ChatGPT i la fi d'un món

EPISODI 2

QUINA LÒGICA HI HA DARRERE D'UN LLM?

Tommaso Flaminio, Lluís Godó
IIIA – Institut d'Investigació en Intel·ligència Artificial
CSIC – Consejo Superior de Investigaciones Científicas

«Es important 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 (a vegades ho seran i altres vegades no).»

Raquel Fernández, episodi 1

Què es la *lògica*?

La lògica és la disciplina que estudia els passos necessaris per generar un raonament vàlid.

Per fer-ho, proporciona principis i regles, anomenades *regles d'inferència*, per analitzar l'estructura d'un argument i assegurar conclusions vàlides a partir d'un punt de partida inicial.

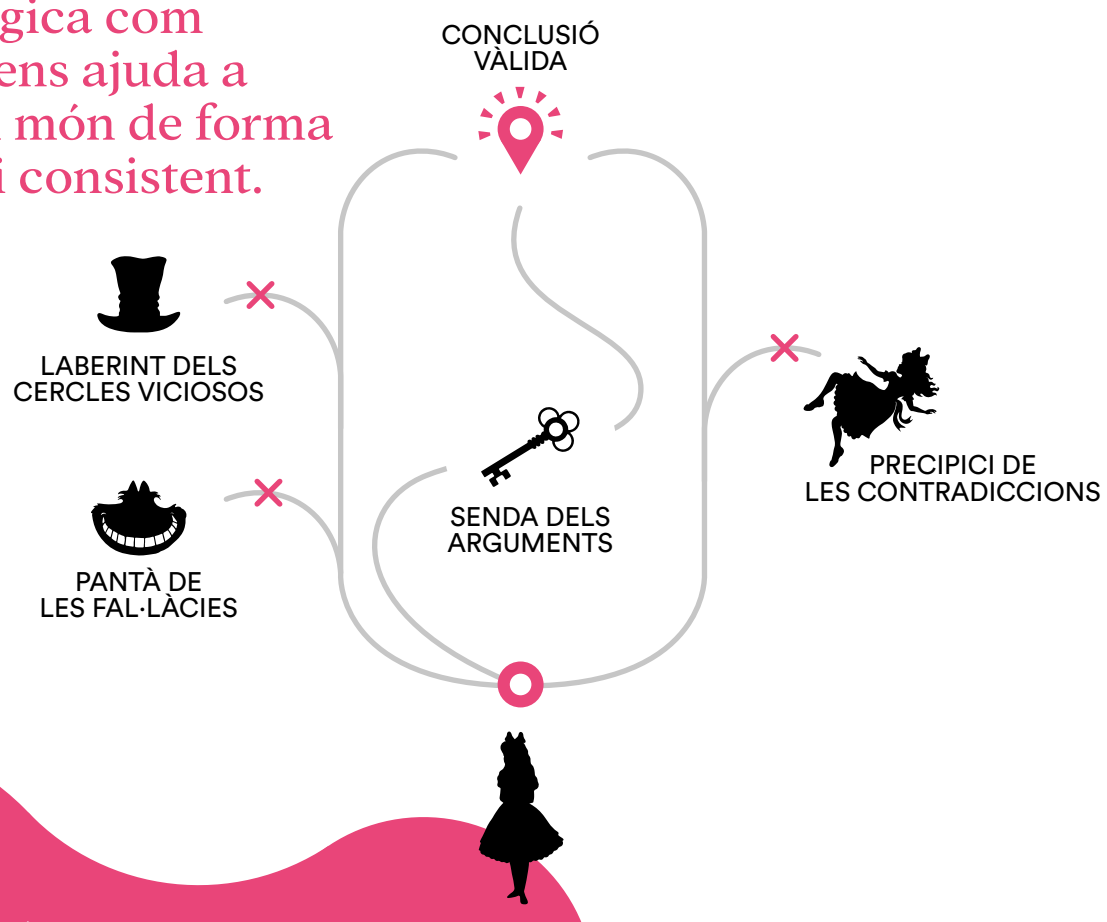
AQUESTES SÓN ALGUNES DE LES REGLES D'INFERÈNCIA DE LA LÒGICA CLÀSSICA

Si és veritat que plou, aleshores no és veritat que no plou.	$A \to no\ (no\ A)$
En qualsevol moment, o plou o no plou.	A o $no\ A$
Si plou i, quan plou, es mulla el terra, aleshores el terra està mullat.	$(A\ \&\ A \to B) \to B$
Si el terra no està mullat i, quan plou, es mulla el terra, aleshores no plou.	$(no\ B\ \&\ A \to B) \to no\ A$
Si, quan plou, es mulla el terra, i quan es mulla el terra, surt en Kim a jugar amb els caragols, aleshores, quan plou, surt en Kim a jugar amb els caragols.	$(A \to B\ \&\ B \to C) \to A \to C$
Si d'allò que diu en Kim es dedueix que plou i, també, que no plou, aleshores en Kim menteix.	$(K \to A\ \&\ K \to no\ A) \to no\ K$

Per a què serveixen *les regles de la lògica?*

