

Un modelo que no olvida

Una bitácora que corrige y un sueño que consolida, sobre un modelo de lenguaje congelado

Maximiliano Speranza

Investigador independiente, Buenos Aires, Argentina

[ORCID 0009-0005-0413-8554](https://orcid.org/0009-0005-0413-8554)

maximiliano.speranza@gmail.com

Octubre de 2026 (versión 2)

Resumen

Un asistente que se usa todos los días necesita recordar lo que el usuario le dijo, aunque esos datos cambien con el tiempo, y necesita decir «no sé» ante lo que nunca le contaron. Estudiamos dos canales de memoria sobre un modelo de lenguaje congelado, Qwen3-0.6B. La *bitácora* guarda cada hecho con su versión fuera de los pesos y se lee con un lector aprendido. El *sueño* consolida periódicamente los hechos recientes en un adaptador LoRA de 10 millones de parámetros, con una receta que repasa los hechos ya consolidados, repasa dos habilidades inventadas junto con sus bordes de abstención, practica el «no sé» y se ancla al modelo con el que empieza cada sueño.

En ocho sueños encadenados de 100 hechos cada uno, con tres semillas, lo aprendido la primera semana se sostiene entre 0,88 y 0,97 de exactitud después de siete sueños más (media 0,94), y ningún paquete cae más de 0,09 por debajo del valor con que se aprendió. Una ablación aísla el ingrediente. Sin el repaso de hechos, con todo lo demás igual, la exactitud sobre la primera semana pasa de 0,97 a 0,01 en un solo sueño. Comparamos con MEMIT y AlphaEdit, dos métodos publicados de edición de conocimiento que escriben cada hecho nuevo directamente en los pesos del modelo, sin entrenarlo. Editando los mismos hechos semana a semana, a la octava semana conservan 0,01 y 0,10 de lo aprendido la primera, y los dos dañan habilidades que la edición nunca tocó. Frente a un RAG entrenado con el mismo presupuesto la bitácora empata en exactitud, y la diferencia está en el costo, 16 tokens por consulta contra 505, o contra 4.514 con un historial de 200 entradas viejas.

El límite que medimos es la composición espuria. Ante una entidad conocida y una relación que nunca se le enseñó para esa entidad, a la octava semana el modelo contesta un valor en el 51 a 61 % de los casos en lugar de abstenerse, y el «no sé» ante entidades inventadas baja de 0,97 a 0,64. Parejar cuántos hechos recibe cada entidad, la intervención más obvia, no cambió nada (40 respuestas de 200 en las dos condiciones). El límite se corta afuera de los pesos. Un portero que busca en la bitácora la entidad de la pregunta por identidad y deja que el lector de la bitácora decida la relación, con la entidad reemplazada por una que conoce, baja la composición a entre 0,000 y 0,007 y deja pasar entre 0,983 y 1,000 de los hechos dichos, con tres semillas del lector, el umbral fijado en la otra cadena y nada entrenado de nuevo.

Con hechos del mundo real la memoria se sostiene. En una cadena de sueños sobre CounterFact, el banco público con que se evalúa la edición de conocimiento, cada semana se corrigen 100 hechos que el modelo ya creía saber. A la cuarta semana lo corregido la primera sigue ganando por probabilidad en 0,99 a 1,00 de los casos, en tres semillas, y sin el repaso la exactitud sobre esos hechos cae a 0,17 a 0,35. Pero la perplejidad sobre texto escrito por personas, que la medida de la versión 1 (frases del propio modelo) no veía, sube desde el primer sueño y crece, hasta multiplicarse por 55 en inglés a la octava semana. Encontramos por intervención tres causas y una receta final, un sueño anclado al origen que enseña sólo la respuesta, ata todo lo demás al modelo original y suma texto humano, reformulaciones de la propia versión 0, otros sujetos anclados y un piso para lo nuevo. A la octava semana, en tres semillas, lo corregido la primera semana sigue ganando en 1,00 de los casos, las paráfrasis de

lo corregido en 0,95 a 0,965, y el daño medido con perplejidad calibrada (cada modelo a su temperatura óptima, porque a temperatura 1 el modelo base está sobreconfiado y el daño se esconde) sobre 150 artículos distintos de Wikipedia queda entre +1,1 y +1,5 % en inglés y entre -0,48 y -0,95 % en español. Ese resto crece despacio con las semanas y está en los nombres propios. Un portero del adaptador, que usa lo consolidado sólo cuando el texto nombra un sujeto aprendido y si no deja hablar al modelo original, lo lleva a entre +0,26 y +0,29 % en inglés sin tocar la memoria. Con una lista de 18.466 sujetos el portero por párrafo se diluye, y limitarlo a los ocho tokens que siguen a la mención lo devuelve a entre +0,08 y +0,19 %. Una ablación preregistrada mostró que el ancla de 2.000 hechos del mundo de una receta anterior sobraba, y la receta final es la misma sin ella. Cuesta especificidad, porque los vecinos bajan de 0,70 a 0,60 a 0,62.

En una charla simulada de dos semanas con tres usuarios, el modelo que durmió con la receta final recuerda al otro día 0,918 de lo que se le dijo, preguntado con frases que nunca vio, contra 0,062 del mismo modelo con la charla entera en el contexto, y su texto humano a temperatura 1 no empeora. Cuando el usuario se corrige gana siempre el valor nuevo. Donde nunca se le dijo nada, en cambio, contesta con el valor de otro dato que sí se dijo en 33 de 60 preguntas, y verificar contra la bitácora el valor que proponen los pesos lleva la abstención de 0,083 a 0,450.

Novedades y correcciones de esta versión (la versión 2 en Research Square). La versión 1 midió el idioma con la perplejidad sobre frases generadas por el propio modelo y concluyó que el sueño no lo empeora. Sobre texto humano el daño llegaba a +1.270 % en inglés a la cuarta semana (§4.9). Esta versión corrige esa afirmación y agrega (1) una cadena de sueños sobre hechos reales de CounterFact, en tres semillas (§4.8), (2) las tres causas del daño, halladas por intervención, y las recetas que lo sacan, medidas con perplejidad calibrada (§4.9 y §4.10), (3) ocho semanas en tres semillas, con una predicción escrita antes que falló para la receta con ancla del mundo, y una tercera trampa de medición, un registro de texto humano sacado de un solo artículo, reemplazado por 150 artículos distintos (§4.11), (4) el portero del adaptador y su límite de escala (§4.12), (5) una ablación preregistrada que sacó el ancla de hechos del mundo y dio la receta final (§4.13), (6) el mismo sujeto con otra relación (§4.14), (7) una charla simulada de dos semanas (§4.15) y (8) lo que falló en el camino, con todas las variantes probadas (§5 y Apéndice A).

1. Introducción

Casi todo el trabajo sobre memoria en modelos de lenguaje apunta al conocimiento del mundo, ya sea enseñarle al modelo un hecho nuevo (edición de conocimiento) o dejarle leer documentos externos (recuperación, RAG). Un asistente que se usa a diario necesita además otra cosa, que es recordar la interacción misma. Qué le dijo el usuario, cuándo, y qué había dicho antes si eso cambió. Esta memoria tiene una propiedad que el conocimiento del mundo tiene poco, y es que se corrige seguido. «Mi perro se llama Toby» puede ser la semana siguiente «tengo dos perros, Toby y Luna», y el sistema tiene que saber cuál versión vale hoy sin confundirlas ni olvidar que la primera existió.

Las herramientas disponibles cubren partes de este problema. La edición de conocimiento escribe un hecho directamente en los pesos del modelo, sin reentrenarlo. Los dos métodos de referencia son MEMIT (Meng et al., 2023), que localiza las capas intermedias donde el modelo guarda asociaciones entre un sujeto y sus atributos y las modifica con una actualización calculada, sin entrenamiento, para muchos hechos a la vez, y AlphaEdit (Fang et al., 2025), una variante pensada para editar hechos en secuencia que limita cada cambio a direcciones que no alteran las respuestas sobre lo que se quiere conservar. Estos métodos se evalúan habitualmente con dos bancos públicos de hechos del mundo real. CounterFact (Meng et al., 2022) reúne unos 22.000 hechos en forma de frase para completar («la lengua materna de Danielle Darrieux es»), y zsRE (Levy et al., 2017; De Cao et al., 2021) reúne preguntas escritas por personas («¿de quién es la

Adoración de la Trinidad?»). La edición de conocimiento sobrescribe el hecho sin guardar qué decía antes y, como mostramos, acumula daño cuando las ediciones se encadenan. El RAG lee el historial como texto, pero su costo crece con el largo del historial en cada consulta. El ajuste fino directo aprende lo nuevo pisando lo viejo. Este trabajo pregunta si un modelo congelado puede recordar cientos de hechos versionados de una interacción, corregirlos sin reentrenar, decir «no sé» ante lo que nunca escuchó y sostener todo eso sesión tras sesión con un costo de lectura que no dependa del largo del historial.

Proponemos separar el problema en dos canales. La *bitácora* resuelve la corrección. Es memoria fuera de los pesos, rápida de escribir y de leer, que guarda cada hecho con su versión. El *sueño* resuelve la permanencia. Es una consolidación periódica que pasa lo reciente a los pesos de un adaptador, para que lo ya aprendido no dependa para siempre de que la bitácora lo siga leyendo. La pregunta central es qué hace falta en cada sueño para que el modelo no olvide lo consolidado mientras sigue aprendiendo hechos nuevos.

Los aportes son los siguientes.

1. Una receta de consolidación que, en ocho sueños encadenados y tres semillas, sostiene lo aprendido la primera semana entre 0,88 y 0,97, sin que ningún paquete caiga más de 0,09 por debajo de su propio valor inicial (§4.1).
2. Una ablación que identifica el ingrediente. Es el repaso explícito de los hechos consolidados, y no el ancla al modelo anterior ni la bitácora, lo que evita el olvido (§4.2).
3. Una comparación con dos editores de conocimiento publicados, MEMIT y AlphaEdit, sobre los mismos hechos y el mismo punto de partida, con el mecanismo por el cual AlphaEdit daña capacidades que la edición nunca tocó verificado por intervención (§4.3).
4. Una medición de la abstención separada por clase de pregunta, que muestra que el «no sé» se sostiene cuando todo es desconocido y se erosiona cuando la pregunta combina partes conocidas (§4.4).
5. Una cadena de sueños con hechos reales, sobre CounterFact y con sus métricas estándar, que reproduce la retención, la ablación y el portero del mundo sintético en tres semillas y ocho semanas (§4.8).
6. Una corrección de la versión 1 y su solución. Medida sobre texto humano, la receta de la versión 1 daña el idioma desde el primer sueño y el daño crece (+246 % en el primero, +1.270 % a la cuarta semana). Identificamos por intervención tres causas, una réplica escrita por un modelo que ya se corrió, una réplica en un solo idioma y un objetivo que enseña la frase entera en lugar de la respuesta, y una receta final que a la octava semana, en tres semillas, deja la memoria intacta, la paráfrasis de lo corregido en 0,950 a 0,965 y el daño calibrado sobre 150 artículos de Wikipedia entre +1,1 y +1,5 % en inglés y por debajo de la versión 0 en español. La pérdida central, la entropía cruzada sólo sobre la respuesta y el resto atado a la versión 0, la usa KLOD (Jeong et al., 2026) para ediciones de una sola vez, y no la reclamamos como propia (§4.9, §4.10, §4.11).
7. Un portero del adaptador que deja el texto que no nombra lo aprendido en manos del modelo original y lleva el daño al inglés a entre +0,26 y +0,29 % sin tocar la memoria, su dilución con decenas de miles de sujetos y una ventana de ocho tokens que la resuelve (§4.12).
8. Una ablación preregistrada de la receta en familias de piezas, que saca el ancla del mundo y muestra qué protege cada familia, las anclas de texto el idioma y las del sujeto la generalización (§4.13).
9. El caso de un mismo sujeto con otra relación sobre hechos reales, donde el sueño contamina las otras relaciones del sujeto corregido y tres intentos de separarlas chocan con el mismo límite, identificar la relación (§4.14).

10. Una charla simulada de dos semanas con un usuario, que mide lo que el objetivo pide, recordar lo que se dijo, preguntado al otro día y sin el contexto, con cuatro recetas del sueño, y el costo que queda con la final, que contesta con el valor de otro dato dicho donde nunca se dijo nada (§4.15).
11. Un portero que corta la composición afuera de los pesos, separando la entidad, que se busca en la bitácora por identidad, de la relación, que decide el lector de la bitácora (§4.5).
12. Un registro de las predicciones escritas antes de cada corrida, con las cumplidas y las falladas, y una intervención con resultado nulo informada como tal (§4.16).
13. Tres trampas de medición, frases del propio modelo, perplejidad sin calibrar y un registro de texto humano de un solo tema, con el control que las detecta, y un relato de lo que falló, con todas las variantes probadas y las agregadas después de ver resultados (§5, Apéndice A).

