Revista científica editada por la Facultad de Información y Medios Audiovisuales de la Universidad de Barcelona
ISSN 1575-5886

BiD 54 (Junio 2025)
DOI: 10.1344/bid2025.54.05

ARTÍCLE ORIGINAL

Desxifrar sigles mèdiques en català amb ChatGPT, Gemini i Copilot: una anàlisi comparativa

Deciphering Medical Acronyms in Catalan with ChatGPT, Gemini, and Copilot: A Comparative Analysis

Descifrar siglas médicas en catalán con ChatGPT, Gemini y Copilot: un análisis comparativo

Adéla Kotatkova
Universitat Jaume I
kotatkov@uji.es

Rebut: 05/05/2025
Acceptat: 28/06/2025

Citación recomendada

Kotatkova, Adéla (2025). Desxifrar sigles mèdiques en català amb ChatGPT, Gemini i Copilot: una anàlisi comparativa. *BiD*, 54. https://doi.org/10.1344/BID2025.54.05.

Resum

Objectius

Aquest estudi té com a objectiu avaluar la capacitat de tres xatbots d'intel·ligència artificial (IA) —ChatGPT, Gemini i Copilot— per desxifrar sigles mèdiques polisèmiques en català, tot analitzant la seva eficàcia en funció del context. L'objectiu és determinar fins a quin punt aquestes eines poden ajudar a desambiguar les sigles per millorar la comprensió de textos mèdics per a professionals sanitaris, de comunicació i pacients.

Metodologia

Es van seleccionar 60 sigles mèdiques amb alta polisèmia. A cada sigla se li van assignar quatre contextos diferents: sense context, context mèdic genèric, una frase real i un paràgraf breu. Es va mesurar la proporció d'encerts de cada IA en cada context, mitjançant un càlcul de potència estadística. Els resultats es van analitzar utilitzant proves no paramètriques, amb l'objectiu de comparar la precisió entre les tres IA.

Resultats

Els resultats obtinguts mostren una gran variabilitat en la capacitat dels xatbots per reconèixer i interpretar les sigles mèdiques en català en funció del context. Els sistemes presenten millor eficiència a mesura que augmenta la informació contextual, especialment quan les sigles apareixen en

frases o paràgrafs reals. ChatGPT obté els millors resultats, amb una desambiguació més precisa en contextos naturals. Aquest estudi demostra el potencial de les eines d›intel·ligència artificial per millorar la comprensió de les sigles mèdiques per a pacients i professionals, una necessitat clau en un llenguatge mèdic català encara poc desenvolupat.

Paraules clau

Sigles, intel·ligència artificial, català, llenguatge mèdic, terminologia, polisèmia, desambiguació.

Abstract

Goals: This study aims to assess the ability of three artificial intelligence (AI) chatbots—ChatGPT, Gemini, and Copilot—to decode polysemous medical acronyms in Catalan, analysing their effectiveness depending on the level of contextual information. The objective is to determine the extent to which these tools can assist in disambiguating acronyms to improve the comprehension of medical texts by healthcare professionals, communicators, and patients.

Methodology: A total of 60 highly polysemous medical acronyms were selected. Each acronym was presented within four different levels of context: no context, general medical context, a real sentence, and a brief paragraph. The accuracy rate of each AI system was measured across all contexts using statistical power analysis. The results were analysed using non-parametric tests to compare the precision of the three AI tools.

Results: The findings reveal significant variability in the chatbots' ability to recognize and interpret medical acronyms in Catalan depending on the context. Performance improved as more contextual information was provided, especially when acronyms appeared within real sentences or paragraphs. ChatGPT achieved the highest overall accuracy, offering more precise disambiguation in natural contexts. The study highlights the potential of AI tools to enhance the understanding of medical acronyms among both patients and professionals—an increasingly important need in the context of Catalan medical language, which remains underdeveloped

Keywords

Acronyms, artificial intelligence, Catalan, medical language, terminology, polysemy, disambiguation.

Resumen

Objetivos: Este estudio tiene como objetivo evaluar la capacidad de tres chatbots de inteligencia artificial (IA) —ChatGPT, Gemini y Copilot— para descifrar siglas médicas polisémicas en catalán, analizando su eficacia en función del contexto. El objetivo es determinar hasta qué punto estas herramientas pueden ayudar a desambiguar las siglas para mejorar la comprensión de textos médicos por parte de profesionales sanitarios, de la comunicación y pacientes.

Metodología: Se seleccionaron 60 siglas médicas con alta polisemia. A cada sigla se le asignaron cuatro contextos diferentes: sin contexto, contexto médico genérico, una frase real y un párrafo breve. Se midió la proporción de aciertos de cada IA en cada contexto mediante un cálculo de potencia estadística. Los resultados se analizaron utilizando pruebas no paramétricas, con el objetivo de comparar la precisión entre las tres IA.

Resultados: Los resultados obtenidos muestran una gran variabilidad en la capacidad de los chatbots para reconocer e interpretar las siglas médicas en catalán en función del contexto. Los sistemas presentan mejor eficiencia a medida que aumenta la información contextual, especialmente cuando las siglas aparecen en frases o párrafos reales. ChatGPT obtiene los mejores resultados, con una desambiguación más precisa en contextos naturales. Este estudio demuestra el potencial de las herramientas de inteligencia artificial para mejorar la comprensión de las siglas médicas por parte de pacientes y profesionales, una necesidad clave en un lenguaje médico catalán aún poco desarrollado.