La lògica no tracta de distingir la veritat de la falsedat. Proposa una sèrie de regles per identificar quan un raonament està ben construït; i ens garanteix que, si donem per vàlides les premisses del raonament, aleshores la conclusió també serà vàlida.

Podem imaginar les regles de la lògica com un mapa que ens ajuda a representar el món de forma estructurada i consistent.

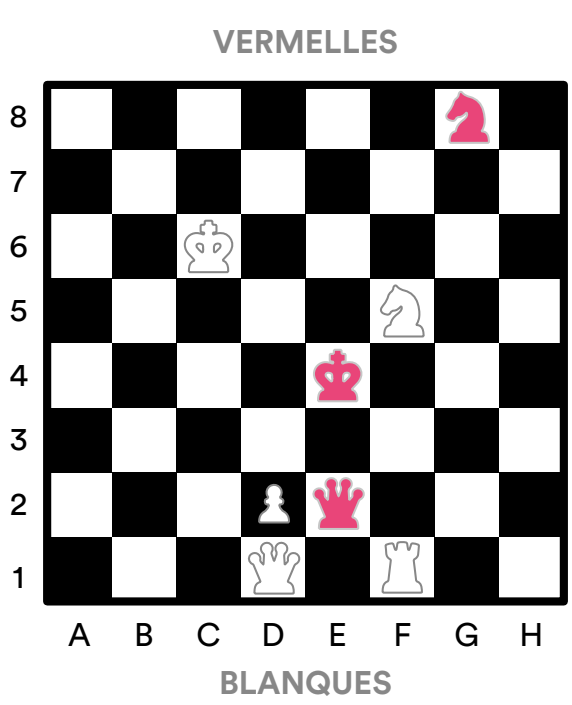


Com establim *aquestes regles?*

Primer, hem de triar l'àmbit sobre el qual volem pensar, el món en el qual desenvolupar els nostres raonaments. Per exemple, **el món dels nombres, el de les lleis, el dels jocs...**

Després, construïm un llenguatge: un conjunt de paraules i símbols que ens permeten representar i descriure els objectes d'aquest àmbit, així com les seves propietats fonamentals.

- Color blanc o vermell.
- Peó, reina, rei, torre, alfil...
- Enroc, escac, doble amenaça...
- 1, 2, 3, 4, 5, 6, 7, 8.
- A, B, C, D, E, F, G, H.



Finalment, definim un conjunt de regles d'inferència que ens permeten obtenir conclusions noves o validar una argumentació.

- No hi pot haver dues peces en la mateixa casella.
- Si una peça d'un color accedeix a una casella ocupada per una peça d'un altre color, aleshores la primera peça captura la segona.
- Guanya qui capturi el rei sense que aquest pugui escapar.
- ...

A TRAVÉS DEL MIRALL I ALLÒ QUE L'ALÍCIA HI VA TROBAR

Lewis Carroll

«El peó blanc (Alícia) juga i guanya en 11 jugades.»

Et convidem a descobrir com, al llibre, l'Alícia es mou seguint les regles dels escacs en les seves aventures al país al qual arriba a través del mirall.

Podem instal·lar aquestes regles *en una màquina*?

