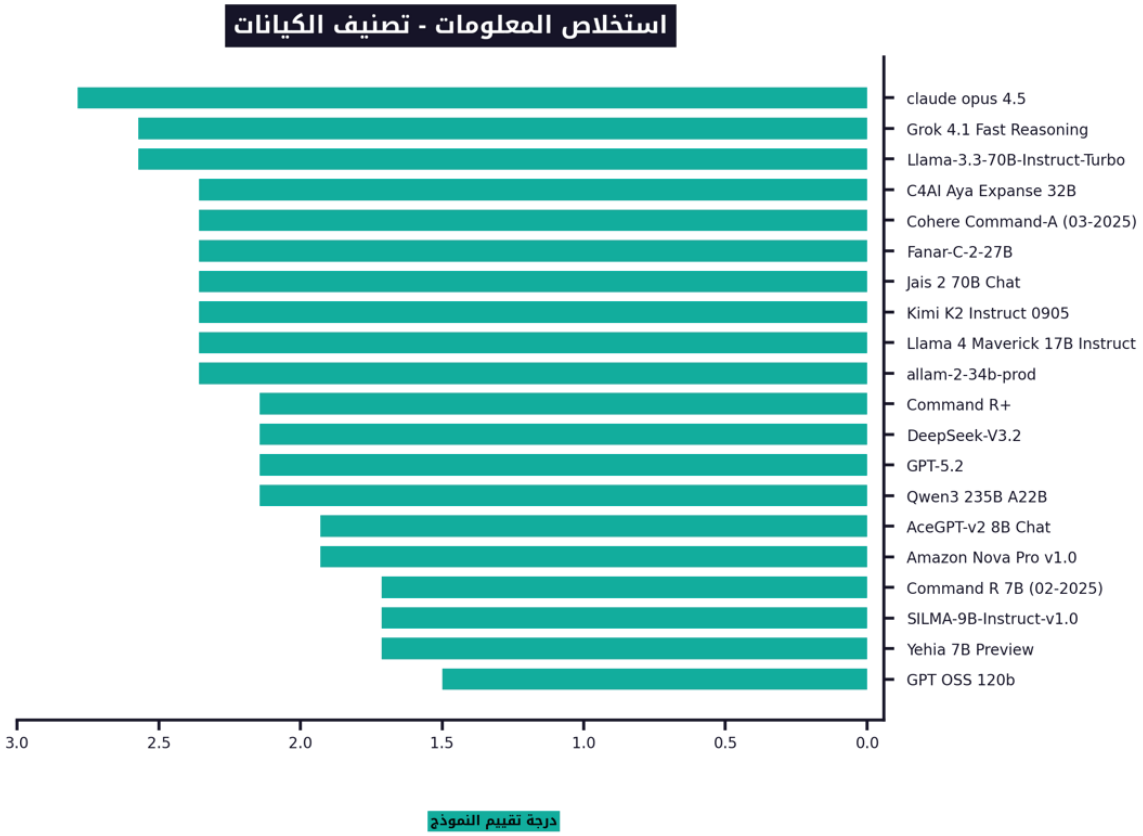
٥. استخلاص المعلومات (Information Extraction).



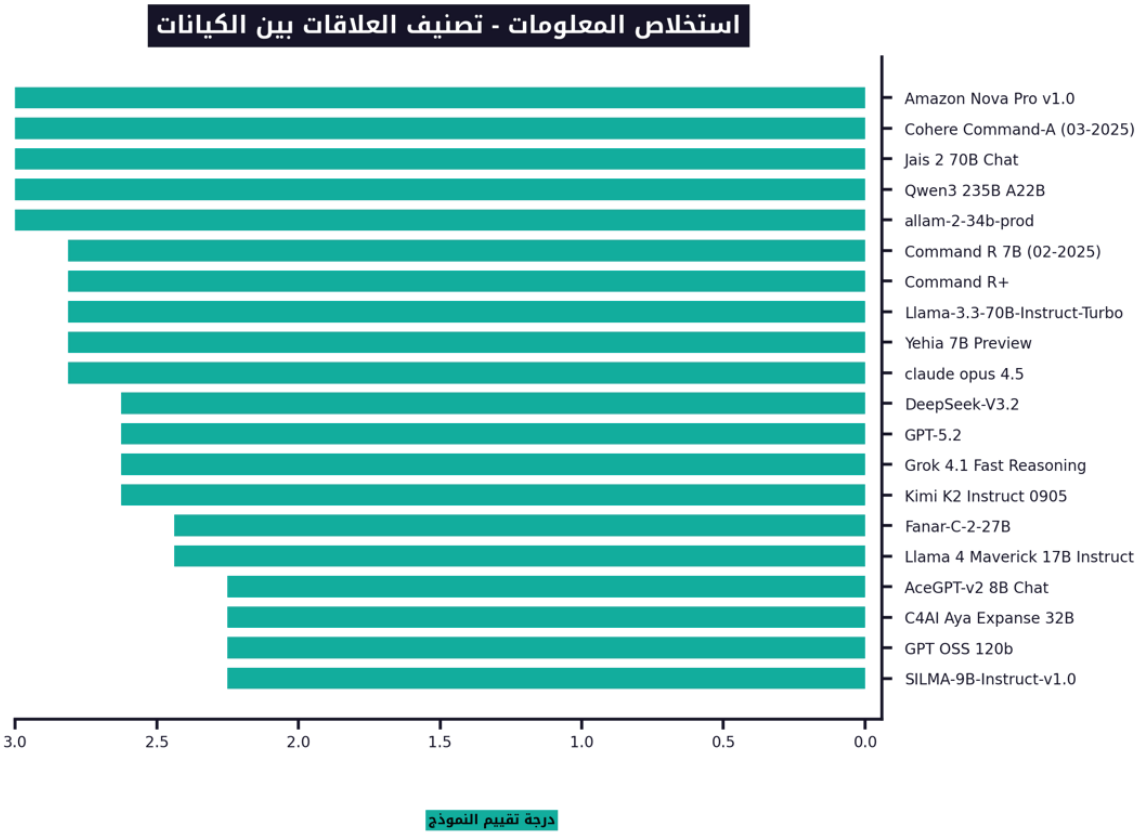
٥.١ تحديد أسماء الأمراض (Disease Mention Identification).



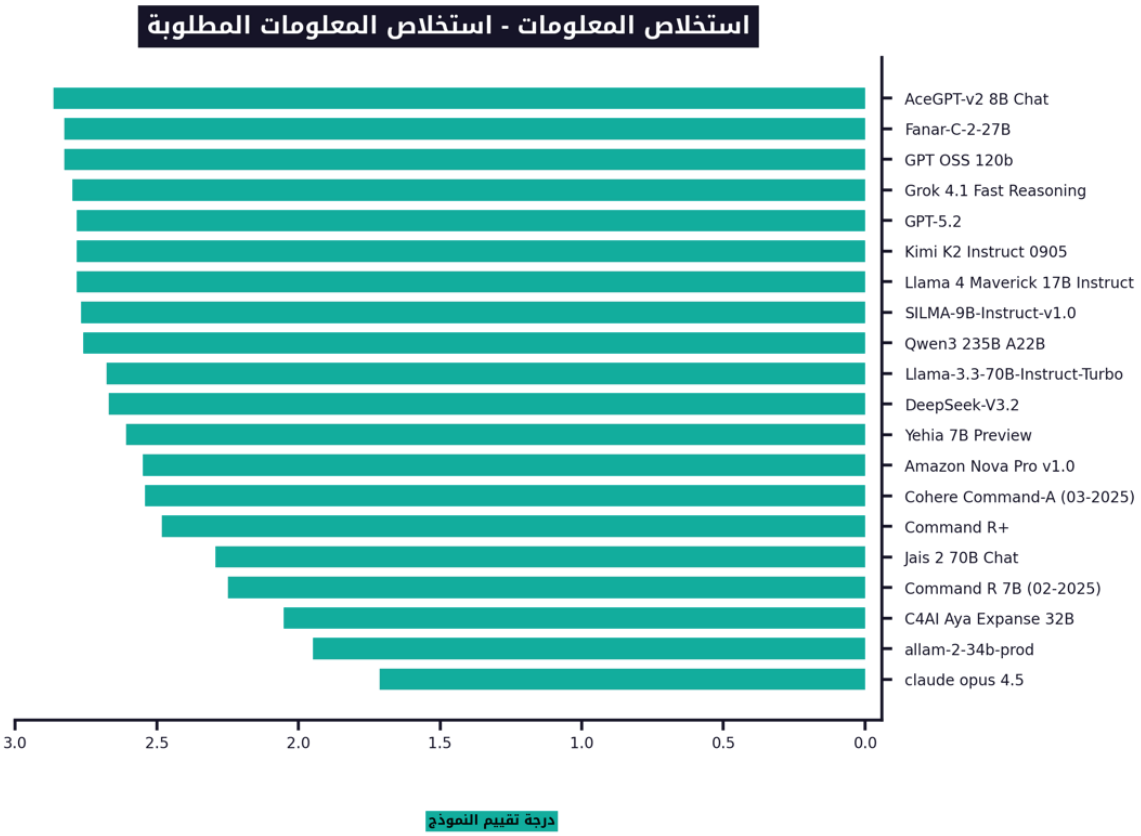
٥.٢ تصنيف الكيانات (Entity Categorization).



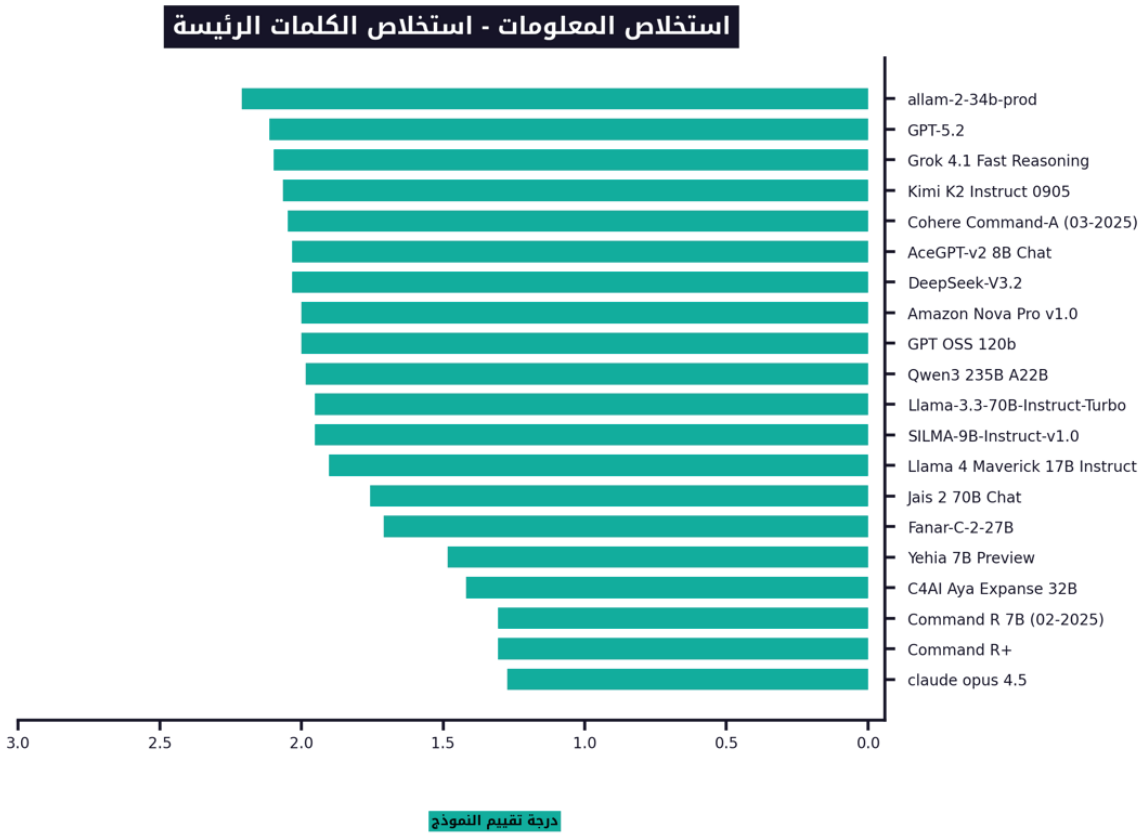
٥.٣ تصنيف العلاقات بين الكيانات (Entity Relation Classification).



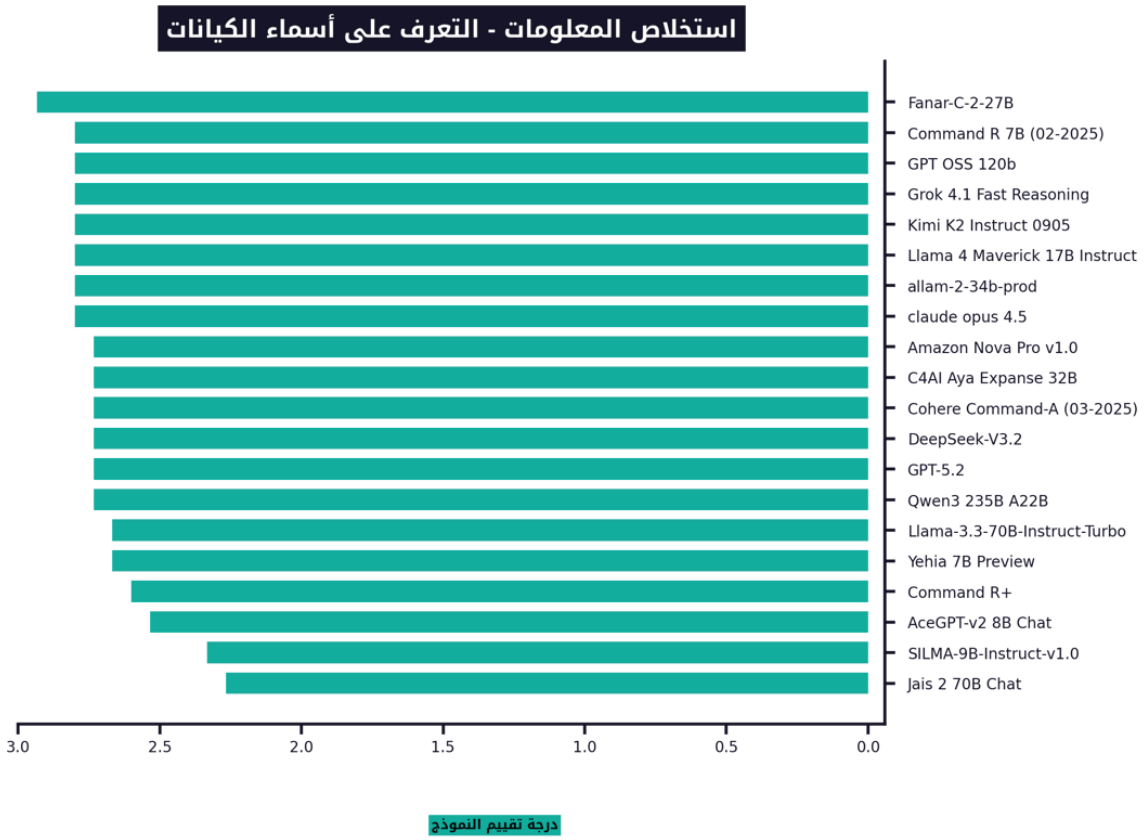
٥.٤ استخلاص المعلومات المطلوبة (Extracting Required Information).



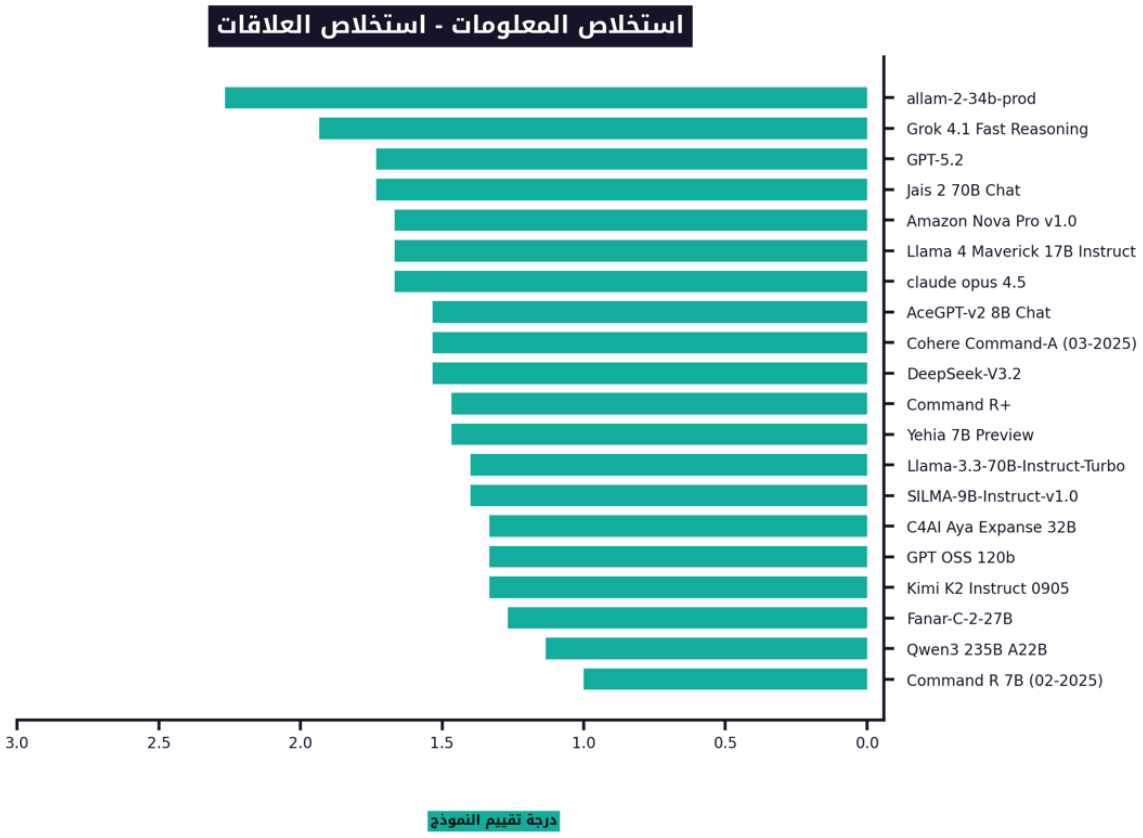
٥.٥ استخلاص الكلمات الرئيسة (Keyword Extraction).



