ChatGPT i la fi d'un món

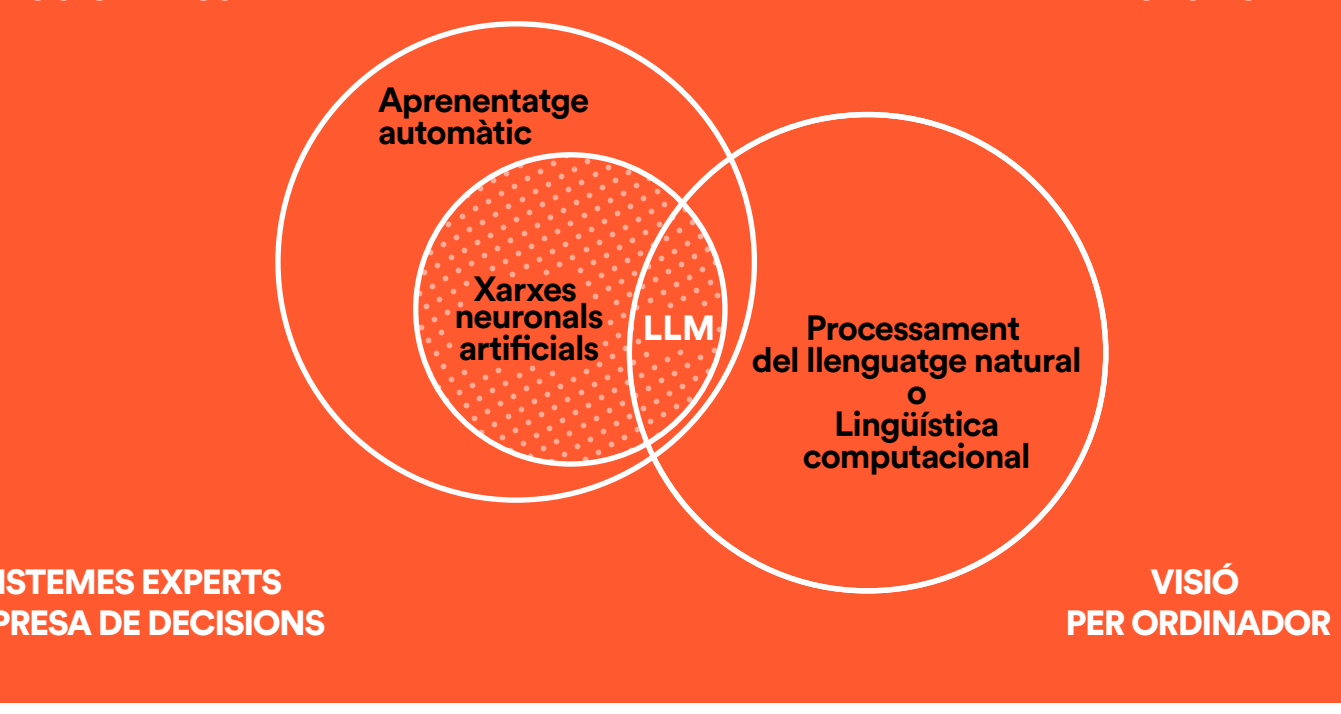
EPISODI 1

QUÈ SÓN ELS LLM?

Raquel Fernández
Catedràtica de Lingüística Computacional
Institute for Logic, Language and Computation
Universitat d'Amsterdam

Eines com ChatGPT han esdevingut molt populars darrerament. ChatGPT és un exemple de «model de llenguatge gran», o *large language model* (LLM) en anglès. Aquest tipus de tecnologia fa uns anys que existeix dins del camp de la lingüística computacional, que és una subdisciplina de la intel·ligència artificial.

Intel·ligència artificial



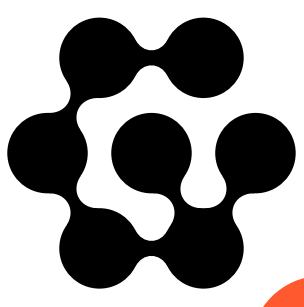
Els models de llenguatge grans més recents són capaços de generar automàticament textos que s'assemblen molt als textos produïts per éssers humans.

COM HO FAN?

Un LLM és

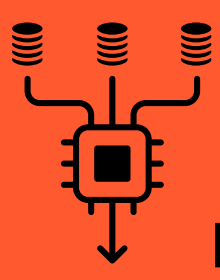
Un model de llenguatge és un programa d'ordinador dissenyat per humans (investigadors en IA, en lingüística computacional, etc.).

Està basat en un model matemàtic anomenat ***xarxa neuronal artificial***.G



Es tracta d'un programa que utilitza la tècnica d'**aprenentatge automàtic** (*machine learning*): un LLM aprèn a generar textos i s'entrena, per generar-los, per mitjà de dades (moltes dades!).

El tipus de xarxa neuronal artificial que fan servir els LLM actuals s'anomena *transformador*.