2. Trabajos relacionados

Consolidación tipo sueño. SEAL (Zweiger et al., 2025) entrena al modelo, con aprendizaje por refuerzo, para escribir sus propias instrucciones de actualización de pesos, e informa que el desempeño sobre tareas anteriores baja de a poco a medida que se acumulan ediciones. Behrouz et al. (2026) proponen una fase de sueño en dos etapas, una destilación de las memorias de un modelo chico hacia uno más grande y una etapa de «soñar» en la que el modelo genera por refuerzo sus propios datos de práctica. Lee et al. (2026) consolidan el contexto reciente en pesos rápidos mediante pasadas recurrentes fuera de línea, en arquitecturas que combinan bloques de espacio de estados con atención. SCoL (Wang et al., 2026) usa meta-aprendizaje por refuerzo para que el modelo decida qué capas actualizar con el contexto. Nuestro trabajo difiere en el objeto, hechos versionados de una interacción con correcciones explícitas, y en que la consolidación se acompaña de un canal aparte, la bitácora, para lo que todavía no se consolidó.

Edición de conocimiento. MEMIT (Meng et al., 2023) edita miles de asociaciones a la vez modificando las proyecciones de salida de varias capas intermedias. AlphaEdit (Fang et al., 2025) proyecta la perturbación al espacio nulo de una covarianza del conocimiento a preservar, para no alterar las salidas sobre ese conocimiento, y está pensado para edición secuencial. WISE (Wang et al., 2024) separa una memoria principal de una memoria lateral de ediciones con un enrutador, y Larimar (Das et al., 2024) agrega una memoria episódica distribuida que se escribe en un paso. Comparamos directamente contra MEMIT y AlphaEdit (§4.3). KLOD (Jeong et al., 2026) edita con entropía cruzada sólo sobre el objetivo y un KL contra el modelo original en las posiciones del prefijo, que es en lo esencial la pérdida de nuestro sueño anclado al origen (§4.9). La usa para 3.000 ediciones optimizadas juntas en una sola capa, sin encadenar semanas, sin repaso y sin medir la perplejidad sobre texto humano. No reclamamos esa pérdida como propia.

Memoria explícita con alcance. SERAC (Mitchell et al., 2022) guarda las ediciones en una memoria explícita y entrena un clasificador de alcance que decide si una entrada cae dentro de alguna edición antes de responder con un modelo contrafáctico. GRACE (Hartvigsen et al., 2023) guarda un libro de códigos en una capa con un radio que decide cuándo aplicar cada edición, y RRDA (Zhang y Huang, 2026) usa un enrutador de relevancia entre un adaptador que escribe la edición y otro que preserva el comportamiento original. MemLLM (Modarressi et al., 2024) guarda el conocimiento como ternas de sujeto, relación y objeto que el modelo consulta de forma explícita. El portero de la §4.5 aplica esa idea general a un problema distinto, la composición espuria que la consolidación crea en los pesos, y la resuelve separando la entidad de la relación. La teoría del índice hipocampal (Teyler y DiScenna, 1986) propone una división

del trabajo parecida en el cerebro, con el hipocampo guardando un índice de lo vivido y la corteza guardando el contenido.

Hechos en los pesos, de forma continua. O'Neill et al. (2026) escriben hechos inventados en Qwen3, uno por vez, con un LoRA de rango 16 que se funde en los pesos antes del siguiente, y sin repasar los anteriores. Tras veinte escrituras el modelo contesta bien el 1 % de las preguntas sobre los hechos anteriores cuando cada escritura repite el enunciado, y el 46 % cuando usa reformulaciones variadas. Nuestra ablación reproduce ese colapso cuando se quita el repaso (§4.2), y con el repaso lo evita. Las reformulaciones como aumento vienen de Allen-Zhu y Li (2024), que muestran que sin ellas un hecho memorizado no se puede extraer, y de O'Neill et al. (2026). Nuestras autoparáfrasis (§4.10) son de esa familia.

Memoria de un asistente personal. Zhu et al. (2026) muestran que un asistente que comprime la conversación en su caché recuerda la alergia del usuario el 0 a 1 % de las veces (97 % con la conversación entera), y concluyen que un caché comprimido no es una capa de persistencia. HypReflect (Hwang et al., 2026) personaliza un modelo de forma continua destilando hipótesis sobre las preferencias del usuario, y SleepGate (Xie, 2026) resuelve la interferencia de valores viejos en el caché de atención con un etiquetador de conflictos y una compuerta de olvido. Ninguno de los tres consolida en los pesos lo que el usuario afirmó, con sus correcciones. Los dos trabajos más cercanos que encontramos en el control de literatura del 6 de octubre son de 2026. Dennis et al. (2026) consolidan cada noche en un adaptador LoRA el conocimiento de conversaciones de desarrollo de software, con reflexión, síntesis de datos y ajuste fino, y retienen 80,4 % del conocimiento contra 36,8 % de tres ciclos de compactación del contexto en cascada. Li (2026) guarda los hechos de cada usuario como ediciones locales en la tabla de memoria de un modelo Engram, que se prenden sólo en el disparador y dejan intacto el resto del texto, la misma idea que el portero del adaptador (§4.12) lleva a un LoRA usando la bitácora. Choi (2026) muestran que en la actualización continua de conocimiento el orden entre métodos cambia con el momento de la evaluación y la capacidad del adaptador, y por eso informamos trayectorias de ocho semanas y no sólo el punto final. Poner el hecho en el contexto, como hace la línea de la bitácora en la charla, tiene su antecedente en IKE (Zheng et al., 2023), y DiSC (Padmanabhan et al., 2026) actualiza el conocimiento paramétrico destilando el contexto con una KL en los tokens compartidos para conservar lo aprendido en el posentrenamiento, un compromiso cercano al del sueño anclado.

Conocimiento nuevo y alucinación. Gekhman et al. (2024) muestran que, al ajustar un modelo con hechos nuevos, esos ejemplos se aprenden más lento que los compatibles con lo que el modelo ya sabe y que, a medida que se aprenden, aumenta de forma lineal la tendencia a alucinar. La composición espuria que medimos es un fenómeno del mismo tipo, en un régimen distinto (adaptador con repaso explícito en lugar de ajuste directo).

Memoria versionada en trabajos previos del autor. La bitácora retoma una línea sobre memoria versionada dentro del modelo desarrollada en modelos chicos entrenados desde cero, que estudió la indexación geométrica (Speranza, 2026a), dónde conviene ubicar la abstención (Speranza, 2026b), un sello de orden que resuelve el conflicto entre versiones (Speranza, 2026c) y el alcance de la consulta en la recuperación (Speranza, 2026d). Este trabajo lleva esa memoria a un modelo preentrenado y le agrega la consolidación.

Tabla 1. Receta del sueño. Probabilidad de que una muestra del lote salga de cada fuente.

fuentes	semana 1	semanas 2 a 8
frases propias generadas por el modelo antes de dormir (réplica)	0,30	0,20
hechos de los paquetes anteriores, con su valor original	no hay	0,15
las dos habilidades, con sus ejemplos de abstención	0,15 o 0,25	0,15 o 0,25
«no sé» ante combinaciones conocidas (70 %) y entidades o relaciones inventadas (30 %)	0,20	0,20
hechos nuevos con una versión más nueva leída de la bitácora	0,15	0,15
hechos nuevos del paquete (enunciado o pregunta)	el resto	el resto

3. Método

3.1. El modelo y el mundo sintético

Usamos Qwen3-0.6B (Yang et al., 2025) con todos sus parámetros congelados en todos los experimentos. Los hechos salen de un idioma sintético en castellano con 102 entidades y 24 relaciones (2.448 pares posibles). Cada hecho es una terna relación, entidad y valor, con un enunciado natural («el gasto de finca es 64») y una pregunta directa («¿cuál es el gasto de finca?»). Que las entidades y los valores sean inventados garantiza que el modelo no los conoce de antemano, así que todo acierto se debe a la memoria que medimos.

3.2. Canal 1, la bitácora

La bitácora escribe una clave por entrada, a partir del enunciado completo con el valor tapado, en la capa 6 del modelo, y la lee en la salida con un lector aprendido. Con la clave derivada del valor, en una versión anterior, la exactitud global quedaba entre 0,61 y 0,65. Con la clave del enunciado tapado, en un estudio preregistrado con tres semillas, el modelo congelado con la bitácora alcanza una exactitud global de 0,900 a 0,938, elige la entrada correcta entre varias versiones del mismo hecho 0,920 a 0,955 de las veces y dice «no sé» ante hechos ausentes entre 0,947 y 0,989 de las veces, con una falsa abstención de 0,000 a 0,009. Un lector entrenado con 500 entradas ajenas en el archivo sostiene entre 0,932 y 0,945 con esas 500 entradas presentes, y uno entrenado y evaluado con 3.000 entradas da 0,900 (una semilla). En la lectura el modelo procesa sólo la pregunta, unos 16 tokens, y el archivo se consulta por fuera del contexto.

3.3. Canal 2, el sueño

Cada sueño entrena un adaptador LoRA (Hu et al., 2022) de rango 16 en las siete proyecciones de las 28 capas (10,09 millones de parámetros), durante 6.000 pasos con lotes de 8 y tasa de aprendizaje $2 \cdot 10^{-4}$. Cada muestra del lote se sortea entre las fuentes de la Tabla 1. Sobre las muestras de repaso (frases propias, hechos anteriores y habilidades) se agrega un ancla KL de peso 0,3 contra el modelo con el que empieza ese sueño, y la pérdida se promedia por ejemplo y no por token, para que los ejemplos largos de una habilidad no dominen a los ejemplos cortos de abstención. Cada sueño parte del adaptador que dejó el anterior. Esta es la receta de la versión 1, con la que se hacen el mundo sintético y las comparaciones de las secciones siguientes. La receta para hechos reales, que cambia el objetivo y las anclas, se describe en la §4.10, y la final, que es la misma sin el ancla del mundo, en la §4.13.

La réplica de habilidades fue 0,25 en las semillas 2 y 3 desde la primera semana. En la semilla 1 fue 0,15 en las semanas 1 a 3 y 0,25 desde la 4, después de que la semana 1 mostrara el código en 0,44 con 0,15. Es una desviación del protocolo que declaramos acá, y también el motivo por el que la semilla 1 dedica más muestras a los hechos nuevos en sus primeras semanas. Con réplica 0,15 los hechos nuevos directos son entre el 15 y el 20 % de las muestras, y con 0,25 entre el 5 y el 10 %.

3.4. Dos habilidades inventadas

Para saber si una consolidación rompe capacidades que no tocó hacen falta capacidades que el modelo no pueda tener por su preentrenamiento. Si no, «se sostiene» y «nunca se rompió» serían indistinguibles. Usamos dos. «Nudo» es un lenguaje de programación de juguete que se traduce a Python, y $\diamond$ es un operador aritmético inventado, $a \diamond b = 2a + b$. Sin entrenamiento el modelo no las resuelve (0,00 en las dos), y explicarle la regla de $\diamond$ en el prompt tampoco alcanza (0,00). Un sueño de 1.500 pasos lleva Nudo de 0,00 a 1,00 en 100 programas nunca vistos, y $\diamond$ de 0,00 a 0,88 (0,87 con otra semilla), o a 0,99 si el procedimiento se escribe paso a paso. Cada habilidad tiene un borde de abstención, decir «no sé» ante un operador que nunca vio ($\star$, que comparte dos de sus tres tokens con $\diamond$) y ante una palabra de Nudo que no existe. El borde de $\diamond$ se perdió por completo (0,00) cuando el sueño siguiente repasó sólo los ejemplos positivos de la habilidad, y se sostuvo en 1,00 cuando repasó también los de abstención. Todas las cadenas de este trabajo parten del mismo adaptador con las dos habilidades ya aprendidas.

3.5. Evaluación

Después de cada sueño se mide, con generación voraz e igualdad exacta, lo siguiente.

- **Lo nuevo**, los 100 hechos del paquete recién dormido.
- **Lo viejo**, cada paquete anterior por separado. La retención de la primera semana es la exactitud sobre el paquete 1.
- **La bitácora gana**, 50 hechos del paquete con una versión nueva en la bitácora. Se lee con un lector de una etapa previa sobre un archivo de una sola entrada, así que mide si una lectura de la bitácora se impone a lo consolidado en los pesos, no la exactitud de la búsqueda.
- **Composición**, 50 pares de una entidad con hechos consolidados y una relación que nunca tuvo para ella, donde la respuesta correcta es «no sé». Se informa la fracción que contesta un valor.
- **Abstención por clase**, 40 preguntas por clase con una entidad inventada, una relación inventada o las dos, todas distintas de las usadas para practicar el «no sé».
- **Habilidades**, Nudo, $\diamond$ (con el cálculo final hecho por una calculadora externa, para separar la regla de la suma) y sus dos bordes de abstención.

Con 100 preguntas el error estándar de una exactitud cerca de 0,9 es de unos 0,03, y con 50 de unos 0,06. Como la medición de composición con 50 pares es ruidosa y esos pares pueden coincidir con combinaciones que algún sueño practicó como «no sé», en las semillas 2 y 3 se reservaron además 300 pares antes del primer sueño, excluidos de todo paquete, control y práctica de abstención durante las ocho semanas. Sobre ellos se mide la composición limpia.