Això va ser el que es va plantejar un matemàtic anomenat **Alan Turing**.

L'any 1936 va idear una màquina imaginària molt especial: un dispositiu programable capaç de rebre informació (per exemple, un nombre), processar-la seguint unes regles (per exemple, dividir aquest nombre per 2 tantes vegades com es pugui sense obtenir decimals) i oferir el resultat (dir si el nombre és parell o senar). La lògica és essencial per a aquesta màquina, ja que li permet definir les regles i operacions que ha de seguir per fer els seus càlculs.

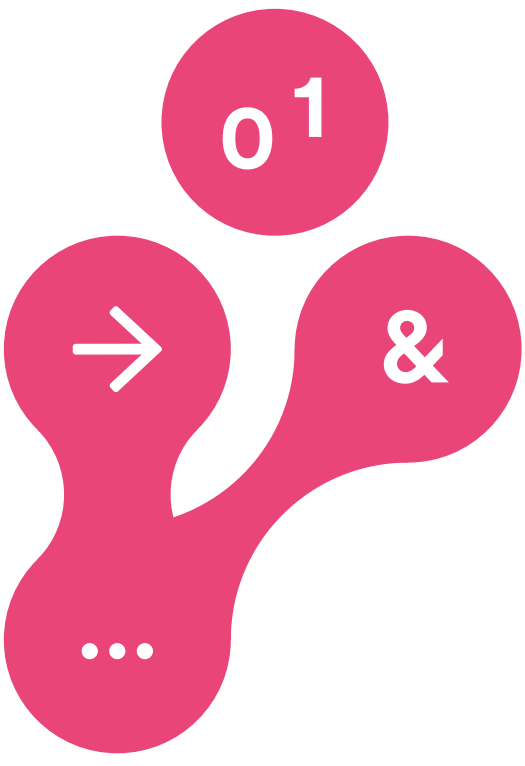
Gràcies a aquest enfocament lògic, Alan Turing va poder plantejar preguntes fonamentals sobre els límits i capacitats dels ordinadors.

«Proposo considerar la pregunta següent: **"Poden pensar les màquines?"**.»

Al llarg dels anys, la lògica bàsica en què es va recolzar Turing s'ha anat desenvolupant i ha donat lloc a variants i extensions.

Amb aquestes potents eines lògiques, es dissenyen màquines amb unes certes **habilitats de «pensament»**:

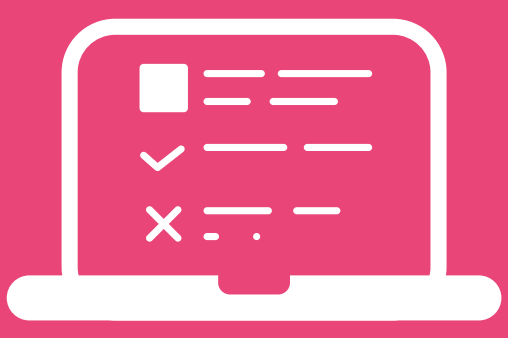
1. poden representar el **coneixement** mitjançant símbols; i
2. poden representar el **raonament** manipulant aquests símbols, sempre seguint les regles lògiques.



Aquest tipus d'intel·ligència artificial, basada en la lògica i la manipulació de símbols, s'anomena *IA simbòlica*.

El camp de la IA que s'ocupa d'aquest tipus de sistemes s'anomena *representació del coneixement i el raonament*.

Els experts en aquest camp estan constantment investigant i desenvolupant noves maneres d'utilitzar la lògica perquè les màquines puguin raonar de manera automàtica.

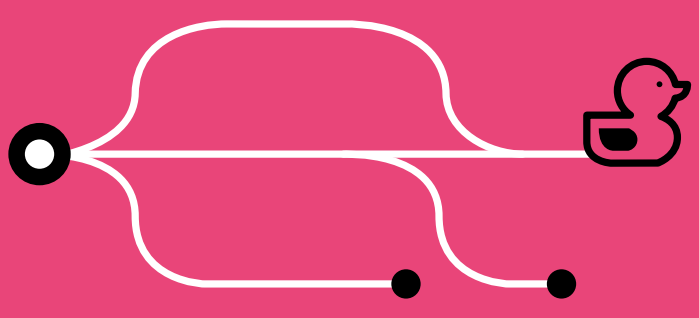


Un exemple d'aquestes màquines són els anomenats *sistemes de decisió basats en el coneixement*.