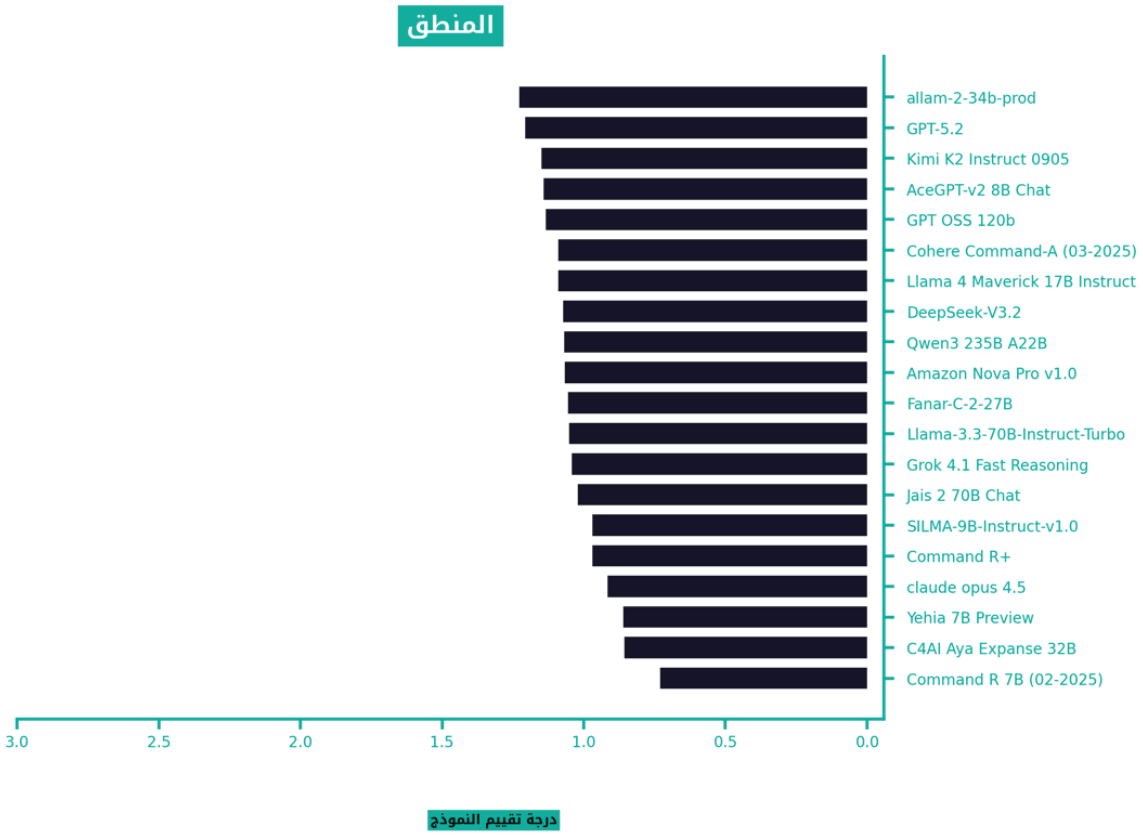
٥.٦ التعرف على أسماء الكيانات (Named Entity Recognition).



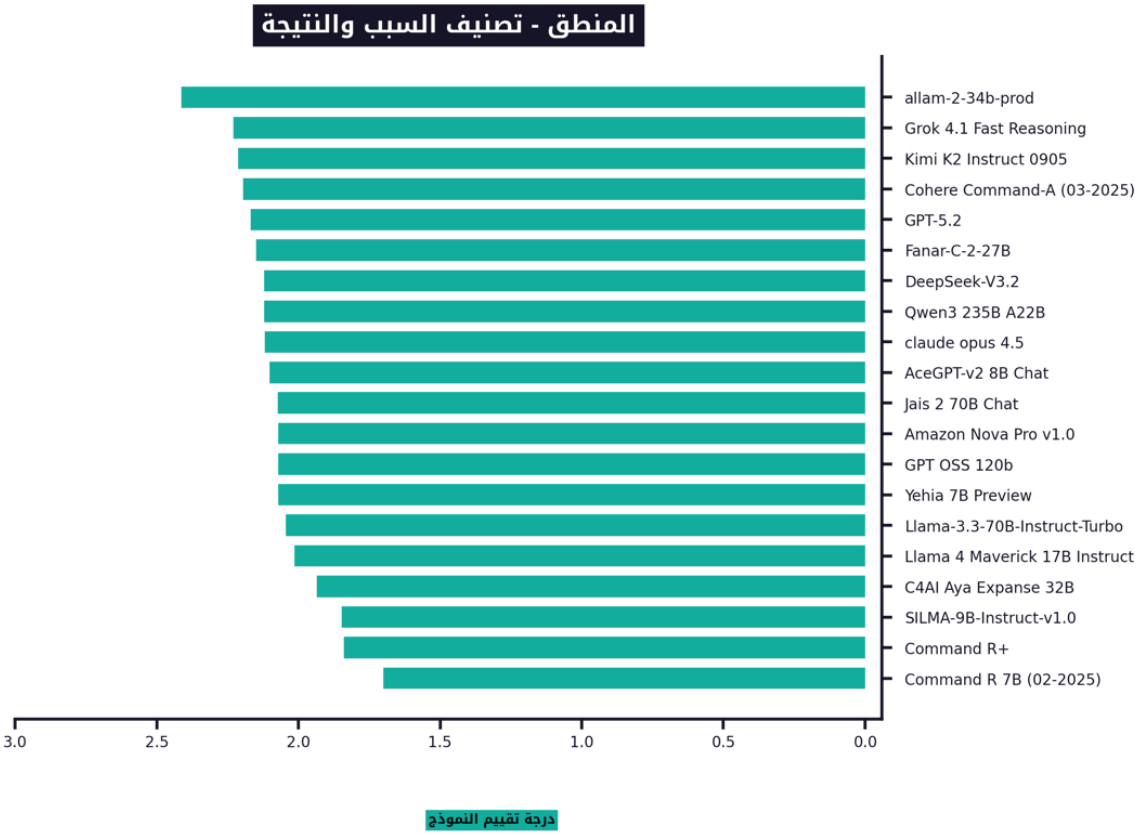
٥.٧ استخلاص العلاقات (Relation Extraction).



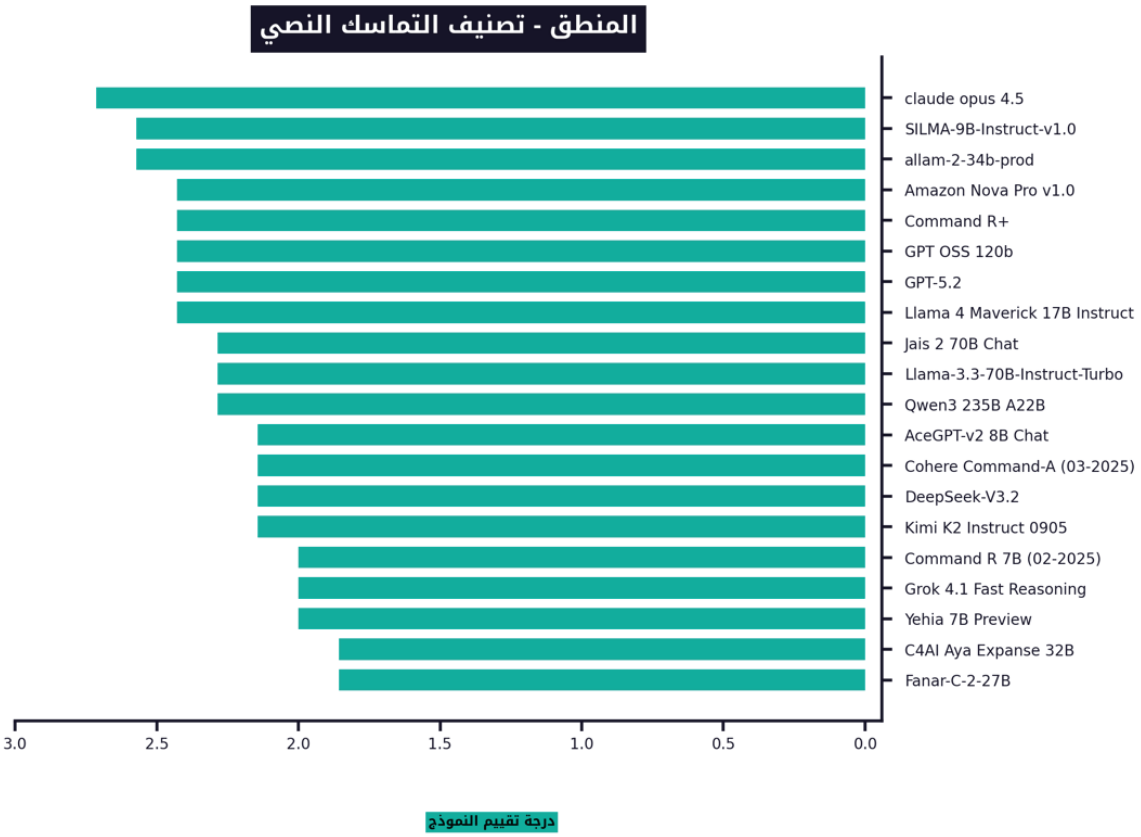
٦. المنطق (Logic).



٦.١ تصنيف السبب والنتيجة (Cause Effect Classification).

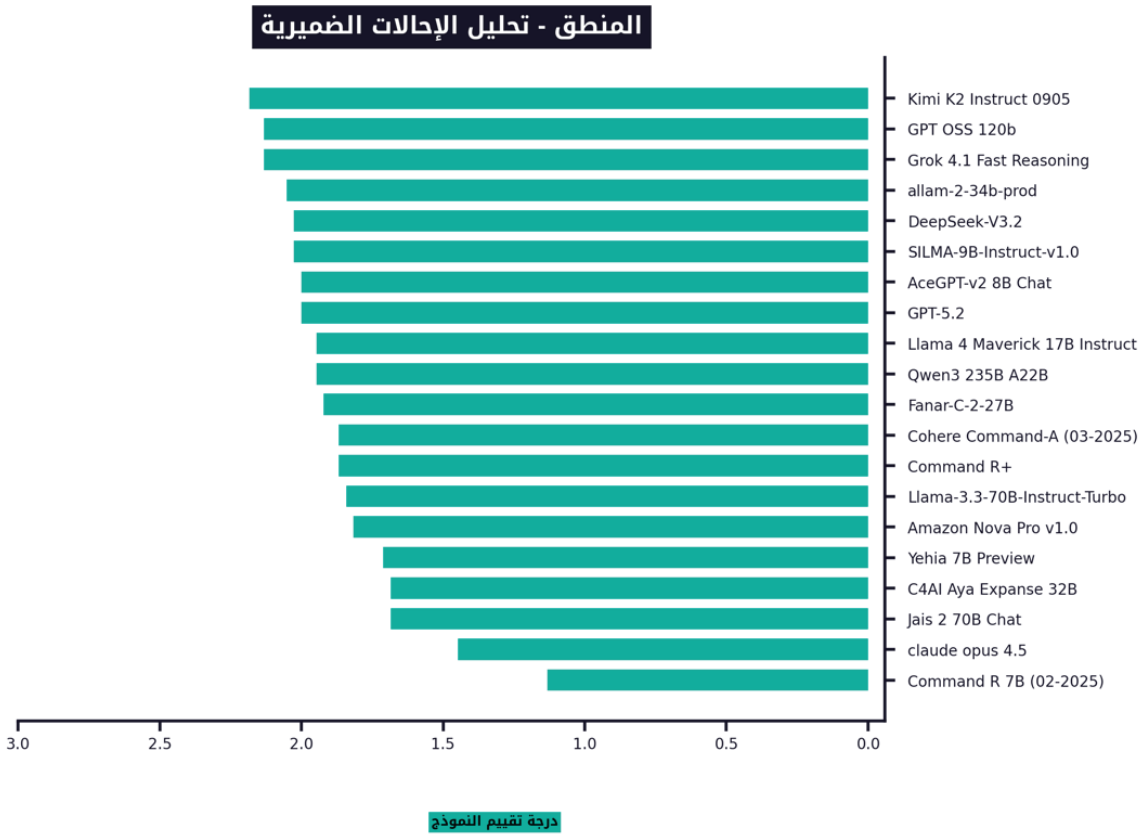


٦.٢ تصنيف التماسك النصي (Coherence Classification).

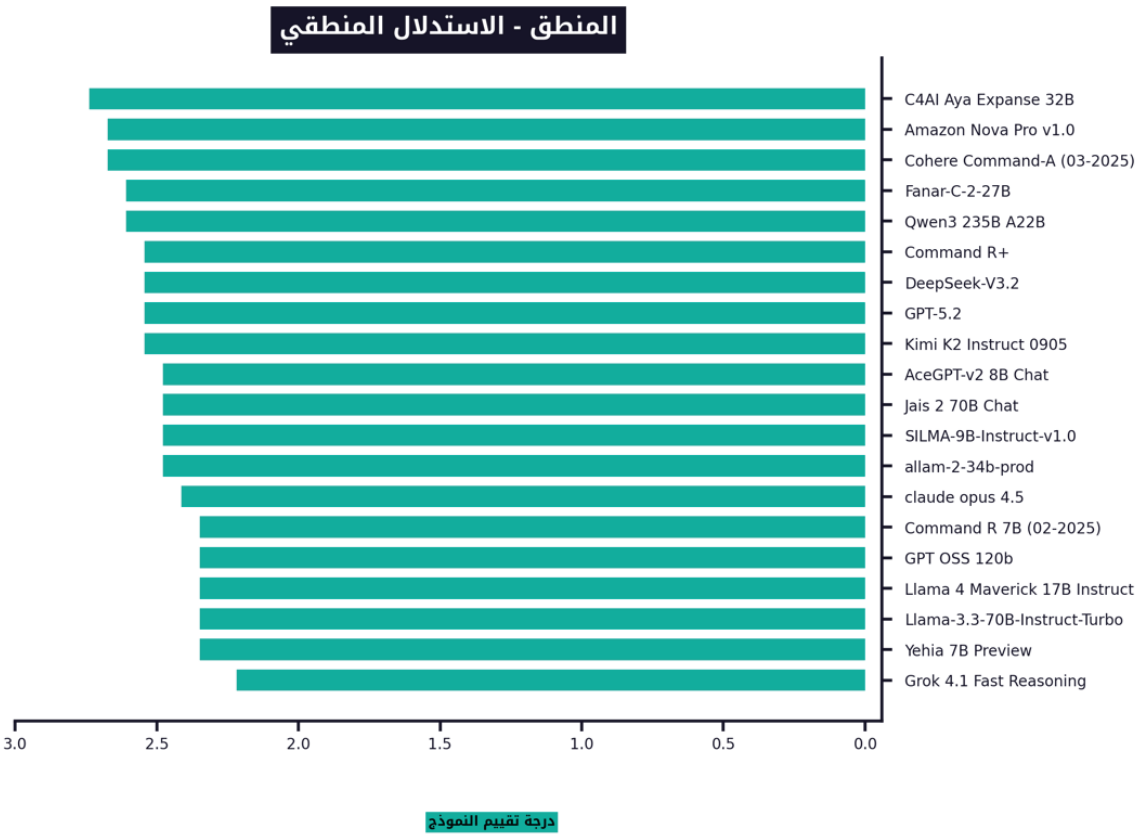


٦.٣ التحقق من المنطق السليم (Commonsense Validation).

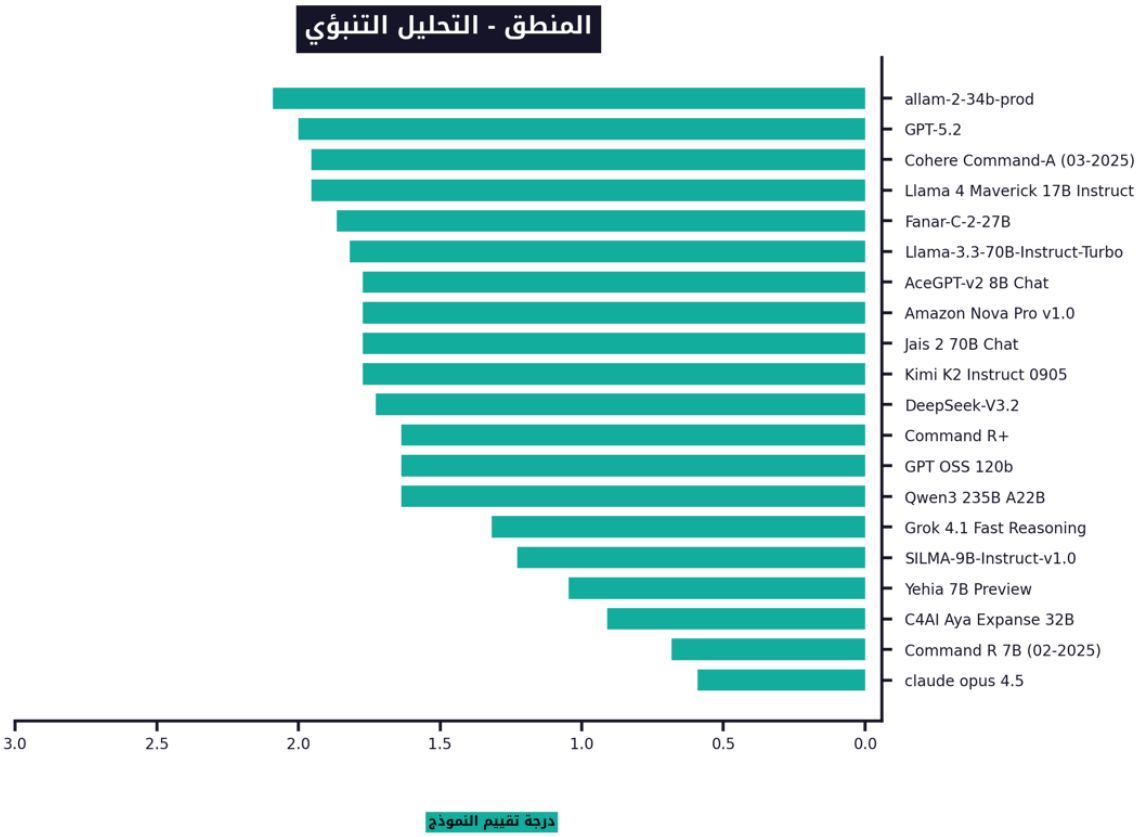
٦.٤ تحليل الإحالات الضميرية (Coreference Resolution).



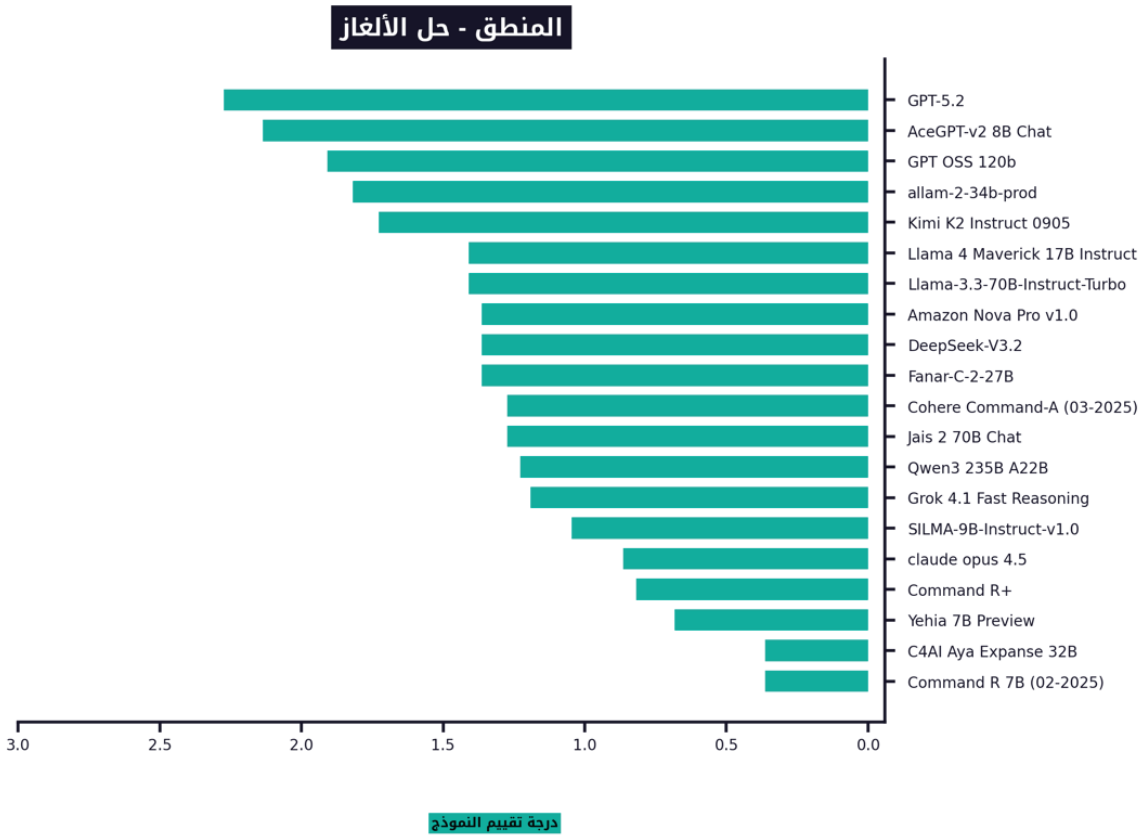
٦.٥ الاستدلال المنطقي (Logical Reasoning).