3.6. Líneas de base y ablaciones

- **Ingenua**. Los mismos paquetes de hechos, sin ancla, sin réplica de nada, sin «no sé» y sin bitácora.
- **Sin repaso de hechos**. La receta de la semilla 1 sin la fuente de hechos anteriores. Su primera semana es la misma corrida que la de la semilla 1.
- **MEMIT y AlphaEdit**. Los mismos paquetes semanales y el mismo punto de partida, editados en secuencia sobre las proyecciones de salida de las capas 4 a 8. Usamos $\lambda = 15.000$ y una covarianza estimada con 3.000 frases generadas por el propio modelo, y para AlphaEdit el umbral de espacio nulo $2 \cdot 10^{-2}$ y la regularización 10 del original. Dos diferencias con los originales quedan declaradas. La covarianza sale de frases propias y no

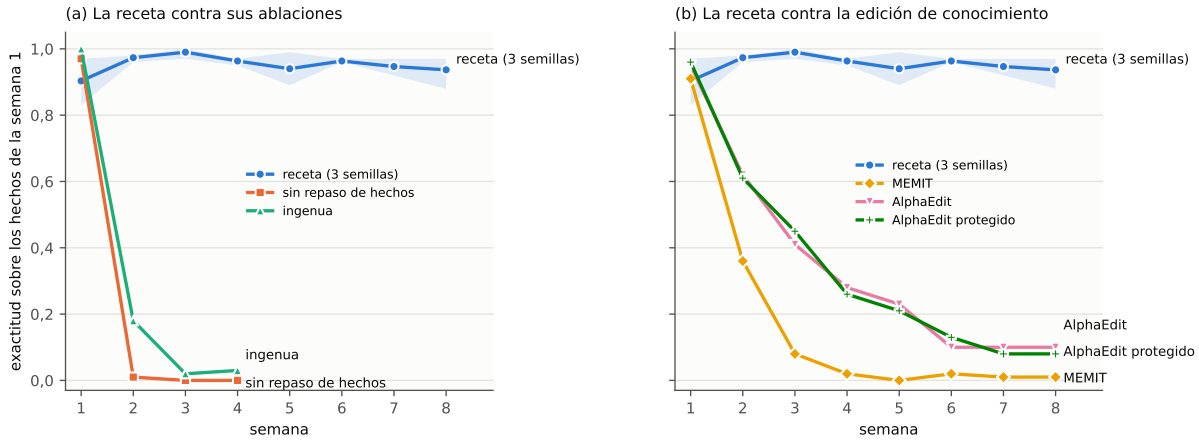


Figura 1. Exactitud sobre los 100 hechos de la primera semana a lo largo de ocho semanas. En la semana 1 cada curva muestra lo recién aprendido. La banda es el rango de las tres semillas de la receta. (a) Sin el repaso de hechos la primera semana desaparece en un sueño. (b) La edición secuencial con MEMIT y AlphaEdit, sobre los mismos hechos, la pierde de a poco.

de Wikipedia, y la evaluación usa la misma plantilla que la edición, lo que favorece a los editores. Como estos métodos no tienen bitácora, en su lugar se mide la corrección por edición, que edita 50 hechos con un valor nuevo, mide y deshace.

- **Sorteo balanceado.** La receta de la semilla 1 en cuatro semanas, pero eligiendo qué entidades reciben hechos nuevos con peso $1/(1+c)^{1,5}$, donde c es cuántos hechos ya tiene la entidad, con el muestreo ponderado sin reemplazo de [Efraimidis y Spirakis \(2006\)](#).

3.7. Registro previo

Antes de cada corrida escribimos en el repositorio del proyecto la predicción numérica y la regla de decisión, con la fecha del commit como marca de orden. La §4.16 las contrasta con los resultados. Las del portero (§4.5) están en un archivo propio, paso por paso, con los pasos que fallaron, y las de la cadena con hechos reales en otro.

4. Resultados

4.1. Ocho sueños seguidos no borran lo viejo

La Tabla 2 y la Figura 1a muestran la exactitud sobre el paquete de la primera semana al final de cada una de las ocho semanas. A la octava semana queda en 0,97, 0,88 y 0,96 según la semilla. La semilla 2 tiene su peor valor justo al final, así que no afirmamos ausencia de tendencia, pero la caída es chica. Otra forma de medirlo es comparar cada paquete con el valor con que se aprendió la semana en que se durmió. Ninguno cae más de 0,09 por debajo de ese valor en ninguna semilla (0,09, 0,01 y 0,02 como peor caso), y varios mejoran en sueños posteriores. El paquete de la semana 7 de la semilla 1, por ejemplo, se aprendió en 0,78 y a la semana siguiente estaba en 0,89.

El resto de las mediciones acompaña. Lo nuevo de cada semana se aprende entre 0,77 y 0,97, más bajo en las semillas 2 y 3, que dedican menos muestras a los hechos nuevos. El peor paquete anterior en cualquier semana queda en 0,86, 0,85 y 0,81 según la semilla. En las semillas 2 y 3 es un paquete que ya se había aprendido bajo (0,81 y 0,77) y que después subió. En la semilla 1 es el de la semana 6, que bajó de 0,95 a 0,86 al dormir la séptima y volvió a 0,95 en la octava. La bitácora se impone a lo consolidado entre 0,94 y 1,00 de las veces en las veinticuatro evaluaciones. Las habilidades se sostienen a la octava semana entre 0,88 y 1,00, con una primera

Tabla 2. Exactitud sobre los hechos de la primera semana al final de cada semana, por semilla. La semana 1 es lo recién aprendido.

semana	1	2	3	4	5	6	7	8
semilla 1	0,97	0,98	1,00	0,97	0,99	0,97	0,97	0,97
semilla 2	0,83	0,96	1,00	0,97	0,94	0,96	0,95	0,88
semilla 3	0,91	0,98	0,97	0,95	0,89	0,96	0,92	0,96
media	0,90	0,97	0,99	0,96	0,94	0,96	0,95	0,94

Tabla 3. La receta contra sus ablaciones en las mismas cuatro semanas (semilla 1). Composición es la fracción que contesta un valor donde corresponde «no sé» (menor es mejor).

	semana 1 retenida en la							«no sé»
	sem. 2	sem. 3	sem. 4	bitácora	composición	$\diamond$	Nudo	entidad nueva
receta completa	0,98	1,00	0,97	0,96	0,34	1,00	1,00	0,90
sin repaso de hechos	0,01	0,00	0,00	1,00	0,12	1,00	1,00	0,97
ingenua	0,18	0,02	0,03	0,24	1,00	0,01	0,00	0,00

erosión de $\diamond$ en la semilla 1 (0,88), y los dos bordes de abstención siguen en 1,00 en las tres semillas.

4.2. Qué pasa sin la receta

Con el ancla, la bitácora, la práctica de «no sé» y la réplica de habilidades intactas, quitar sólo el repaso de hechos hace que la primera semana pase de 0,97 a 0,01 en un sueño y a 0,00 en los dos siguientes (Tabla 3, Figura 1a). La bitácora y las habilidades, que sí se siguen repasando, no se degradan. El ancla contra el modelo anterior, que sigue actuando sobre las frases propias y las habilidades, no alcanza para proteger hechos que no se repasan. Hace falta repasarlos. La composición es incluso menor sin repaso (0,12 contra 0,34), lo que es coherente con que dependa de cuántos hechos sigue sabiendo el modelo. La línea ingenua colapsa en todo a la vez. Pierde lo viejo, pierde Nudo desde la primera semana y $\diamond$ desde la segunda, y desde la primera deja de decir «no sé» en todas las clases y contesta un valor ante cualquier combinación.

4.3. Contra la edición de conocimiento

Los dos editores aprenden bien lo nuevo de cada semana (entre 0,84 y 0,97) y corrigen un hecho por edición entre 0,80 y 1,00 de las veces, pero pierden lo editado antes (Tabla 4, Figura 1b). A la octava semana MEMIT retiene 0,01 de la primera semana y AlphaEdit 0,10. Ajustar λ no resuelve el compromiso. Con $\lambda = 40.000$ MEMIT edita sólo 0,55 de los hechos nuevos y retiene 0,18 de la primera semana a la segunda, menos que con 15.000, y con 150.000 edita 0,07. Ninguno de los dos tiene una noción de «no sé», y contestan un valor en el 98 a 100 % de las combinaciones sin hecho.

Los editores también dañan capacidades que la edición nunca tocó. Con MEMIT, Nudo cae a 0,00 desde la quinta semana. Con AlphaEdit el daño empieza antes, Nudo en 0,25 a la segunda semana y 0,00 desde la cuarta, y $\diamond$ en 0,00 desde la sexta. AlphaEdit sólo puede editar en el espacio nulo de la covarianza que estima, y esa covarianza sale de frases en prosa. Medimos cuánta energía ponen en ese espacio nulo las entradas de la capa que corresponden a cada tipo de texto, con la misma covarianza del experimento (2.798 frases). En las capas 4 a 6 las de los programas de Nudo ponen entre el 41 y el 48 %, las de $\diamond$ entre el 18 y el 22 % y las de prosa nueva entre el 10 y el 12 %, y en las capas 7 y 8 el orden es el mismo con valores menores. AlphaEdit edita justamente donde viven las habilidades, y el orden coincide con el del daño, primero Nudo y después $\diamond$.

Tabla 4. Retención de la primera semana con la receta (semilla 1) y con edición secuencial de los mismos 100 hechos por semana, y estado de las habilidades a la octava semana.

método	semana 1 retenida en la semana							a la semana 8			
	2	3	4	5	6	7	8	◇	Nudo	«no sé» ★	«no sé» Nudo
receta	0,98	1,00	0,97	0,99	0,97	0,97	0,97	0,88	0,97	1,00	1,00
MEMIT	0,36	0,08	0,02	0,00	0,02	0,01	0,01	1,00	0,00	0,90	0,00
AlphaEdit	0,62	0,41	0,28	0,23	0,10	0,10	0,10	0,00	0,00	0,00	0,00
AlphaEdit protegido	0,61	0,45	0,26	0,21	0,13	0,08	0,08	0,97	0,66	0,65	0,07

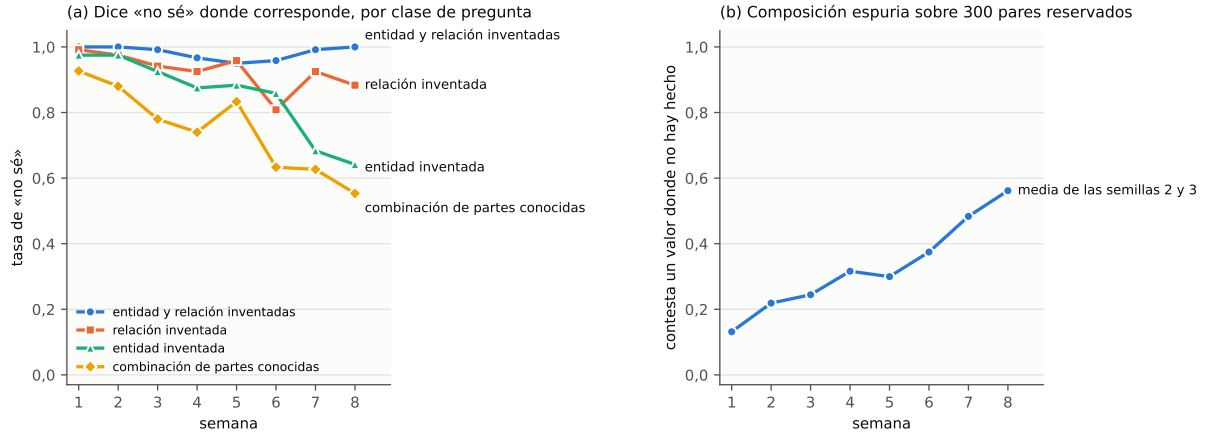


Figura 2. (a) Tasa de «no sé» por clase de pregunta, media de las tres semillas. Con una entidad y una relación inventadas la abstención se sostiene. Con una entidad inventada y una relación conocida, o con una entidad y una relación conocidas que nunca se combinaron, se erosiona. (b) Fracción de los 300 pares reservados en la que el modelo contesta un valor en lugar de «no sé», media de las semillas 2 y 3.

Para probar que ese es el mecanismo repetimos la cadena agregando a la covarianza los 200 ejemplos de entrenamiento de cada habilidad (AlphaEdit protegido). Nudo pasa de 0,25 a 1,00 en la segunda semana, y a la octava semana ◇ sigue en 0,97 y Nudo en 0,66. La retención de hechos no cambia (0,61, 0,45, 0,26, 0,21, 0,13, 0,08 y 0,08 contra 0,62, 0,41, 0,28, 0,23, 0,10, 0,10 y 0,10), así que la protección cuida lo que se pone en la covarianza y nada más. Lo que sí cae con la protección son los dos bordes de abstención, el «no sé» de Nudo en 0,07 y el de ★ en 0,65. La covarianza protegida incluía sólo los ejemplos positivos de cada habilidad, no los de abstención, y el resultado repite lo que vimos con el sueño en la §3.4. Proteger lo que el modelo sabe hacer no protege lo que sabe que no sabe hacer.

