

Sparse Autoencoders para la recuperación interpretable de memoria en LLMs

Rompiendo la caja negra: un prisma semántico

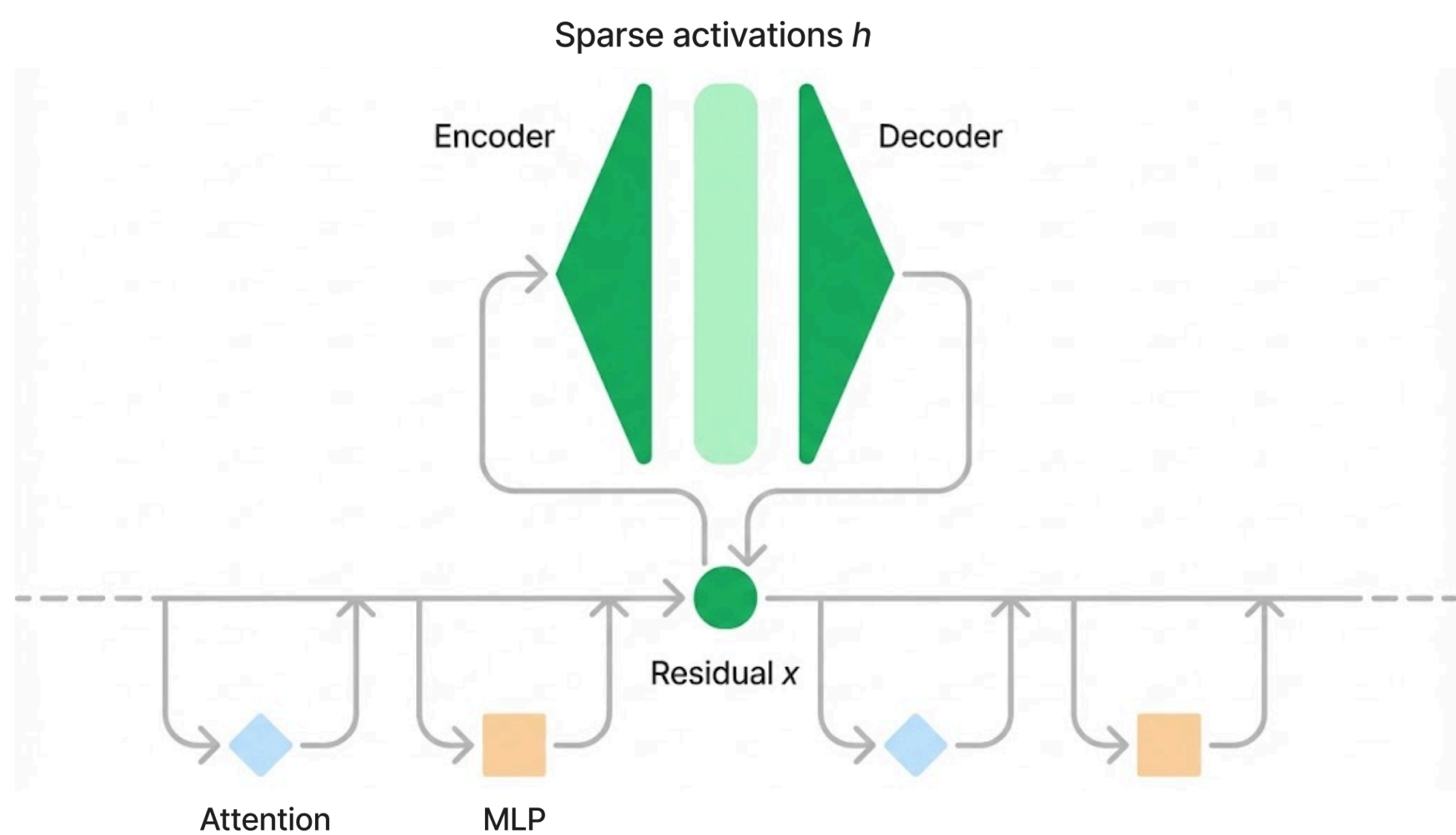
Waehner Nicolás^[1], Szereszewski Julián^[1], Smith Martina^[2]