Palabras clave

Siglas, inteligencia artificial, catalán, lenguaje médico, terminología, polisemia, desambiguación.

1. Introducció

Les sigles mèdiques són un recurs habitual en la redacció de textos sanitaris, tant en els especialitzats adreçats a col·legues com en els divulgatius dirigits a pacients. Tot i que compleixen una funció d'economia textual i poden facilitar la fluïdesa de l'escriptura, sovint actuen com a barreres per a la comprensió, especialment per part de persones no expertes. Una de les principals causes d'aquesta opacitat és la polisèmia (Navarro, 2008), la qual representa un repte tant per a la lexicografia com per a les tecnologies de processament del llenguatge natural. Aquesta inestabilitat lexicològica contribueix a la dificultat de desambiguació, especialment en llengües com el català que no disposen de repertoris específics. Una mateixa sigla pot tenir diversos significats segons el context, l'especialitat mèdica o la llengua de referència. De fet, molts textos mèdics en català inclouen sigles que prenen com a base el terme original en castellà o en anglès (Koťátková, 2023), cosa que incrementa la complexitat terminològica i redueix la transparència del discurs sanitari.

La magnitud del problema és notable. Segons el *Siglas médicas en español. Repertorio de siglas, acrónimos, abreviaturas y símbolos utilizados en los textos médicos en español* (Navarro, 2023), més de 15.000 sigles tenen més d'un significat. D'aquestes, 4.882 tenen dues accepcions, 2.558 en tenen tres, 1.673 en tenen quatre i 5.967 en tenen cinc o més —fins a gairebé 50 accepcions diferents en alguns casos.

La dificultat per desxifrar aquestes sigles s'agreuja encara més en un entorn comunicatiu digitalitzat, on tant pacients com professionals recorren cada vegada més a cercadors i a xatbots d'intel·ligència artificial (IA) per obtenir aclariments terminològics. Tanmateix, la capacitat real d'aquests sistemes per interpretar correctament sigles mèdiques en català i oferir respostes contextualment adequades encara no ha estat objecte d'una anàlisi sistemàtica. En particular, cal explorar fins a quin punt aquestes eines poden desambiguar sigles polisèmiques en funció del context en què apareixen.

Aquest article ofereix una primera aproximació empírica a aquesta qüestió en català mitjançant l'anàlisi comparativa de tres xatbots d'ús generalitzat. L'objectiu principal de l'estudi és avaluar la precisió de ChatGPT, Gemini i Copilot en la desambiguació de sigles mèdiques polisèmiques en català, en funció de quatre nivells de context que simulen situacions comunicatives reals en què una persona no experta intenta interpretar una sigla. Per completar aquesta anàlisi, l'estudi també persegueix dos objectius específics: (1) quantificar la consistència estadística de les diferències observades entre sistemes, i (2) descriure la qualitat lingüística i terminològica de les respostes generades, amb especial atenció a l'ús d'altres llengües i als errors terminològics detectats.

Amb aquest estudi, pretenem determinar si algun d'aquests xatbots ofereix un rendiment suficientment fiable i lingüísticament adequat que en justifiqui l'ús com a eina de suport per a professionals de la salut, experts lingüístics o usuaris no especialitzats en contextos reals de lectura mèdica. Aquesta qüestió és especialment rellevant en català, on no es disposa d'un repertori complet i accessible de sigles mèdiques. A diferència del castellà, que compta amb obres com el *Diccionario de siglas médicas*, de la Sociedad Española de Documentación Médica (2023), amb 6.690 sigles registrades, o el *Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias* (Yetano i Alberola, 2003), en català només es disposa d'algunes publicacions menors o molt especialitzades, sense un recurs general de referència. El cercador de termes mèdics més eficient és, sens dubte, el *Diccionari enciclopèdic de medicina* (DEMCAT): Versió de treball (2023) del Centre de Terminologia (TERMCAT), però, com el *Cercaterm* (2023), només inclou algunes de les sigles més habituals dins de la fitxa del terme i no permet fer cerques específiques per sigla.

En aquest context, els xatbots basats en intel·ligència artificial generativa han experimentat una expansió notable a partir del 2022, amb la popularització de sistemes com

ChatGPT, desenvolupat per OpenAI. Aquests models, entrenats amb grans quantitats de textos, tenen la capacitat de generar respostes coherents, contextuals i en múltiples llengües, i s'estan consolidant com a eines d'assistència en múltiples àmbits, inclòs el sanitari. Tot i que el seu ús es difon ràpidament entre pacients, professionals i institucions, el seu comportament davant de fenòmens terminològics complexos encara és poc conegut. Aquest estudi s'inscriu en aquest marc d'exploració, amb l'objectiu de valorar fins a quin punt aquests sistemes poden ser útils per facilitar la comprensió terminològica i a la millora de la comunicació sanitària.