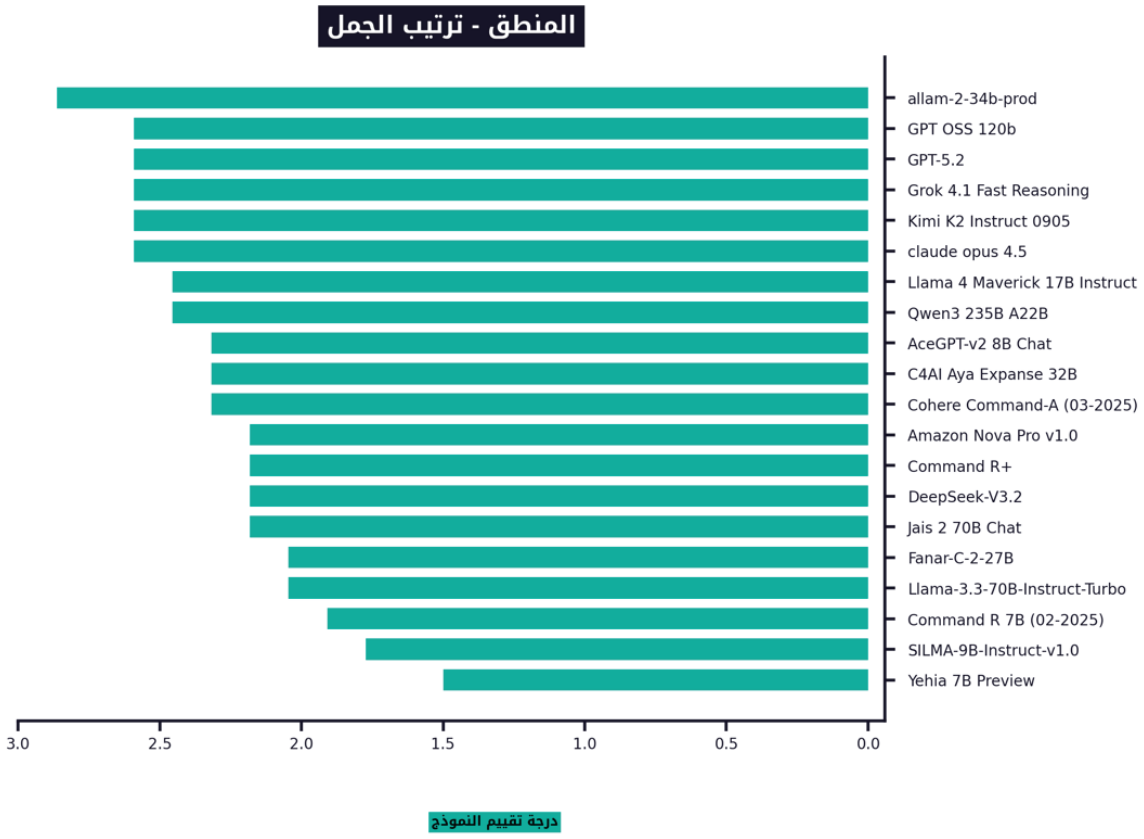
٦.٦ التحليل التنبؤي (Predictive Analysis).



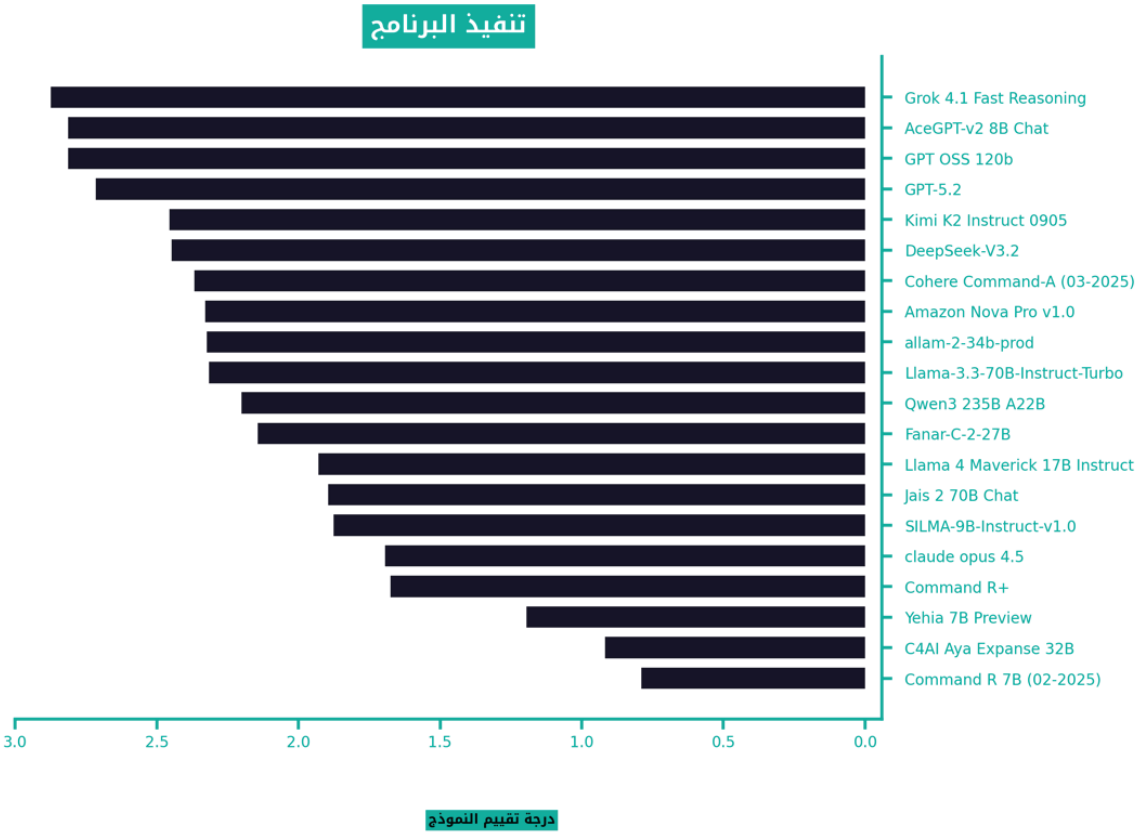
٦.٧ حل الألغاز (Riddle Solving).



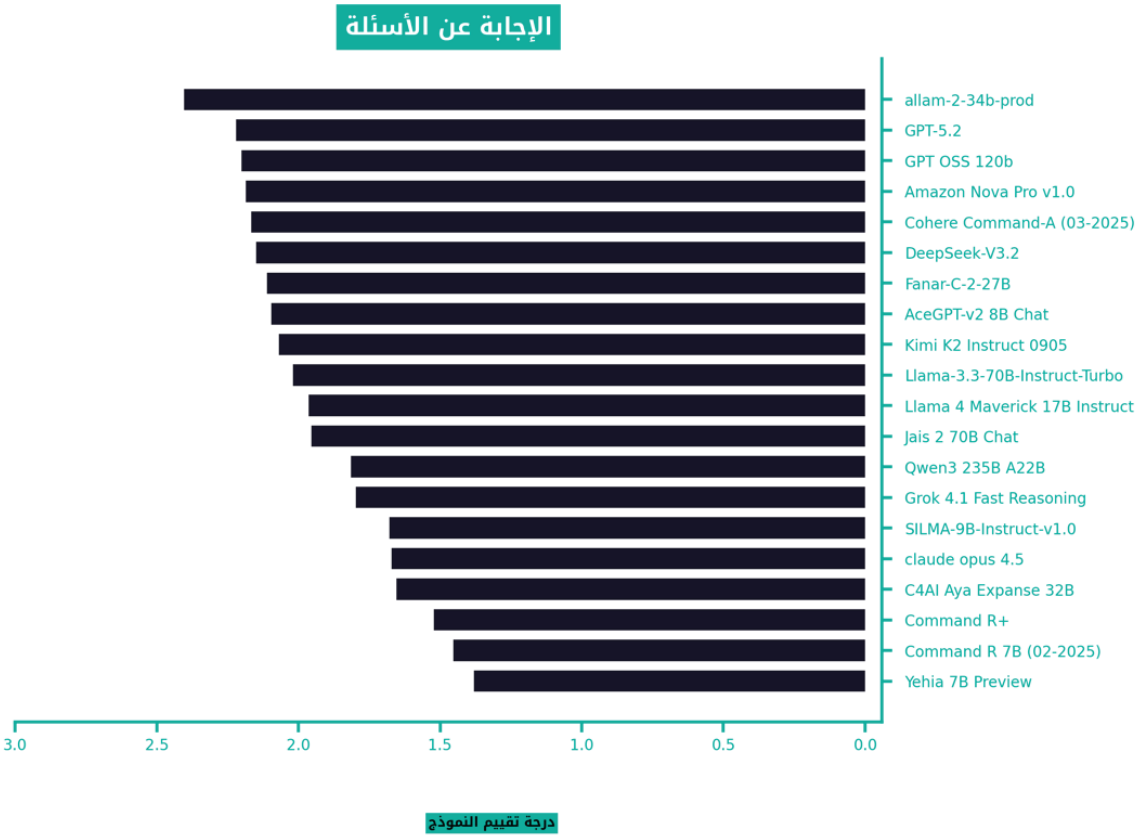
٦.٨ ترتيب الجمل (Sentence Ordering).



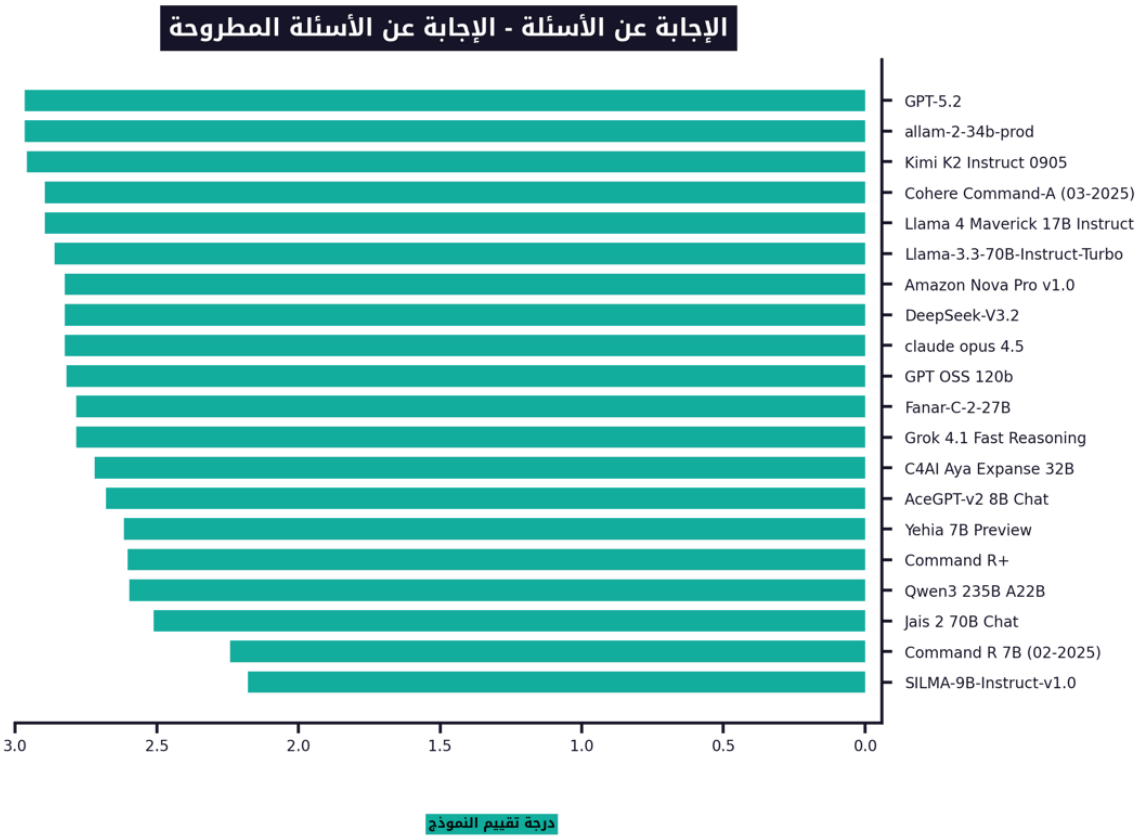
٧. تنفيذ البرامج (Program Execution).



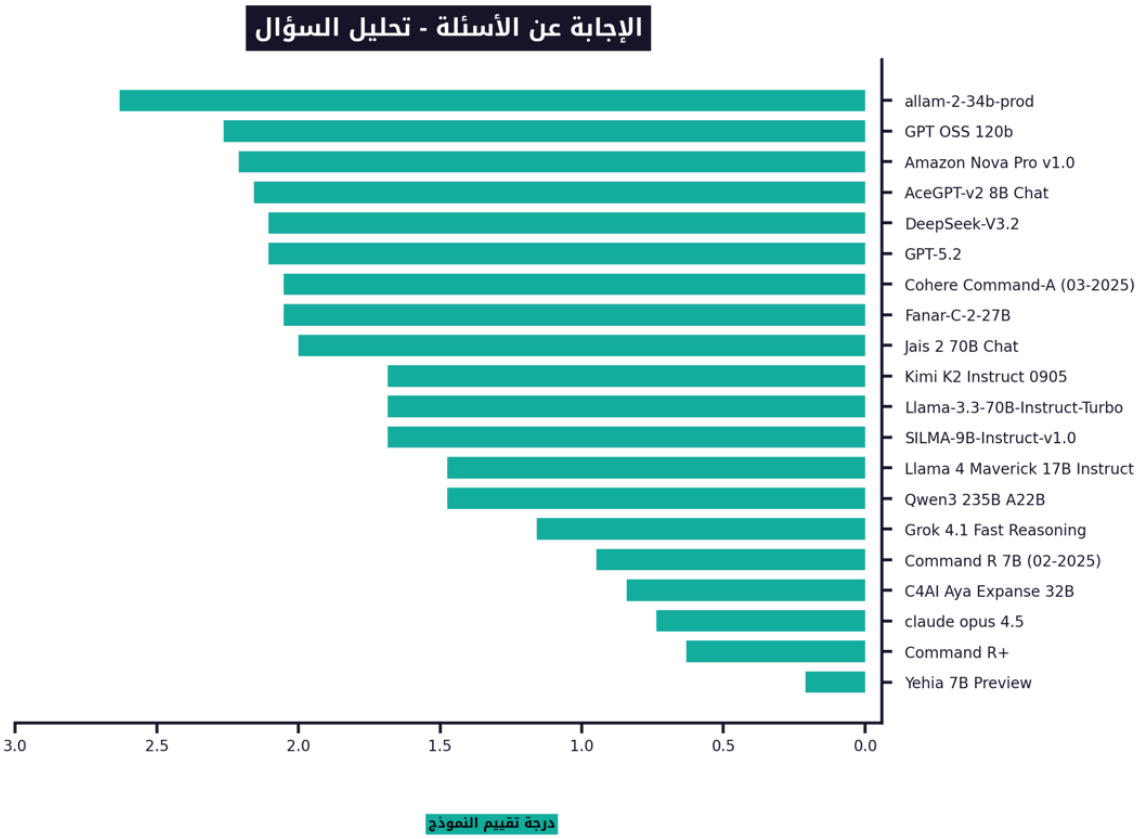
٨. الإجابة عن الأسئلة (Question Answering).



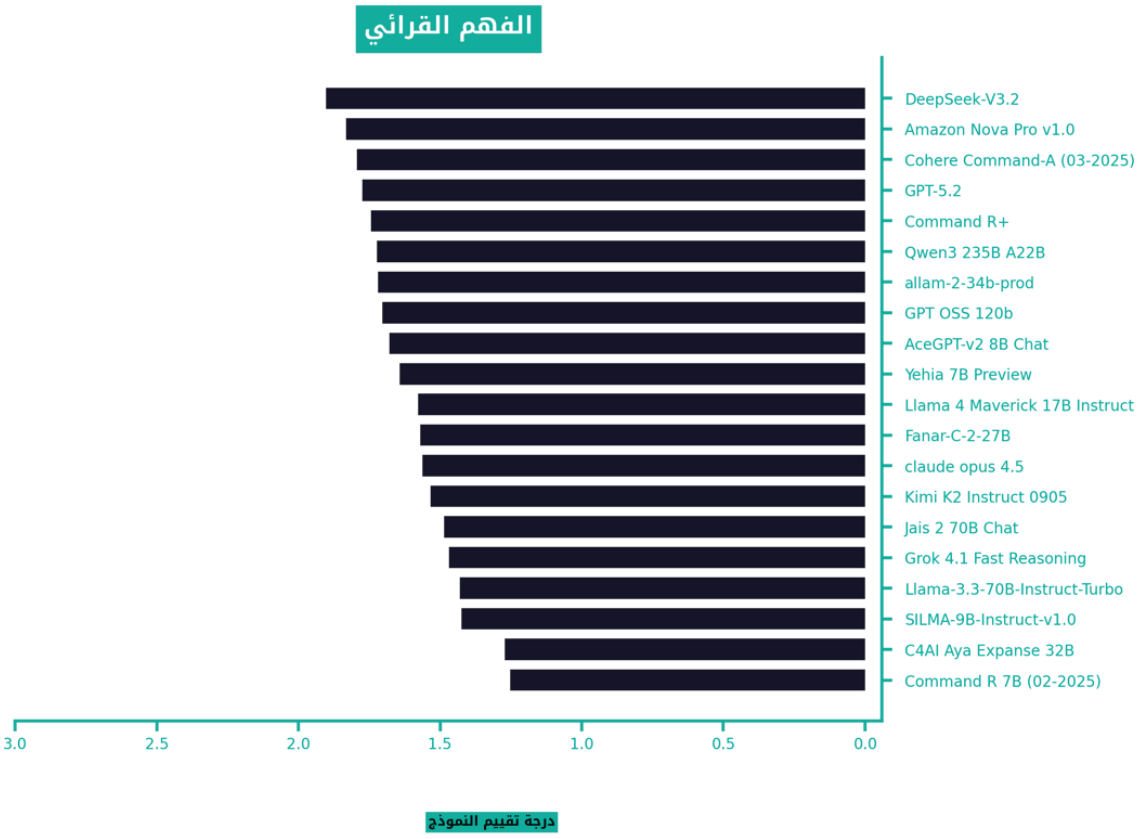
٨.١ الإجابة عن الأسئلة المطروحة (Answering Given Question).



