# Avaluació de l'Aprenentatge en Context en Models de Llenguatge per a Tasques de Manipulació de Text en Català

Nils Duran

# 1 Introducció i Motivació

Els models grans de llenguatge (LLM) han demostrat una capacitat peculiar per a resoldre tasques mitjançant l'ús de l'aprenentatge en context (ICL), descrit formalment per Brown et al. [1]. A diferència de l'ajust fi (*fine-tuning*), l'ICL permet que un model aprengui a executar una tasca a partir de poques instruccions o exemples inclosos directament en el prompt. No obstant això, sorgeixen dubtes raonables sobre si aquesta capacitat es deu a un aprenentatge algorítmic en temps d'inferència o bé a la memorització de dades contingudes en el seu gegantí corpus de preentrenament.

Per a estudiar aquest fenomen, és essencial dissenyar tasques que siguin clarament de fora de distribució (OOD), és a dir, tasques absurdes, arbitràries o amb regles combinades que difícilment s'hauran observat en la fase de preentrenament. Avaluar aquestes capacitats en llengües de recursos mitjans com el català [2] afegeix interès, ja que permet estudiar el comportament dels models en un context lingüístic menys representat en les dades de preentrenament en comparació amb l'anglès.

Aquest treball té com a objectiu analitzar de quina manera la mida del model, la dificultat algorítmica de la tasca i la tècnica de prompting afecten la capacitat d'executar manipulacions de cadenes de caràcters de complexitat incremental en català.

# 2 Descripció Experimental

L'experiment s'ha dissenyat a partir d'un conjunt de 100 frases sintàcticament diverses en català. Per a evitar problemes de codificació i que caràcters especials poguessin distreure els models o introduir ambigüitats en la correspondència de caràcters de l'abecedari, he modificat les frases d'entrada de tal manera que no tinguin ni accents, ni dièresis, ni la ç ni la l·l. Sense errors ortogràfics, simplement escollint paraules que de per si ja no contenen aquests caràcters.

## 2.1 Tasques d'Avaluació

S'han programat de forma algorítmica quatre transformacions de text en català amb nivells de dificultat creixent (hipòtesi de dificultat: Aritmètica < Idioma < Atbash < Combinada):

1. **Tasca 1: Idioma de la 'i' (Baixa)**: Consisteix a substituir totes les vocals (*aeou*) per la lletra *i* (o *I* si està en majúscula), respectant la resta de consonants, puntuació i espais.

2. **Tasca 2: Atbash (Mitjana)**: Substitueix cada lletra per la seva recíproca o complement a l'abecedari ($a \leftrightarrow z$, $b \leftrightarrow y$, etc.). Els números i símbols queden intactes.

3. **Tasca 3: Aritmètica (Mitjana)**: Detecta els números dins de la frase i els multiplica per 2 i al resultat li resta 3. Si no hi ha números, la frase es manté idèntica.

4. **Tasca 4: Combinada (Alta)**: Aplica seqüencialment les tres regles anteriors en l'ordre següent: primer l'aritmètica, després el complement Atbash, i finalment l'idioma de la 'i'.

## 2.2 Paràmetres de Prompting i Avaluació

Cada tasca s'ha avaluat sota tres modalitats de prompting:

- **Zero-Shot (0-shot)**: Només s'indiquen les instruccions detallades de la tasca en català (com l'apartat 2.1).
- **Few-Shot (5-shot)**: S'inclouen les instruccions acompanyades de 5 exemples aleatoris d'inputs i outputs correctes de la tasca.
- **Examples-Only**: No s'inclou cap instrucció textual; el prompt només conté els 5 exemples d'entrenament en context amb el missatge d'identificar el patró i aplicar-lo a la nova frase.

Vaig voler avaluar sis models de pesos oberts (Gemini 3.1 Flash-Lite m'he adonat que no ho és, però l'he inclòs igualment) mitjançant les APIs de Google AI Studio i Groq: Llama 3.1 8B Instant [3], Llama 3.3 70B Versatile [3], Gemma 4 31B IT [4], Gemini 3.1 Flash Lite Preview, GPT-OSS 20B [5] i GPT-OSS 120B [5]. L'avaluació calcula dues mètriques: l'exactitud exacta (*Exact Match*, EM), on la resposta ha de coincidir al 100% caràcter per caràcter amb la referència, i la similitud per paraules (*Word Match Percentage*, WM) mitjançant *SequenceMatcher* per a tolerar talls menors.

