Com llegeixen els LLMs?

El paper dels tokens en els models de llenguatge

Sant Albert 2026

Facultats de Ciències i de Biociències

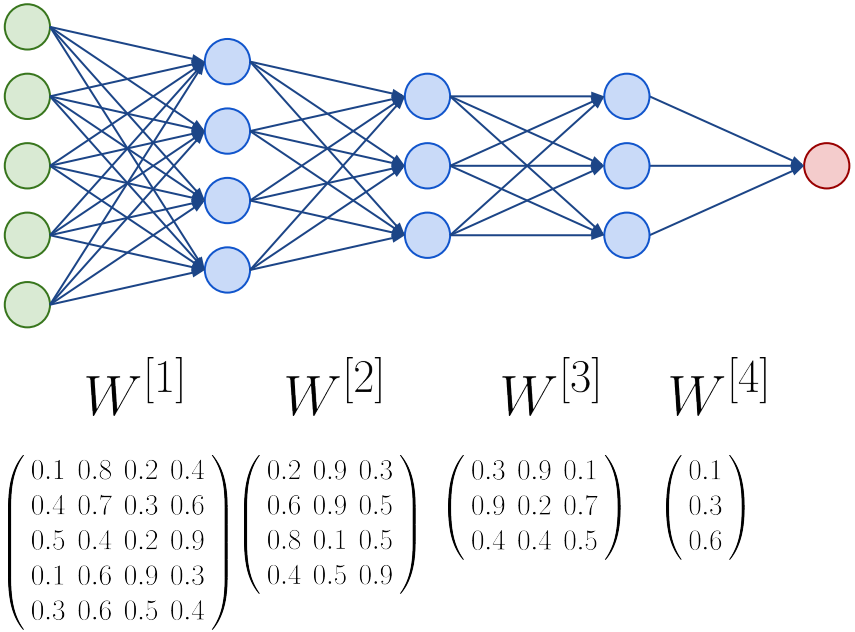
UAB
Universitat Autònoma de Barcelona

Per què necessitem la tokenització?

Els **models de llenguatge** estan basats en **xarxes neuronals.**

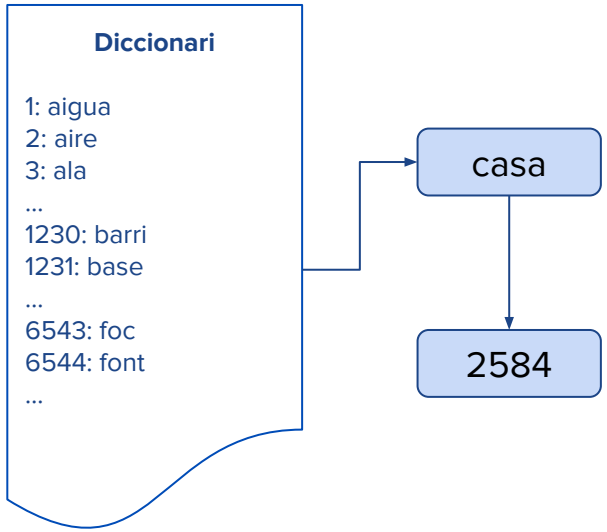
Les xarxes neuronals són composicions de funcions que **transformen vectors numèrics.**

Necessitem llavors **codificar el text** amb valors numèrics.



Diccionaris de paraules

Una primera idea és **crear un diccionari** i **assignar a cada paraula l'índex corresponent** on apareix al diccionari.



Problemes:

- No totes les paraules possibles apareixen al diccionari.
- Difícil de gestionar el format del text, com signes de puntuació.
- Desaprofitem la similaritat conceptual entre paraules semblants.

Diccionari de tokens

Fixem un alfabet $\Sigma = \{a, b, c, ...\}$ i denotem per Σ^* el conjunt de seqüències de caràcters que podem formar amb els caràcters de Σ.

