

دراسة وتحليل إحدى الدراسات الحديثة في مجال البحث من حيث المنهجية المستخدمة والنتائج المتوصل إليها

مهمة رقم 3

إعداد: أماني عبدالله العتيبي
رقم جامعي: 241002204
بإشراف الدكتورة: دعاء محمد

عنوان الرسالة

EmoBench: Evaluating the Emotional Intelligence of Large Language Models

أسماء الباحثين

Sahand Sabour Siyang Liu Zheyuan Zhang June M. Liu Jinfeng Zhou Alvionna S. Sunaryo
Juanzi Li Tatia M.C. Lee Rada Mihalcea Minlie Huang

معهد الذكاء الاصطناعي، جامعة تسينغها بكين، الصين

تم النشر في عام 2024م

عدد صفحات الدراسة 19 صفحة

المقدمة

تتناول هذه الدراسة تقييم الذكاء العاطفي للنماذج اللغوية الكبيرة LLMs مثل GPT-4 يهدف البحث إلى سد الفجوة في الأدبيات الحالية التي تركز بشكل رئيسي على التعرف على العواطف، مع إهمال القدرات الأساسية الأخرى للذكاء العاطفي مثل تنظيم العواطف وفهمها وتوظيفها

EmoBench هي أداة متعددة الاستخدامات يمكن أن تكون مفيدة بعدة طرق. ويمكن استخدامه لتقييم وتحسين قدرات الذكاء العاطفي لنماذج الذكاء الاصطناعي، وهو أمر بالغ الأهمية في المجالات المتعلقة بالذكاء الاصطناعي. يمكن أيضاً استخدام EmoBench في برامج تطوير القيادة لتدريب وتعزيز الذكاء العاطفي لدى القادة، مما يؤدي إلى تحسين مهارات اتخاذ القرار وإدارة العلاقات

بالإضافة إلى ذلك، يمكن تكييف السيناريوهات والأسئلة في EmoBench مع برامج التدريب التي تهدف إلى تحسين الذكاء العاطفي بين الموظفين، مما يؤدي إلى موظفين أكثر ذكاءً عاطفياً. يمكن أن يساعد فهم EmoBench أيضاً في التنمية الشخصية من خلال توفير رؤى حول إدارة العواطف وتحسين العلاقات بين الأشخاص.

تقدم EmoBench إطاراً منظماً لاستكشاف التقاطع بين الذكاء الاصطناعي والذكاء العاطفي، مما يؤدي إلى تطبيقات مبتكرة في مختلف المجالات. ويمكن استخدامه أيضاً لتعزيز تفاعلات الذكاء الاصطناعي في خدمة العملاء أو أنظمة تفاعل المستخدم، مما يضمن تجارب مستخدم أكثر استجابة عاطفية وإرضاءً. يعد EmoBench أيضاً مصدراً قيماً للأغراض الأكاديمية، حيث يعمل كأساس لدراسة تأثير الذكاء العاطفي في التكنولوجيا والتفاعلات بين الإنسان والحاسوب. بشكل عام، يمكن أن يؤدي الاستفادة من EmoBench إلى تعزيز فهم وتطبيق الذكاء العاطفي في كل من السياقات المهنية والشخصية، مما يؤدي إلى نتائج أفضل وتفاعلات أكثر فعالية.

أهداف الدراسة

تطوير معيار شامل لتقييم الذكاء العاطفي في النماذج اللغوية: تقديم معيار EMOBENCH الذي يعتمد على نظريات نفسية راسخة.

اختبار للنماذج اللغوية الكبيرة عبر سيناريوهات معقدة: إنشاء مجموعة من الأسئلة اليدوية التي تتطلب فهمًا عميقًا للعواطف وتوظيفها.

تحديد الفجوات بين أداء التطبيقات والإنسان: مقارنة أداء التطبيقات اللغوية مع الأداء البشري لتحديد الفجوات في الذكاء العاطفي.

المنهج المستخدم

البحث استخدم منهجًا تحليليًا يعتمد على تقييم الذكاء العاطفي للتطبيقات اللغوية الكبيرة (LLMs- Large Language Models) من خلال تطوير معيار .EMOBENCH

هذا المعيار يستند إلى نظريات نفسية قائمة ويشمل تعريفًا شاملاً للذكاء العاطفي في الآلات، والذي يتضمن الفهم العاطفي Emotional Understanding والتطبيق العاطفي Emotional Application

أساليب جمع البيانات

تم جمع البيانات من خلال إنشاء مجموعة من 400 سؤال يدويًا، مصممة بعناية لتتطلب تفكيرًا وفهمًا عميقين. الأسئلة شملت سيناريوهات معقدة تتطلب تقييمًا شاملاً للحالات العاطفية للأفراد المعنيين

طرق تحليل البيانات

تم تحليل البيانات باستخدام نماذج لغوية كبيرة متعددة (مثل GPT-4، GPT-3.5، وغيرها). تم تقييم الأداء عبر سيناريوهات مختلفة واستخدام التصويت بالأغلبية لتحديد الإجابات النهائية للنماذج

نقاط القوة في المنهجية

شمولية المعيار Emobench: يغطي المعيار أبعادًا متعددة للذكاء العاطفي، بما في ذلك فهم وتطبيق العواطف

تنوع السيناريوهات: يشمل سيناريوهات معقدة ومتنوعة مما يزيد من دقة التقييم

التقييم المتعدد اللغات: يتم تقييم النماذج باللغتين الإنجليزية والصينية، مما يعزز من موثوقية النتائج