[1] UBA – FCEN – Departamento de Física, CABA, Argentina

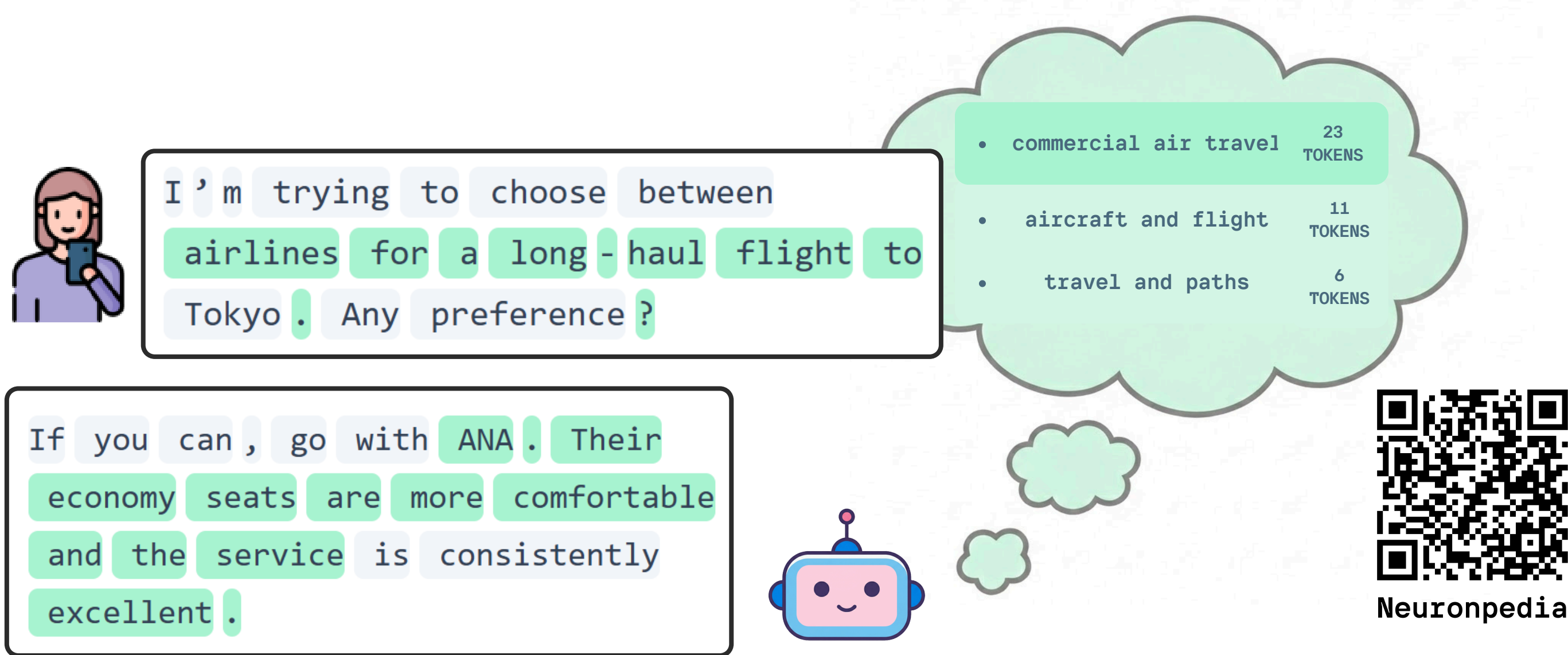
[2] UBA – FCEN – Departamento de Ciencias de la Atmósfera y los Océanos, CABA, Argentina

Introducción

Un **Sparse Autoencoder** (SAE) es un tipo de autoencoder diseñado para aprender una **representación latente dispersa** (sparse).



- En este espacio latente, se pueden asociar conceptos a activaciones [1]:
- Podemos hacer la memoria interpretable y no una caja negra.
 - Encuentra similitudes semánticas más que similitudes sintácticas