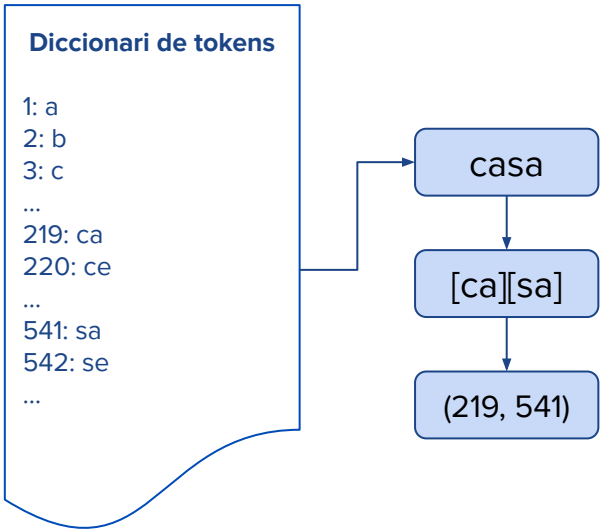
$$\Sigma^* = \{casa, cotxe, Toni, ...\}$$

Un **diccionari de tokens** és un subconjunt $\mathcal{D} \subseteq \Sigma^*$ tal que, **per a tota seqüència** de caràcters $w \in \Sigma^*$, **existeixen tokens** $(t_1, t_2, \dots, t_n)$ amb $t_i \in \mathcal{D}$ satisfent

$$w = t_1 t_2 \cdots t_n$$

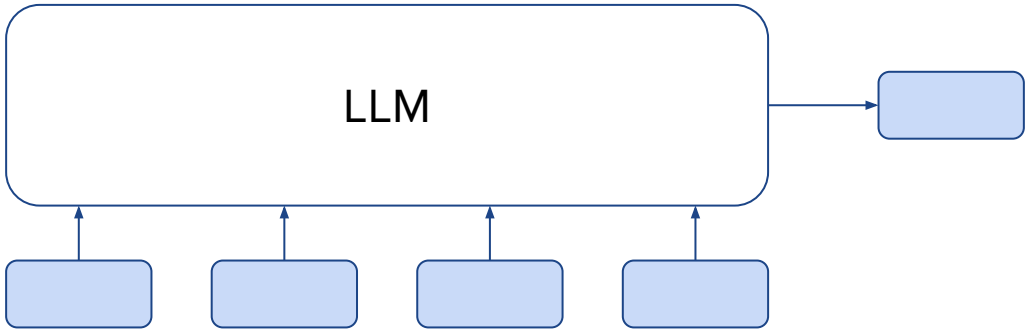
Un **tokenitzador** és una funció $T : \Sigma^* \to \mathcal{D}^*$ tal que

$$\text{concat}(T(w)) = w$$



Funcionament d'un LLM

La base d'un LLM és un model que prediu el següent token més probable donada una seqüencia de tokens.

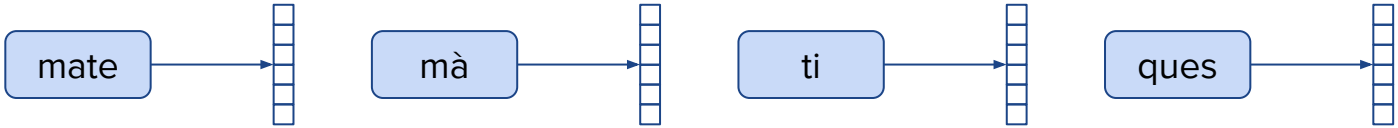


Per **entrenar el model** que prediu el següent token se segueixen **tres passos:**

- Pre-entrenament no supervisat
- Ajust supervisat
- Aprenentatge amb retroacció humana / recompenses verificables

Representació vectorial del text

Matemàticament, **el que busquem és trobar una bona representació vectorial per a cada token**, de forma que el model de llenguatge pugui **predir les probabilitats de forma acurada.**

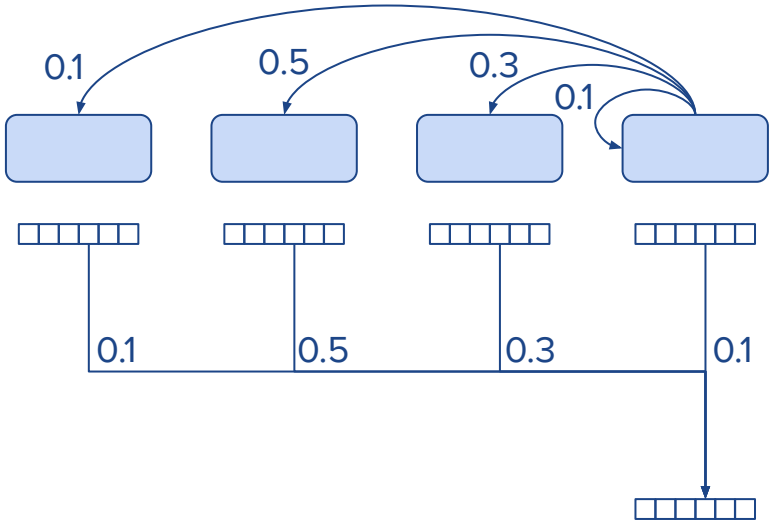


Els models de llenguatge actuals utilitzen **xarxes neuronals profundes per transformar l'identificador de cada token** en una representació vectorial.