4.4. La abstención por clase y la composición espuria

La abstención no se degrada igual en todas las clases (Figura 2a). Con la entidad y la relación inventadas se sostiene en 1,00 de punta a punta. Con una relación inventada baja de 0,99 a 0,88. Con una entidad inventada y una relación conocida baja de 0,97 a 0,64 (0,70, 0,53 y 0,70 según la semilla a la octava semana). Y ante una entidad y una relación conocidas que nunca se combinaron baja de 0,93 a 0,55. El patrón sugiere que el modelo decide si contestar mirando por separado si conoce la relación y si conoce la entidad, y que con suficientes hechos acumulados reconocer una de las dos partes alcanza para que conteste. Revisando las respuestas guardadas de una corrida anterior, con un mundo más chico, en 43 de 43 errores de composición la entidad y la relación preguntadas aparecían por separado entre los hechos consolidados, y el valor contestado era el de otro hecho.

Sobre los 300 pares reservados, que ningún sueño tocó nunca, la composición crece de 0,06 a 0,61 en la semilla 2 y de 0,21 a 0,51 en la semilla 3 (Figura 2b). Es el límite principal de este

Tabla 5. El portero a la octava semana de las semillas 2 y 3, sobre los 300 pares reservados y los 800 hechos dichos. «Pasan» es la fracción de hechos dichos que el portero deja contestar a los pesos. «Compone» es la fracción de pares reservados en la que el sistema contesta un valor. El área es la que queda bajo la curva que separa hechos de composiciones por el margen entre la mejor entrada y el slot nulo. Salvo la última fila, el lector es el de la semilla preregistrada.

	hechos que pasan	compone	área
pesos solos, sin portero	todos	0,613 y 0,510	
lector de la bitácora tal como está	0,31 y 0,29	0,013 y 0,010	0,65 (0,995 vistas, 0,52 nuevas)
sin entrenar, medias del modelo congelado			0,58 y 0,59
sin entrenar, token contra token			0,81 y 0,80
entidad por identidad y relación por el lector	0,948 y 0,965	0,000 y 0,000	1,000
lo mismo, umbral fijado en la otra cadena, tres semillas	0,983 a 1,000	0,000 a 0,007	1,000

trabajo. En la semilla 1, con 50 pares por semana, la serie es 0,08, 0,10, 0,28, 0,34, 0,22, 0,48, 0,24 y 0,44.

Dentro de cada semana, las entidades con más hechos acumulados componen más. Sobre las primeras cuatro semanas de la semilla 1 (200 preguntas) la correlación entre hechos por entidad y contestar un valor es $r = 0,37$ ($p = 6 \cdot 10^{-8}$), 0,34 contra 0,09 entre entidades con más de dos hechos y el resto, y dentro de cada semana r va de 0,12 a 0,34. Para ver si la relación es causal corrimos el sorteo balanceado, que reparte los hechos nuevos entre las entidades con menos hechos. Sumando las cuatro semanas, el modelo contesta un valor en 40 de 200 preguntas con el sorteo uniforme y en 40 de 200 con el balanceado. La correlación es real y la intervención no la confirma como causa. Hay que notar que las dos cadenas usan semillas distintas. En una comparación exploratoria, con 3,5 veces más pares posibles la composición a 400 hechos fue la misma (0,34 contra 0,32 y 0,34 en un mundo de 696 pares, con 50 preguntas cada una y recetas no idénticas), lo que tampoco apunta al tamaño del mundo.

4.5. El portero, la composición se corta afuera de los pesos

Los pesos no pueden saber qué no se les dijo, pero la bitácora sí, porque guarda por construcción todo lo que se dijo. El portero usa eso. Antes de responder, la bitácora decide si el par de la pregunta existe, y sólo si existe contestan los pesos. Si no, la respuesta es «no sé». Nada se entrena. Las respuestas de los pesos son las que ya guardó cada evaluación de las semillas 2 y 3, y la bitácora de la octava semana tiene los 800 hechos de la cadena, uno por entrada. Cada paso tuvo su predicción y su regla escritas antes de correrlo (Tabla 5).

Con el lector de la bitácora tal como está, la composición baja de 0,613 y 0,510 a 0,013 y 0,010, y la bitácora de la otra semilla, usada como bitácora ajena, deja pasar sólo 0,046 y 0,028 de los hechos que no tiene. Pero de los hechos que sí se dijeron pasan 0,31 y 0,29, y por la regla escrita eso no es una solución. El diagnóstico lo explica. Con las 29 entidades con las que se entrenó el lector, los hechos pasan entre 0,97 y 0,98 y el área es 0,995. Con las 73 entidades que el lector nunca vio el área es 0,52, el azar. El lector aprendió a reconocer sus entidades, no a comparar una entidad cualquiera.

Sin nada entrenado tampoco alcanza. Comparando la pregunta reescrita como afirmación contra las entradas, por coseno sobre la media de los estados del modelo congelado y sin dejar que la pregunta use la misma plantilla que la entrada, el área es 0,58 y 0,59. Comparando token contra token, con la interacción tardía de ColBERT (Khattab y Zaharia, 2020), sube a 0,81 y 0,80 en los embeddings y cae en las capas siguientes. El desglose separa las dos piezas. La entidad se empareja sola por identidad, igual con las entidades vistas y con las nuevas, y lo que falla es la relación entre plantillas distintas («la altura de bahía» contra «bahía mide»).

El portero final separa esas piezas. La entidad es un conjunto abierto y se busca por identidad. Las candidatas son las entradas de la bitácora que nombran la misma entidad que la pregunta, lo que supone que al escribir se marca la entidad de cada entrada, como hace cualquier base

Tabla 6. La bitácora contra RAG sobre el mismo banco de preguntas. Exactitud por generación libre. La bitácora tiene tres semillas y 128 preguntas por semilla. El RAG tiene una semilla, y con 200 entradas viejas su banco es de 64 preguntas, con menos de diez en algunas clases.

	exactitud	«cuál era antes»	«no sé» de relación	tokens por consulta	n
bitácora, historial corto	0,953 a 0,984	0,95 a 1,00	1,00	16	3×128
bitácora, con 500 entradas viejas	0,938 a 0,961	0,95 a 1,00	1,00	16	3×128
RAG entrenado, historial corto	0,943	0,935	0,77	505	512
RAG entrenado, con 200 entradas viejas	0,953	0,67	1,00	4.514	64
RAG sin entrenar, historial corto	0,334	0,065	0,35	505	512
RAG sin entrenar, con 200 entradas viejas	0,109	0,33	0,00	4.514	64

de conocimiento con un reconocedor de entidades. La relación es un conjunto cerrado que el lector ya distingue. En la pregunta y en las candidatas la entidad se reemplaza por una que el lector conoce, y el lector decide si alguna candidata tiene la relación preguntada. Con la semilla preregistrada del lector pasan 0,948 y 0,965 de los hechos (0,944 y 0,965 con las entidades nuevas), la composición queda en 0,000 en las dos cadenas y el área es 1,000. La bitácora ajena deja pasar 0,000 de lo que no tiene y 0,96 de lo que tiene. Con otra entidad de reemplazo el resultado se repite (0,989 y 0,985 de los hechos, composición 0,000), y la plantilla que expresa la relación con un verbo, sin la palabra de la pregunta, pasa entre 0,94 y 1,00, así que no es coincidencia de palabras.

Con las otras dos semillas del lector el área sigue en 1,000, pero el corte que da el slot nulo cambia. Con una es estricto y deja pasar entre 0,62 y 0,64 de los hechos, y con la otra es laxo y deja componer 0,027. Es calibración, no separación. Fijando el umbral del margen en una cadena y aplicándolo a la otra, en las tres semillas y en los dos sentidos pasan entre 0,983 y 1,000 de los hechos y la composición queda entre 0,000 y 0,007.

El alcance hay que leerlo junto con el número. Con la entidad reemplazada y el valor tapado, la decisión depende sólo de la relación preguntada y de las relaciones y plantillas de las candidatas, y esas cuatro plantillas son las mismas con las que se entrenó el lector. Lo que se muestra que generaliza es la entidad nueva, no una forma nueva de decir una relación. Y el portero corta la composición porque no deja contestar a los pesos cuando la bitácora no tiene el par, así que la bitácora tiene que conservarse, aunque sea como índice de lo que se dijo. Con hechos reales el mismo principio repara el daño que el sueño le hace a otros sujetos (§4.8).

4.6. La bitácora contra RAG

Un RAG que pega el historial en el prompt, entrenado con un adaptador del mismo tamaño durante 1.500 pasos, empata en exactitud con la bitácora (Tabla 6), y con un historial de 200 entradas viejas también. Por clase, el RAG entrenado iguala o supera a la bitácora en la versión vigente de un hecho (0,978 contra 0,913 a 0,986), y la bitácora es mejor en el «no sé» ante una relación que no existe (1,00 contra 0,77) cuando el historial es corto. Sin entrenamiento el RAG no funciona (0,334 y 0,109).

La diferencia de fondo es el costo. La bitácora procesa 16 tokens por consulta, lo que mide la pregunta, y el RAG 505 con el historial corto y 4.514 con 200 entradas viejas, 31 y 276 veces más. El tiempo por consulta del RAG pasa de 0,25 a 4,1 segundos al agregar esas 200 entradas. Con 500 entradas viejas, unos 19.000 tokens, el RAG deja de entrar en la GPU usada en todo el trabajo (una T4 de 16 GB se queda sin memoria al procesar el prompt, en fp32 y en fp16), mientras que la bitácora con esas 500 entradas sigue entre 0,938 y 0,961. En calidad la bitácora no gana, y así lo informamos.

4.7. Un control fuera del mundo sintético

Todo lo anterior usa hechos inventados. Como control de que la receta no depende de eso, dormimos un único sueño sobre 100 hechos reales de CounterFact (Meng et al., 2022) y otros

Tabla 7. Un sueño sobre hechos reales. Exactitud por generación, con la respuesta cortada en el primer salto de línea. Una semilla.

	hechos entrenados	hechos no entrenados	perplejidad sobre frases propias no vistas
CounterFact	0,10 $\rightarrow$ 1,00	0,10 $\rightarrow$ 0,28	10,9 $\rightarrow$ 7,1
zsRE	0,03 $\rightarrow$ 1,00	0,01 $\rightarrow$ 0,03	10,9 $\rightarrow$ 7,2

Tabla 8. La cadena con hechos reales (CounterFact). En cada celda de retención, la fracción en que el valor corregido supera al original por probabilidad y la fracción que se contesta exacta. Los vecinos son otros sujetos con la misma relación, y se informa la fracción en que el valor verdadero sigue ganando. Semilla 0; las semillas 1 y 2 y las semanas 5 a 8, en el texto.

	semana 1	semana 2	semana 3	semana 4
lo corregido esa semana, con repaso	1,00 y 0,99	1,00 y 0,99	1,00 y 1,00	1,00 y 1,00
lo corregido la semana 1, con repaso		1,00 y 0,99	0,99 y 0,93	0,99 y 0,97
lo corregido la semana 1, sin repaso		0,91 y 0,50	0,87 y 0,47	0,75 y 0,17
paráfrasis de la semana 1, con repaso y sin repaso		0,81 y 0,69	0,85 y 0,64	0,87 y 0,62
vecinos, modelo base	0,74	0,75	0,77	0,79
vecinos, pesos consolidados	0,34	0,42	0,42	0,46
vecinos, con el portero	0,74	0,75	0,77	0,79

100 de zsRE (Levy et al., 2017), en la versión para edición de De Cao et al. (2021), elegidos entre los que el modelo base no completa bien. Es una receta reducida (adaptador, ancla KL, réplica de frases propias y 1.500 pasos, sin práctica de «no sé» ni bitácora). Se mide también la exactitud sobre otros 100 hechos de cada banco que nunca se entrenaron y la perplejidad sobre 74 frases propias que el sueño no vio (Tabla 7).

Los hechos reales se aprenden por completo en los dos bancos, y la perplejidad sobre frases propias que el sueño no vio no empeora. *Corrección.* Esas frases las generó el propio modelo, y esa medida no ve el daño al idioma. Con la receta completa, un solo sueño sobre CounterFact sube la perplejidad sobre texto escrito por personas un 246 % en inglés y un 18 % en español (§4.9). Este control reducido no se midió sobre texto humano, así que la frase de la versión 1 («sin empeorar la perplejidad sobre texto no visto») vale sólo para frases del propio modelo. En CounterFact el sueño además hace acertar 18 hechos que nunca vio sin hacer fallar ninguno que ya acertaba. Los 18 son respuestas verdaderas, y 17 son valores que también aparecen como respuesta entre los hechos entrenados (París, francés, latín, Antártida). La explicación está en la forma de la respuesta. Antes del sueño, con la plantilla de pregunta y respuesta, el modelo base repetía la pregunta en 64 de los 100 hechos no entrenados de CounterFact y daba un objeto corto sólo en 36. Después da un objeto corto en los 100. El sueño enseña, además de los hechos, la forma de la tarea, y eso destapa respuestas que el modelo ya tenía. En zsRE, cuyas preguntas son naturales, el mismo cambio de forma (de 28 a 98 respuestas cortas) deja los hechos no entrenados casi igual, de 0,01 a 0,03, así que ahí el paso de 0,03 a 1,00 en los hechos entrenados es conocimiento nuevo. Por eso tomamos zsRE como el caso limpio. Este control es un solo sueño, sin práctica de abstención. La retención a lo largo de varios sueños con hechos reales la mide la sección siguiente, y el «no sé» con datos reales queda sin medir.