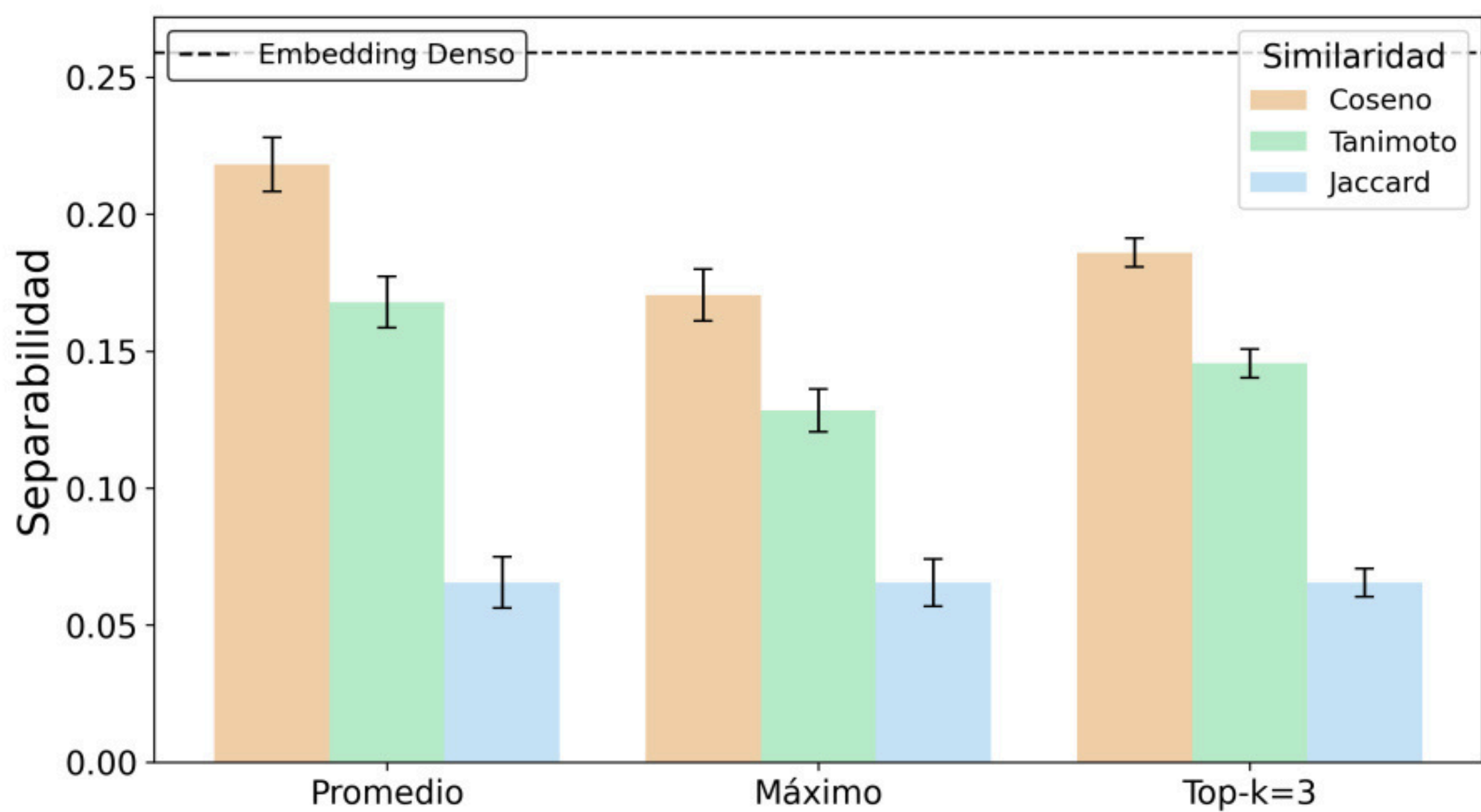
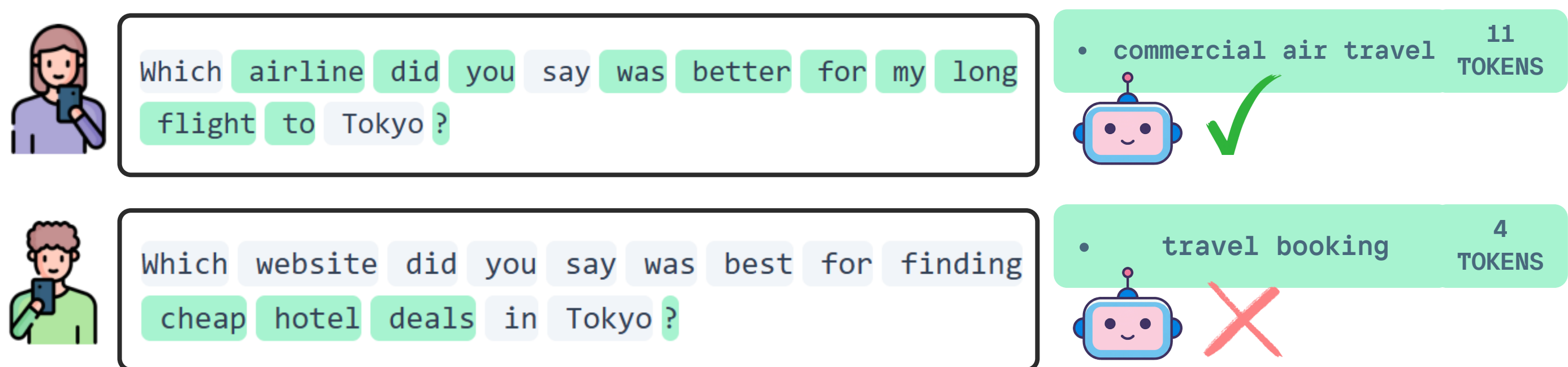