El punt més important és que **la representació vectorial de cada token depèn de la frase** que estem processant en cada moment. D'aquesta forma, la representació d'un mateix token pot canviar segons el context.

Mecanisme d'atenció

El pas més important per generar una **representació vectorial que depengui del context** és el **mecanisme d'atenció.**



Per **calcular el percentatge d'atenció** a cada paraula de la seqüència s'utilitzen diverses **matrius de paràmetres.**

$$v_m = \beta_v + \Omega_v x_m$$

$$q_n = \beta_q + \Omega_q x_n$$

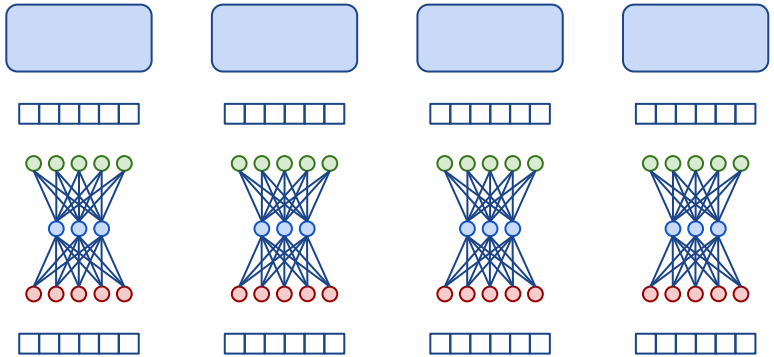
$$k_l = \beta_k + \Omega_k x_l$$

$$w(x_n, x_i) = \text{softmax}_i(k_\bullet^T q_n)$$

$$\tilde{x}_n = \sum_{i=1}^{N} w(x_n, x_i) v_i$$

Transformació independent

Per **afegir flexibilitat al model,** també s'apliquen **transformacions de forma independent** a les representacions vectorials de cada paraula.



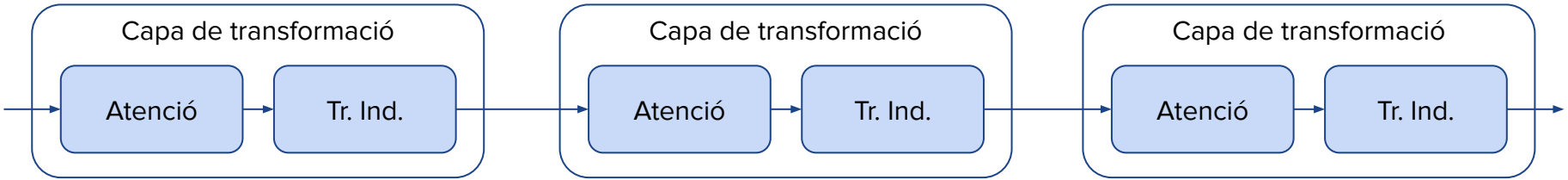
Aquestes transformacions també contenen **matrius de paràmetres.**

$$FFN(x) = \max(0, xW^{[1]} + b^{[1]})W^{[2]} + b^{[2]}$$

Arquitectura de transformació

Si ajuntem el **mecanisme d'atenció** amb la **transformació independent** obtenim una **capa de l'arquitectura de transformació.**