4.8. Una cadena de sueños con hechos reales

El control anterior es un solo sueño. Para ver si la retención transfiere a hechos reales repetimos la cadena sobre CounterFact completo (Meng et al., 2022) (19.728 casos) con la misma receta, réplica de frases propias 0,20, repaso de las semanas anteriores 0,15, ancla KL 0,3 al punto de partida, adaptador de rango 16 y 3.000 pasos. Cada semana se corrigen 100 hechos reales que el modelo ya cree saber, de sujetos distintos, reemplazando el valor verdadero por el

Tabla 9. Cambio de perplejidad contra el modelo base, en porcentaje, a lo largo de la cadena con hechos reales. Las frases del propio modelo son 120 frases generadas por el modelo base (la medida de la versión 1). El texto humano son 150 párrafos de Wikipedia en inglés, tomados de SQuAD (Rajpurkar et al., 2016), y 150 oraciones en español escritas fuera del modelo. La réplica de la versión 0 parte de la cuarta semana de la receta; la versión 0 y bilingüe, del modelo base. Semilla 0.

semana	1	2	3	4	5	6	7	8
frases del propio modelo, receta	-30	-30	-15	+6	+43	+91	+172	+256
Wikipedia en inglés, receta	+246	+372	+878	+1.270	+1.610	+1.891	+3.768	+5.412
prosa en español, receta	+18	+36	+84	+150	+263	+412	+698	+917
frases del propio modelo, réplica de la versión 0					-4	-4	-11	-10
Wikipedia en inglés, réplica de la versión 0					+1.156	+1.201	+872	+1.027
prosa en español, réplica de la versión 0					+147	+197	+181	+217
Wikipedia en inglés, versión 0 y bilingüe desde la semana 1	+72	+276	+353	+559				
prosa en español, versión 0 y bilingüe desde la semana 1	+6	+38	+58	+97				

de CounterFact. Medimos con las métricas estándar de la edición, que comparan probabilidades. La eficacia pide que el valor nuevo supere al viejo en la frase corregida, la paráfrasis lo mismo en frases reformuladas con un prefijo al azar, y los vecinos que en otros sujetos con la misma relación el valor verdadero siga ganando. Informamos además la respuesta exacta por generación. La ablación sin repaso parte de la misma primera semana (Tabla 8).

Con repaso, lo corregido la primera semana se sostiene a la cuarta en 0,99 por probabilidad y 0,97 contestado exacto, y sus paráfrasis, que nunca se entrenaron, en 0,87. Lo de las semanas 2 y 3 queda en 0,99 y 0,96 y en 1,00 y 1,00. Sin repaso, la primera semana cae a 0,75 y 0,17 y sus paráfrasis a 0,62. La caída por probabilidad es menor que la que habíamos predicho (0,60) y la exacta mayor. Es el patrón del mundo sintético, ahora con hechos reales. Con dos semillas más, la primera semana queda a la cuarta en 1,00 y 0,98 y en 1,00 y 0,99 con repaso, y en 0,74 y 0,29 y en 0,74 y 0,35 sin él. Extendida a ocho semanas, lo corregido la primera sigue en 0,98 por probabilidad y 0,92 exacto, y lo de las semanas 2 a 7 entre 0,97 y 0,99 por probabilidad. La perplejidad sobre frases del propio modelo, que daba un 6 % más a la cuarta semana, se trata en la sección siguiente con texto humano.

El sueño solo tiene un costo, el mismo que en el mundo sintético aparecía como composición. Lleva el valor nuevo a otros sujetos con la misma relación, y los vecinos en que el valor verdadero gana bajan de entre 0,74 y 0,79 con el modelo base a entre 0,34 y 0,46 con los pesos consolidados. Con hechos reales el portero no contesta «no sé», porque el modelo base sabe cosas del mundo. Si la bitácora no tiene el sujeto de la pregunta contesta el modelo base, y si lo tiene contestan los pesos consolidados. Las correcciones y sus paráfrasis pasan en todas las semanas (1,000) y de los vecinos pasan a lo sumo 0,016, así que los vecinos vuelven al nivel del modelo base sin que se pierda ninguna corrección. El portero decide por la identidad del sujeto. En CounterFact un mismo sujeto puede tener más de un hecho. De 20.391 sujetos, 785 tienen más de una relación, y el caso de un mismo sujeto con otra relación se mide en la §4.14.

4.9. El idioma, medido con texto humano

La versión 1 midió el idioma con la perplejidad sobre frases que el sueño no vio, pero generadas por el propio modelo base. Esas frases se parecen a la réplica con que el sueño se entrena, y la medida resultó ciega al daño. Sobre texto escrito por personas la imagen es otra (Tabla 9). La perplejidad sobre Wikipedia en inglés sube un 246 % en el primer sueño y crece sueño a sueño hasta multiplicarse por 55 a la octava semana, y sobre prosa en español sube un 18 % en el primer sueño y se multiplica por 10 a la octava. La memoria, mientras tanto, no se resiente (lo corregido la primera semana sigue en 0,98 y 0,92 a la octava). El modelo recuerda todo y habla cada vez peor.

Hay dos daños distintos. El primero se *acumula*. La réplica que protege el idioma la escribe, en cada sueño, el modelo con que ese sueño empieza, que ya se corrió en los sueños anteriores,

Tabla 10. El sueño anclado al origen. Cambio de perplejidad sobre texto humano contra el modelo base, en inglés y en español, en porcentaje. La entropía cruzada va sólo sobre la respuesta y el ancla KL ata todas las demás posiciones a la versión 0, con el peso indicado. A la derecha, lo corregido la primera semana a la cuarta (por probabilidad y exacto) y las paráfrasis de lo corregido la cuarta semana. Semilla 0, los mismos hechos en todas las filas.

	semana 1	semana 2	semana 3	semana 4	semana 1 a la cuarta	paráfrasis
receta	+246 y +18	+372 y +36	+878 y +84	+1.270 y +150	0,99 y 0,97	0,83
réplica de la versión 0, bilingüe	+72 y +6	+276 y +38	+353 y +58	+559 y +97	1,00 y 1,00	0,78
anclado, peso 1	+27,6 y -1,2	+19,2 y -3,3	+15,5 y -2,3	+43,5 y +1,4	1,00 y 1,00	0,84
anclado, peso 10	+5,1 y -4,9	+25,9 y -2,3	+15,6 y -1,9	+22,0 y -3,8	1,00 y 0,98	0,78
anclado, peso 3, tasa en coseno y recorte	+2,1 y -3,3	+6,0 y -3,5	+3,3 y -3,6	+5,9 y -3,9	1,00 y 1,00	0,60

así que cada semana el modelo se entrena sobre un texto propio un poco más degradado que el anterior. Lo probamos por intervención, con la misma receta y la réplica escrita siempre por el modelo original, la versión 0, desde la quinta semana. La medida sobre frases del propio modelo queda en -10% a la octava semana contra $+256\%$ con la receta, sin perder memoria (lo corregido la primera semana, 0,98 y 0,91 contra 0,98 y 0,92). Sobre texto humano el crecimiento también se frena ($+1.027\%$ en inglés y $+217\%$ en español a la octava, contra $+5.412\%$ y $+917\%$), aunque no se repara lo que la cuarta semana ya traía.

El segundo daño aparece en el *primer* sueño y es trece veces mayor en inglés que en español. Los hechos de CounterFact están en inglés y la réplica que ancla el idioma son frases escritas a partir de veinte consignas en español, de modo que el español queda anclado y el inglés no. Lo probamos con la misma receta y la mitad de la réplica escrita a partir de las mismas veinte consignas traducidas al inglés, por el modelo original. En el primer sueño la perplejidad sobre Wikipedia en inglés sube un 72% en lugar de un 246% , la de la prosa en español un 6% en lugar de un 18% , y los hechos se aprenden igual (1,00 por probabilidad y exactos). El daño del primer sueño era, en su mayor parte, el idioma que la réplica no cubría.

Con las dos correcciones desde el primer sueño, la réplica escrita por la versión 0 y en los dos idiomas, el daño sobre texto humano baja a menos de la mitad del de la receta a la cuarta semana ($+559\%$ en inglés y $+97\%$ en español, contra $+1.270\%$ y $+150\%$), sin costar memoria (lo corregido la primera semana, 1,00 y 1,00 a la cuarta). Pero no se detiene, y sigue creciendo sueño a sueño (en inglés $+72$, $+276$, $+353$ y $+559\%$).

La tercera causa estaba en el objetivo. La receta aplica la entropía cruzada a la frase entera de cada hecho («la lengua materna de X es Y»), que el sueño repite unas 150 veces, así que además de la respuesta le enseña al modelo a escribir plantillas de CounterFact, y el ancla sólo mira la réplica y el repaso. Lo único que tiene que cambiar, sin embargo, es la respuesta. El *sueño anclado al origen* aplica la entropía cruzada sólo a los tokens de la respuesta y ata con el ancla KL todas las demás posiciones, las del enunciado de cada hecho y las de la réplica, a la versión 0 (el adaptador apagado) en lugar del modelo con que empieza el sueño. Cada semana se mide contra el mismo origen, y el daño no puede acumularse por construcción. La réplica la escribe la versión 0 en los dos idiomas. Registramos antes de correr que una variante es la solución si a la cuarta semana deja el texto humano bajo el $+10\%$ en los dos idiomas, sin subir más de 5 puntos desde la primera semana, con lo de cada semana y lo corregido la primera sobre 0,95 por probabilidad y 0,90 exacto.

En el primer sueño el peso del ancla mostró una dosis y respuesta limpia, $+27,6\%$ en inglés con peso 1, $+9,0\%$ con peso 3 y $+5,1\%$ con peso 10, sin costar memoria (Tabla 10). En la cadena apareció un problema del optimizador. Con la tasa de aprendizaje fija y la pérdida de la respuesta en cero durante miles de pasos, la corrida salta de golpe. En los registros, la pérdida de la respuesta pasa de 0,000 a 0,979 en el último tramo de la segunda semana con peso 3, a 0,242 con peso 10 y a 0,461 en la cuarta semana con peso 1, y en esas corridas el daño al idioma salta (a $+84,1$, $+25,9$ y $+43,5\%$ en inglés). Con la higiene estándar contra esos saltos, la tasa que baja en coseno hasta el 10% y el recorte de la norma del gradiente en 1, agregada después

Tabla 11. Del sueño anclado a la receta con ancla del mundo, cuarta semana. Memoria es lo de la semana y la peor retención de las anteriores, paráfrasis las formas de preguntar que nunca se entrenan, vecinos la especificidad sobre otros sujetos con la misma relación (base 0,786). El texto humano es el cambio de perplejidad a temperatura 1 contra la versión 0. Abajo, esa receta calibrada en tres semillas, con la Wikipedia en inglés estrecha (un solo artículo, §4.11).

	memoria	paráfrasis	vecinos	Wikipedia en inglés	prosa en español
sueño anclado (peso 3, coseno y recorte), semilla 0	1,00 y 1,00	0,600	0,482	+5,9	-3,9
más 600 párrafos de texto humano por idioma, semilla 0	1,00 y 1,00	0,505	0,502	-1,6	-2,9
más autoparáfrasis de la versión 0, semillas 0 a 2	1,00 y 1,00	0,725 a 0,765	0,516 a 0,558	-1,2 a -0,6	-4,1 a -3,8
receta con ancla del mundo, semillas 0 a 2	1,00 y 0,99 a 1,00	0,945 a 0,950	0,606 a 0,638	-0,9 a -0,2	-3,7 a -2,5
receta con ancla del mundo, perplejidad CALIBRADA (cada modelo a su temperatura óptima), semillas 0, 1 y 2					
Wikipedia en inglés +0,79, +0,62 y +1,13 · usuarios en inglés +0,29, +1,59 y +1,35 · literatura en inglés +0,68, +0,69 y -0,76					
usuarios en español -2,71, -1,11 y -1,58 · literatura en español -2,14, -1,05 y -1,95					

de ver el primer salto, la corrida es lisa las cuatro semanas y cumple lo escrito. El texto humano en inglés queda en +2,1, +6,0, +3,3 y +5,9 % y el español entre -3,3 y -3,9 %, sin crecer, con lo de cada semana en 1,00 por probabilidad y 1,00 exacto y lo corregido la primera semana en 1,00 y 1,00 a la cuarta. Con las semillas 1 y 2 el primer sueño da +1,1 y +1,5 % en inglés y -3,5 y -3,1 % en español, con la memoria en 1,00. Un adaptador puesto sólo en las MLP de las capas 3 a 8, donde MEMIT edita, también protegía el idioma en el primer sueño (+17,7 %) pero generalizaba peor (0,49 en las paráfrasis) y lo detuvimos.