2. Metodologia

L'estudi parteix de la selecció de sigles mèdiques polisèmiques. Els criteris d'inclusió establiren que cada sigla havia de tenir almenys cinc significats diferents dins de l'àmbit sanitari i un mínim de tres significats addicionals en altres àmbits. Per aquest motiu, es van prioritzar les sigles curtes, especialment les de dues lletres i, en menor mesura, les de tres, ja que són les que presenten una polisèmia més elevada.

A banda del potencial polisèmic, es va procurar que les sigles seleccionades cobrissin una àmplia gamma de conceptes mèdics rellevants en la pràctica clínica. Aquestes inclouen noms de malalties i síndromes, procediments diagnòstics, paràmetres i indicadors clínics, signes i símptomes, estructures anatòmiques, microorganismes i altres termes habituals en la documentació sanitària. La selecció combina sigles àmpliament consolidades, com *CAP* (centre d'atenció primària), amb altres de menys freqüents però presents en textos mèdics especialitzats, com *EP* (empiema pleural).

La cerca de sigles es va dur a terme a través de dos recursos principals: el *Diccionario de siglas médicas* (2023), de la Sociedad Española de Documentación Médica i el *Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias*, de Javier Yetano Laguna i Vicent Alberola Cuñat (2003).

En primer lloc, es van seleccionar aleatòriament 100 sigles mèdiques polisèmiques a partir de recursos terminològics en castellà. Com que aquests recursos no estaven en català, es van traduir els termes, i es van excloure aquelles sigles que, a conseqüència de la traducció, modificaven la lletra inicial (per exemple, *enfermedad* → *malaltia*). Paral·lelament, es van incorporar significats addicionals que eren vàlids en català, però no apareixien en castellà. A partir d'aquest procés, es van obtenir 84 sigles candidates, de les quals dos experts en van seleccionar les 60 més representatives, prioritzant les que presentaven una polisèmia més elevada, una major variació semàntica i una presència potencial més rellevant en textos mèdics o d'interès per a pacients.

La validació de les respostes va ser realitzada per una lingüista especialitzada en llenguatge mèdic i un metge clínic amb experiència en documentació sanitària.

La mida de la mostra (60 sigles) es va determinar mitjançant un càlcul de potència estadística. L'anàlisi va indicar que, per detectar una diferència del 15 % entre dues intel·ligències artificials amb una potència del 80 % i un nivell de significació del 5 %, calia avaluar com a mínim 59 sigles.

Amb l'objectiu de simular diferents situacions en què una persona podria intentar desxifrar el significat d'una sigla, es van establir quatre nivells de context per a cadascun dels casos:

1. Sense cap context → *„Què significa XX?"*
2. Context mèdic genèric → *„En el meu informe mèdic posa XX. Què és?"*
3. En una frase real → *„Què significa XX en aquesta frase?: FRASE"*
4. En un paràgraf breu → *„Què significa aquí XX?: TEXT"*

Per evitar que el xatbot pogués inferir des del primer moment que l'àmbit era sanitari, es va obrir un xat diferent per a cada context i per a cada sigla. Això assegurava que cada prova es feia en condicions independents i no contaminades.

Els contextos 3 (frase) i 4 (paràgraf) es van construir a partir de textos reals de l'àmbit sanitari especialitzat, on apareixia el terme en qüestió. Aquests textos es van extreure principalment de fullets informatius per a pacients, pàgines web hospitalàries, articles científics, casos clínics, tesis doctorals. Les frases escollides tenen una mitjana de 22,3 ± 19,33 paraules, i els paràgrafs de 64,24 ± 20,60 paraules.

La comparació entre sistemes es va centrar en els tres xatbots més utilitzats (ONTSI, 2025) i reconeguts en el moment de l'estudi: Gemini 2.0 (Google), ChatGPT 4o (OpenAI) i Copilot GPT-4 (Microsoft)[2]. Cal destacar que els tres sistemes seleccionats funcionen com a xatbots conversacionals, entrenats amb grans volums de dades textuals, però amb diferències pel que fa a l'accés a informació en temps real. En el moment de l'estudi, Copilot i Gemini disposaven de capacitats de cerca integrades, mentre que ChatGPT-4o no accedia a internet per defecte. Aquesta diferència pot influir en la naturalesa i la justificació de les respostes proporcionades.

Per a cada combinació de sistema i context, es va calcular la proporció d'encert, acompanyada del seu corresponent interval de confiança (IC) al 95 % i de l'error estàndard (EE), amb l'objectiu de quantificar la precisió de les estimacions i avaluar la consistència dels resultats.

Atès que les dades recollides són de naturalesa dicotòmica (encert/error) i que el disseny de l'estudi implica mesures repetides sobre les mateixes sigles i sistemes, es van emprar proves no paramètriques específiques per a dades categòriques.

Per determinar si hi havia diferències significatives entre les tres IA (ChatGPT, Gemini i Copilot) dins de cada context, es va aplicar la prova Q de Cochran. En els contextos en què la prova va resultar significativa, es van realitzar comparacions per parells mitjançant la prova de McNemar amb correcció de Bonferroni, per tal de controlar el risc d'error de tipus I associat a comparacions múltiples. Tots els càlculs estadístics es van dur a terme amb R Studio (R Core Team, 2024), i es van considerar estadísticament significatius els valors de $p < 0,05$.

