Divulga UAB
Revista de difusió de la recerca de la Universitat Autònoma de Barcelona

29/01/2025

El teu assistent virtual preferit et respon, però realment t'entén?

Les conversacions amb assistents virtuals són cada cop més freqüents i s'incorporen fins i tot en empreses privades. Molta gent troba que les respostes són prou convincents per a semblar humanes. Ara bé, desenvolupen realment el llenguatge com les persones? Un nou article examina la responsabilitat lingüística i el raonament en la construcció d'oracions a través d'un estudi que compara 400 persones i 7 models de llenguatge d'última generació.

Imatge generada per GPT

Els Models de Llenguatge Extensos ("Large Language Models", LLM) suposen un dels avanços tecnològics més impressionants que hem presenciat en els últims anys. S'utilitzen amb èxit en diverses aplicacions que inclouen la predicció de la següent paraula, com ara autocompletar el text i respostes automàtiques a preguntes mitjançant agents artificials (bots o assistents virtuals). Actualment, aquestes habilitats lingüístiques es presenten com a prou convincents per l'ull no entrenat que molta gent argumenta que els resultats d'aquests models semblen una resposta humana.

Però, els LLM desenvolupen realment el llenguatge com les persones? Hauríem de fer que ho fessin? Una companyia aèria va ser requerida a pagar-li els danys a un passatger que va rebre informació incorrecta mentre mantenia una conversa amb el bot de la

companyia. Segons un representant de l'empresa, el bot va incloure "paraules enganyoses" a les respostes de les preguntes del client. Finalment, el jutge va emetre una sentència de representació distorsionada negligent a favor del passatger, tot i que la companyia va continuar argumentant que el bot **és** i **hauria de ser el responsable** de les seves pròpies paraules. Aleshores, els xats amb bots, assistents virtuals i altres aplicacions interactives relacionades amb aquestes tecnologies predictives tenen una comprensió del llenguatge semblant a la humana o la seva habilitat està inherentment limitada?

Per poder respondre a aquesta pregunta, personal investigador de la Universitat Rovira i Virgili, la Universitat de Pavia, la Universitat Humboldt de Berlín, la Universitat de Nova York, la Universitat Autònoma de Barcelona i la Institució Catalana de Recerca i Estudis Avançats (ICREA) han comparat 400 persones i 7 models d'última generació en un nou punt de referència que implica indicis lingüístics molt simples. L'objectiu era oferir als models les millors condicions possibles per respondre correctament. El test involucrava processar i respondre oracions com "El John va enganyar la Mary i la Lucy va enganyar la Mary. En aquest context, la Mary va enganyar la Lucy?"

Com era d'esperar, les persones van completar la tasca amb èxit. Els LLM, en canvi, van presentar molta varietat en les seves respostes, de manera que alguns models van respondre millor que d'altres. Així, els LLM com a classe resulten pitjor que les persones. Més notable és el fet que els models van cometre tipus d'errors que eren completament absents de les respostes humanes. Per exemple, a l'enunciat "El Franck es llegeix a ell mateix i el John llegeix a ell mateix, a l'Anthony i al Franck. En aquest context, el Franck va ser llegit?", algun dels models testats van respondre que "en aquest context, és impossible dir amb seguretat qui llegia al Franck" i que, per contestar, necessitaríem saber "informació addicional sobre la situació específica, com el material de lectura del John".

El resum d'aquests resultats per l'autora principal d'aquest estudi, la Prof. Evelina Leivada (UAB i ICREA), revela que quan es rasca la superfície d'un aparent bon rendiment lingüístic, el rendiment lingüístic dels LLM pot amagar defectes inherents a la modelització del llenguatge com a mètode. El missatge a destacar és que la intel·ligència, el raonament i l'ancoratge de paraules en les condicions del món real no pot emergir com un producte secundari de la inferència estadística. Com remarca un altre autor de l'estudi, el Prof. Gary Marcus, en el seu llibre publicat el 2024: "Taming Silicon Valley. How We Can Ensure that AI Works for Us", els sistemes d'Intel·ligència Artificial són indeferents a la veritat que s'amaga darrere de les seves paraules, fet que genera preocupacions sobre la desinformació massiva, la difamació, la contaminació del mercat i la magnificació dels biaixos que es produeixen a gran escala.

Evelina Leivada
Personal investigador ICREA, Departament de Filologia Catalana
Universitat Autònoma de Barcelona
Evelina.Leivada@uab.cat

Referències

Dentella, V., Günther, F., Murphy, E., Marcus, G. & Leivada, E. **Testing AI on language comprehension tasks reveals insensitivity to underlying meaning.** *Scientific Reports* 14, 28083 (2024). https://doi.org/10.1038/s41598-024-79531-8

View low-bandwidth version