نقاط الضعف في المنهجية

العينة: العينة محدودة كانت على 48 مشارك فقط، وتم اختبارهم بـ 400 سؤال، مما قد يؤثر على تمثيل البيانات

التركيز على النصوص فقط: يتم تقييم النماذج بناءً على النصوص دون تضمين إشارات عاطفية أخرى مثل نبرة الصوت أو التعبيرات الوجهية

الاعتماد على الأمثلة المولدة بواسطة التطبيق: قد يؤدي إلى تحيز في تصميم السيناريوهات

دراسة النتائج

مكونات الدراسة: النتائج تشمل مقارنة أداء التطبيقات المختلفة في فهم وتطبيق العواطف عبر سيناريوهات متعددة. تم تقييم الأداء باستخدام معايير محددة وتحليل الفجوات بين أداء التطبيقات والإنسان.

نقاط القوة في النتائج:

كشف الفجوات في الأداء: النتائج تظهر فجوة واضحة بين أداء التطبيقات الحالية وأداء الإنسان، مما يشير إلى مجالات للتحسين

تحديد جوانب التحسين: توفر النتائج معلومات قيمة حول نقاط الضعف في التطبيقات والتي يمكن العمل عليها لتحسين الأداء المستقبلي

نقاط الضعف في النتائج:

تأثير التحيزات: قد تتأثر النتائج بتحيزات تصميم السيناريوهات التحيز الثقافي:

السيناريوهات المصممة قد تعكس التحيزات الثقافية للمصممين، على سبيل المثال، ما يعتبره شخص من ثقافة معينة كاستجابة عاطفية طبيعية قد يختلف تمامًا عن استجابة شخص من ثقافة أخرى التحيز الشخصي:

المصممون قد يعكسون تجاربهم الشخصية ومعتقداتهم في تصميم السيناريوهات التحيز المعرفي:

السيناريوهات المصممة قد تكون متأثرة بتحيزات معرفية مثل التحيز التأكيد، حيث يتم تصميم السيناريوهات لدعم افتراضات معينة بدلاً من اختبار قدرات النماذج بموضوعية التحيز في اللغة:

إذا كانت السيناريوهات مكتوبة بلغة معينة أو تعتمد على تعبيرات لغوية محددة، فقد تتأثر النماذج التي تم تدريبها على مجموعة بيانات تختلف في السياق أو الأسلوب اللغوي

نقاط الضعف في النتائج:

محدودية التقييم البشري: عدد المشاركين البشريين كان محدودًا، مما قد يؤثر على تمثيل الأداء البشري

مدى تطبيق النتائج في السياق البحثي: النتائج تعتبر ذات قيمة كبيرة في توجيه الأبحاث المستقبلية نحو تطوير نماذج أكثر ذكاءً عاطفيًا. من خلال فهم الفجوات الحالية في أداء التطبيقات، يمكن التركيز على تحسين الجوانب المحددة في الذكاء العاطفي مثل الفهم العاطفي والتطبيق العملي للعواطف

Emotional Understanding Taxonomy

Main Category	Secondary Category	Example
Complex Emotions	Emotion Transition	<i>I was annoyed for burning my food, but felt joy after he enjoyed it.</i>
	Mixture of Emotions	<i>I'd be happy and disappointed if my friend got a raise instead of me.</i>
	Unexpected Outcome	<i>Laughing hysterically after a tragic event shows devastation.</i>
Personal Beliefs and Experiences	Cultural Value	<i>A waiter getting tipped in Japan would likely not feel gratitude.</i>
	Sentimental Value	<i>I would be unbothered by losing an item without sentimental value.</i>
	Persona	<i>As an introvert, I would feel joy from having to spend a night alone.</i>
Emotional Cues	Vocal cues	<i>A sigh after finishing a hard task shows relief, not disappointment.</i>
	Visual cues	<i>A white face usually indicates fear, unless it's from low blood sugar.</i>
Perspective-taking	Faux pas	<i>Henry criticized the painting on the wall, not knowing I drew it.</i>
	Strange story	<i>Getting an F, where it is the highest mark, leads to pride.</i>
	False belief	<i>Seeing her cry, I thought she was sad, but she'd received great news.</i>

Emotional Application Taxonomy

Relationship	Problem	Example
Social	Self	<i>I work better than my coworkers and I feel my workload is higher because of it.</i>
	Others	<i>I missed a deadline because my coworker didn't tell me about it in time.</i>
Personal	Self	<i>My best friend is going on a road trip with his other friends, leaving me behind.</i>
	Others	<i>My younger brother is being bullied but he begged me not to tell our parents.</i>

Identify Emotions and Their Causes

Category: Complex Emotions (Unexpected Outcome)

Scenario: After a long day of terrible events, Sam started laughing hysterically when his car broke down.

Q: What emotion would Sam feel in this situation?

A) Joy B) Amusement **C) Sadness** D) ...

Q: Why would Sam feel Sadness in this situation?

A) He is overwhelmed by everything happening..

B) He finds the whole situation very entertaining.

...



EmoBench

Relationship: Personal **Problem:** Interpersonal

Scenario: Rebecca's son lost his soccer game and is feeling very upset and blames himself for the loss.

Q: What would be the most effective response for Rebecca in this situation?

A) "Losses teach us important lessons. Let's train harder for next time. "

B) "You should've played better. Can't change it now."

...

Use Understanding to Manage Thoughts

Figure 2: Overview of Our Benchmark (EMOBENCH).

الحمد لله رب العالمين