3. Resultats

Els resultats obtinguts sobre el reconeixement de les sigles mèdiques es presenten en la taula 1, on es detallen els encerts per cada sigla en diferents contextos. Tot i les diferències globals entre les IA, s'hi detecta una gran variabilitat segons la sigla i el context. Per exemple, sigles com *CAP* o *OME* són encertades per totes les IA en tots els contextos, mentre que altres, com *AT* (arteritis de Takayasu), no han estat identificades per cap IA.

Taula 1: Encerts detallats per sigla, IA i context

	TERME COMPLET	GEMINI_1	GEMINI_2	GEMINI_3	GEMINI_4	ChatGPT_1	ChatGPT_2	ChatGPT_3	ChatGPT_4	Copilot_1	Copilot_2	Copilot_3	Copilot_4
AA	aneurisma aòrtic	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗	✓	✓
AC	artèria caròtide	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗	✓	✓
AM	aplàsia medul·lar	✗	✓	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
AP	atrèsia pulmonar	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓
AT	arteritis de Takayasu	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
BI	bilirrubina indirecta	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓

	TERME COMPLET	GEMINI_1	GEMINI_2	GEMINI_3	GEMINI_4	ChatGPT_1	ChatGPT_2	ChatGPT_3	ChatGPT_4	Copilot_1	Copilot_2	Copilot_3	Copilot_4
CAP	centre d'atenció primària	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CC	càncer de còlon	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓
CE	circulació extracorpòria	✗	✗	✓	✓	✗	✗	✓	✓	✗	✓	✓	✓
CI	cardiopatia isquèmica	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓
CID	carcinoma intraductal	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
CV	ciclofosfamida i vincristina	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
DA	diverticulitis aguda	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
DC	displàsia cel·lular	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
DM	dermatomiositis	✗	✗	✓	✓	✗	✗	✓	✓	✗	✓	✓	✓
DP	decúbit pron	✗	✗	✓	✓	✗	✗	✓	✓	✗	✓	✗	✓
EA	extrasístole auricular	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
EC	*Escherichia coli*	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
EE	ecoendoscòpia	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓
EM	encefalomielitis	✗	✗	✗	✓	✗	✗	✓	✓	✗	✓	✗	✓
EP	empiema pleural	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
EV	enterovirus	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
HC	hipotiroïdisme congènit	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
HF	hepatitis fulminant	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗
HV	hàl·lux valg / galindó	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
IC	índex cardíac	✗	✗	✓	✓	✓	✗	✓	✓	✗	✗	✓	✓
IP	índex de protrombina	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
IR	insulinoresistència	✗	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
IT	isquèmia transitòria	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
LD	laparoscòpia diagnòstica	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
LL	limfoma limfoblàstic	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
LP	limfadenectomia pèlvica	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓
LT	lepra tuberculoide	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓
MB	meningitis bacteriana	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
MF	micosi fungoide	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
MI	mononucleosi infecciosa	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
MM	mieloma múltiple	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓
MT	membrana timpànica	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
NL	nefropatia lúpica	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗
OC	osteoclasts	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
OME	otitis mitjana exsudativa (o amb efusió).	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
PA	pancreatitis aguda	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
PAM	poliangeïtis microscòpica	✗	✓	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
PAP	pressió d'artèria pulmonar.	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
PAS	pressió arterial sistèmica	✗	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓
PC	pancreatitis crònica	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓
PCA	pancreatitis crònica alcohòlica	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
PE	potencials evocats	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	✓

	TERME COMPLET	GEMINI_1	GEMINI_2	GEMINI_3	GEMINI_4	ChatGPT_1	ChatGPT_2	ChatGPT_3	ChatGPT_4	Copilot_1	Copilot_2	Copilot_3	Copilot_4
PP	polineuropatia progressiva	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
PT	prova de la tuberculina	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
PTI	púrpura trombocitopènica immune	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
SC	síndrome de Cushing	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
SF	síndrome de Fanconi	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
SM	síndrome de Marfan	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓
SP	síndrome paraneoplàsica	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗	✓	✓
SS	síndrome de Sjögren	✗	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓
TC	teixit connectiu	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓
TP	trastorn de la personalitat	✗	✗	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗
TT	traumatisme toràcic	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓
TV	taquicàrdia ventricular	✗	✗	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓
	Encerts	6	11	44	52	5	7	56	58	2	10	40	48
	Encerts %	10%	18%	73%	87%	8%	12%	93%	97%	3%	17%	67%	80%

En els dos primers nivells (sense context i context mèdic), els encerts fan referència al fet que la IA, sense disposar d'un context concret, sol desplegar una llista d'opcions quan se li pregunta pel possible significat d'una sigla (exemple 1). En el context 1, aquestes opcions acostumen a estar agrupades per àmbits, entre els quals sovint –però no sempre– s'inclou l'àmbit mèdic. En el context 2, tot i que també ofereix més d'una possibilitat, les propostes es limiten a l'àmbit sanitari. Si en aquests contextos (1 i 2) una de les opcions coincideix amb el terme que realment es buscava, la resposta es considera un encert. En canvi, en els nivells 3 i 4 (frase i paràgraf), les IA quasi sempre proporcionen una única resposta i, per tant, només es considera encert si aquesta és la correcta.