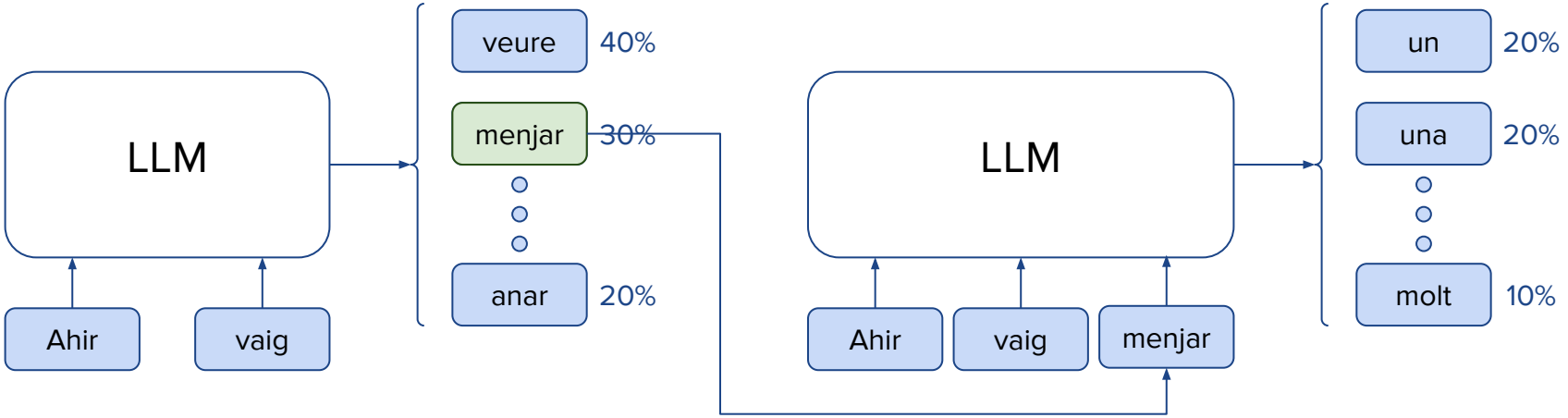
Concatenant diverses capes de transformació podem generar representacions vectorials de cada token que **capturen molt bé la informació semàntica del context.**



Els models actuals més potents tenen **desenes de capes i centenars de mil·lers de milions de paràmetres.**

Generació de text

Quan el **model de llenguatge està entrenat** podem generar **nou text seguint un procés iteratiu que inclou una part aleatòria.**



El **procés continua** fins que el model **genera un token special** que representa el final de la frase.

Observacions

- Un **LLM és un model autoregressiu,** on cada element de la seqüència a processar és un token.
- Segons la **mida del diccionari de tokens,** les **seqüències** a processar poden ser **més llargues o més curtes.**
- Un **token** es representa inicialment per un únic **índex numèric.** Llavors, si el mateix token forma part de diferents paraules, el model ha de **modificar la representació del token segons el seu context.**
- Si **una paraula és un sol token,** el model parteix d'**una sola representació.** Si en són **diversos tokens,** ha de **combinar-los per donar sentit a la paraula** completa.

Mida del diccionari i longitud de la seqüència

Diccionari de tokens petit

- Valors dels índexos són números petits.
- Cada frase s'ha de codificar amb una seqüència llarga de tokens.

Exemple

(2, 31, 9, 17, 2, 8, 25, 5, 14, 9, 17, 22)

Diccionari de tokens gran

- Valors dels índexos són números grans.
- Cada frase s'ha de codificar amb una seqüència curta de tokens.

Exemple

(1485, 57412, 33678, 1457, 22346, 8953)

Algoritme Byte Pair Encoding

Suposem que tenim un **corpus de documents** $\mathcal{C} = \{w_1, w_2, \dots, w_n\}$ i **considerem l'alfabet** $\sum$ utilitzat en el nostre corpus. Fixem un **número màxim de tokens** N.

Llavors, **definim** $\mathcal{D}_0 = \sum$ i **procedim de forma iterativa.** En la iteració k cada text del corpus té una representació en el diccionari de tokens $\mathcal{D}_k$.

Aleshores, definim la **freqüència d'una parella ordenada de tokens** de $\mathcal{D}_k$ com

$f_k(u, v)$ = número d'aparicions de la seqüència $[u][v]$ en el corpus

on $[u], [v] \in \mathcal{D}_k$.

Considerem els **tokens** $[u_k]$ **i** $[v_k]$ **que maximitzen la funció** $f_k(u, v)$ i definim un **nou token** com

$$[z_k] = [u_k][v_k]$$

Algoritme Byte Pair Encoding

Amb el nou token **actualitzem el diccionari**

$$\mathcal{D}_{k+1} = [z_k] \cup \mathcal{D}_k$$

I **actualitzem també les representacions** de tots els textos del corpus utilitzant el nou diccionari de tokens.

Repetim aquesta operació fins arribar a tenir el **número màxim de tokens** fixat.