El precio es la generalización. Las paráfrasis de lo corregido, que nunca se entrenan, ganan en 0,54 a 0,60 de los casos contra 0,67 a 0,86 con la receta, y los vecinos quedan igual o apenas mejor (0,47 a 0,50 contra 0,34 a 0,46). Leído por generación, en el primer sueño la paráfrasis nombra el valor nuevo tanto como los vecinos (0,16 contra 0,16), así que lo que queda de paráfrasis es sobre todo un sesgo de la relación y no el hecho. No lo produce la fuerza del ancla, porque con peso 10 y la tasa fija las paráfrasis dan 0,645 en la primera semana y con peso 3 y la tasa fija 0,745. A igual peso lo que cambia es la higiene, con la que el modelo se mueve menos, y con ella bajan a la vez el daño (de +9,0 a +2,1 %) y la paráfrasis (de 0,745 a 0,54). Es una curva de compromiso entre generalizar y dañar el idioma, y el sueño anclado al origen elige el extremo que menos daña. Deja dos límites abiertos, el daño que todavía queda, que en un uso de años no es aceptable, y la generalización a otras formas de preguntar. La subsección siguiente cierra el primero y reduce el segundo.

4.10. Generalizar sin dañar el idioma, cuatro semanas

Una segunda trampa de medición. Con el ancla sobre texto humano la perplejidad a temperatura 1 *mejora* respecto de la versión 0 en los seis registros, y eso no puede ser una ganancia de conocimiento. La explicación es que el modelo de base está sobreconfiado y cualquier cosa que suavice su distribución baja la perplejidad sin que el modelo sepa más. La comparación justa pone a cada modelo en su propia temperatura óptima (perplejidad calibrada, 40 textos por registro en cinco registros que ninguna ancla ve, Wikipedia en inglés, mensajes de usuarios a un asistente en inglés y en español, y literatura en inglés y en español). Con esa medida, un doctorado que a temperatura 1 mejora un 0,9 % queda un 0,79 % peor. Desde acá informamos el daño calibrado. Esta sección cuenta la receta con ancla del mundo a cuatro semanas. La receta final es la misma sin el ancla del mundo (§4.13), y la §4.11 la lleva a ocho.

Las piezas. Partiendo del sueño anclado, agregamos de a una las piezas de la Tabla 11. El ancla sobre texto humano (600 párrafos por idioma de Wikipedia, en inglés tomados de SQuAD (Rajpurkar et al., 2016), atados enteros con KL a la versión 0) baja el daño por debajo de cero a temperatura 1 pero la paráfrasis cae a 0,505. Las autoparáfrasis, en las que la versión 0 reescribe cada corrección de cinco maneras (Allen-Zhu y Li, 2024; O’Neill et al., 2026) y sólo el valor lleva entropía cruzada, la suben a 0,74. La receta con ancla del mundo suma 2.000 hechos verdaderos de CounterFact (el ancla del mundo, que la §4.13 saca) y tres más por hecho corregido de otros

Tabla 12. La receta final a la cuarta y a la octava semana, en tres semillas, contra la receta con el ancla de 2.000 hechos del mundo. Memoria es la probabilidad con que gana lo corregido la primera semana y, entre paréntesis, la peor retención exacta de las semanas anteriores. Vecinos es la especificidad (base 0,786 a la cuarta semana y 0,700 a la octava, porque cada semana evalúa sujetos distintos). Wikipedia amplía el cambio de perplejidad calibrada contra la versión 0 en 150 párrafos de 150 artículos distintos por idioma, sin portero y con el portero del adaptador (§4.12).

	semana	memoria	paráfrasis	vecinos	inglés, sin y con portero	español, sin y con portero
receta final, semilla 0	4	1,00 (1,00)	0,945	0,584	+0,77 y +0,14	-1,30 y -0,18
receta final, semilla 1	4	1,00 (1,00)	0,950	0,584	+0,94 y +0,05	-1,17 y -0,19
receta final, semilla 2	4	1,00 (0,99)	0,960	0,582	+0,93 y +0,03	-1,48 y -0,19
receta final, semilla 0	8	1,00 (0,97)	0,960	0,600	+1,11 y +0,28	-0,48 y -0,09
receta final, semilla 1	8	1,00 (0,99)	0,965	0,604	+1,23 y +0,29	-0,95 y -0,21
receta final, semilla 2	8	1,00 (0,99)	0,950	0,618	+1,49 y +0,26	-0,94 y -0,11
receta con ancla del mundo, semillas 0 a 2	8	1,00	0,920 a 0,965	0,608 a 0,646	+1,50 a +1,89 y +0,20 a +0,36	-1,04 a -0,73 y -0,18 a -0,17

sujetos con la misma relación, atados enteros a la versión 0, un prefijo ajeno (la mitad de las veces el hecho va precedido por una frase de la versión 0) y un piso de 0,35 del lote para lo nuevo. La paráfrasis llega a 0,945 a 0,950 en tres semillas, con la memoria en 1,00 por probabilidad y la retención de las semanas anteriores entre 0,99 y 1,00. La receta se eligió entre dieciséis variantes corridas sobre la semilla 0 (Apéndice A), y las semillas 1 y 2, que no se usaron para elegirla, la confirman.

Lo que queda. Calibrado, el español mejora en las tres semillas (-1,05 a -2,71 %) y el inglés queda entre -0,76 y +1,59 %. El criterio escrito antes, a lo sumo +1,0 % calibrado en todos los registros, se cumple en la semilla 0 y no en las semillas 1 y 2, por los usuarios en inglés (+1,59 y +1,35 %) y, en la semilla 2, la Wikipedia en inglés (+1,13 %). Lo declaramos como límite abierto. El otro precio está en los vecinos, que bajan de 0,786 a 0,606 a 0,638. Con hechos de CounterFact, falsos a propósito, generalizar la corrección a otras formas de preguntar y no tocar lo que el modelo sabe de otros sujetos tiran en direcciones opuestas. La §4.15 muestra cómo se paga ese mismo intercambio cuando lo que se consolida es lo que dijo un usuario.

4.11. Ocho semanas y una tercera trampa de medición

Hasta acá las recetas llegaban a cuatro semanas. Antes de correr la octava escribimos cuatro predicciones para la receta con ancla del mundo en tres semillas. Que lo corregido la primera semana siguiera sobre 0,95 por probabilidad y 0,90 exacto, que la paráfrasis siguiera sobre 0,90 y los vecinos sobre 0,55, y que el daño calibrado de la octava semana no superara al de la cuarta en más de un punto en ningún registro. También predijimos que el criterio de la §4.10, a lo sumo +1,0 % calibrado en todos los registros, seguiría fallando en inglés, y siguió fallando. Las de memoria y paráfrasis se cumplieron (memoria 1,00 en las tres semillas, paráfrasis 0,920 a 0,965, vecinos 0,608 a 0,646). La del idioma falló. En la Wikipedia en inglés el daño calibrado subió de +0,79, +0,62 y +1,13 % a +2,74, +3,80 y +2,58 %, y la regla escrita antes dice que no podemos afirmar que el daño deje de crecer más allá de la cuarta semana. Crece mucho menos que con la receta de la versión 1, que fue de +1.270 a +5.412 % entre las mismas semanas, pero no se detiene.

Una tercera trampa. Al buscar qué palabras se dañaban aparecieron términos de fútbol americano. Los 40 párrafos de Wikipedia en inglés de la perplejidad calibrada venían todos del artículo sobre el Super Bowl 50, y los 150 de la medida a temperatura 1 de sólo tres artículos (el mismo, Varsovia y los normandos), porque SQuAD ordena sus párrafos por artículo y tomamos los primeros. Armamos una medida nueva con un párrafo de cada uno de 150 artículos distintos por idioma, disjunta del ancla de texto humano. Con ella el inglés de la receta con ancla del mundo sube de +0,64 a +0,93 % en la cuarta semana a +1,50 a +1,89 % en la octava, un crecimiento de 0,67 a 1,25 puntos contra 1,45 a 3,18 en la medida estrecha, y el español queda

por debajo de la versión 0 ($-0,73$ a $-1,04$ %) también en texto enciclopédico. En la medida estrecha, repartido por token en la semilla 0, el daño a los nombres propios es un nivel que no cambia entre las semanas 4 y 8 ($+0,042$ y $+0,044$ nats), y lo que crece despacio son las palabras comunes ($-0,002$ a $+0,006$ nats). Desde acá informamos la medida amplia. La lección se suma a las dos trampas anteriores. Un registro de texto humano tiene que cubrir muchos temas, porque un solo artículo puede dañarse bastante más que el promedio.

La receta final a ocho semanas. La receta final, que la §4.13 elige con una ablación, es la de la §4.10 sin el ancla de hechos del mundo. A la octava semana, en tres semillas, lo corregido la primera semana gana en 1,00 de los casos, la peor retención exacta de cualquier semana anterior es 0,97, la paráfrasis queda entre 0,950 y 0,965 y los vecinos entre 0,600 y 0,618, contra 0,700 del modelo original sobre los mismos sujetos (Tabla 12). Sin portero, el daño calibrado en el inglés amplio sube de $+0,77$ a $+0,94$ % en la cuarta semana a $+1,11$ a $+1,49$ % en la octava, un crecimiento de 0,29 a 0,56 puntos que cumple la regla de un punto que la receta con ancla del mundo no cumplía, y el español sigue por debajo de la versión 0 ($-1,29$ a $-2,44$ %), los mensajes de usuarios en inglés entre $+0,28$ y $+1,14$ %, la literatura en inglés entre $+1,06$ y $+2,16$ % y la Wikipedia estrecha, el artículo del Super Bowl, entre $+2,93$ y $+3,43$ % (a la cuarta semana eran $+1,01$ a $+1,45$ %). La literatura en inglés también crece entre las semanas 4 y 8.

4.12. El portero del adaptador

El daño que queda está en el texto que habla de otra cosa. La idea del portero de la §4.5, decidir afuera de los pesos si algo se sabe, se puede aplicar al adaptador entero. El portero del adaptador mira si el texto nombra a alguno de los sujetos que la bitácora tiene anotados. Si nombra alguno, el modelo habla con el adaptador, y si no, habla la versión 0 exacta. No se entrena nada. A la octava semana la bitácora tiene 800 sujetos, y el portero se prende en todas las preguntas sobre lo corregido y en todas sus paráfrasis, en 0,019 de las preguntas sobre los vecinos, que por eso vuelven al nivel del modelo original, en 10 de los 150 párrafos de Wikipedia en inglés y en 4 de los 150 en español. Con la receta final el inglés amplio pasa a entre $+0,26$ y $+0,29$ % y el español a entre $-0,09$ y $-0,21$ % en las tres semillas, sin cambiar la memoria (Tabla 12). Lo que queda viene de los 10 artículos en inglés que nombran un sujeto corregido (París, por ejemplo). Prender la parte editada sólo cerca de lo editado no es una idea nueva, GRACE (Hartvigsen et al., 2023) lo decide por token con un radio en el espacio de activaciones y WISE (Wang et al., 2024) con una memoria lateral y un enrutador, y aquí lo decide la bitácora por identidad.

El límite de escala, y una ventana. Con 800 sujetos el portero casi nunca se prende por error. Una bitácora de años no tendría 800 sino decenas de miles. Con los 18.466 sujetos de CounterFact como lista, el equivalente a tres años y medio de 100 hechos por semana, el portero por párrafo se prende en el 46 % del inglés amplio y en el 16 % del español, y con la receta con ancla del mundo el inglés vuelve a $+0,88$ %. Prenderlo sólo en los W tokens que siguen a la mención resuelve la dilución. Con $W = 8$ el token de la respuesta cae dentro de la ventana en todas las preguntas y paráfrasis de CounterFact, y en el inglés amplio con los 18.466 sujetos el daño baja a $-0,01$ % con la receta con ancla del mundo (se prende el 8,6 % de los tokens), y con la receta final pasa de $+0,76$, $+0,82$ y $+0,87$ % por párrafo a $+0,19$, $+0,15$ y $+0,08$ % en las tres semillas, con el español entre $-0,02$ y $-0,07$ %. El portero no reemplaza a las anclas. La receta mínima sin anclas (§4.13) queda en $+1,46$ % con la misma ventana. Dos límites. En CounterFact la respuesta va pegada al sujeto, y en una charla real no. En la charla de la §4.15, 52 de las 58 preguntas sobre datos dichos son sobre el propio usuario sin nombrarlo, así que un portero por nombre no aplica. Probamos dos sustitutos, por palabras de la pregunta y por similitud de frases, y ninguno llega a la memoria completa sin prenderse casi siempre (cubren

Tabla 13. Ablación de la receta a la octava semana, semilla 0 salvo que se indique. Memoria es la probabilidad con que gana lo corregido la primera semana. Inglés y español son la Wikipedia amplia calibrada, sin y con el portero del adaptador. Las predicciones (Q1 a Q4 y M1 a M4) se escribieron antes de correr cada familia.

	memoria	paráfrasis	vecinos	inglés, sin portero	inglés, con portero	español, con portero
receta con ancla del mundo (cfq)	1,00	0,965	0,612	+1,86	+0,20	-0,18
sin anclas de texto, humano y del mundo (cts)	1,00	0,950	0,626	+8,60	+0,99	+0,04
sin anclas del sujeto, vecinos, prefijo y réplica en inglés (ctt)	1,00	0,690	0,608	+1,02	+0,34	-0,16
sin el ancla del mundo (csm), receta final	1,00	0,960	0,600	+1,11	+0,28	-0,09
receta mínima, sin ninguna ancla (cmn), semillas 0 y 1	1,00	0,835	0,460 y 0,476	+12,5 y +13,9	+1,22 y +1,64	0,00 y -0,03
sólo lo nuevo, cmn sin autoparáfrasis (cpu)	1,00	0,640	0,500	+8,33	+0,44	-0,08

Tabla 14. Un mismo sujeto con otra relación, primera semana, 100 sujetos de CounterFact que tienen otro registro con otra relación. La fuga es la fracción de preguntas sobre otra relación del sujeto corregido en que la generación nombra el valor nuevo. NS dentro del sujeto es la fracción en que el valor verdadero de esa otra relación le gana al valor nuevo. PS son las paráfrasis de la relación corregida. 255 preguntas de otra relación. Semilla 0.

	fuga	NS dentro del sujeto	PS
modelo base	0,004	0,867	
receta	0,051	0,831	0,720
réplica dirigida al sujeto	0,008	0,890	0,465
autoparáfrasis	0,090	0,776	0,810
las dos	0,012	0,831	0,650

0,91 y separan con un área de 0,85). En la charla, igual, la receta final no daña el idioma sin portero (§4.15).