# 3 Resultats i Anàlisi

A la Taula 1 es recullen els resultats d'Exact Match (EM) i Word Match (WM) obtinguts globalment.

**Taula 1:** Rànquing global de models i modalitats de prompting (mètriques mitjanes en %)

| **Model** | **Zero-Shot** | | **Few-Shot** | | **Examples-Only** | |
|---|---|---|---|---|---|---|
| | **EM** | **WM** | **EM** | **WM** | **EM** | **WM** |
| `llama-3.1-8b` | 1.25 | 27.01 | 8.75 | 40.77 | 6.25 | 37.40 |
| `gpt-oss-20b`* | 37.50 | 47.31 | 39.75 | 48.55 | **65.71** | **79.83** |
| `gemma-4-31b` | **74.75** | **96.79** | 72.00 | **96.38** | 58.50 | 92.07 |
| `gemini-3.1-lite` | 38.75 | 70.15 | 42.50 | 78.87 | 24.50 | 72.46 |
| `llama-3.3-70b` | 23.50 | 45.02 | 26.50 | 55.96 | 9.00 | 48.14 |
| `gpt-oss-120b`* | 72.25 | 83.99 | **76.75** | 83.84 | 72.38 | 91.19 |

*Nota: Aquests models no tenen tots els resultats per límits de l'API.

## 3.1 Comportament dels Models i Efecte Escala

Es confirmen de forma molt clara les lleis d'escala. El model més petit, `llama-3.1-8b-instant`, fracassa rotundament en l'avaluació zero-shot, obtenint un Exact Match marginal de l'1.25%. Per contra, els models amb major nombre de paràmetres com `gemma-4-31b-it` i `openai/gpt-oss-120b` aconsegueixen un rendiment molt elevat.

També hi ha una gran diferència entre les diferents famílies de models, i els models més antics són molt pitjors als més recents. Però dins de les mateixes "famílies", el model de major escala sempre és clarament superior al de menor escala, com es pot veure en els casos de Llama 3.1 vs Llama 3.3 i GPT-OSS 20B vs GPT-OSS 120B.

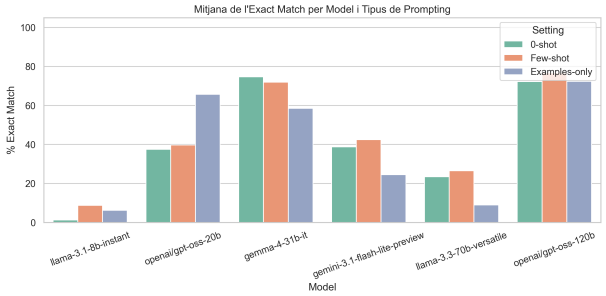


**Figura 1:** Mitjana global de l'Exact Match per Model i tipus de Prompting.

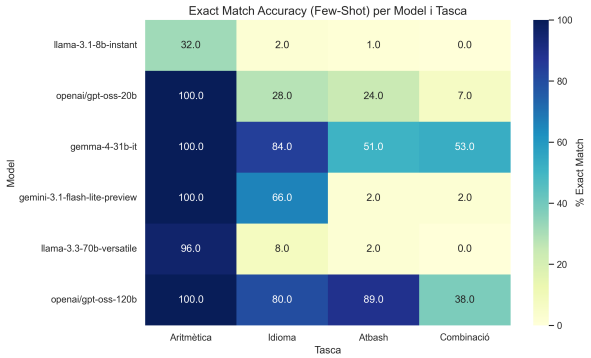


**Figura 2:** Exact Match (Few-Shot) per Model i Tasca.

Un resultat particularment curiós s'observa en el model mitjà `openai/gpt-oss-20b`. Aquest model obté només un 37.50% d'EM en zero-shot i un 39.75% en few-shot, pero quan se li retira la instrucció textual i només rep exemples directes (*Examples-Only*), el seu Exact Match puja fins a un sorprenent **65.71%** i el seu Word Match fins al **79.83%**.

## 3.2 Dificultat de les Tasques i Avaluació Numèrica

Les tasques segueixen un ordre diferent a la hipòtesi inicial. La tasca més senzilla amb diferència és l'aritmètica, seguida de l'idioma de la 'i', després l'atbash (intercanvi de caràcters) i finalment la combinació.