Al **final de T iteracions** tenim que el diccionari de tokens és

$$\mathcal{D}_{T+1} = \Sigma \cup \{z_1, z_2, \dots, z_T\}$$

I tenim una sèrie de **regles de fusió:**

$$[z_1] = [u_1][v_1], [z_2] = [u_2][v_2], \dots [z_T] = [u_T][v_T]$$

Algoritme Byte Pair Encoding

Diccionari de tokens

```
a b c e è g i l n o r s ü
```

Tokenització del text

```
ciència       [c][i][è][n][c][i][a]
llengües      [l][l][e][n][g][ü][e][s]
ciències      [c][i][è][n][c][i][e][s]
llibres       [l][l][i][b][r][e][s]
boca          [b][o][c][a]
```

Freqüència de parelles

Parella	Freqüencia
[c] + [i]	4
[e] + [s]	3
[l] + [l]	2
[i] + [è]	2
...	...

Regles de fusió

```
[c] + [i] = [ci]
```

Algoritme Byte Pair Encoding

Diccionari de tokens

```
a b c e è g i l n o r s ü ci
```

Tokenització del text

ciència	[ci][è][n][ci][a]
llengües	[l][l][e][n][g][ü][e][s]
ciències	[ci][è][n][ci][e][s]
llibres	[l][l][i][b][r][e][s]
boca	[b][o][c][a]

Freqüència de parelles

Parella	Freqüencia
[e] + [s]	3
[l] + [l]	2
[ci] + [è]	2
[è] + [n]	2
...	...

Regles de fusió

[c] + [i] = [ci], [e] + [s] = [es]

Algoritme Byte Pair Encoding

Diccionari de tokens

```
a b c e è g i l n o r s ü ci es
```

Tokenització del text

```
ciència       [ci][è][n][ci][a]
llengües      [l][l][e][n][g][ü][es]
ciències      [ci][è][n][ci][es]
llibres       [l][l][i][b][r][es]
boca          [b][o][c][a]
```

Freqüència de parelles

Parella	Freqüencia
[l] + [l]	2
[ci] + [è]	2
[n] + [ci]	2
[è] + [n]	2
...	...

Regles de fusió

`[c] + [i] = [ci], [e] + [s] = [es],`

`[l] + [l] = [ll]`

Algoritme Byte Pair Encoding

Diccionari de tokens

```
a b c e è g i l n o r s ü ci es ll
```

Tokenització del text

```
ciència       [ci][è][n][ci][a]
llengües      [ll][e][n][g][ü][es]
ciències      [ci][è][n][ci][es]
llibres       [ll][i][b][r][es]
boca          [b][o][c][a]
```

Freqüència de parelles

Parella	Freqüencia
[ci] + [è]	2
[n] + [ci]	2
[è] + [n]	2
[ci] + [a]	1
...	...

Regles de fusió

[c] + [i] = [ci], [e] + [s] = [es],

[l] + [l] = [ll], [ci] + [è] = [ciè]

Algoritme Byte Pair Encoding

Diccionari de tokens

```
a b c e è g i l n o r s ü ci es ll ciè
```

← Límit de tokens establert

Tokenització del text

```
ciència       [ciè][n][ci][a]
llengües      [ll][e][n][g][ü][es]
ciències      [ciè][n][ci][es]
llibres       [ll][i][b][r][es]
boca          [b][o][c][a]
```

← Tokenització definitiva

Tokenització de noves paraules

Diccionari de tokens

a b c e è g i l n o r s ü ci es ll ciè

Regles de fusió

[c] + [i] = [ci], [e] + [s] = [es],

[l] + [l] = [ll], [ci] + [è] = [ciè]

biociències

↓

[b][i][o][c][i][è][n][c][i][e][s]

↓

[b][i][o][ci][è][n][ci][e][s]

↓

[b][i][o][ci][è][n][ci][es]

↓

[b][i][o][ciè][n][ci][es]

Tokenització de noves paraules

Diccionari de tokens

Token	Id	Token	Id
[a]	0	[o]	9
[b]	1	[r]	10
[c]	2	[s]	11
[e]	3	[ü]	12
[è]	4	[ci]	13
[g]	5	[es]	14
[i]	6	[ll]	15
[l]	7	[ciè]	16
[n]	8		

biociències



[b][i][o][ciè][n][ci][es]

(1, 6, 9, 16, 8, 13, 14)

Observacions

L'algoritme BPE és un algoritme que construeix un diccionari de tokens **basat únicament en les dades** disponibles. Això comporta algunes **implicacions:**

- El **volum i la distribució de dades** de cada idioma **té un gran impacte** en el diccionari de tokens que es construeix.
- El diccionari de tokens **no inclou coneixement expert** sobre cap matèria en particular, només reflecteix les regularitats del corpus.