Exemple 1: Resposta i encert de Gemini sobre „Què és MF?":

En medicina:

- Micosis Fungoide: És el tipus més comú de limfoma cutani de cèl·lules T, un càncer que afecta la pell.
- Medicina Familiar: Especialitat mèdica que proporciona atenció mèdica contínua i integral a individus i famílies.

En física:

- Milifarad: Unitat de capacitat elèctrica equivalent a una mil·lèsima part d'un farad (mF). En textos antics, també es podia utilitzar „mf" o „MFD" per abreujar „microfarad" (µF).

En finances:

- Pot referir-se a una Maestría en Finanzas.
- També pot aparèixer com a símbol de cotització de l'empresa Missfresh Limited a la borsa NASDAQ.
- Mujer Financiera Platinum: Un programa educatiu per a dones per aprendre a invertir.

En altres àmbits:

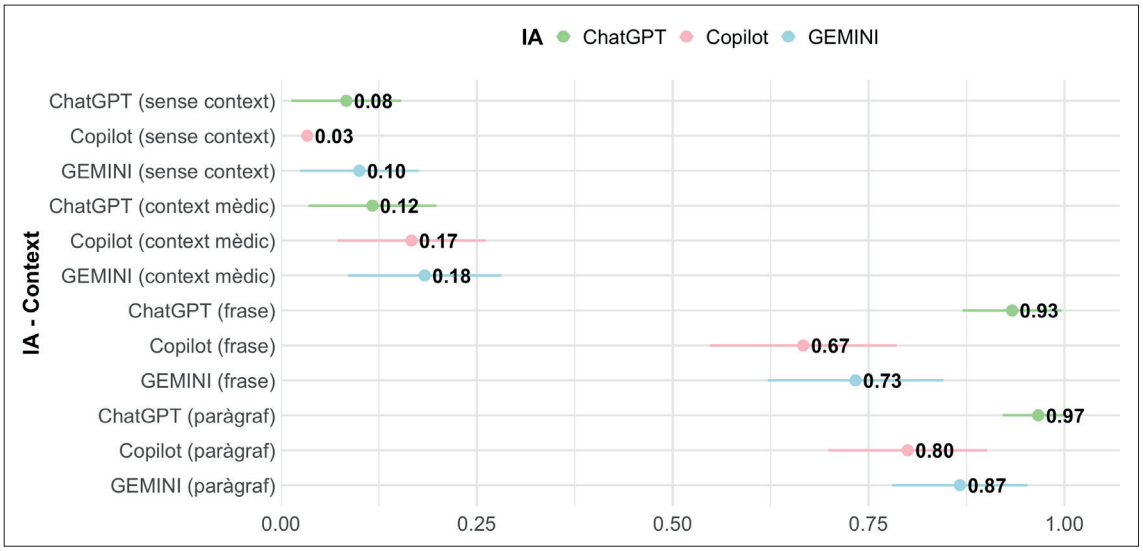
- Maestro FIDE: Un títol d'escacs atorgat per la Federació Internacional d'Escacs (FIDE).
- Mezzoforte: Un terme musical que indica un grau d'intensitat del so moderadament fort.
- .mf: El domini de nivell superior geogràfic (ccTLD) per a Saint-Martin (part francesa del Carib).
- En alguns jocs d'ordinador, com ara *Path of Exile*, MF pot abreujar „Magic Find", una estadística que augmenta la probabilitat de trobar objectes màgics.

Pel que fa a la comparació global, la taula 2 resumeix la proporció d'encert, l'interval de confiança al 95 % i l'error estàndard per a cadascuna de les tres IA en els quatre contextos analitzats, mentre que la figura 1 n'ofereix una representació visual. Els resultats mostren que l'eficàcia de les IA augmenta progressivament a mesura que es disposa de més informació contextual, especialment quan aquesta prové d'una frase o d'un paràgraf real. S'observa també que ChatGPT assoleix els millors resultats en els contextos reals, tant de frase com de paràgraf, mentre que en els dos primers contextos (sigla sola i amb context mèdic breu) les diferències entre sistemes no són gaire rellevants.

Taula 2: Els valors de proporció d'encerts, interval de confiança al 95 % i error estàndard per a cadascuna de les tres IA en els quatre contextos

IA	Context	Proporció d'encert	IC 95%	Error estàndard
Gemini	1. Sense context	0,1	[0,047 – 0,201]	0,0387
	2. Context mèdic	0,183	[0,106 – 0,299]	0,05
	3. Frase	0,733	[0,610 – 0,829]	0,0571
	4. Paràgraf	0,867	[0,758 – 0,931]	0,0439
ChatGPT	1. Sense context	0,083	[0,036 – 0,181]	0,0357
	2. Context mèdic	0,117	[0,058 – 0,222]	0,0414
	3. Frase	0,933	[0,870 – 0,996]	0,0322
	4. Paràgraf	0,967	[0,921 – 1,000]	0,0232
Copilot	1. Sense context	0,033	[0,000 – 0,079]	0,0232
	2. Context mèdic	0,167	[0,072 – 0,261]	0,0481
	3. Frase	0,667	[0,547 – 0,788]	0,0609
	4. Paràgraf	0,8	[0,699 – 0,901]	0,0516

Figura 1: Proporció d'encerts amb interval de confiança al 95% per IA i context



ChatGPT presenta el millor rendiment global, amb una proporció d'encerts del 93,3 % en el context de frase i del 96,7 % en el de paràgraf. A més, els seus errors estàndard (0,0322 i 0,0232) indiquen una precisió excel·lent i una gran consistència en les respostes. Gemini, per la seva part, mostra també un rendiment destacat en contextos més amplis, amb un 86,7 % d'encerts en el context de paràgraf i un error estàndard de 0,0439, que indica una bona precisió. Copilot presenta una evolució positiva amb l'ampliació del context, però amb un comportament més irregular. Aconsegueix un 80 % d'encerts en el context de paràgraf, però només un 66,7 % en el context de frase. En aquests dos casos, els errors estàndard (0,0516 i 0,0609) reflecteixen una precisió moderada.