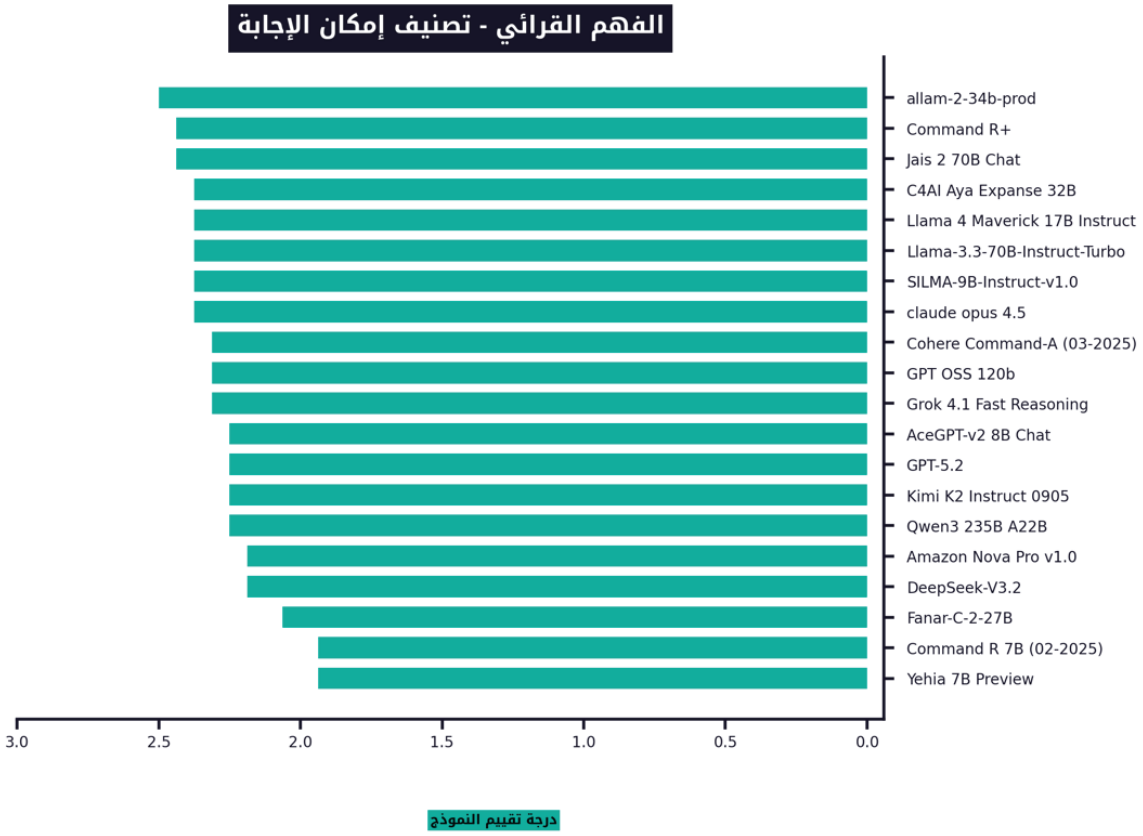
٨.٢ تحليل السؤال (Question Decomposition).



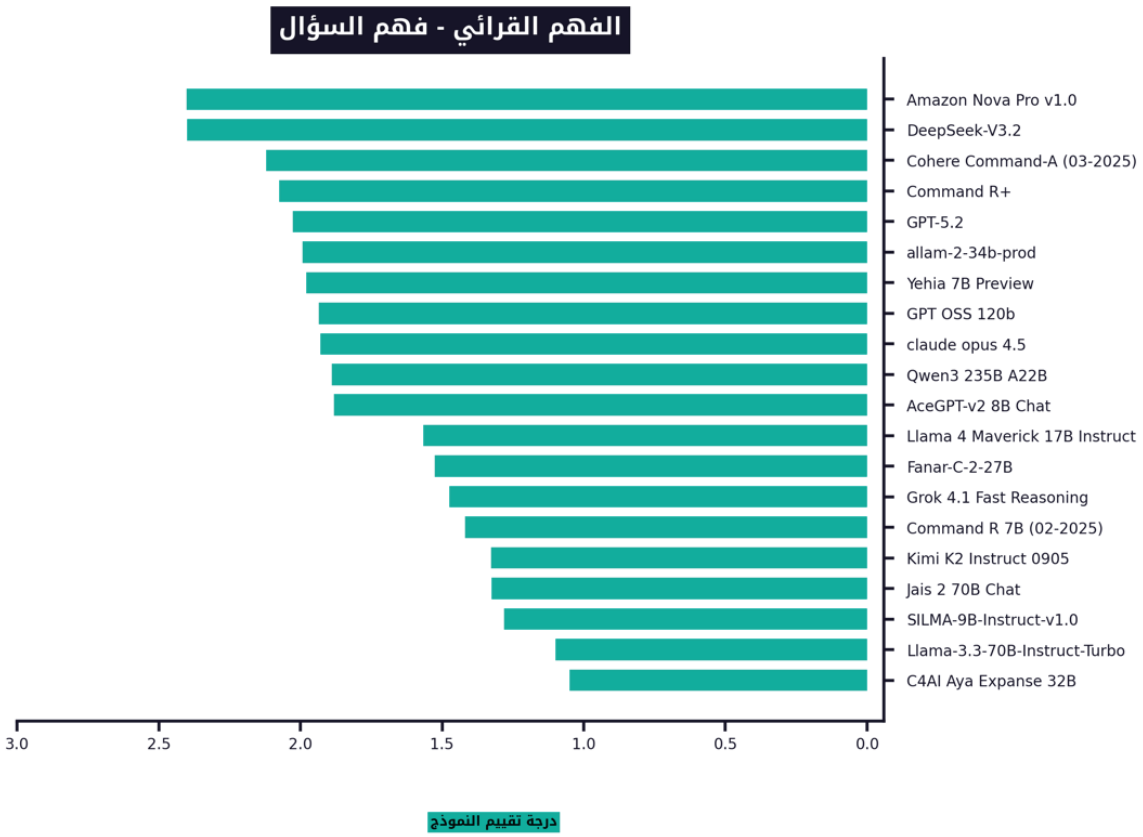
٩. الفهم القرائي (Reading Comprehension).



٩.١ تصنيف إمكان الإجابة (Answerability Classification).

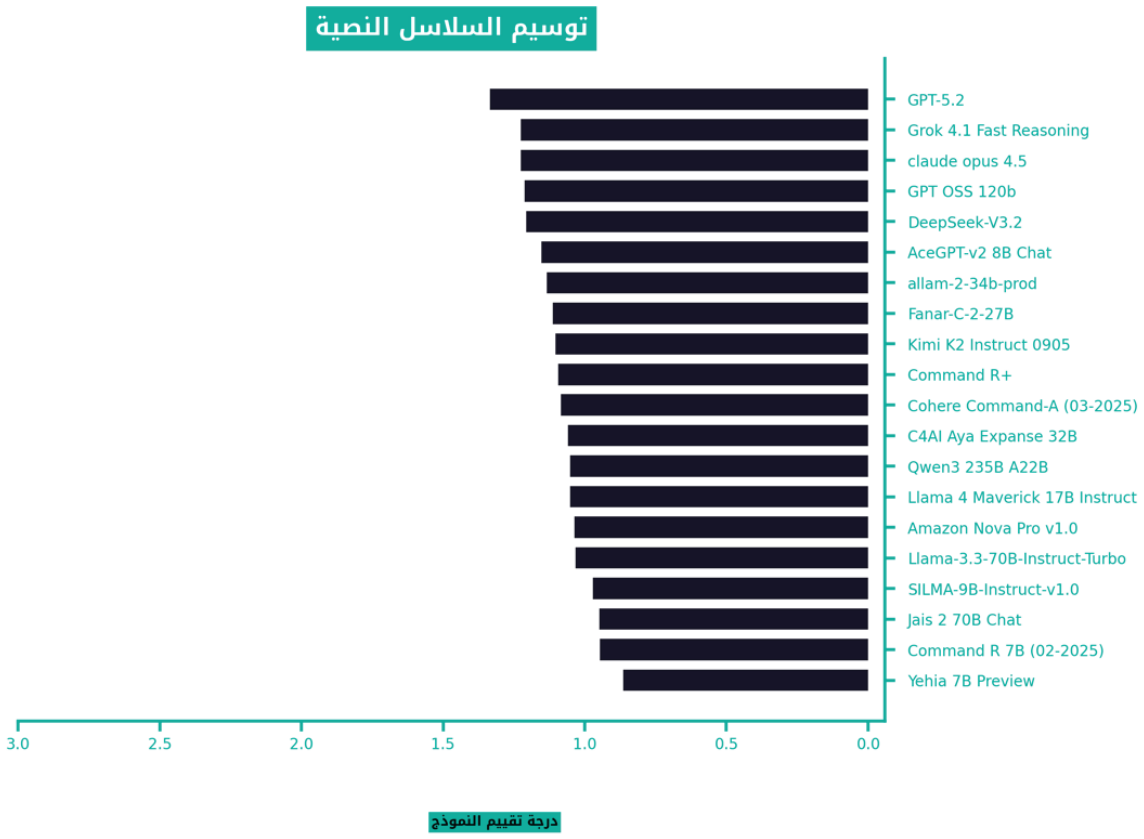


٩.٢ فهم السؤال (Question Understanding).

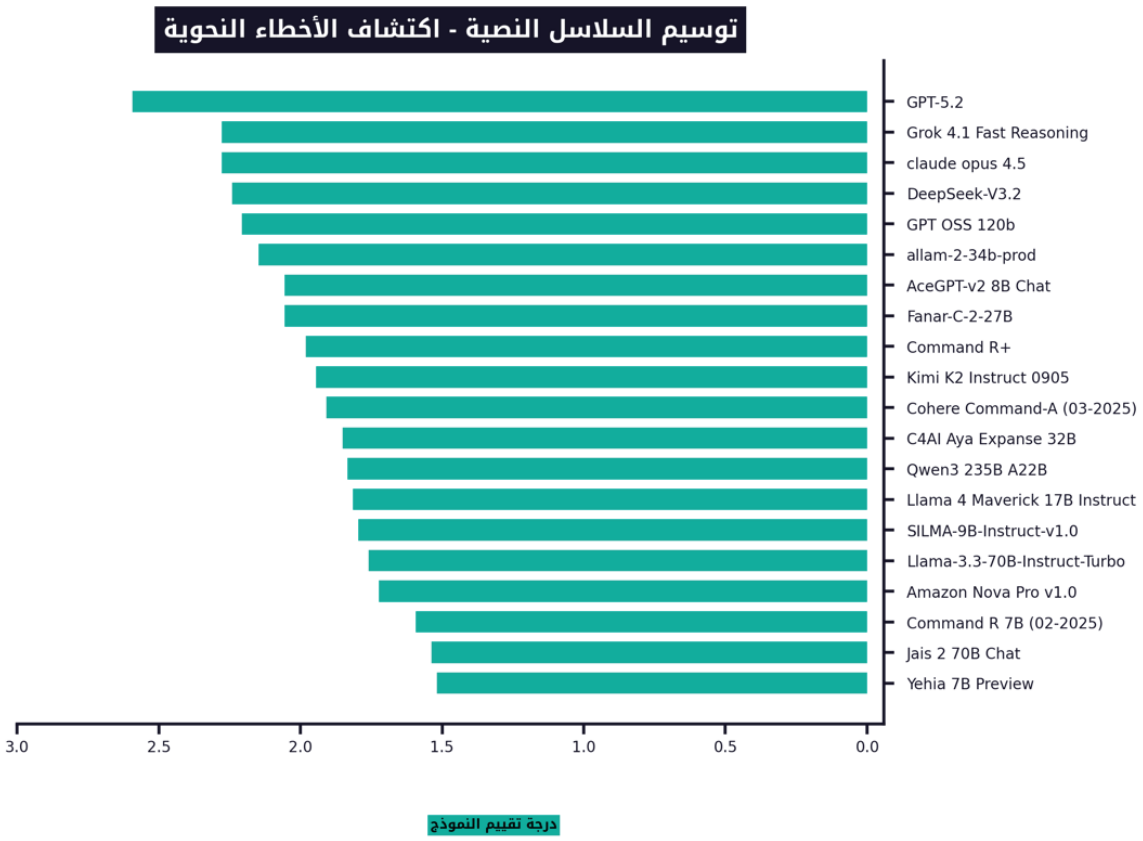


٩.٣ التحقق من الاستدلال النصي (Textual Inference Validation).

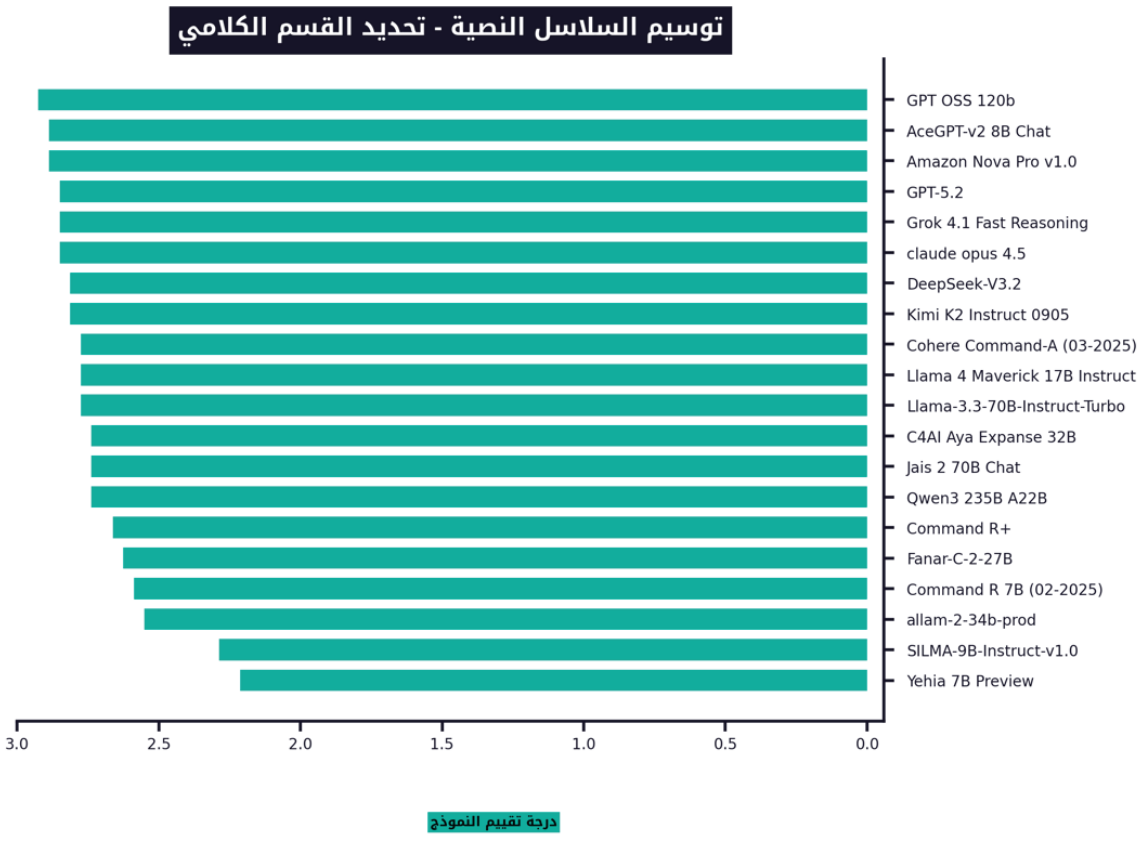
١٠. توسيم السلاسل النصية (Sequence Tagging).



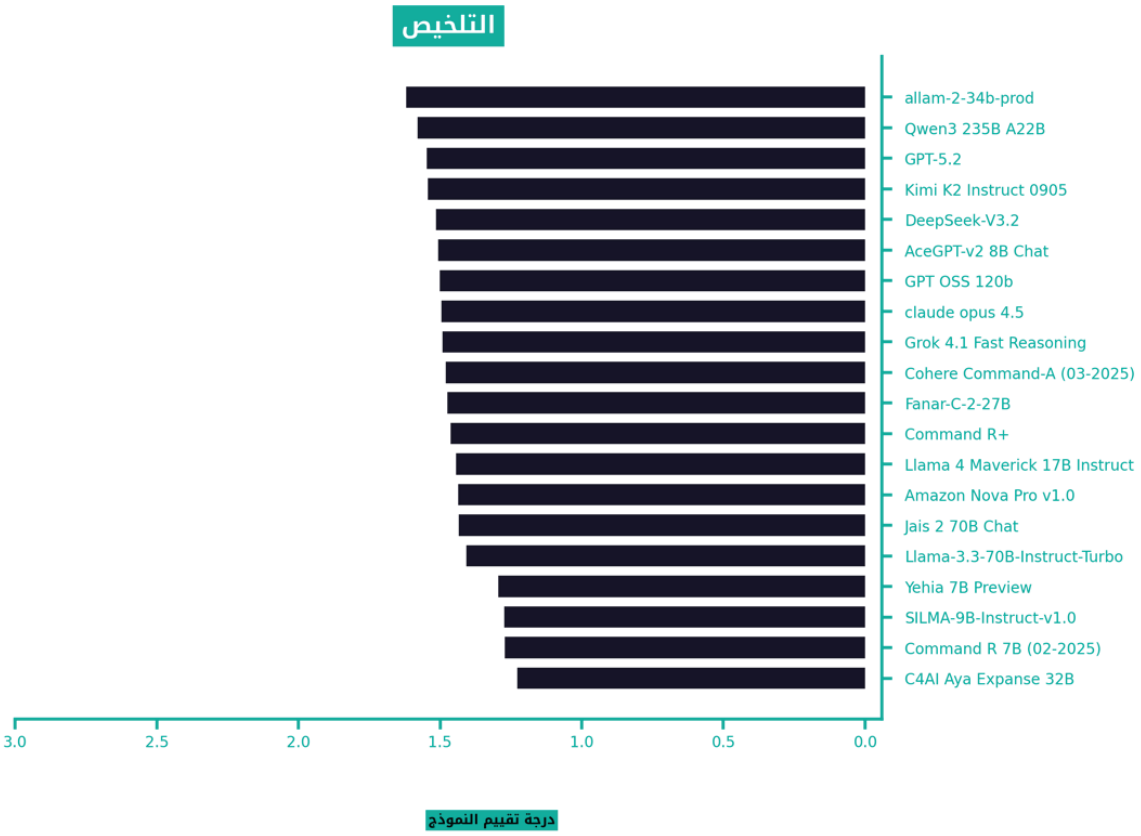
١٠.١ اكتشاف الأخطاء النحوية (Grammar Detection).