LA SIGLA ***GPT*** FA REFERÈNCIA A L'EXPRESSIÓ ANGLESA *GENERATIVE PRE-TRAINED TRANSFORMER*

Per què s'anomenen *models de llenguatge «grans»*?

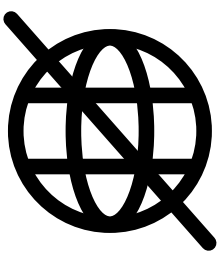
S'anomenen «grans» perquè les xarxes neuronals artificials que hi ha darrere dels LLM tenen una estructura complexa, amb milers de milions de **paràmetres**, G que són com les palanques que controlen la resposta del sistema. A més a més, necessiten grans quantitats de material d'aprenentatge per poder aprendre a produir textos de bona qualitat.

Per entrenar un model d'aquest tipus, calen milers d'ordinadors treballant en paral·lel durant mesos. Cal tenir en compte que això requereix un consum energètic considerable, que pot tenir conseqüències negatives per al medi ambient.

Quin és el *material d'aprenentatge?*

El **material d'aprenentatge** (*training data*) és un conjunt de textos que ha de ser tan gran i divers com sigui possible. El més habitual és utilitzar textos descarregats d'internet, per exemple de Vikipedia, de llibres i diaris digitals, de blogs o de xarxes socials, com Twitter, Instagram o Facebook.

En principi, un model del llenguatge no està directament connectat a internet: durant la fase d'entrenament, només té accés als textos descarregats que formen el material d'aprenentatge.



Cal tenir en compte que l'ús d'aquest tipus de dades pot ser problemàtic: per una banda, hi pot haver problemes relacionats amb privacitat, consentiment i propietat intel·lectual; per altra banda, pot ser que les dades estiguin esbiaixades i no representin correctament certs sectors de població.

Què aprèn a fer un LLM?

La idea fonamental és molt senzilla: un LLM aprèn a **predir la paraula següent** en una frase o text, donat un context inicial o *prompt*. G

Per exemple, no totes les paraules del diccionari tenen la mateixa probabilitat d'aparèixer després d'un fragment de text com «Avui fa...»: les paraules *sol*, *exactament*, *quinze* o *molt* són possibles continuacions, mentre que paraules com *casa*, *ballar* o *és* tenen una probabilitat molt baixa (gairebé nul·la) de ser la paraula següent.

A partir de totes les frases i textos que formen part del **material d'aprenentatge**, el model aprèn a assignar les probabilitats adequades a cada paraula possible i, per extensió, a cada frase. És a dir, aprèn els patrons que determinen quines frases són naturals i raonables en un context donat o com a resposta a una consulta.

Predir una paraula darrere de l'altra permet generar textos, respostes i converses senceres. És per això que els LLM es consideren un exemple d'**intel·ligència artificial generativa.**

Aprenentatge no supervisat

Exemple de text d'entrenament:

Vaig llegir un **còmic** **molt entretingut**

Donat aquest *prompt* o context

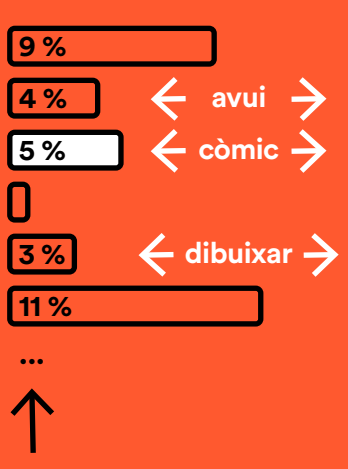
Volem que el model aprengui a predir aquesta paraula

Vaig llegir un...

còmic? **dibuixar?** **avui?**

Exemple de predicció del model quan encara està aprenent
(el model prediu una distribució de probabilitat)

xarxa neuronal

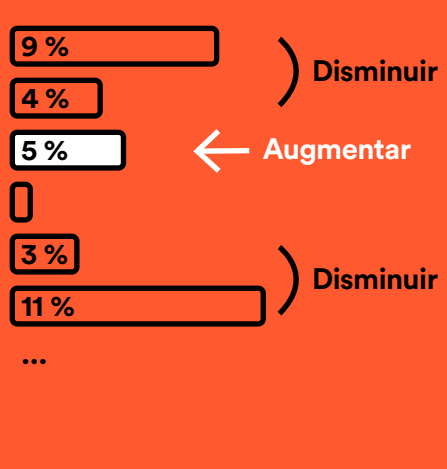


Llista de totes les paraules possibles
(la llista és tan llarga com el diccionari sencer).
Cada barra representa la probabilitat assignada a cada paraula, donat el context o *prompt*.

Predicció correcta en aquest cas

0 %
0 %
100 %
0 %
0 %
0 %
...

Correcció de la distribució de probabilitat durant l'aprenentatge



Com s'entrena un LLM com ChatGPT?

ChatGPT ha estat creat per una empresa privada, OpenAI, que no ha revelat tots els detalls sobre com ha entrenat aquest model concret. Tanmateix, sabem que l'entrenament inclou tres fases:

En una primera fase, tots els LLM s'entrenen a predir la paraula següent donat un context d'entrada.