4.13. La receta más chica

Con el portero cuidando el texto que no nombra lo aprendido, cabía que sobraran piezas. Probamos dos caminos con predicciones escritas antes. El primero fue la receta mínima, que conserva sólo el sueño anclado, la higiene del optimizador y las autoparáfrasis. La memoria se sostuvo (1,00 a la octava semana en dos semillas), pero sin anclas el sueño daña mucho el inglés (+12,5 y +13,9 % sin portero), el portero no alcanza a esconderlo (+1,22 y +1,64 %, sobre el +0,5 % fijado) y la paráfrasis baja a 0,835, bajo el 0,85 pedido. Las anclas también ayudaban a generalizar. Soñar sólo el valor nuevo, sin autoparáfrasis, deja el inglés con portero en +0,44 % pero la paráfrasis cae a 0,640.

El segundo camino quitó a la receta con ancla del mundo una familia de piezas por vez (Tabla 13). La regla escrita antes elegía la receta más chica con memoria de al menos 0,95, paráfrasis de al menos 0,90 y el inglés y el español amplios con portero en a lo sumo +0,5 %. Las cuatro predicciones se cumplieron. Las anclas de texto protegen el idioma (sin ellas el inglés sube a +8,60 % y con portero queda en +0,99), las anclas del sujeto sostienen la generalización (sin ellas la paráfrasis cae a 0,690) y el ancla de 2.000 hechos del mundo sobraba (sin ella paráfrasis 0,960 e inglés con portero +0,28 %). Con una semilla era una prueba de selección. Antes de adoptarla escribimos que en las semillas 1 y 2 la receta sin el ancla del mundo tenía que cumplir lo mismo, y lo cumplió (memoria 1,00, paráfrasis 0,965 y 0,950, inglés con portero +0,29 y +0,26 %, español -0,21 y -0,11 %). Es la receta final. Sin portero, además, su inglés amplio a la octava semana (+1,11 a +1,49 %) no es peor que el de la receta con el ancla del mundo (+1,50 a +1,89 %), lo que coincide con lo que ya habían mostrado las variantes a cuatro semanas, que esa ancla no movía el nivel (§5).

Tabla 15. Lo que el usuario dijo en la charla, preguntado al otro día con frases que el sueño nunca vio. Tres personas por dos semanas, 291 preguntas sobre datos dichos. Tokens de la consulta, con el contexto que cada brazo agrega. Las recetas del sueño son la de la versión 1 (ancla KL 0,3 al modelo de la semana anterior), la intermedia (sueño anclado al origen con texto humano, §4.9), la del ancla del mundo (§4.10) y la final, que es la misma sin ella (§4.13).

	acierta	tokens por consulta
el modelo sin memoria	0,003	22
el mismo modelo con la charla entera en el contexto	0,062	1.446
el mismo modelo con la bitácora entera en el contexto	0,320	381
el mismo modelo con la línea justa de la bitácora	0,591	59
después del sueño, receta de la versión 1	0,887	22
después del sueño, receta intermedia	0,856	22
después del sueño, receta con ancla del mundo	0,914	22
después del sueño, receta final	0,918	22
sueño con ancla del mundo y línea justa de la bitácora	0,959	59
sueño final y línea justa de la bitácora	0,966	59

4.14. Un mismo sujeto con otra relación

De los 400 sujetos corregidos en la cadena real, 36 tienen en CounterFact otro registro con otra relación. A la cuarta semana el sueño los contamina. En las 93 preguntas de otra relación independiente, la fracción en que el valor verdadero le gana al valor corregido baja de 0,806 con el modelo base a 0,667, y el valor corregido aparece en la generación en 0,140 de los casos contra 0,011 del base. En los sujetos todavía no corregidos no pasa, de modo que el daño es del sujeto que se tocó. El portero por identidad no lo ve, porque deja pasar toda pregunta sobre un sujeto de la bitácora.

Probamos tres maneras de separar la relación corregida de las otras, las tres registradas antes de correr. La primera, sin entrenar nada, deja pasar la pregunta si el modelo original le daría al valor viejo una probabilidad comparable a la del enunciado guardado. Falló (área 0,75 contra 0,85 pedida) porque el lenguaje es ambiguo entre relaciones vecinas. «Leonhard Euler took up work in» admite «physics», el campo, tanto como «Berlin», el lugar. La segunda cambia el sueño en lugar de la puerta. Repasa, con el ancla, lo que el propio modelo original dice de cada sujeto corregido, sin el valor viejo. Con un banco de 100 sujetos que tienen otra relación, la fuga baja de 0,051 a 0,008, pero la corrección generaliza peor a sus paráfrasis (0,720 a 0,465) (Tabla 14). La tercera le agrega al sueño la corrección reescrita por el propio modelo con sus palabras, para enseñarle qué contextos *sí* son la misma relación. Juntas, las dos piezas dejan la fuga en 0,012 y la paráfrasis en 0,650, mejor que la suma de sus efectos por separado (0,047 y 0,555), aunque sin llegar a los umbrales escritos. En una cadena de cuatro sueños la protección no escala. Corta la fuga un 84 % en la primera semana y un 54 % en la cuarta (0,114 a 0,052), mientras la réplica guardada crece de 250 a 1.049 frases con la misma franja del lote.

Una observación lateral de esta cadena. Las reescrituras válidas que el modelo produce a pedido caen semana a semana (199, 162, 134 y 70 de 300 pedidas), con el mismo filtro. Cada sueño erosiona también la capacidad del asistente de seguir una instrucción, algo que ninguna perplejidad mide.

4.15. Una charla simulada de dos semanas

Todo lo anterior carga los hechos como frases de plantilla. El objetivo de este trabajo es otro, recordar lo que el usuario dijo. Generamos una charla de dos semanas por persona, con 38 atributos posibles (la ciudad, el trabajo, el nombre del perro, la pareja, una alergia). En la primera semana el usuario cuenta al pasar 18 datos propios y algunos de sus parientes, con hipótesis («si tuviera un gato le pondría Rocco») y negaciones («no tengo auto») que no son

Tabla 16. La charla por clase de pregunta, con las cuatro recetas (tres personas, dos semanas). El texto humano es el cambio de perplejidad a temperatura 1 contra el modelo base en la segunda semana, y para las dos últimas recetas también el calibrado (cada modelo a su temperatura óptima, cinco registros, seis corridas). Sin dato son las 60 preguntas sobre lo que nunca se dijo (atributos nunca mencionados, negaciones e hipótesis). G3a y G3b son los porteros que verifican la respuesta contra la bitácora.

	versión 1	intermedia	ancla del mundo	final
lo corregido, valor nuevo (n 42)	0,881	0,833	0,952	0,929
lo corregido, valor viejo	0,000	0,000	0,000	0,000
lo de la semana 1 que no cambió (n 66)	0,939	0,894	0,939	0,939
lo de un pariente (n 27)	0,704	0,778	0,778	0,741
hechos del mundo que el modelo base sabía	0,956	0,986	0,986	0,985
Wikipedia en inglés, T=1, semana 2	+44 y +45 %	-1,7 a -2,2 %	-1,1 a -1,4 %	-1,2 a -1,8 %
prosa en español, T=1, semana 2	+17 y +29 %	-2,4 a -2,9 %	-1,1 a -1,7 %	-1,1 a -1,9 %
cinco registros humanos, calibrado	no medido	no medido	-1,56 a +0,62 %	-2,05 a +0,35 %
sin dato, se abstiene sin portero	0,117	0,233	0,100	0,083
sin dato, contesta con el valor de otro dato dicho	15 de 60	17 de 60	31 de 60	33 de 60
sin dato, se abstiene con G3a (acierto con dato igual)	0,750	0,717	0,483	0,450
sin dato, se abstiene con G3b (acierto con dato)	0,917 (0,770)	0,883 (0,739)	0,650 (0,770)	0,650 (0,787)

datos, entre charla de relleno, en 26 a 29 turnos por semana. En la segunda corrige 7 datos («me mudé a Rosario»), agrega 8 y suma parientes. Cada fin de semana el modelo duerme con la receta en formato de chat y un escritor perfecto de la bitácora, y al otro día se le pregunta sin la charla en el contexto, con frases de pregunta disjuntas de las del sueño. Tres personas, las predicciones escritas antes de correr cada receta.

El sueño recuerda lo que se dijo mejor que tenerlo en el contexto. Con la receta con ancla del mundo el modelo que durmió acierta 0,914 de lo dicho, contra 0,062 del mismo modelo con la charla entera en el contexto, que cuesta 1.446 tokens por consulta, y 0,320 con la bitácora entera, que cuesta 381 (Tabla 15). Un modelo de este tamaño casi no sabe usar lo que se le pone en el contexto, y después del sueño lo sabe sin él. Ni siquiera la línea exacta de la bitácora, la que contesta la pregunta, alcanza sin el sueño (0,591), mientras que el sueño más esa línea llega a 0,959. Cuando el usuario se corrige gana el valor nuevo (0,952) y el viejo no aparece nunca en las cuatro recetas, lo de la primera semana que no cambió se sostiene (0,939) y se conservan 0,986 de los hechos del mundo que el modelo base sabía (Tabla 16).

El idioma. Con la receta de la versión 1 la charla repite lo que pasó en la cadena real. El primer sueño casi no toca el texto humano (+10,5 % en inglés, sin cambio en español) y el segundo lo daña (+44 y +45 % en inglés, +17 y +29 % en español), porque la réplica de esta charla también la escribía el modelo de cada semana. Con la receta intermedia y con la final la perplejidad sobre texto humano a temperatura 1 baja en las seis corridas, y calibrada la receta con ancla del mundo queda entre -1,56 y +0,62 % en los cinco registros que se calibran (la prosa en español sólo se mide a temperatura 1, donde baja entre 1,1 y 1,7 %). El criterio registrado antes de correrla pedía, en las seis celdas de persona y semana, a lo sumo +1,0 % de daño calibrado y un recuerdo (lo dicho, lo corregido y lo que no cambió) que no cayera más de 0,05 respecto de la receta intermedia. El daño calibrado máximo es +0,62 % y el recuerdo pasa de 0,864 a 0,928 en 264 preguntas, así que se cumple.

El costo, la abstención. Donde nunca se le dijo nada el modelo inventa, con las tres primeras recetas, y casi nunca contesta el valor de una hipótesis (2, 2 y 0 veces de 28). La receta con ancla del mundo inventa más. Se abstiene en 0,100 de las 60 preguntas sin dato, contra 0,233 de la intermedia, y en 31 de ellas contesta con el valor de *otro* dato que el usuario sí dijo, contra 17 de la intermedia y 15 de la versión 1 (a «¿a qué me dedico?» contesta la empresa, a «¿qué instrumento toco?» el deporte). Ya pasa en la primera semana (15 de 30 contra 7 de 30), y los

datos con autoparáfrasis se filtran apenas más que los que no tienen (0,15 contra 0,11), así que es un efecto global de la receta y no de cada dato. Las respuestas filtradas no son valores al azar. En 28 de las 31 el valor es de otro atributo del mismo usuario (a «¿en qué ciudad vivo?» contesta el barrio, a «¿a qué me dedico?» la empresa) y en 3 es del mismo atributo de otra persona (la empresa del mejor amigo). Con la receta intermedia eran 15 de 17 y 2 de 17, y siete de los pares de atributos se repiten en las dos recetas, de modo que la receta con ancla del mundo amplía una confusión que ya existía. Se concentra en las preguntas sobre algo que el usuario mencionó sin afirmarlo, las hipótesis, donde contesta con el valor de otro dato en 20 de 28 casos contra 11 de 28 con la intermedia. Es el mismo intercambio que en la cadena real, donde la paráfrasis sube y los vecinos bajan, y la misma pared que en el resto del trabajo, saber qué relación pregunta una frase.