En els contextos més restrictius (sense context i context mèdic), totes les IA obtenen resultats molt baixos, amb taxes d'encert inferiors al 20 %. Els errors estàndard, baixos o moderats, indiquen una variabilitat interna limitada i evidencien un predomini d'errors sistemàtics.

Per tal d'analitzar si hi havia diferències significatives entre les tres IA en la seva capacitat de desxifrar sigles mèdiques, es va aplicar la prova de Q de Cochran per a cada context. Els resultats (taula 3) mostren que no s'observen diferències significatives entre les IA en els contextos més restrictius (sense context i context mèdic). En canvi, tant en el context de frase com en el de paràgraf, les diferències sí que són estadísticament significatives ($p < 0,01$), la qual cosa indica que el rendiment relatiu de les IA varia quan es disposa de més informació contextual.

Taula 3: Resultats de la prova Q de Cochran per context

Context	Q estadístic	p-valor	Interpretació
1 (sense context)	3,25	0,197	No hi ha diferències significatives
2 (context mèdic)	2	0,368	No hi ha diferències significatives
3 (frase)	18,9	0,000078	Hi ha diferències significatives ($p < 0,01$)
4 (paràgraf)	9,5	0,00865	Hi ha diferències significatives ($p < 0,01$)

Per identificar quines parelles d'IA presentaven diferències significatives en aquests contextos, es va utilitzar la prova de McNemar amb correcció de Bonferroni (taules 4 i 5). En el context de frase, ChatGPT va mostrar un rendiment significativament superior al de GEMINI ($p < 0,01$) i al de Copilot ($p < 0,001$), mentre que no es van observar diferències entre GEMINI i Copilot. En el context de paràgraf, només es va detectar una diferència significativa entre ChatGPT i Copilot ($p < 0,05$). Això suggereix que les diferències més marcades es produeixen en contextos intermedis com la frase, mentre que en el context de paràgraf, tot i que ChatGPT manté el millor rendiment, la distància amb les altres IA es redueix.

Taula 4: Resultats de la prova de McNemar (Context 3 - Frase) amb correcció de Bonferroni

Comparació	*Khi-quadrat*	*p-valor*	*p-valor ajustat*	*Interpretació*
ChatGPT vs. GEMINI	10,1	0,0015	0,0045	Significatiu ($p < 0,01$)
ChatGPT vs. Copilot	14,1	0,00018	0,00054	Molt significatiu ($p < 0,001$)
GEMINI vs. Copilot	0,562	0,453	1,000	No significatiu ($p \geq 0,05$)

Taula 5: Resultats de la prova de McNemar (Context 4 - Paràgraf) amb correcció de Bonferroni

Comparació	*Khi-quadrat*	*p-valor*	*p-valor ajustat*	*Interpretació*
ChatGPT vs. GEMINI	3,12	0,0771	0,2313	No significatiu ($p \geq 0,05$)
ChatGPT vs. Copilot	6,75	0,0094	0,0281	Significatiu ($p < 0,05$)
GEMINI vs. Copilot	0,75	0,386	1,00	No significatiu ($p \geq 0,05$)

En conjunt, els resultats mostren que ChatGPT-4o és el sistema més efectiu en la desambiguació de sigles mèdiques en català, especialment en contextos reals (frase i paràgraf), on obté els índexs d'encert i consistència més alts. Gemini 2.0 mostra un rendiment intermedi, amb una millora notable quan disposa de més informació contextual, mentre que Copilot presenta un comportament més irregular, amb resultats menys consistents i més sensibles al tipus de context.

4. Discussió i conclusions

Aquest estudi ha permès analitzar de manera sistemàtica la capacitat de tres sistemes d'intel·ligència artificial (Gemini, ChatGPT i Copilot) per interpretar sigles mèdiques polisèmiques en quatre nivells de context, amb l'objectiu de reproduir situacions reals de lectura de textos mèdics. Els resultats obtinguts ofereixen una primera aproximació al comportament d'aquestes eines davant les ambigüitats terminològiques, habituals en informes mèdics, webs d'informació sanitària i altres documents de l'àmbit de la salut.

Diversos estudis recents han explorat la capacitat de models d'IA per desambiguar sigles mèdiques en anglès. Kuġić, Schulz i Kreuzthaler (2023, 2024) han demostrat que ChatGPT millora significativament el seu rendiment quan disposa de més context, i que pot competir amb altres sistemes, fins i tot en casos amb sigles poc freqüents. Altres treballs, com el de Wagh i Khanna (2023), han obtingut resultats notables amb variants especialitzades de BERT entrenades amb dades biomèdiques. Tot i això, aquests estudis se centren exclusivament en l'anglès i no comparen diferents xatbots generals accessibles al públic. Aquest estudi és el primer que aborda la qüestió en català i contribueix a omplir el buit existent amb un enfocament sistemàtic centrat en la qualitat de les respostes i el rendiment dels xatbots en contextos d'ambigüitat terminològica.