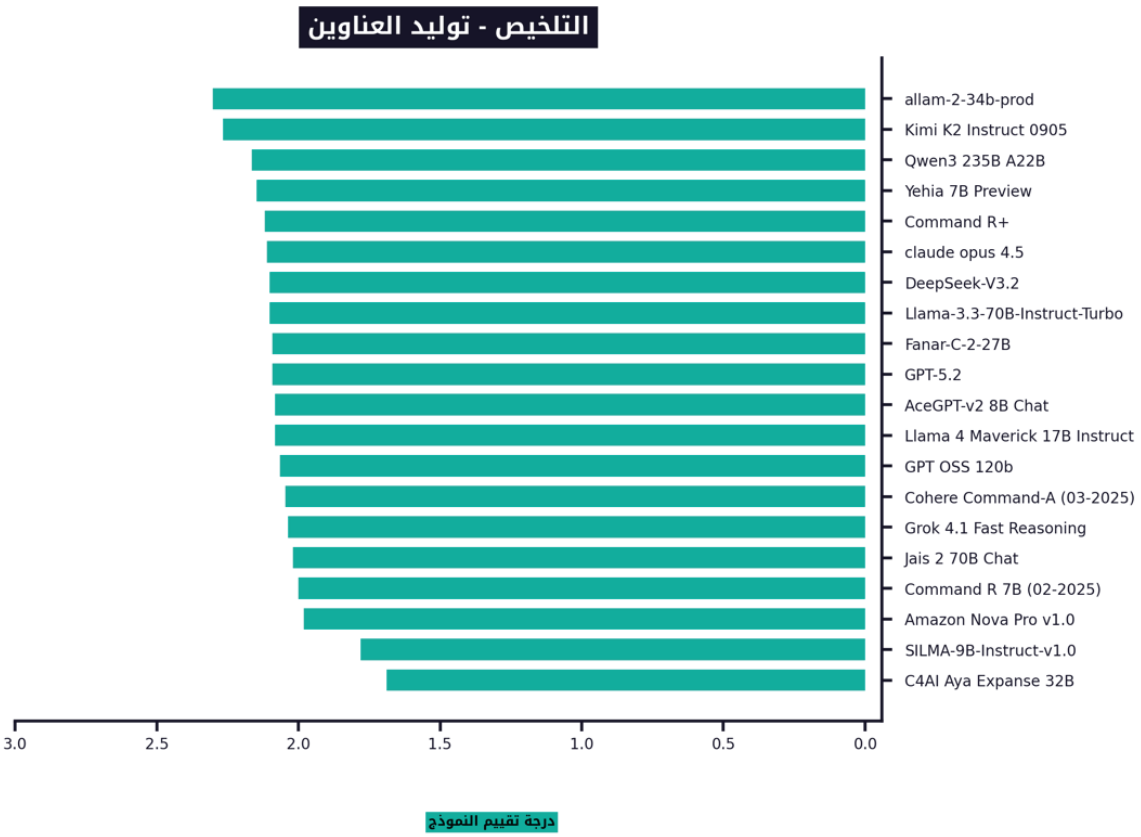
١٠.٢ تحديد القسم الكلامي (Part Of Speech Tagging).



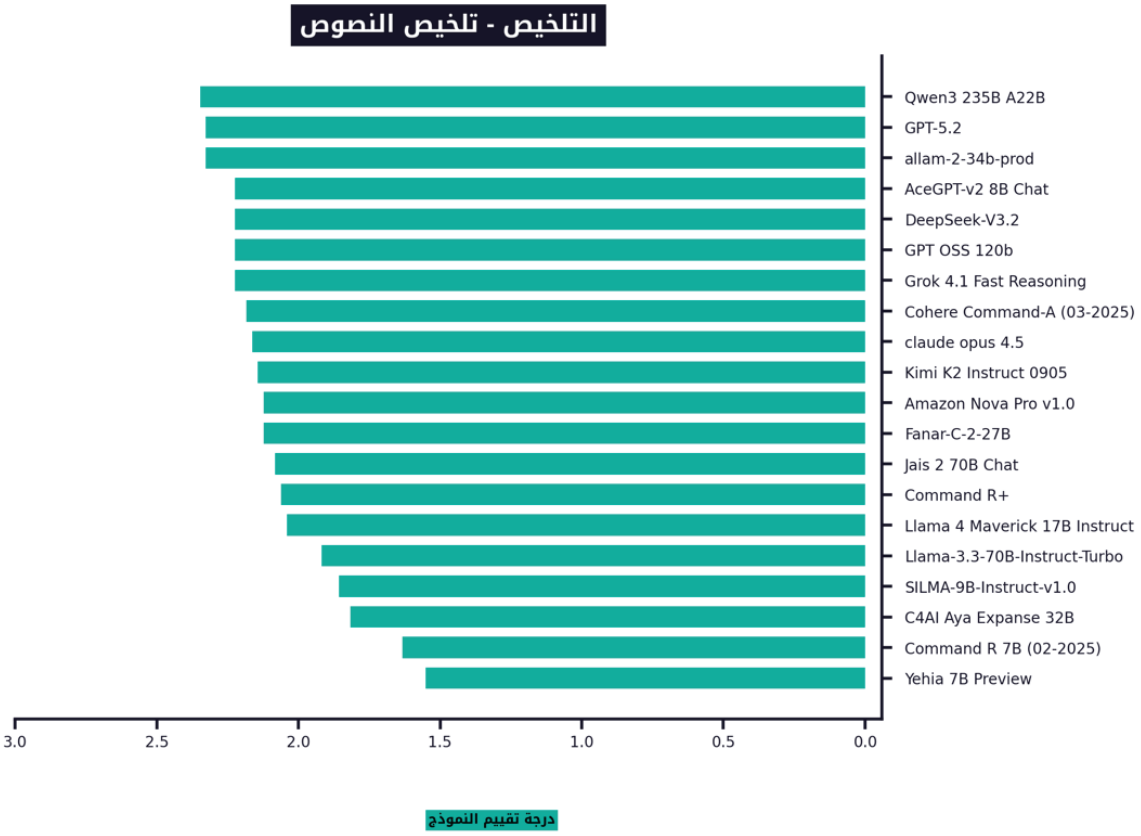
١١. التلخيص (Summarization).



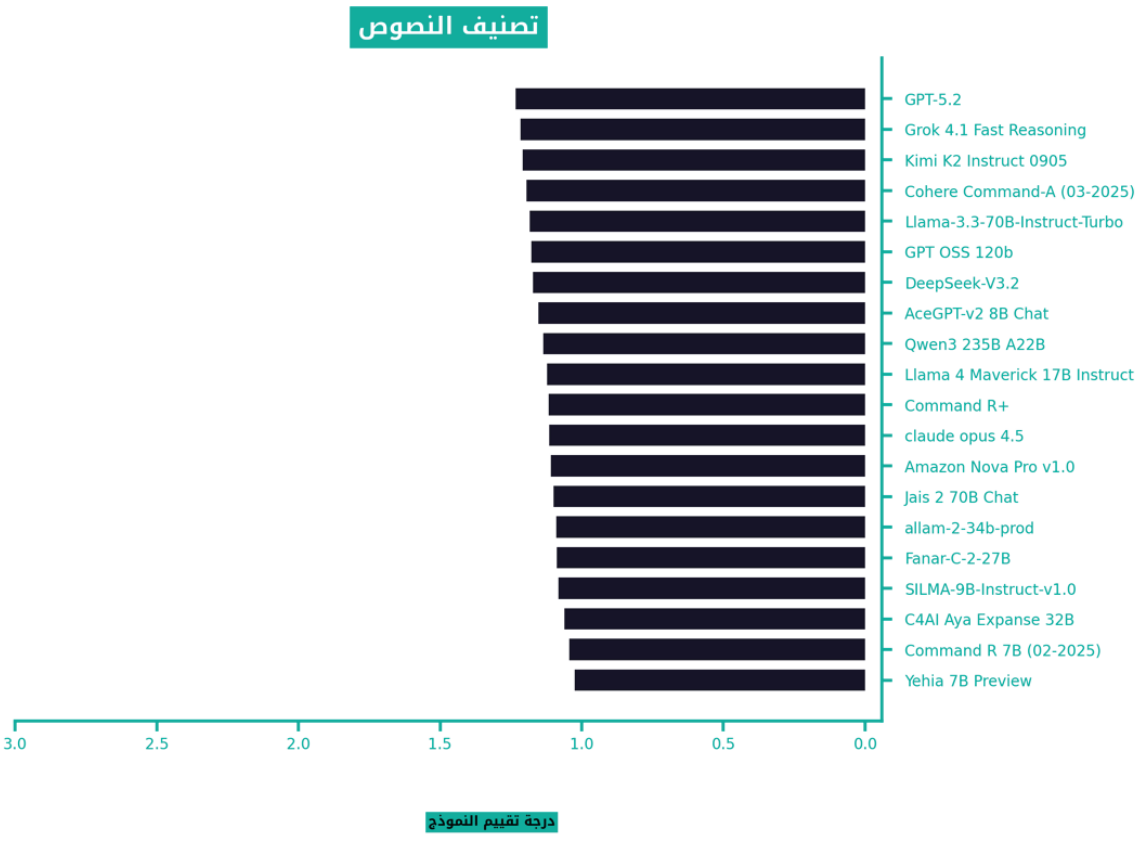
١١.١ توليد العناوين (Subject Generation).



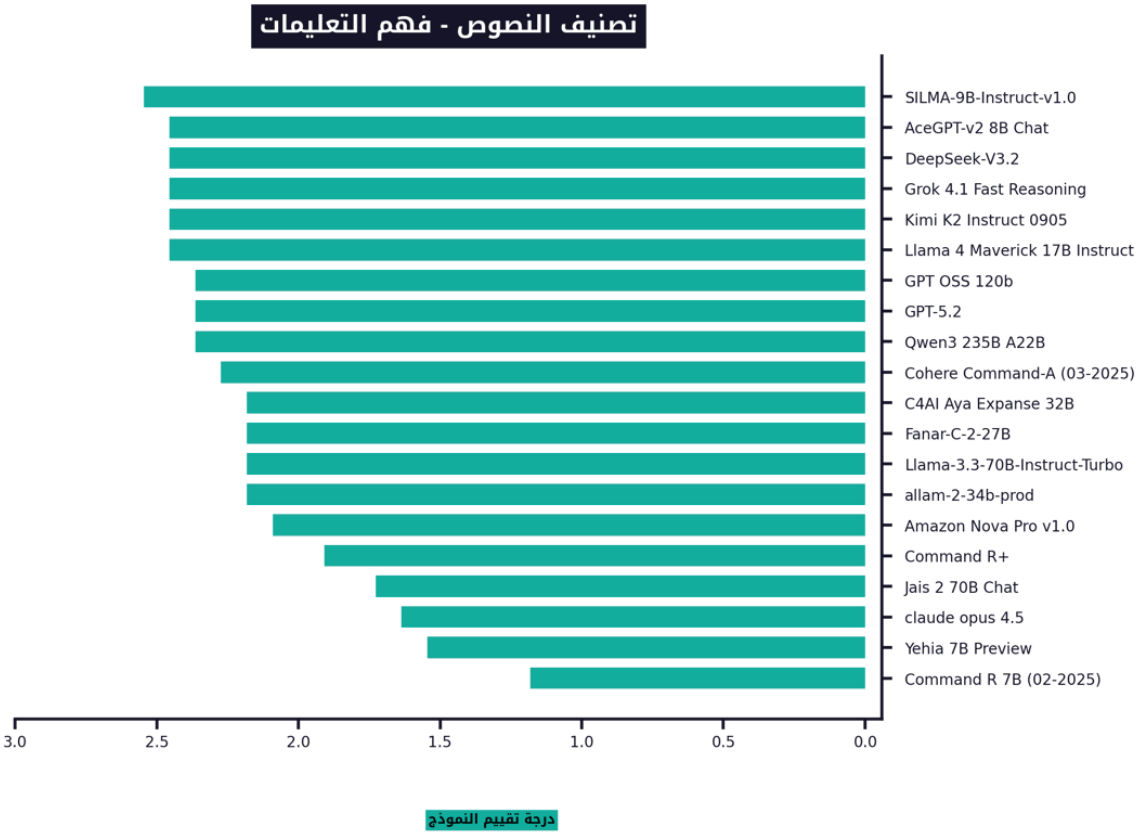
١١.٢ تلخيص النصوص (Text Summarization).



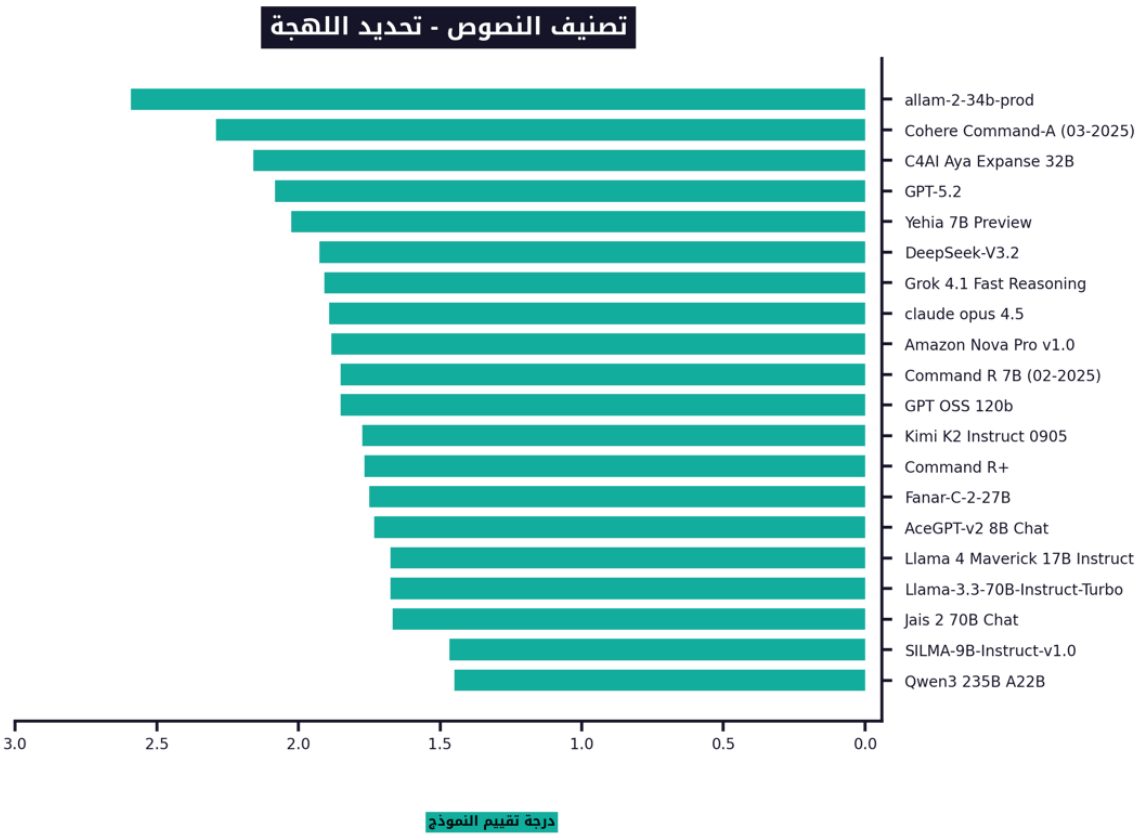
١٢. تصنيف النصوص (Text Classification).



١٢.١ فهم التعليمات (Command Interpretation).

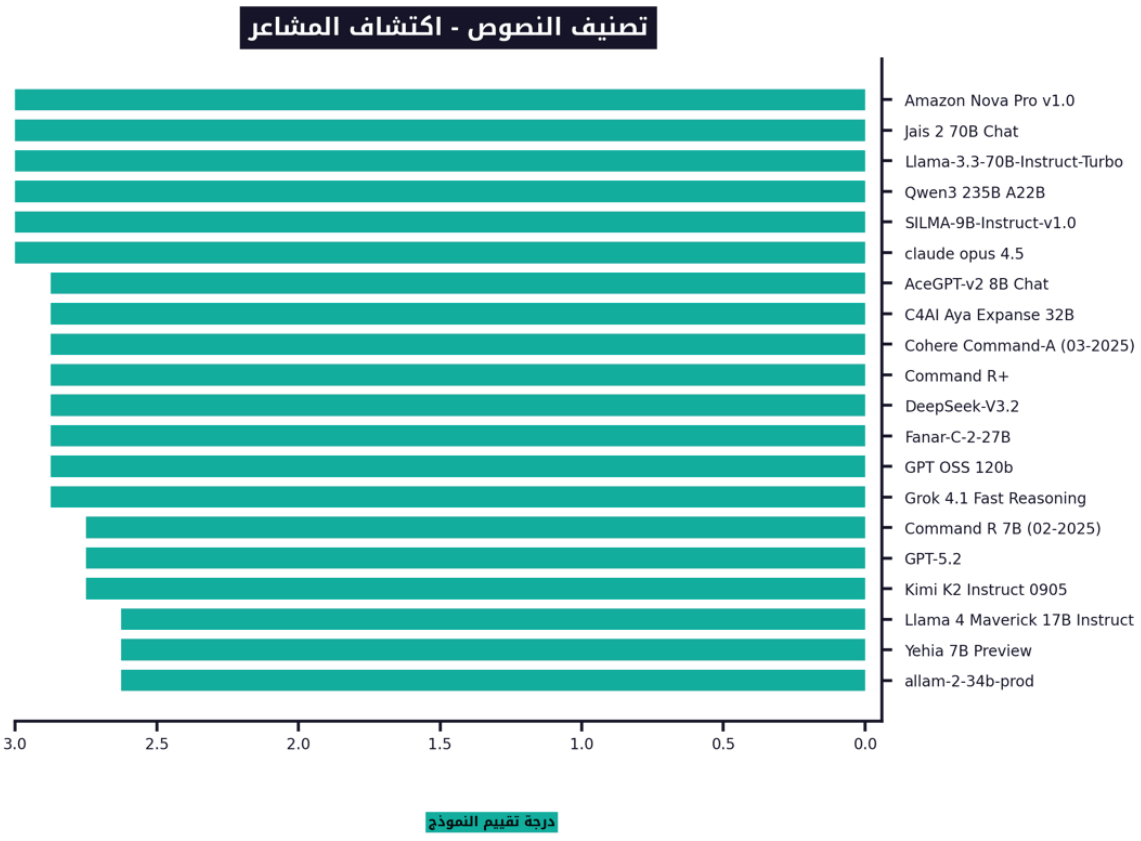


١٢.٢ تحديد اللهجة (Dialect Identification).