BASE DE CONEIXEMENT → **DECISIÓ DEL SISTEMA**

- Sembla un ànec.
- Camina com un ànec.
- Neda com un ànec.
- Claca como un ànec.

- Probablement és un ànec.

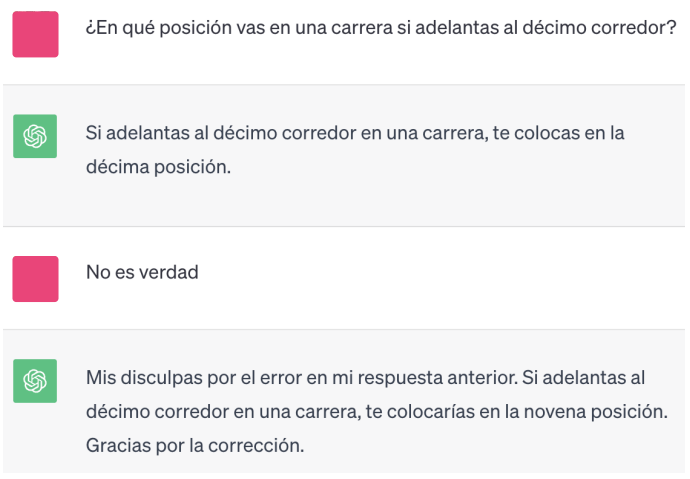


Un LLM, funciona com una màquina lògica i simbòlica?

No. Un LLM té un enfocament diferent, està dins de les aplicacions de la IA conegudes com a «generatives».

1

La generació d'informació de ChatGPT no és simbòlica; recordeu-ho: es basa en la predicció o la generació de la paraula següent. Aquesta predicció es fonamenta en tècniques i models de la *IA subsimbòlica*, com ara les xarxes neuronals profundes.



Dins d'un LLM, no hi ha cap mecanisme per deduir, mitjançant regles i manipulació de símbols, noves veritats a partir d'un conjunt d'informació prèvia.

I això planteja un desafiament: la nova informació generada per la IA no sempre és fiable.

2

Aquí és on la lògica té un paper crucial en la detecció de veritats o falsedats, o, més ben dit, de conclusions vàlides, consistents amb la informació que considerem veritable.

La pràctica i la familiaritat de l'usuari amb la lògica seran fonamentals per aprofitar al màxim aquests sistemes d'IA generativa.

En l'exemple anterior, **pots demostrar que ChatGPT està equivocat?**

?

Aleshores, *un LLM pot «pensar»?*

Quan Turing es pregunta si les màquines poden pensar, proposa un joc per determinar-ho.

CURIOSAMENT, NO TÉ A VEURE AMB LA DESTRESA LÒGICA, SINÓ AMB LA NATURALITAT D'UNA CONVERSA.

El joc de Turing, que ell va anomenar *el joc de la imitació*, proposa acceptar que una màquina és intel·ligent si, després d'interaccionar per escrit amb un humà, és capaç d'«enganyar-lo» i fer-se passar per humà.

!

El test de Turing no és un test homologat i formal, ni té unes preguntes prefixades, ni una durada determinada. És, més aviat, un joc filosòfic per explorar conceptes profunds relacionats amb la ment i la tecnologia.

Test de Turing

Aquestes són algunes possibles preguntes d'un test de Turing:

Si poguessis tenir una conversa amb qualsevol figura històrica, qui seria i per què?

Pots parlar-me d'alguna ocasió en què vas intentar fer alguna cosa nova i no va sortir com esperaves? Què en vas aprendre, d'aquesta experiència?