Els resultats mostren clarament que el rendiment de totes tres IA augmenta quan el context disponible és més ampli, resultats que són coherents amb l'estudi de Liu et al. (2024). Les diferències entre sistemes esdevenen especialment rellevants a partir del moment en què es proporciona un context mínimament elaborat, com ara una frase (context 3). En aquest nivell, ChatGPT obté un rendiment significativament superior respecte a la resta, fet que en reforça el potencial com a eina de suport a la comprensió terminològica en situacions comunicatives reals. Aquest avantatge es manté, i fins i tot s'accentua, en el context més extens (context 4, paràgraf), tot i que estadísticament només s'ha pogut confirmar una diferència significativa amb Copilot. Aquesta capacitat pot ser especialment útil per a professionals sanitaris i del llenguatge (com traductors, correctors o terminòlegs), així com per a pacients i familiars que consulten informes mèdics o busquen informació en línia.

Quant als desencerts, alguns no s'allunyen gaire del terme correcte, però d'altres constitueixen exemples clars d'al·lucinació per part de la IA. Per exemple, davant la frase "El CID es diagnostica per biòpsia i es confirma amb l'estudi postoperatori del tumor extirpat", Copilot interpreta *CID* com a "Classificació Internacional de Malalties", quan en realitat es tracta de "carcinoma intraductal". Aquest tipus d'error pot tenir conseqüències rellevants si les respostes del xatbot es prenen com a font fiable d'informació en l'àmbit sanitari. Aquesta manca de fiabilitat terminològica de l'àmbit mèdic ja ha estat assenyalada en estudis sobre l'ús de la IA en tasques lexicogràfiques, com el de de Schryver (2023), que mostra com els xatbots pot generar definicions i respostes aparentment convincents però sovint imprecises o inventades, sense cap sistema de verificació integrada.

Pel que fa a la qualitat lingüística de les respostes, Copilot tendeix a respondre en anglès amb bastant freqüència, fins i tot quan la pregunta s'ha formulat en català (*Thoracic Trauma*). Aquest fet pot suposar un obstacle per a pacients que no dominen aquesta llengua i que cerquen respostes clares i accessibles. A més, en algunes ocasions respon amb el terme en castellà o en català, però amb errors (com **micosis*, *síndrome *paraneoplàsic*, **polyangiitis microscòpica*). Per la seva banda, Gemini tendeix a generar respostes en català, també amb alguns errors (**micosis*), seguit de respostes en castellà, i només ocasionalment fa ús de l'anglès. ChatGPT és la IA que manté amb més consistència el català com a llengua de resposta, amb menys errors que la resta (**pàncreatitis crònica*). Tot i que en alguns casos recupera desplegaments de sigles en castellà o anglès, és la que comet menys errors terminològics en català i la que utilitza un llenguatge més natural.

Una altra aportació rellevant d'aquest treball és l'anàlisi de l'estabilitat i la precisió de les respostes. El càlcul de l'error estàndard ha permès valorar la consistència dels resultats i la fiabilitat de la proporció d'encerts en cada context. En aquest sentit, ChatGPT no sols és la IA amb millor rendiment, sinó també la que ofereix més precisió estadística, especialment en els contextos 3 (frase) i 4 (paràgraf).

Cal assenyalar algunes limitacions potencials de l'estudi. Tot i que la grandària de la mostra (60 sigles) és suficient per detectar diferències del 15 % entre IA, podria no serho per identificar efectes més subtils o per extreure conclusions generalitzables a totes les sigles mèdiques. A més, l'estudi es basa en interaccions d'un únic torn per sigla i context, mentre que en una situació real l'usuari podria interactuar de manera més flexible, amb preguntes successives o peticions d'aclariment. Així mateix, cal tenir en compte que les respostes de les IA poden variar amb el temps, en funció de les actualitzacions dels models o dels canvis en el seu entrenament.

Finalment, convé destacar que totes tres IA mostren una tendència clara a recuperar significats de sigles formades també en castellà o anglès. Aquest comportament pot resultar positiu, ja que en textos en català és habitual que la sigla faci referència a un terme en una altra llengua, especialment en castellà o anglès (les anomenades sigles al·locentrades, tal com s'ha descrit en Kotátková, 2023).