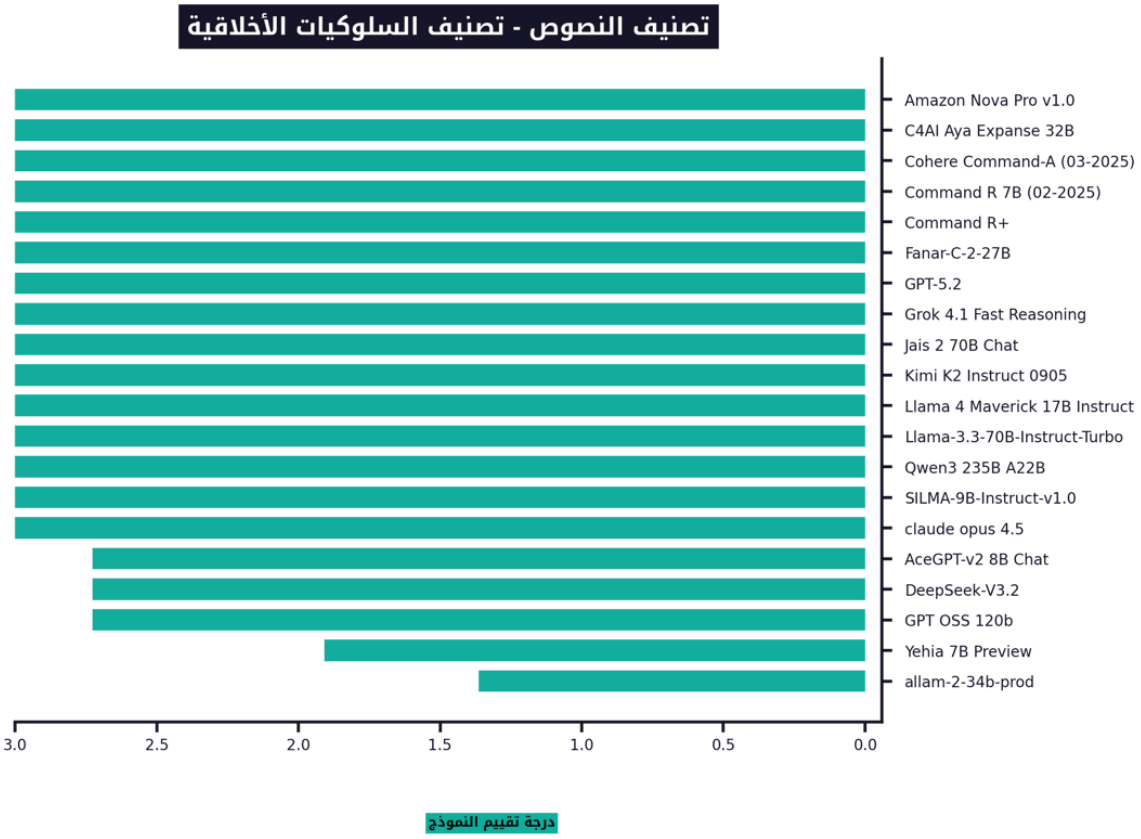


١٢.٣ تعرّف أفعال الحوار (Dialogue Act Recognition).

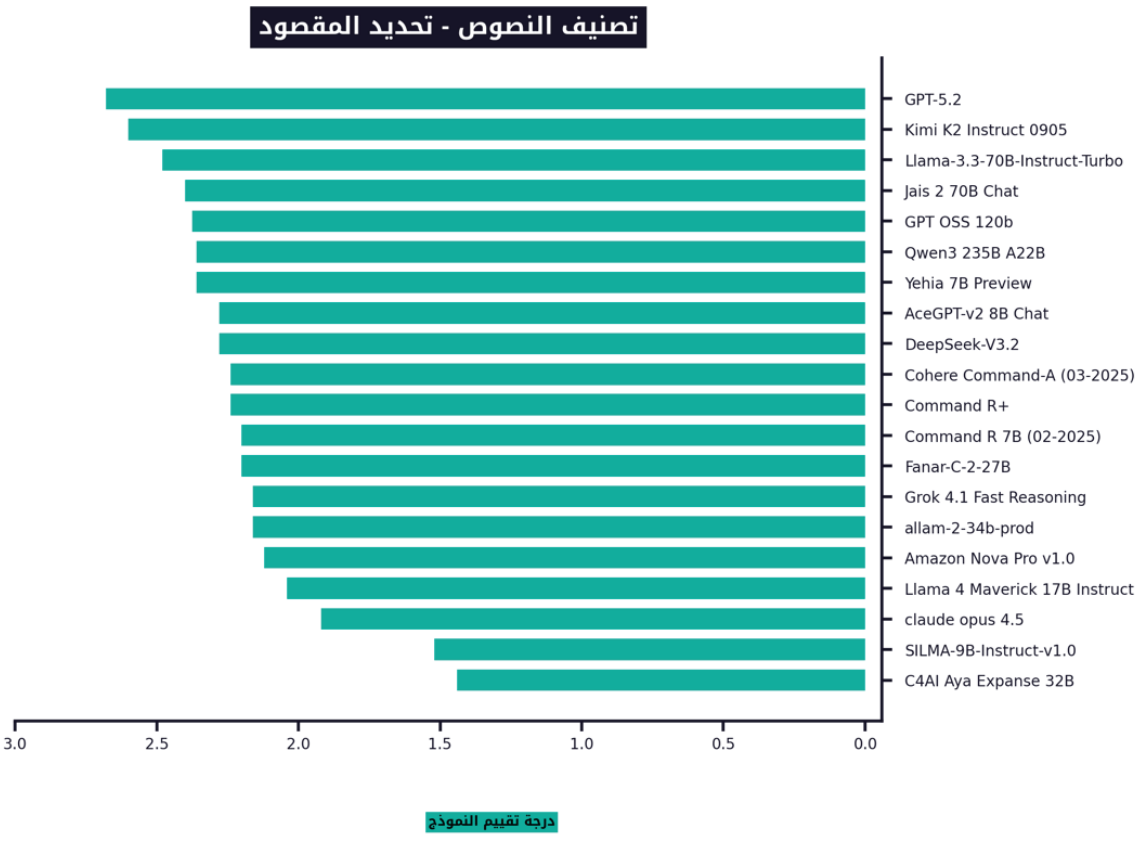
١٢.٤ اكتشاف المشاعر (Emotion Detection).



١٢.٥ تصنيف السلوكيات الأخلاقية (Ethics Classification).

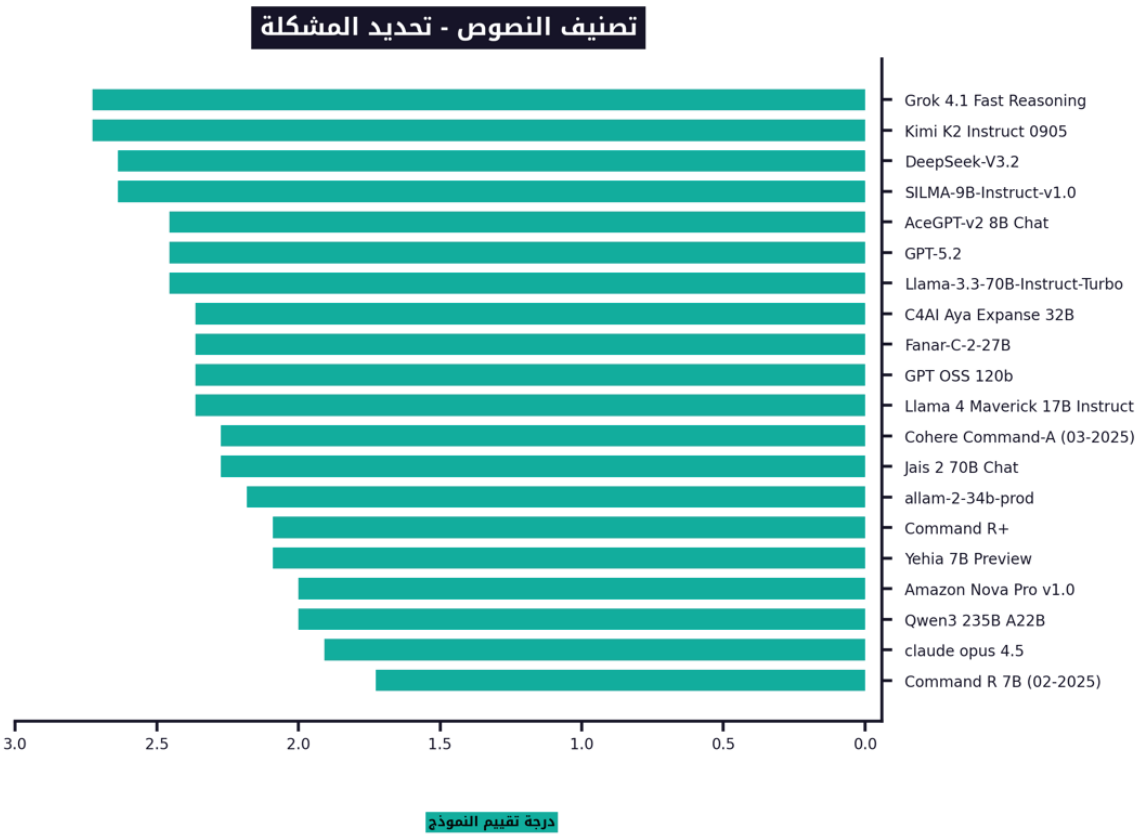


١٢.٦ تحديد المقصود (Intent Classification).

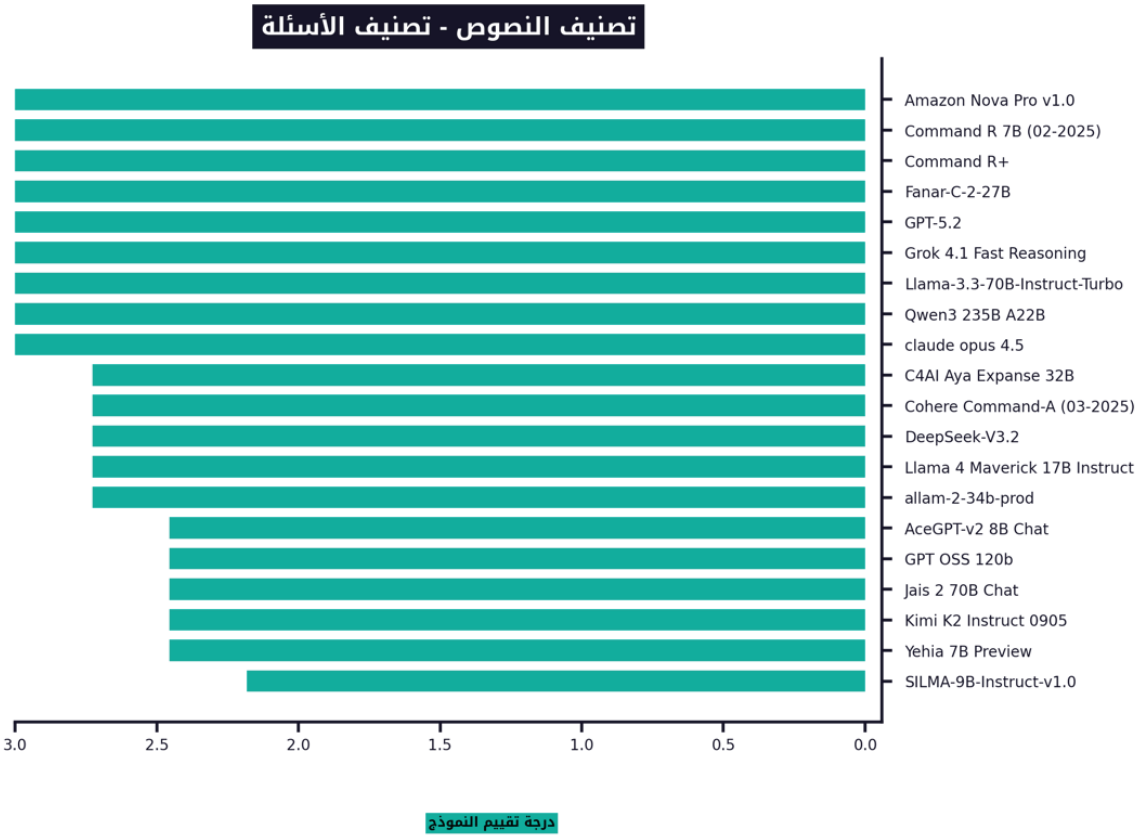


١٢.٧ اكتشاف الإساءة (Offensive Language Detection).

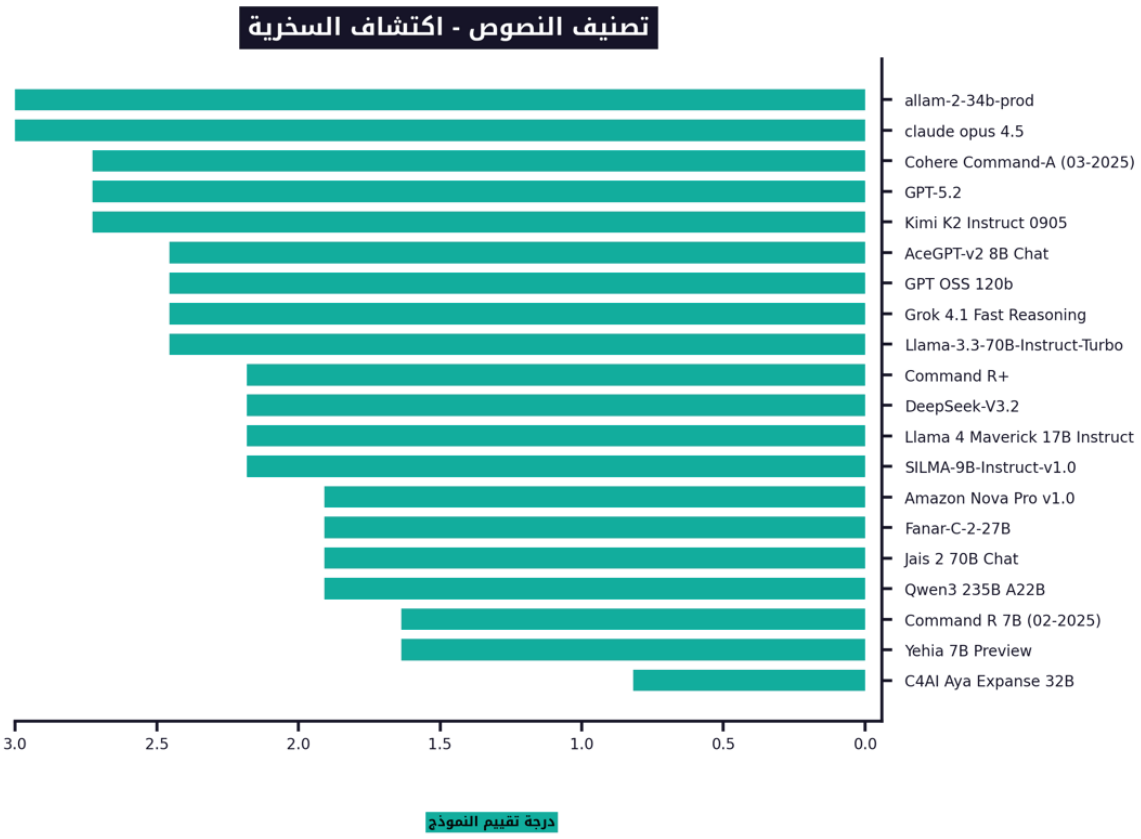
١٢.٨ تحديد المشكلة (Problem Identification).



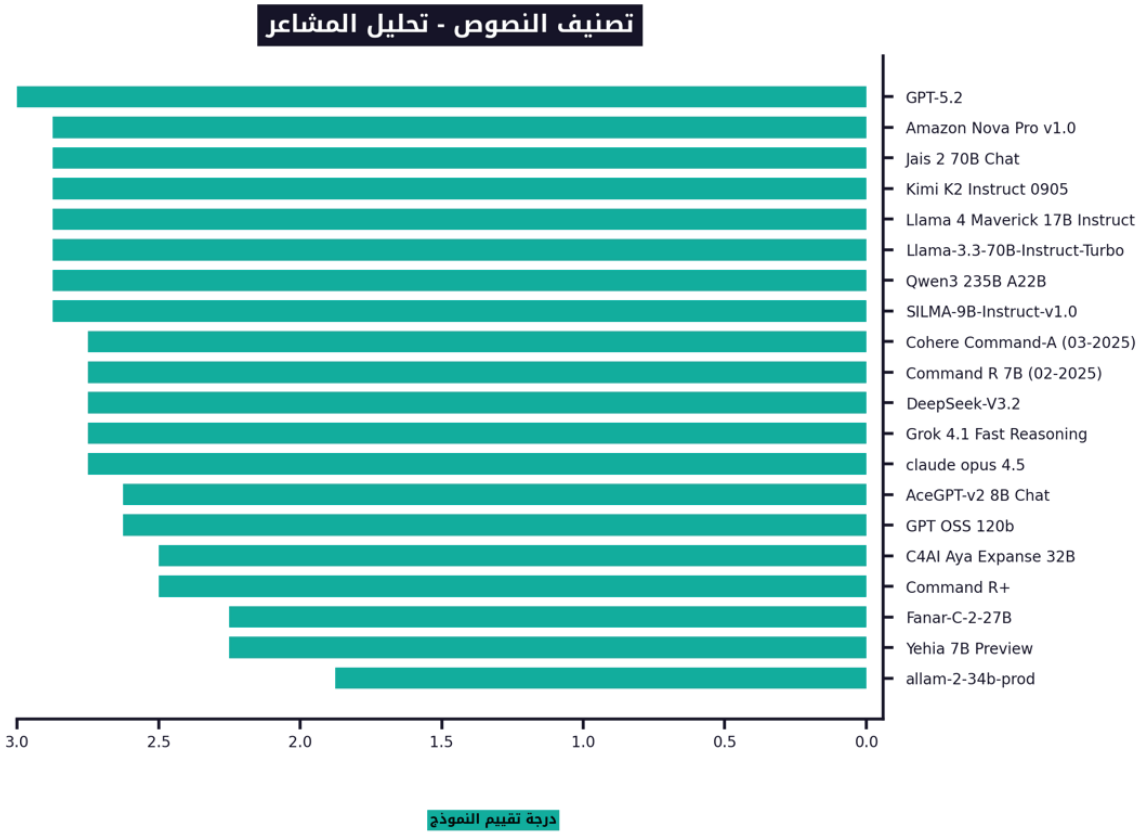
١٢.٩ تصنيف الأسئلة (Question Categorization).



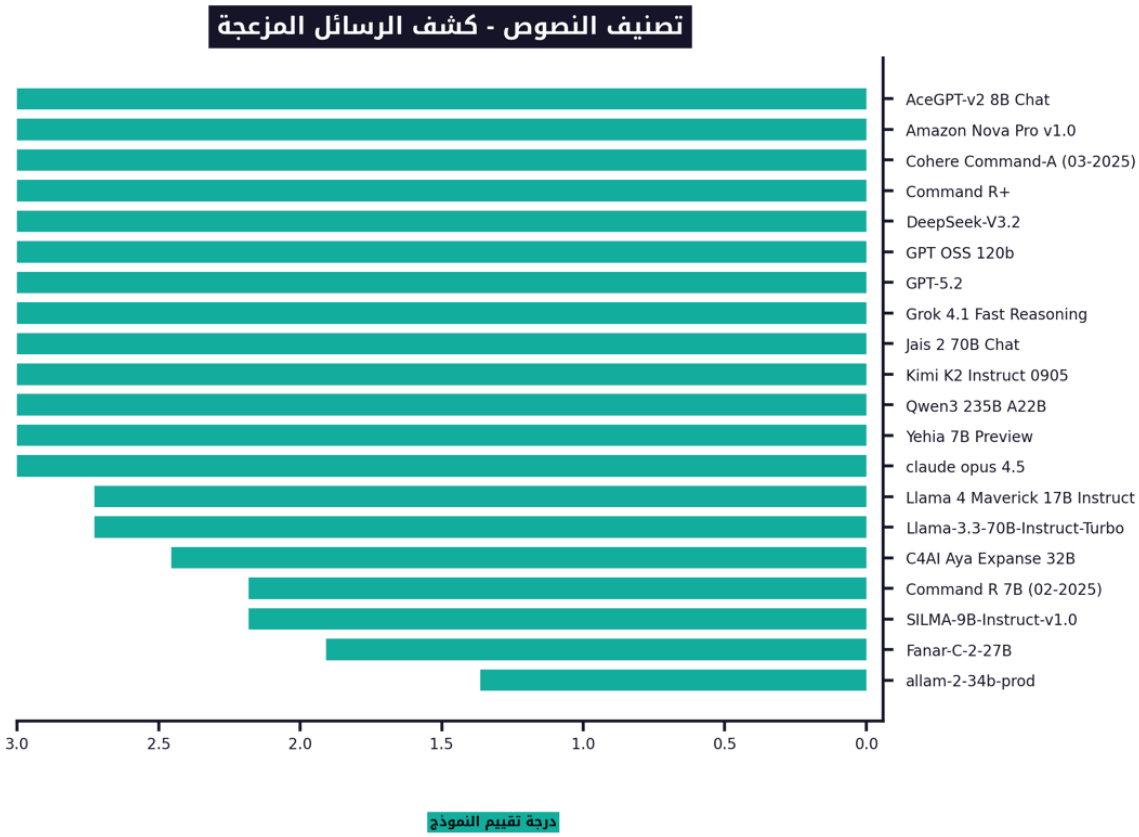
١٢.١٠ اكتشاف السخرية (Sarcasm Detection).