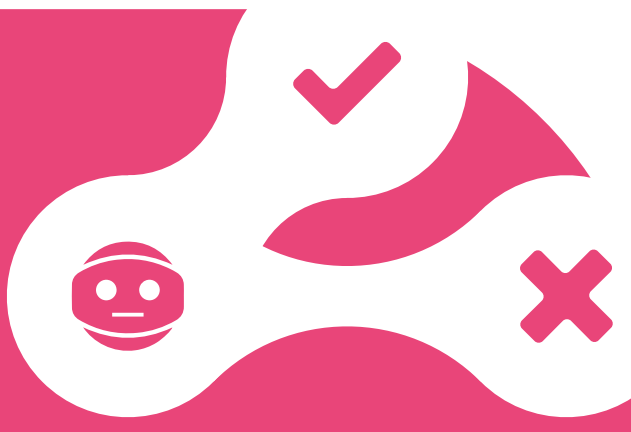
Dona un exemple d'una fal·làcia lògica i explica per què és enganyosa.

Si poguessis tenir un superpoder, quin seria i per què?

Dona dues interpretacions de la frase: «Vaig veure un home amb el telescopi».

Què li preguntaries tu?

Algun LLM ha passat el *test de Turing*?



Sí! El LLM de Google, LaMDA, va passar el test. Però això no significa res en concret, només fa més apassionant el camp de la intel·ligència artificial i ens empeny a seguir investigant i preguntant.

1. Si entenem el raonament com la capacitat de generar nova informació a partir d'una base de dades o coneixements previs, de manera que superi el test de Turing, aleshores podem afirmar que, en general, un LLM raona i exhibeix un comportament intel·ligent en la majoria de les tasques.

2. No obstant això, el seu procés de generació d'informació no segueix una lògica deductiva i simbòlica tradicional, sinó que es basa en models de llenguatge avançats. Això pot considerar-se com una «caixa negra» en la qual, encara que la informació generada sembli plausible, no sempre és necessàriament vàlida.

3. Aquesta observació planteja un desafiament interessant per als experts en lògica, que consisteix a comprendre quin tipus de raonament (que no és necessàriament deductiu) hi ha al darrere d'una intel·ligència artificial que mostra un comportament intel·ligent, com aquells sistemes basats en models de llenguatge avançats.

I aquest desafiament ens convida a seguir aprofundint en la pregunta fonamental: **què hi passa, al nostre cervell, quan pensem?**

Preguntes de reflexió

1. **Creus que la capacitat de «pensar» d'una màquina és igual a la d'un ésser humà? Per què sí o per què no?**

2. **Quina és la diferència entre la lògica deductiva i el raonament basat en models de llenguatge, i com poden coexistir en la presa de decisions?**

3. **Quines implicacions ètiques podrien sorgir quan utilitzem sistemes d'intel·ligència artificial per prendre decisions importants?**

4. **Quina és la teva opinió sobre la idea que les màquines poden «enganyar» els humans en una conversa escrita? Creus que això és una forma d'intel·ligència?**

REFERÈNCIES PER SABER-NE MÉS

a. Turing, Alan (1936). «On Computable Numbers, with an Application to the Entscheidungsproblem (*Decision Problem*)». *Proceedings of the London Mathematical Society, 2*, 42.

b. Turing, Alan (1950). «Computing Machinery and Intelligence». *Mind, 59*, 236.

c. Carroll, Lewis (1998). «Alicia Anotada» (Edició de Martin Gardner). Akal.

B BIBLIOGRAFIA

- Deaño, Alfredo (1999). «Introducción a la lógica formal». Alianza Editorial.
- Petzold, Charles (2008). «The Annotated Turing: A Guided Tour Through Alan Turing's Historic Paper on Computability and the Turing Machine». Wiley.
- Cerf, Vinton G. (2023). «Large Language Models». *Cerf. Communications of the ACM 7*, vol. 55, núm. 8.
- Dickson, Ben (2022). «Large Language Models have a Reasoning Problem». TechTalks. https://bdtechtalks.com/2022/06/27/large-language-models-logical-reasoning
- Wolfram, Stephen (2023). «What Is ChatGPT Doing… and Why Does It Work?». https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/#what-really-lets-chatgpt-work?

Text: Tommaso Flaminio, Lluís Godó / Disseny: La Puput Gràfica Coop V
Fundació "la Caixa", 2023