**Taula 2:** Rànquing de dificultat de les tasques (mitjana global de models en règim Few-Shot)

| Tasca | Exact Match (EM) % | Word Match (WM) % |
|---|---|---|
| Idioma | 44.67 | 76.89 |
| Atbash | 28.17 | 56.75 |
| Aritmètica | **88.00** | **98.50** |
| Combinació | 16.67 | 37.42 |

A la Tasca Aritmètica, que és la més senzilla (en part perquè les frases que no tenen nombres es mantenen idèntiques i sumen encerts directes), és interessant analitzar la capacitat dels models per a extreure i calcular operacions numèriques lineals ($f(n) = 2n - 3$).

Aquí la versió d'exemples sols pateix més, segurament perquè hi ha frases que no tenen números, i amb només una o dues transformacions, hi pot haver més ambigüitat en la identificació de l'equació a aplicar.

S'observa que els models grans (`gemma-4-31b-it`, `openai/gpt-oss-120b` i `openai/gpt-oss-20b`) assoleixen un perfecte **100%** de precisió numèrica tant en Zero-Shot com en Few-Shot,

demostrant que la seva capacitat d'extracció i càlcul és impecable. En canvi, el model més petit `llama-3.1-8b-instant` té problemes de càlcul matemàtic severs, assolint només un 5.88% d'encerts en Zero-Shot, que millora fins al 21.18% amb Few-Shot.

**Taula 3:** Exactitud en l'extracció i càlcul numèric correcte per a la Tasca Aritmètica (%)

| Model | Zero-Shot | Few-Shot | Examples-Only |
|---|---|---|---|
| `llama-3.1-8b-instant` | 5.88 | 21.18 | 17.65 |
| `gpt-oss-20b`* | **100** | **100** | 81.18 |
| `gemma-4-31b-it` | **100** | **100** | **91.76** |
| `gemini-3.1-lite` | 98.82 | **100** | 24.71 |
| `llama-3.3-70b-versatile` | 85.88 | 96.47 | 22.35 |
| `gpt-oss-120b`* | **100** | **100** | 77.65 |

*Nota: Aquests models no tenen tots els resultats per límits de l'API.

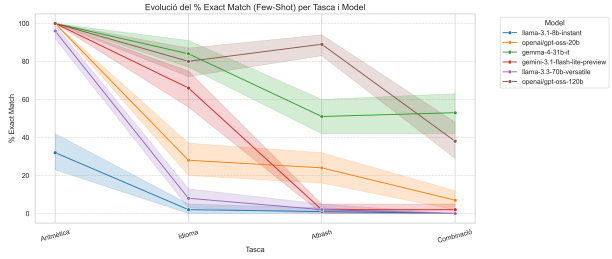


**Figura 3:** Evolució de l'Exact Match (Few-Shot) per a cada Tasca i Model.

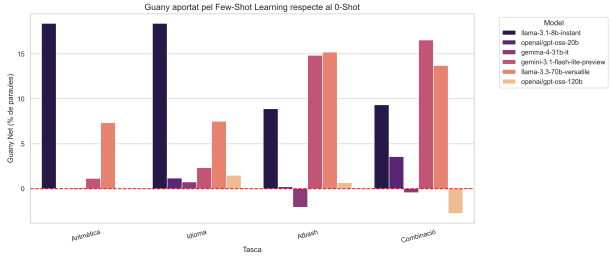


**Figura 4:** Guany net de Word Match gràcies al Few-Shot respecte al Zero-Shot.

## 3.3 Anàlisi de la Similitud per Paraules (Word Match)

Mentre que l'Exact Match (EM) avalua la coincidència exacta de la cadena, la similitud per paraules Word Match (WM) actua com una mètrica contínua o suau. Aquesta mètrica permet valorar si el model s'apropa al comportament desitjad o si comet errades menors de format, espais o caràcters.

Com es reflecteix a la Figura 5, models de major escala com `gemma-4-31b-it` i `openai/gpt-oss-120b` tenen distribucions de Word Match extremadament tancades prop del 100%, demostrant que, fins i tot quan fallen en EM, el seu text generat és gairebé perfecte. En canvi, el model `llama-3.1-8b-instant` quan falla acostuma a fallar en més d'una paraula.