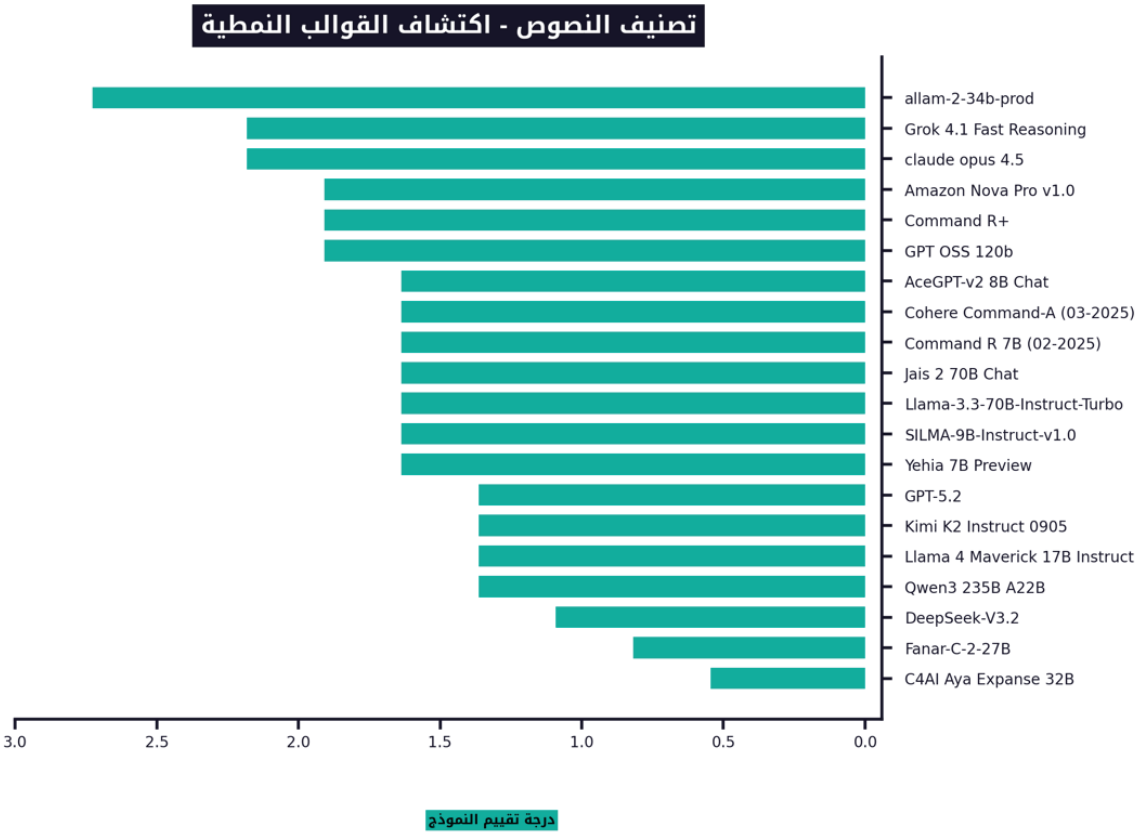
١٢.١١ تحليل المشاعر (Sentiment Analysis).



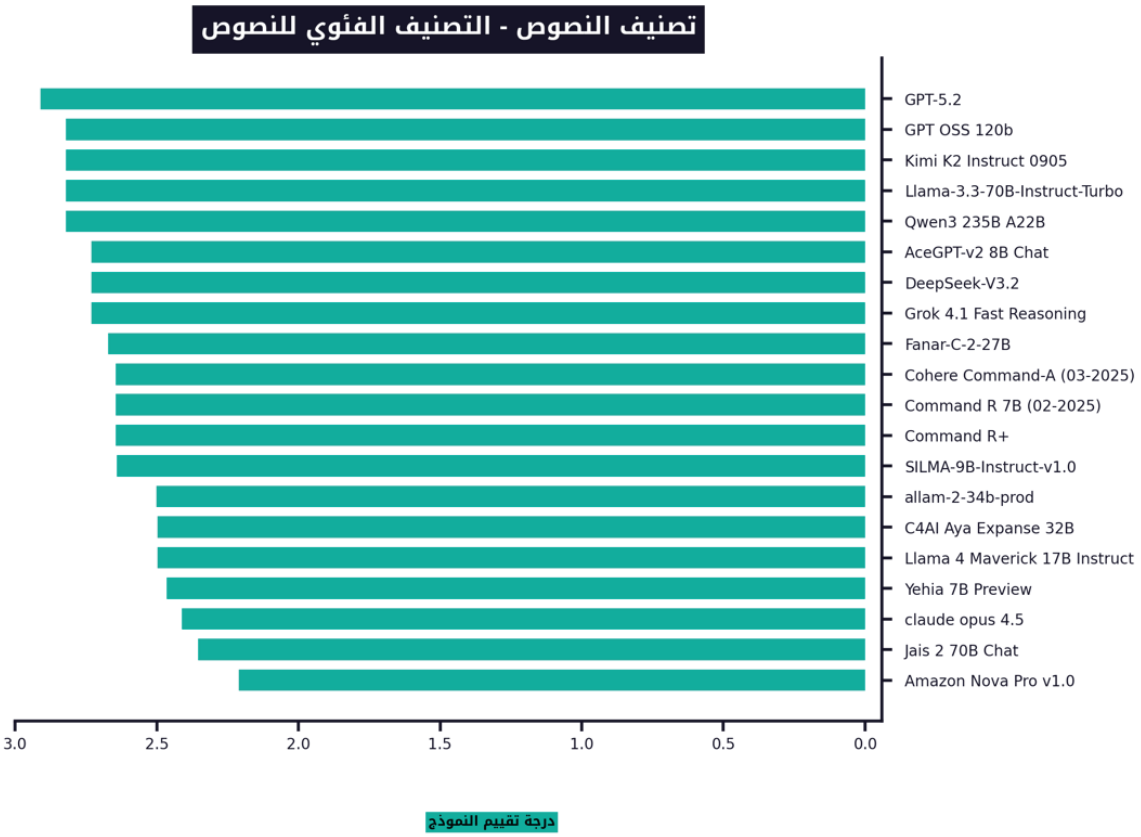
١٢.١٢ كشف الرسائل المزعجة (Spam Detection).



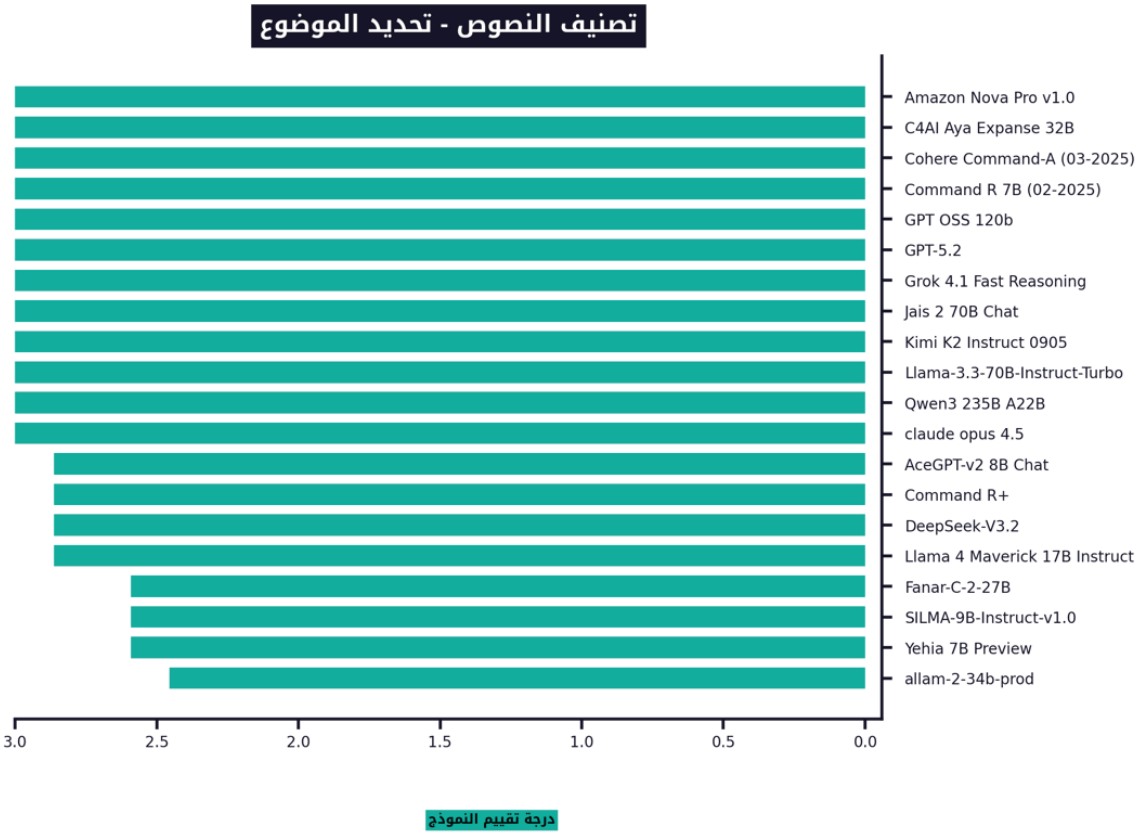
١٢.١٣ اكتشاف القوالب النمطية (Stereotype Detection).



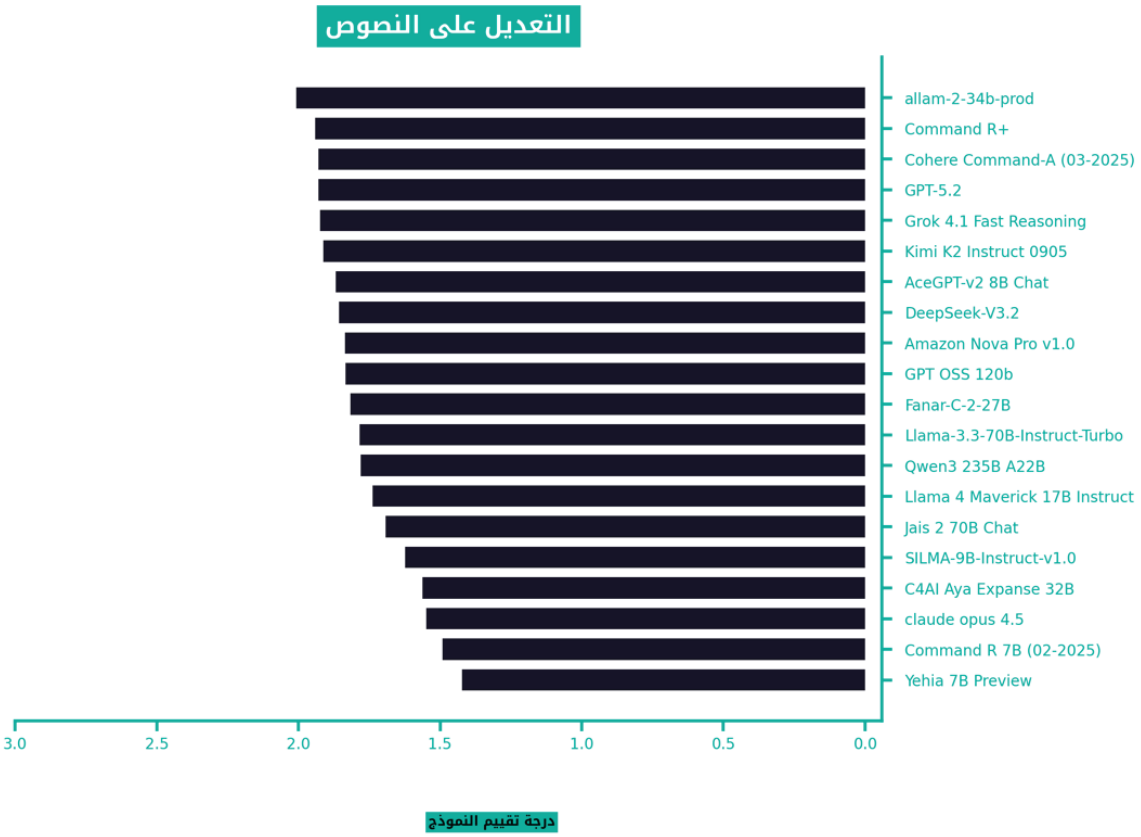
١٢.١٤ التصنيف الفئوي للنصوص (Text Categorization).



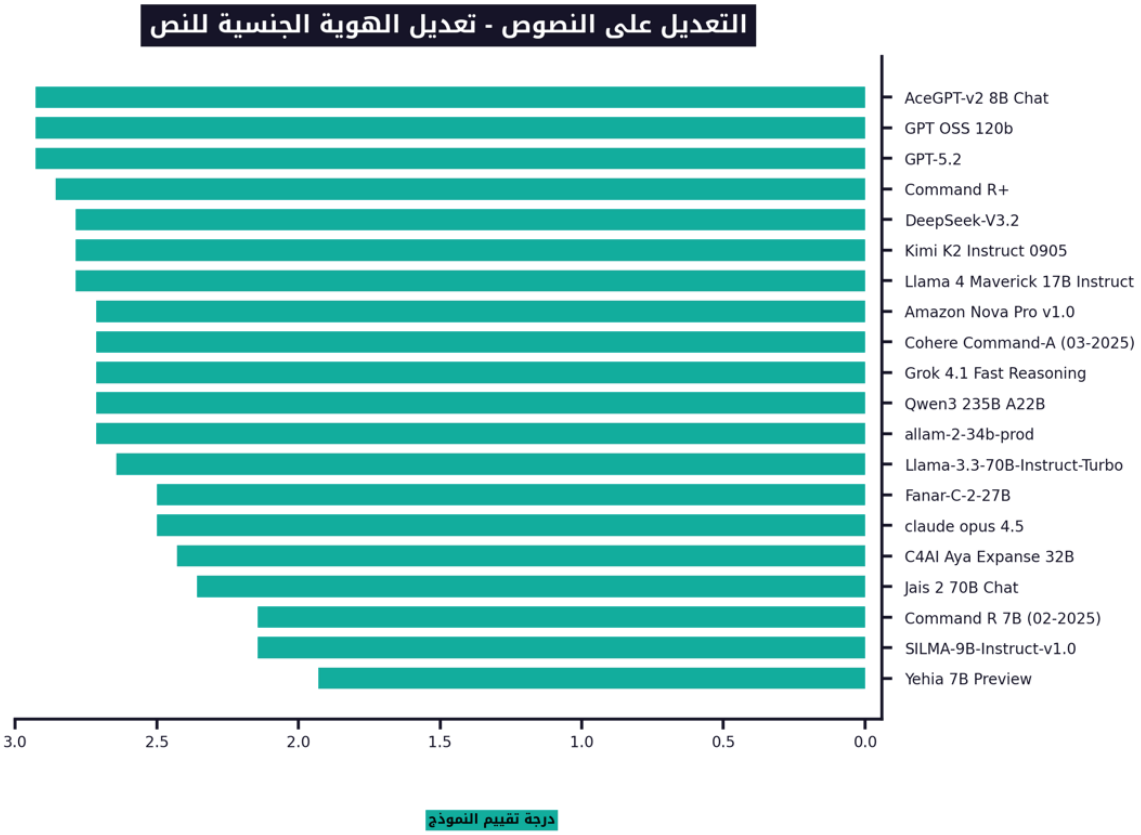
١٢.١٥ تحديد الموضوع (Topic Identification).



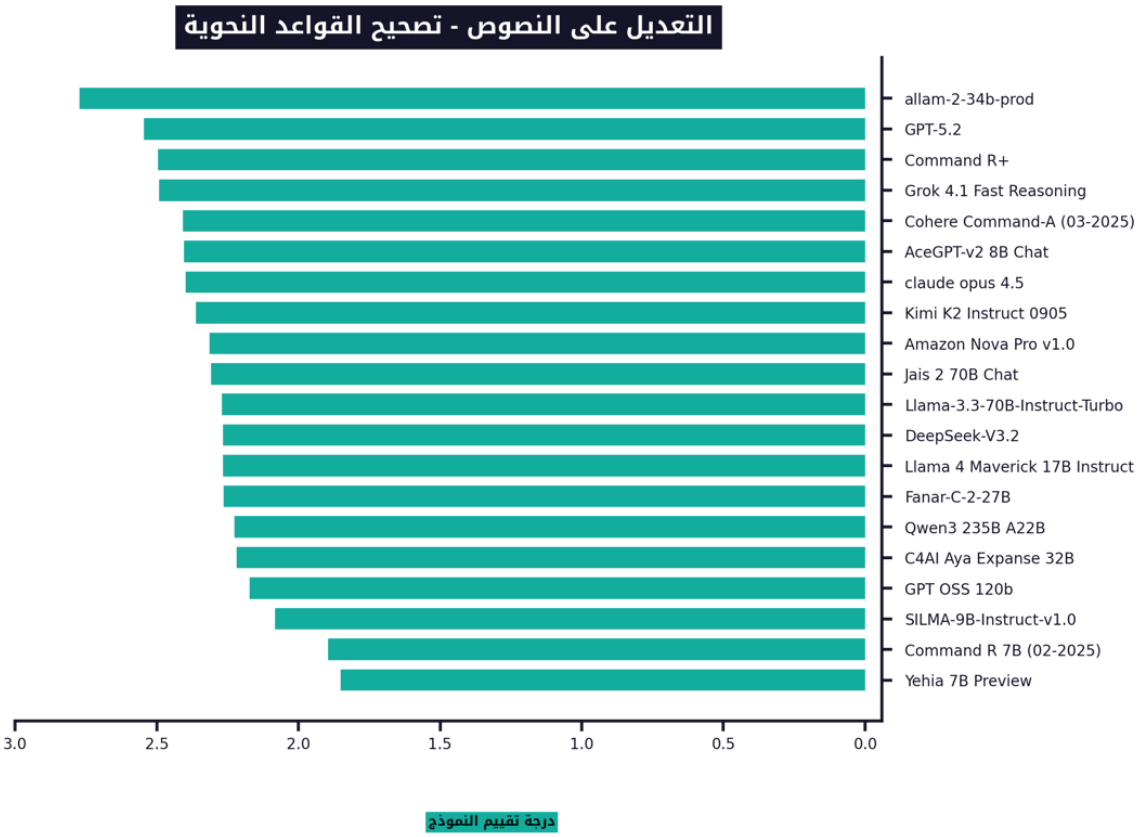
١٣. التعديل على النصوص (Text Manipulation).