Tokenització segons l'idioma

Explain why increasing the temperature can accelerate a chemical reaction

```
[Explain][ why][ increasing][ the][ temperature][ can]
[ accelerate][ a][ chemical][ reaction]
```

73 caràcters
10 tokens

+20%

Explica por qué aumentar la temperatura puede acelerar una reacción química

```
[Exp][lica][ por][ qué][ aumentar][ la][ temperatura]
[ puede][ acelerar][ una][ reacción][ química]
```

75 caràcters
12 tokens

+60%

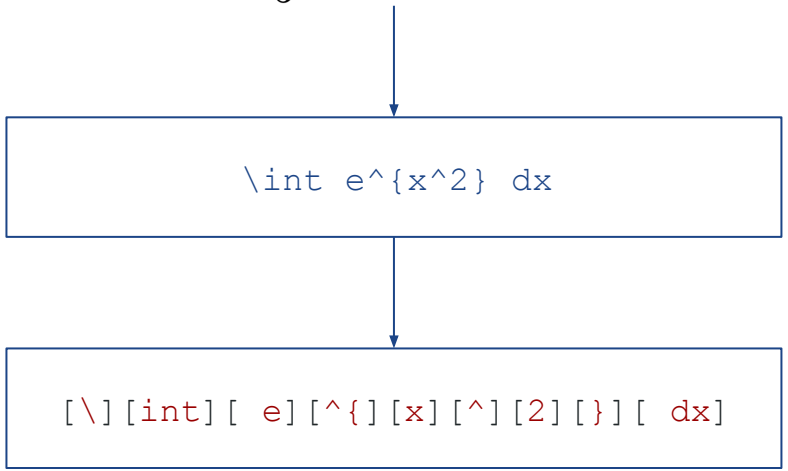
Explica per què augmentar la temperatura pot accelerar una reacció química

```
[Exp][lica][ per][ què][ augment][ar][ la][ temperatura]
[ pot][ acceler][ar][ una][ re][acc][ió][ química]
```

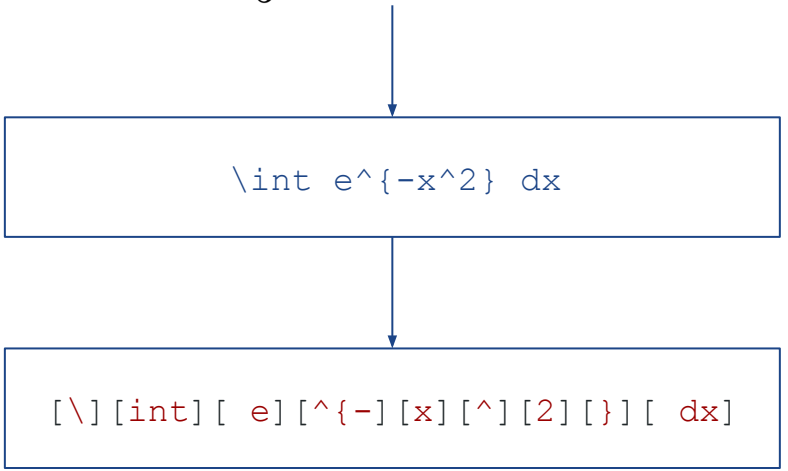
74 caràcters
16 tokens

Interpretació dels tokens

$$\int e^{x^2} dx$$



$$\int e^{-x^2} dx$$



Conclusions

- Com en moltes àrees de la ciència, el que fem depèn de **com representem l'objecte d'estudi.** En els LLMs, la unitat bàsica no és la paraula, sinó el token.
- La **representació** del text en tokens **no sempre està alineada** amb el criteri que podrien tenir humans experts.
- **Entendre els tokens ens dóna perspectiva:** les limitacions dels LLMs no són només "no saben", sinó sovint "ho veuen fragmentat" o "els hi costa més processar-ho".

Com llegeixen els LLMs?

El paper dels tokens en els models de llenguatge

Sant Albert 2026

Facultats de Ciències i de Biociències

UAB
Universitat Autònoma de Barcelona