La tasca de Combinació és el major repte, on només els models d'escala superior aconsegueixen mantenir-se per sobre del 80% de similitud, mentre que els models petits baixen a menys del 30%.

És possible que tasques com l'aritmètica o fins i tot alguna altra estiguessin lleugerament representades en el corpus de preentrenament, però la tasca de combinació, que requereix aplicar tres regles seqüencials, és clarament OOD i mostra que només els models més grans tenen la capacitat d'aprendre a combinar regles de forma algorítmica en context.

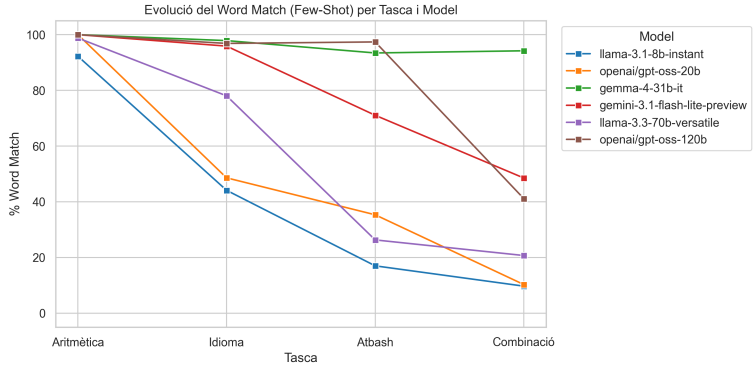


**Figura 5:** Evolució del % de Word Match (Few-Shot) per a cada tasca i model.

Hi ha una caiguda molt regular i previsible en el rendiment de les tasques; no hi ha models que facin bé la tasca de combinació però malament l'aritmètica, per exemple.

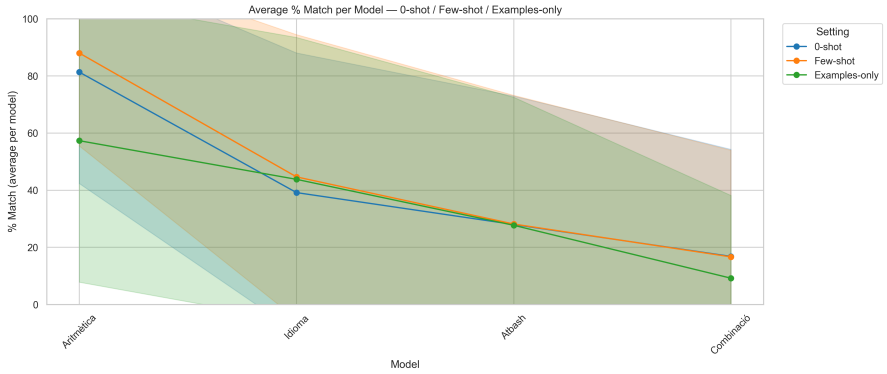


**Figura 6:** Evolució del % de Exact Match (Few-Shot) per a cada model.

El few-shot és el millor tipus de prompting en general, però els exemples sols no es queden tan enrere com m'imaginava, i en alguns casos com l'idioma de la i fins i tot supera el zero-shot.

# 4 Conclusions i Aprenentatges

Aquest estudi proporciona informació molt valuosa sobre l'aprenentatge en context dels LLMs:

- **La capacitat d'ICL depèn de la mida**: Els models de mida petita (<10B) no tenen prou capacitat per a realitzar tasques algorítmiques complexes.
- **Instruccions vs Exemples**: El prompting basat purament en exemples pot ser superior al zero-shot i few-shot clàssics segons l'alineació.
- **Errors de tokenització**: Les tasques de caràcters són complicades per als LLMs a causa de la fragmentació de paraules, motiu pel qual el Few-Shot és crucial.

# Referències

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1877–1901.

[2] J. Armengol-Estapé, C. P. Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, and M. Villegas, "Are multilingual models the best choice for catalan? an evaluation of pre-trained language models and benchmarks," in *Proceedings of the EACL Workshop on Catalan Language Technologies*, 2021, pp. 1–10.

[3] Meta Llama Team, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[4] Gemma Team, Google, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[5] OpenAI, "Introducing gpt-oss: Open-weight mixture-of-experts models," https://openai.com/index/gpt-oss, 2025.