١٣.١ تعديل الهوية الجنسية للنص (Gender Rewriting).

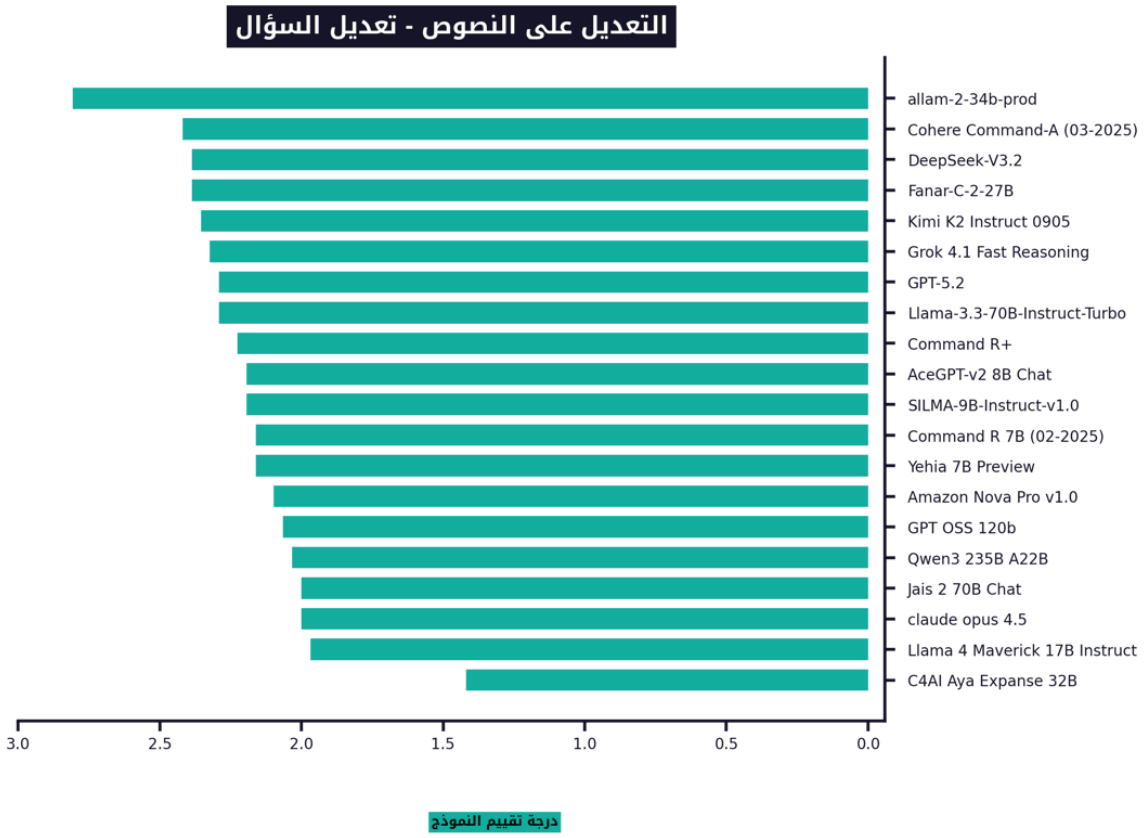


١٣.٢ تصحيح القواعد النحوية (Grammar Correction).

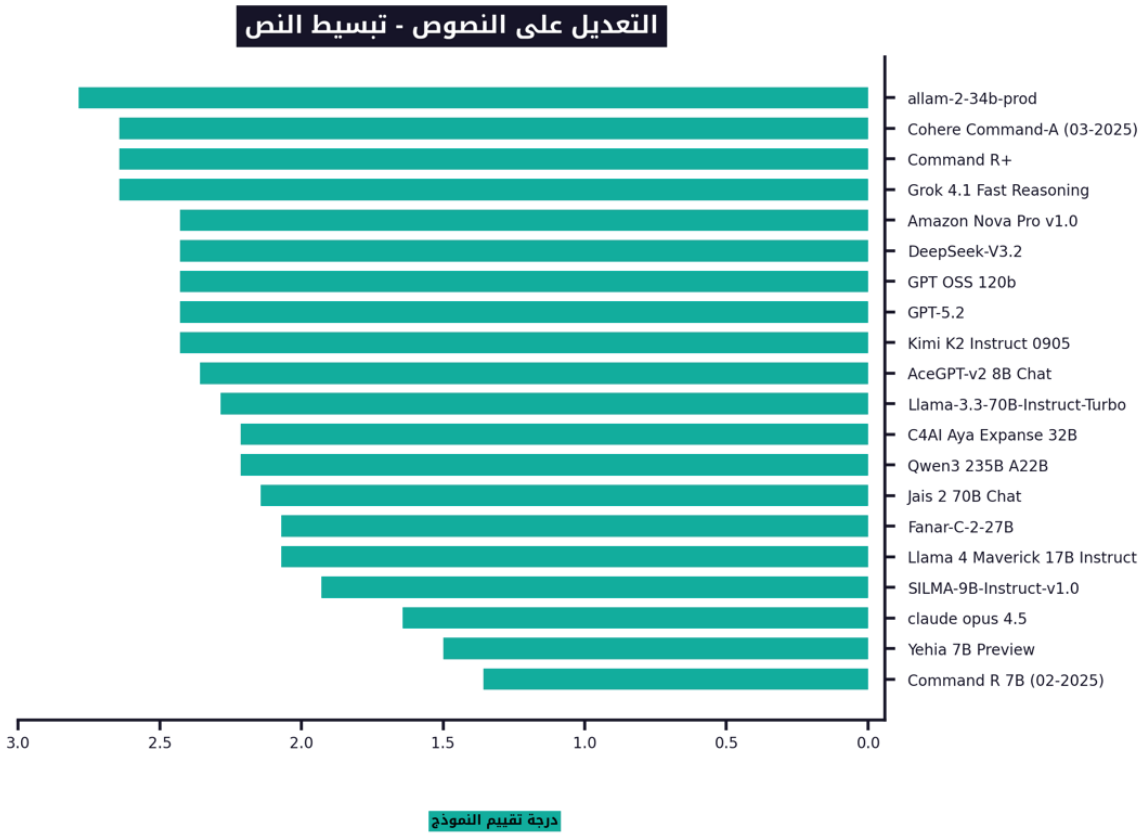


١٣.٣ إعادة الصياغة (Paraphrasing).

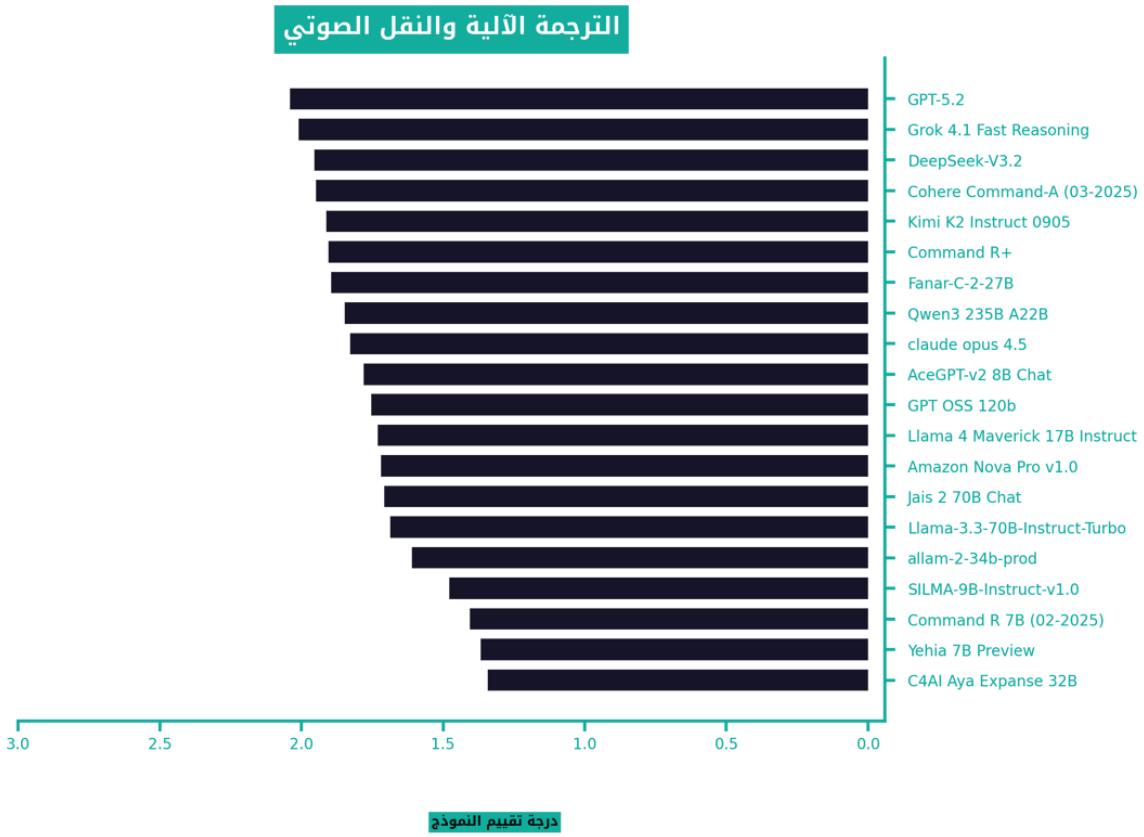
١٣.٤ تعديل السؤال (Question Rewriting).



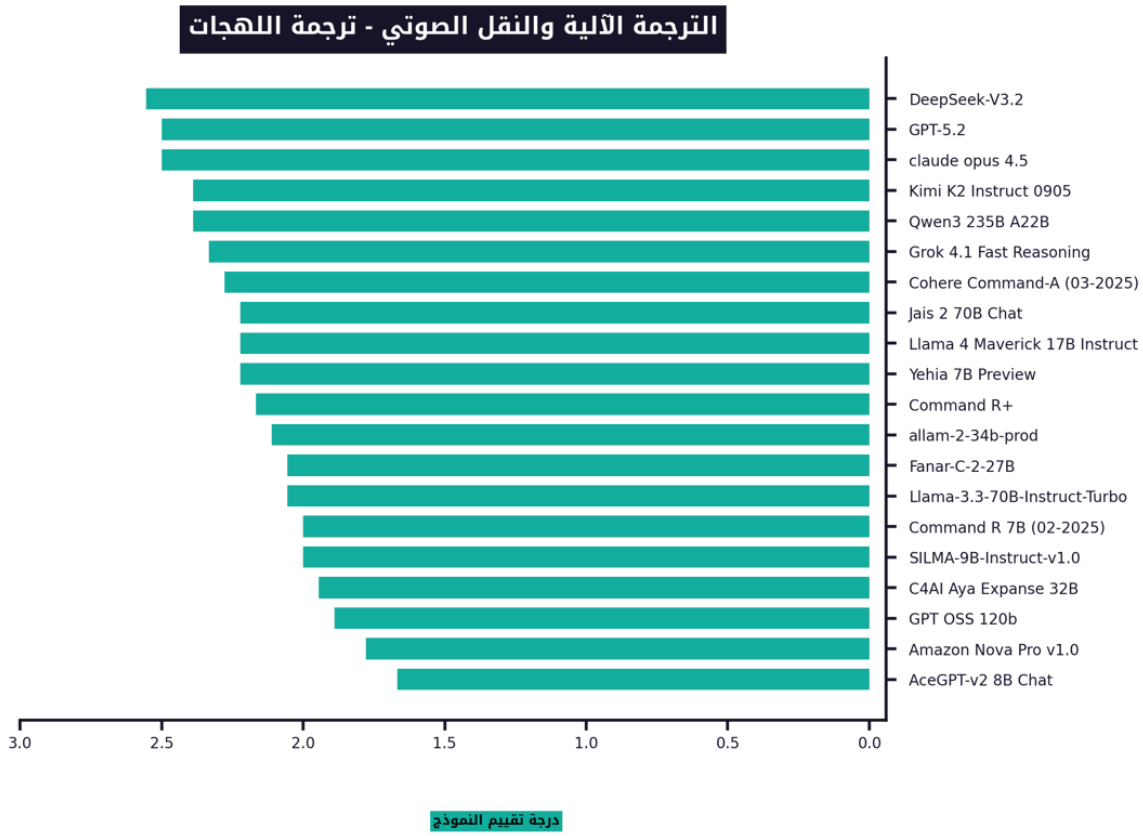
١٣.٥ تبسيط النص (Text Simplification).



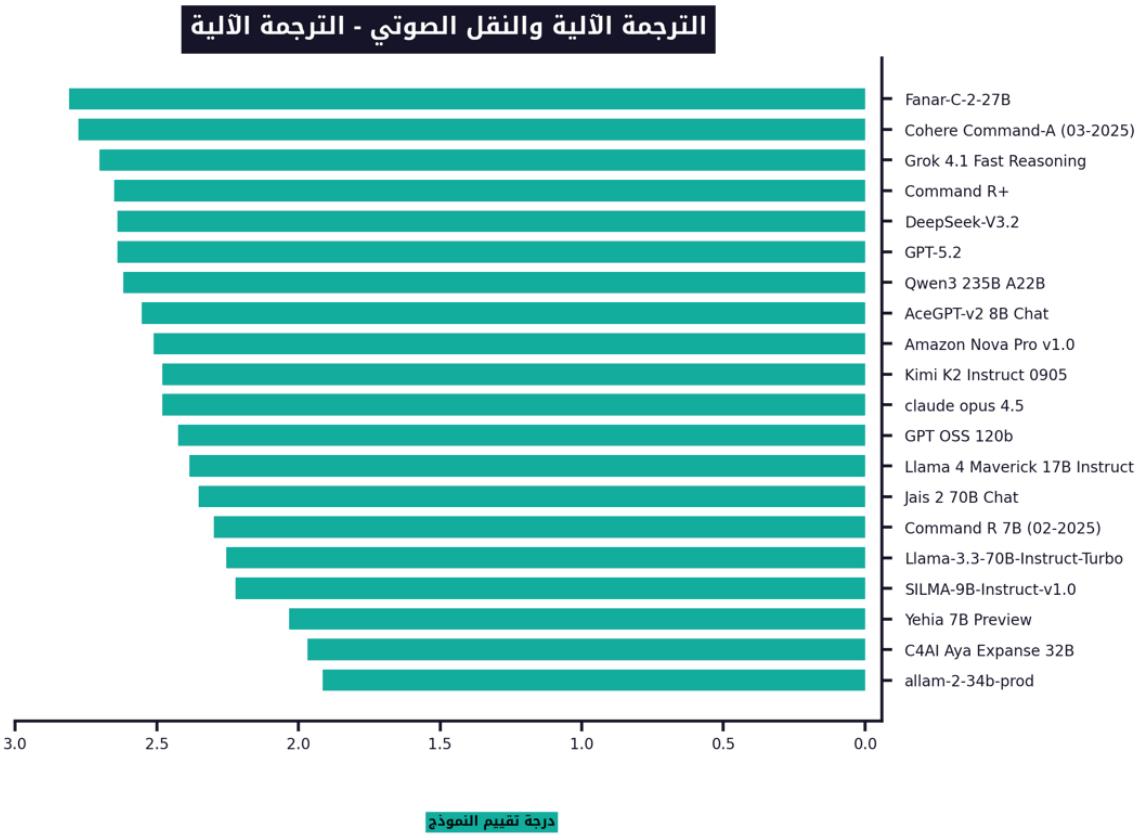
١٤. الترجمة الآلية والنقل الصوتي (Translation/Transliteration).



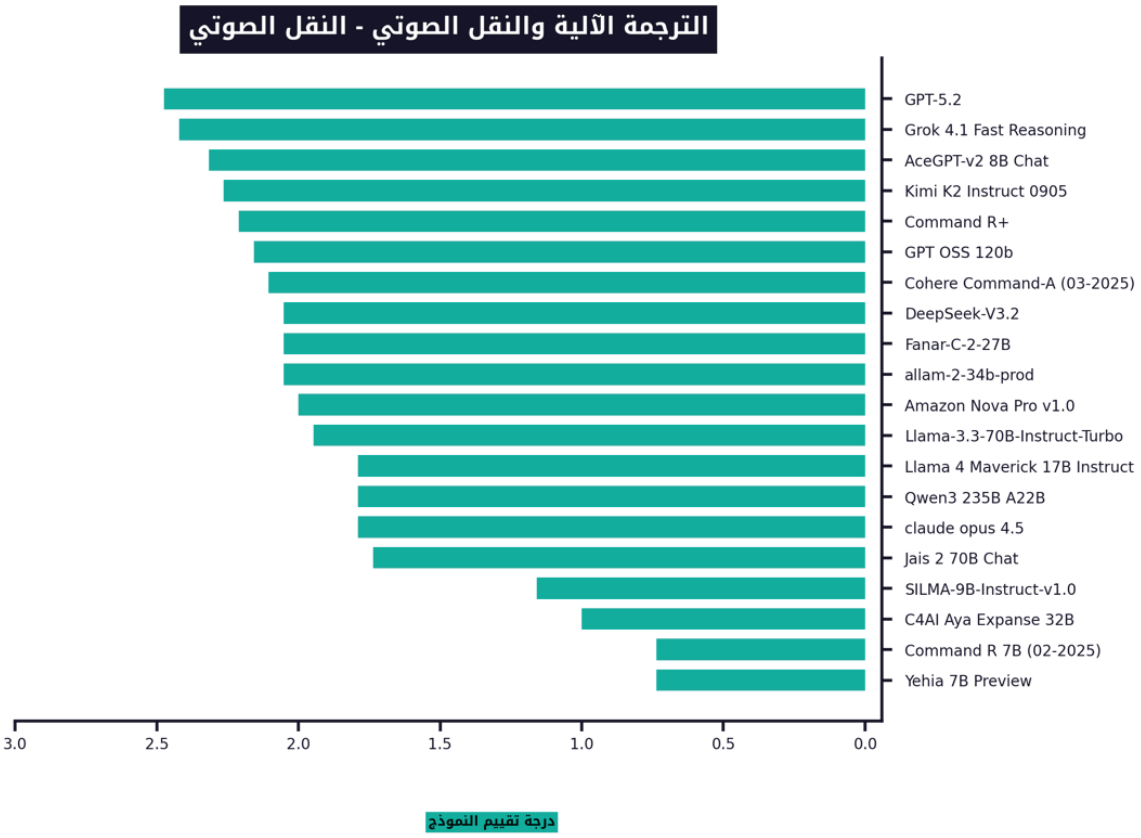
١٤.١ ترجمة اللهجات (Dialect Translation).



١٤.٢ الترجمة الآلية (Machine Translation).



١٤.٣ النقل الصوتي (Transliteration).




رؤية VISION 2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA
برنامج تنمية القدرات البشرية
Human Capability Development Program
مجمع الملك سلمان العالمي للغة العربية
King Salman Global Academy for Arabic Language
BALSAM