Aquest estudi aporta una primera aproximació empírica al comportament de tres sistemes d'IA davant la polisèmia de les sigles mèdiques en català, amb un enfocament sistemàtic. Els resultats mostren que el context és un factor clau per a la desambiguació, i que ChatGPT s'hi adapta millor que la resta (Kuġić, Kreuzthaler i Schulz, 2023), tant en termes de rendiment com de precisió i adequació comunicativa. De fet, la seva proporció d'encerts del 96,7 % en contextos de paràgraf destaca no sols per ser la més alta, sinó també per la seva estabilitat i fiabilitat estadística, fet que reforça el seu potencial com a eina de suport terminològic en situacions reals. Les diferències observades entre les IA, especialment a partir del nivell de frase, poden orientar professionals i institucions en la tria d'eines per millorar la comprensió de textos mèdics, en un entorn en què la claredat i l'accessibilitat de la informació poden tenir un impacte directe en l'atenció i la seguretat del pacient. Cal afegir que aquestes diferències inclouen també una certa inestabilitat pel que fa a l'ús del català, amb respostes que no sempre respecten la normativa ni els termes estandarditzats. Tot i les limitacions del disseny, l'estudi obre la porta a línies futures de recerca que aprofundeixin en l'impacte dels xatbots en la mediació lingüística i terminològica en contextos sanitaris.

Els resultats d'aquest estudi poden ser útils per orientar professionals de la salut, comunicadors i institucions en la selecció d'eines d'intel·ligència artificial aplicables a l'àmbit mèdic en català. En particular, ChatGPT-4o s'ha mostrat com el sistema més eficaç en la desambiguació de sigles mèdiques polisèmiques en contextos reals, amb una proporció d'encerts molt elevada i una qualitat lingüística superior, especialment pel que fa a l'ús correcte del català i la naturalitat de les respostes. Aquesta fiabilitat el fa especialment indicat per a serveis d'atenció al pacient, consultes terminològiques o suports lingüístics per a professionals. Així mateix, els xatbots poden actuar com a eines complementàries per fer més accessibles informes mèdics i altres documents sanitaris, especialment si s'integren en plataformes digitals de salut. Perquè aquest ús sigui realment útil i segur, cal fomentar l'alfabetització digital en salut i garantir que les respostes generades per la IA siguin contrastades i interpretades críticament, amb el suport de professionals humans.

5. Notes

1. La cerca no es va poder fer directament en català, ja que actualment no es disposa d'un repositori o diccionari complet de sigles mèdiques en aquesta llengua que permeti consultar els diferents significats associats a cada sigla.

2. Totes les proves amb els xatbots es van dur a terme entre març i abril de 2025.

6. Bibliografia

De Schryver, Gilles-Maurice. 2023. "Generative AI and lexicography: The current state of the art using ChatGPT". *International Journal of Lexicography*, 36(4), 355–387. https://doi.org/10.1093/ijl/ecad021.

Navarro, Fernando A. 2008. "Repertorio de siglas, acrónimos, abreviaturas y símbolos utilizados en los textos médicos en español". *Panace@. Revista de Medicina, Lenguaje y Traducción*, IX(27), 55–59. https://www.tremedica.org/wp-content/uploads/n27_Panacea27_junio2008.pdf

Koťátková, Adela. 2023. "Les sigles mèdiques en català: autocentrades o al·locentrades?". *Terminàlia*, 28, 14–26. https://doi.org/10.2436/20.2503.01.194.

Kuǧić, Amila, Markus Kreuzthaler, Stefan Schulz. 2023. "Clinical acronym disambiguation via ChatGPT and Bing". *Studies in Health Technology and Informatics*, 309, 78–82. https://doi.org/10.3233/SHTI230743.

Kuǧić, Amila, Stefan Schulz, Markus Kreuzthaler. 2024. "Disambiguation of acronyms in clinical narratives with large language models". *Journal of the American Medical Informatics Association*, 31(9), 2040–2046. https://doi.org/10.1093/jamia/ocae157.

Navarro, Fernando A. 2023. *Siglas médicas en español: Repertorio de siglas, acrónimos, abreviaturas y símbolos utilizados en los textos médicos en español* (2a ed.). Cosnautas. https://www.cosnautas.com/es/catalogo/diccionario-siglas-medicas.

Liu, Ying, Genevieve B. Melton, Rui Zhang. 2024. "Exploring large language models for acronym, symbol sense disambiguation, and semantic similarity and relatedness assessment". *AMIA Joint Summits on Translational Science Proceedings*, 2024, 324–333. https://pmc.ncbi.nlm.nih.gov/articles/PMC11141821

ONTSI. 2025. *Indicadores de uso de inteligencia artificial en España*. Observatori Nacional de Tecnologies i Societat de la Informació.

R Core Team. 2024. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/.

Sociedad Española de Documentación Médica. 2023. *Diccionario de siglas médicas*. http://diccionario.sedom.es/.

TERMCAT. 2023. *Cercaterm*. Centre de Terminologia. https://www.termcat.cat/cercaterm.

TERMCAT. 2023. *Diccionari enciclopèdic de medicina (DEMCAT): Versió de treball*. Centre de Terminologia. https://www.termcat.cat/ca/diccionaris-en-linia/192.

Wagh, Atharwa, Manju Khanna. 2023. "Clinical abbreviation disambiguation using clinical variants of BERT". En: Morusupalli, R.; Dandibhotla, T. S.; Atluri, V. V.; Windridge, D.; Lingras, P.; Komati, V. R. (Eds.), *Multi-disciplinary trends in artificial intelligence: MIWAI 2023*. Lecture Notes in Computer Science, 14078, 214–224. Springer. https://doi.org/10.1007/978-3-031-36402-0_19.

Yetano Laguna, Javier, Vicent Alberola Cuñat. 2003. *Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias*. Ministerio de Sanidad y Consumo.