Criterio de similaridad

Utilizamos Gemma-2b para el embedding disperso y Nomic para el embedding denso.

Evaluamos en 100 casos distintos la separabilidad (SEP) a partir de una similitud (SIM) entre un mismo contexto (C) y dos preguntas: una relacionada (R) y otra no relacionada (NR).

$$SEP(C, R, NR) = SIM(C, R) - SIM(C, NR)$$



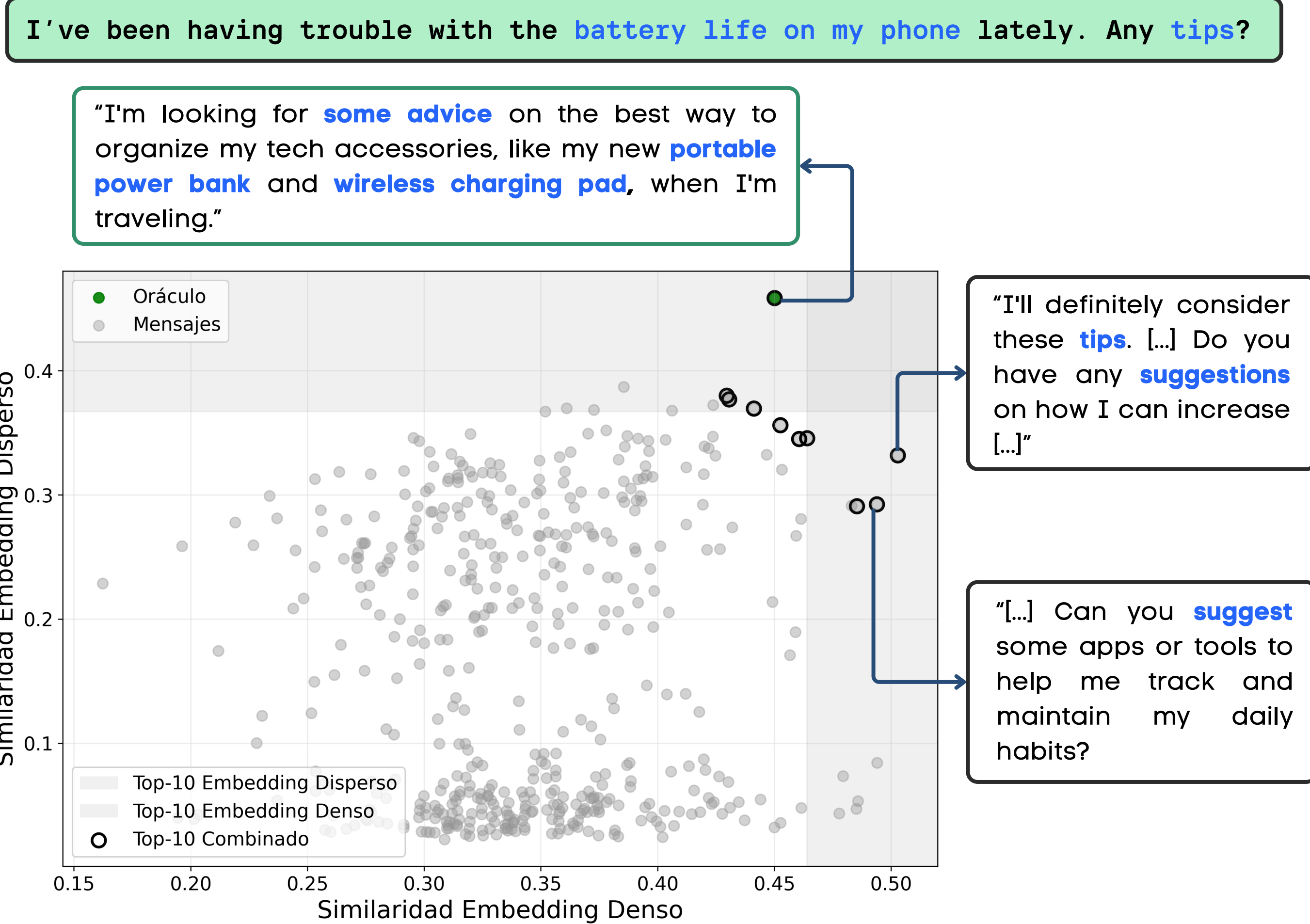
Referencias

- [1] Bricken et al. (2023). Sparse Autoencoders Find Interpretable Features in Language Models.
[2] Wu et al. (2025). LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory.

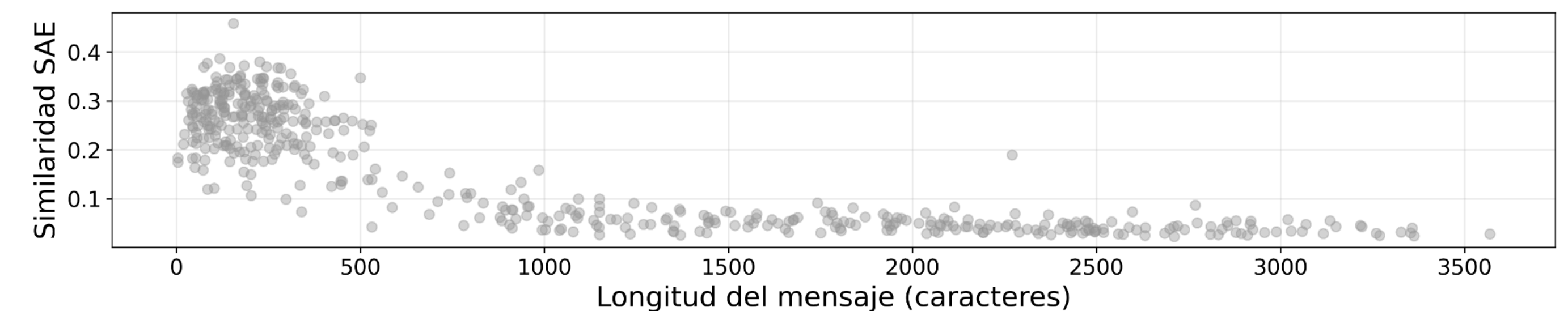
Criterio de recuperación: definiendo el RAG

Evaluamos la distribución de similitudes para cada conversación y definimos una métrica que utilice los dos embeddings:

$$SIM_{combinado} = \sqrt{(SIM_{denso})^2 + (SIM_{disperso})^2}$$



Nuestra implementación tiene algunas limitaciones:



Resultados

Evaluamos el desempeño sobre LongMemEval [2] con ambos embeddings y su combinación:

Métricas	Disperso	Denso	Combinado
Precisión en el contexto*	32%	68%	73%
Precisión en la respuesta	24%	41%	45%
Latencia (s)**	47 ± 18	8 ± 4	55 ± 19
Longitud del contexto (caracteres)	2194 ± 1056	10548 ± 3598	5346 ± 2652

* Calculado según la proporción de casos en los que el criterio de recuperación obtuvo el conjunto completo de mensajes relevantes para responder la pregunta.

Conclusiones

- El embedding generado por SAE se puede utilizar como mecanismo de recuperación de memoria interpretable.
- SAE permite una recuperación basada en activaciones conceptuales, a diferencia de embeddings tradicionales.
- Optimizamos la recuperación combinando un embedding denso con uno disperso.



GitHub