Això es considera **aprenentatge no supervisat**, perquè el model aprèn a partir de grans quantitats de textos existents, sense intervenció humana addicional. Els LLM com ChatGPT passen per dues fases d'aprenentatge més.

1

En una segona fase, treballadors humans produeixen exemples de converses entre dues persones que simulen converses entre una persona i ChatGPT.

2

Aquestes converses s'utilitzen com a material d'aprenentatge addicional. L'objectiu segueix sent aprendre a predir la paraula següent, però ara en un context de diàleg interactiu.

Cal mencionar que les condicions de treball i les remuneracions d'aquests treballadors no són òptimes.

Finalment, en una tercera fase es fa servir l'aprenentage supervisat: quan ja és possible interactuar amb el model, les respostes generades automàticament per ChatGPT són avaluades per més treballadors, els quals les etiqueten com a «bones» o «dolentes» (per exemple, respostes amb llenguatge ofensiu o sobre temes controvertits s'etiquetaran com a «dolentes»).

3

Aquesta informació s'utilitza com a *feedback* amb l'objectiu de potenciar respostes que s'adeqüin a les preferències humanes. Aquest tipus d'aprenentatge s'anomena ***aprenentatge de reforç a partir de les preferències humanes,*** o *reinforcement learning with human feedback* (RLHF) en anglès.

Ens podem refiar dels textos i respostes generats per un LLM?

Cal tenir present que els models de llenguatge **no incorporen cap mecanisme que asseguri la veracitat del contingut** que produeixen: les frases que generen poden sonar totalment plausibles i raonables, però no hi ha garantia que siguin certes (de vegades ho seran, i de vegades, no).

El fet que els textos generats siguin gramaticals i tinguin una forma natural pot donar lloc a l'**antropomorfització** (G) dels *chatbots* basats en models de llenguatge. És a dir, sovint assignem qualitats humanes (com intel·ligència, raonament lògic o emocions) a màquines que només aparenten tenir-ne. Tanmateix, aquests trets no formen part de les capacitats d'un model de llenguatge.

Preguntes de reflexió

1. Quin aspecte de tot allò après aquí sobre els LLM t'ha cridat més l'atenció?
2. ChatGPT és un exemple d'LLM. En coneixes més? Bing AI, Bard... Et sonen? Què poden fer? La tecnologia avança ràpidament. Mentre treballem en la publicació d'aquest capítol, ChatGPT ha anunciat la seva pròxima connexió a internet (en principi, un LLM no està connectat a internet) i Bing ha anunciat una extensió per a la generació d'imatges. En quin punt estem en el moment en què llegeixes això?
3. Quin paper tenen la lògica i la coherència en l'avaluació de les respostes generades per ChatGPT? Què faries per distingir entre una resposta vàlida i una d'incorrecta?
4. Algunes veus alerten sobre el perill d'assignar característiques humanes a una intel·ligència artificial. A tu et sembla perillós? Per què?
5. La intervenció de les persones en la creació, entrenament i resposta dels LLM és importantíssima. Pots assenyalar-ne alguns exemples? Consideres que la implicació de les persones/agents humans és suficientment agraïda, valorada, apreciada? Per què creus que és així? Quines implicacions té que sigui així?
6. Si fas servir LLM per als teus deures, consideres una obligació mencionar-ho? Per què?

GLOSSARI

- **ANTROPOMORFITZACIÓ**: Atorgar, a un animal o cosa, característiques o motivacions humanes.
- **PARÀMETRES**: Valors numèrics que defineixen el comportament del model.
- ***PROMPT***: Instrucció, pregunta o frase inicial proporcionada al model de llenguatge per guiar-ne la resposta o generació de text.
- **XARXA NEURONAL ARTIFICIAL**: Model inspirat en el funcionament del cervell humà. Està format per un conjunt de nodes, coneguts com a *neurones artificials*, que estan connectats i transmeten senyals els uns als altres. Aquests senyals es transmeten des de l'entrada fins a generar una sortida.

REFERÈNCIES PER SABER-NE MÉS

a. Bender, Emily M. et al. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?» en *Proceedings of FAccT 2021*, p. 610-623 (2021, dl.acm.org)
b. Brown, Tom B. et al. «Language Models are Few-Shot Learners» a *Advances in Neural Information Processing Systems* (p. 1877-1901). Curran Associates (2020, neurips.cc)
c. Pointon, Chris. «The carbon footprint of ChatGPT» en *Blogpost sobre l'impacte mediambiental de ChatGPT* (2022, medium.com)
d. «Circuit OEIAC sobre l'ús responsable i sostenible del ChatGPT» (2023, udg.edu)
e. «Beijing Consensus on Artificial Intelligence and Education» (2019, unesco.org)
f. Sabzalieva, Emma i Valentini, Arianna. «ChatGPT e Inteligencia Artificial en la educación superior. Guía de inicio rápido» (2023, unesco.org)

Text: Raquel Fernández / Disseny: La Puput Gràfica Coop V
Fundació "la Caixa", 2023