Qué pieza lo causa. Lo preguntamos con una ablación preregistrada de la semana 1 (30 preguntas sin dato por brazo, tres personas, el recuerdo sobre 117 preguntas con dato). Quitar una sola pieza de la receta con ancla del mundo deja la filtración en 13, 12 y 14 de 30 (sin autoparáfrasis, sin prefijo ajeno y sin las anclas de CounterFact), contra 15 con la receta entera y 7 con la intermedia, diferencias bajo el umbral de 6 que fijamos antes de correr. Quitarlas las tres juntas la baja a 5 de 30, con un recuerdo de 0,855 (0,829 con la intermedia). La segunda tanda cambió lo que además distingue a la receta con ancla del mundo de la intermedia. Sin el piso para lo nuevo la filtración queda en 0 de 30, pero por un colapso y no por una mejora, porque el recuerdo es 0,000 y se abstiene en 29 de 30 (las anclas ocupan casi todo el lote y lo nuevo no se aprende), y con 300 párrafos de texto humano y sin la réplica en inglés queda en 14. Ninguna pieza sola explica la filtración, y las tres de la primera tanda juntas sí. La lectura que fijamos antes de correr es que la causa está en la combinación de las autoparáfrasis, el prefijo ajeno y las anclas de CounterFact, sin que el efecto sea aditivo (las bajas por separado, de 2, 3 y 1 preguntas, suman 6, y la conjunta es de 10). Con 30 preguntas por brazo y una corrida por persona sólo se detectan diferencias grandes, así que ni siquiera esa falta de aditividad se distingue del ruido.

La receta final en la charla. Al sacar el ancla del mundo repetimos la charla entera, tres personas por dos semanas, con una predicción escrita antes. Cada clase tenía que quedar a $\pm 0,05$ de la receta con ancla del mundo en la segunda semana, y si lo dicho, lo corregido o lo que no cambió caían más que eso, el ancla del mundo sí servía en la charla. Lo dicho quedó en 0,96 (0,96 con ancla del mundo), lo corregido en 0,93 (0,95) y lo que no cambió en 0,94 (0,94), así que la conclusión escrita antes se sostiene. La predicción, tomada al pie de la letra, falla, porque lo de los parientes en la segunda semana quedó por encima de la banda (0,89 contra 0,78, con 6 preguntas por persona). Sumando las dos semanas los parientes quedan en 0,741 contra 0,778, de modo que lo leemos como ruido de una clase chica en las dos direcciones. Con todo, el recuerdo es 0,918 contra 0,914, el valor viejo no aparece nunca, se conservan 0,985 de los hechos del mundo y el texto humano a temperatura 1 baja en las seis corridas (el peor registro de la segunda semana queda entre $-0,14$ y $-0,92$ %), y calibrado queda entre $-2,05$ y $+0,35$ % en los cinco registros de las seis celdas, con 27 de los 30 valores por debajo de la versión 0. El costo de abstención no cambia. Contesta con el valor de otro dato dicho en 33 de 60 preguntas sin dato (31 con la receta con ancla del mundo), 29 de ellas con otro atributo del mismo usuario, y el portero G3a lleva la abstención de 0,083 a 0,450 sin perder ningún dato dicho.

El portero que verifica. Tres porteros registrados antes de correr, que clasificaban la pregunta, no cumplieron lo escrito. Los rasgos que habíamos elegido para compararlos resultaron degenerados, porque en la capa 16 de Qwen3-0.6B el segundo token de cada frase tiene una norma de 390 a 508 contra unos 60 del resto, una activación masiva (Sun et al., 2024) que arrastra cualquier media, y dos preguntas sin relación entre sí dan un coseno de 0,975. Sin ese token, un

clasificador de qué se pregunta llega a 0,80, y con él como portero el modelo se abstiene en 0,96 de lo nunca dicho pero pierde 8 puntos de lo dicho. Lo que funcionó sin costo fue invertir los papeles. Los pesos ya saben de qué se les pregunta, así que el portero no clasifica la pregunta sino que verifica la respuesta, y si el valor que proponen los pesos no está en la bitácora contesta «no me lo contaste». El portero G3a, que pide que la respuesta nombre algún valor anotado, lleva la abstención de 0,117 a 0,750 con la receta de la versión 1 y de 0,233 a 0,717 con la intermedia, sin perder ningún dato dicho. Con la receta con ancla del mundo llega sólo a 0,483, porque deja pasar justamente la filtración, ya que el valor de otro dato sí está anotado. G3b, que además pide la plantilla del dato con su valor vigente, la frena (0,650) a costa de 14 puntos de lo dicho (de 0,914 a 0,770). El escritor, que en esta charla es perfecto, lo medimos aparte con el propio modelo congelado. Es conservador, no anota ninguna hipótesis ni negación, y recupera sólo 0,27 de los datos.

4.16. Predicciones cumplidas y falladas

Las predicciones que siguen se escribieron antes de correr cada experimento (§3.7).

- **Ablaciones.** Se cumplieron todas. La línea ingenua retiene menos de 0,60 la primera semana a la cuarta (0,03), la bitácora queda bajo 0,50 (0,24), la composición sobre 0,60 (1,00) y una habilidad bajo 0,70. Sin repaso, la primera semana queda bajo 0,80 a la cuarta (0,00) con la bitácora y las habilidades como en la receta.
- **Semillas 2 y 3.** Se cumplió que la bitácora supera 0,95 en la semilla 2 y que $\diamond$ supera 0,95 hasta la cuarta semana. Fallaron que todos los paquetes anteriores superen 0,85 en todas las semanas (la semilla 3 bajó a 0,83 y 0,81 en las semanas 6 y 7, por paquetes aprendidos bajo) y que la bitácora supere 0,95 en la semilla 3 (0,94 en la semana 7). La composición limpia creció como se predijo, pero más de lo previsto, a 0,51 y 0,61 contra un 0,4 esperado.
- **Semilla 1.** Se cumplieron que la primera semana termine sobre 0,85 (0,97), que ningún paquete caiga más de 0,10 debajo de su valor inicial y que la bitácora supere 0,85. Fallaron que cada semana aprenda lo suyo sobre 0,90 (0,78 en la semana 7), que la abstención por clase se mantenga sobre 0,85 (entidad inventada 0,70 a la octava), que las habilidades se mantengan sobre 0,90 (Nudo 0,44 en la primera semana, $\diamond$ con calculadora 0,88 a la octava) y que la composición quede bajo 0,15 a 400 hechos (0,34).
- **Sorteo balanceado.** La regla escrita pedía composición en la cuarta semana menor que 0,34. Dio 0,30, lo que cumple la letra por dos preguntas de cincuenta, dentro del ruido, y sumando las cuatro semanas el empate es exacto. Lo leemos como un resultado nulo.
- **MEMIT.** Se cumplieron que aprenda lo nuevo sobre 0,90 en las primeras cuatro semanas, que retenga la primera semana bajo 0,90 a la octava y que conteste un valor ante la composición sobre 0,60. La corrección por edición superó 0,90 en siete de ocho semanas (0,86 en la séptima). Falló que las habilidades se mantengan sobre 0,90 a la cuarta semana (Nudo 0,17). También falló la predicción de que un λ mayor retendría más a costa de editar menos, porque retuvo menos.
- **Cadena con hechos reales.** Se cumplieron que cada semana aprenda lo suyo (1,000 por probabilidad, 0,99 a 1,00 exacto), que con repaso la primera semana quede a la cuarta sobre 0,90 por probabilidad (0,99) y sobre 0,70 exacta (0,97), que la perplejidad suba menos de un 10 % (6 %, cumplida en una medida que resultó no ver el daño al idioma, §4.9) y que el portero devuelva los vecinos al nivel del base dejando pasar las correcciones (1,000). Fallaron que sin repaso la primera semana caiga bajo 0,60 por probabilidad (0,75, aunque la exacta cae a 0,17), que la paráfrasis de la semana supere 0,70 en todas (0,67 en

la segunda) y que el sueño solo pierda menos de 0,10 en los vecinos (pierde entre 0,33 y 0,40).

- **Portero.** Fallaron que el lector de la bitácora tal como está deje pasar 0,90 de los hechos (0,31 y 0,29), que la comparación sin entrenar llegue a un área de 0,90 (0,59 con medias y 0,81 token contra token) y que un lector entrenado con 3.000 entidades, sin buscar la entidad por identidad, baje la composición de 0,10, tanto con el archivo de entrenamiento al azar (0,46 y 0,39, área 0,62) como con mundos rotativos que devuelven los distractores de la misma entidad (0,13 a 0,25 en tres semillas, área 0,60 a 0,64). Se cumplieron que el portero baje la composición de 0,10 y que la bitácora ajena no deje pasar lo que no tiene, todas las predicciones del portero con la entidad por identidad (hechos sobre 0,90 con entidades vistas y nuevas, composición bajo 0,10, área sobre 0,95) y las de la calibración cruzada en las tres semillas (hechos sobre 0,95, composición bajo 0,05).
- **Tres semillas y ocho semanas con hechos reales.** Con repaso, la primera semana a la cuarta supera 0,90 y 0,70 en las semillas 1 y 2 (1,00 y 0,98, 1,00 y 0,99). Sin repaso, la exacta queda bajo 0,40 como se predijo (0,29 y 0,35), y la probabilidad otra vez no baja de 0,60 (0,74 en las dos). A la octava semana se cumplieron que la primera semana siga sobre 0,90 y 0,70 (0,98 y 0,92), que la perplejidad de la receta siga subiendo y pase el 10 % (256 %) y que con la réplica de la versión 0 quede bajo el 10 % sin perder retención (−10 %, 0,98 y 0,91). Con la réplica bilingüe se cumplió que el primer sueño quede bajo el +100 % en inglés y el +25 % en español (+72 y +6 %), y falló que con las dos correcciones la cuarta semana quede bajo el +100 % en inglés y el +30 % en español (+559 y +97 %), con la memoria intacta (1,00).
- **Sueño anclado al origen.** Una variante es la solución si a la cuarta semana cumple la memoria (lo de cada semana y lo corregido la primera sobre 0,95 por probabilidad y 0,90 exacto) y el idioma (texto humano bajo el +10 % en inglés y en español, sin subir más de 5 puntos desde la primera semana). Con peso 1 y con peso 10, esta última agregada después de ver el primer sueño, se cumplió la memoria y falló el idioma (+43,5 y +22,0 % en inglés) por los saltos del optimizador. Con peso 3, tasa en coseno y recorte, agregada después del salto de la segunda semana con peso 3, se cumplieron las dos (+5,9 % en inglés, −3,9 % en español, 3,8 puntos sobre la primera semana, memoria 1,00 y 1,00).
- **Otra relación del mismo sujeto.** Se cumplió que el sueño la contamine (la fuga de 0,011 a 0,140). Fallaron que la huella del valor viejo separe las relaciones con un área de 0,85 (0,75), que la réplica dirigida proteja sin perder más de 0,05 de paráfrasis (pierde 0,255), que réplica dirigida y autoparáfrasis juntas cumplan los dos umbrales (0,831 contra 0,847 de NS y 0,650 contra 0,670 de paráfrasis), que la protección llegue a un cuarto de la fuga en cuatro sueños (la mitad) y, por 0,005, que la paráfrasis supere 0,60 todas las semanas (0,595 en la primera). Se cumplió que la autoparáfrasis sola suba la paráfrasis sin proteger (0,810, fuga 0,090).
- **Charla simulada.** Se cumplieron que en la primera semana recuerde lo dicho sobre 0,75 (0,889) con el modelo sin memoria bajo 0,05 (0,009), que la corrección gane sobre 0,70 con el viejo bajo 0,15 (0,881 y 0,000), que lo anterior no caiga más de 0,10 (0,939), que los parientes no se confundan (0,667 y 0,083), que se conserven los hechos del mundo (0,954) y que el modelo que durmió no quede más de 0,15 debajo de la bitácora en el contexto (la supera por 0,47). La conservación del idioma se había dado por cumplida con la medida fija, y sobre texto humano falla en la segunda semana. Falló que afirme un valor bajo 0,30 donde nunca se dijo nada (0,40 y 0,50). Los tres porteros de la charla fallaron lo escrito (qué se pregunta 0,24 con los rasgos elegidos, 0,80 y 0,70 con dos rasgos corregidos después del diagnóstico, y la verificación estricta pierde 0,107 de lo dicho). La verificación del valor, sin costo, se informa como resultado descriptivo.

- **Receta con ancla del mundo en la cadena, cuatro semanas.** Los criterios D1 a D5, escritos antes de las corridas de confirmación, dieron lo siguiente. D1 se cumplió (memoria en tres semillas y cuatro semanas, lo de cada semana sobre 0,95 por probabilidad y 0,90 exacto, y lo retenido igual, con 0,99 y 0,98 como mínimos). D2 falló por poco. Los seis registros quedan bajo +0,5 % a temperatura 1, pero calibrados el inglés llega a +1,13 % en Wikipedia y a +1,59 y +1,35 % en los mensajes de usuarios en dos semillas, sobre el +1,0 % fijado, y la medida por token, hecha sólo en la semilla 0, dio +0,6 % contra el +0,5 % pedido. D3, la charla con la receta final, se cumplió (§4.15). D4 se cumplió a medias. El portero que verifica, medido sobre respuestas guardadas y no integrado en un sistema en línea, no pierde ningún dato dicho, pero con la receta final deja pasar la filtración. D5, el control de literatura del 1 de octubre, se cumplió y encontró un trabajo que ocupa la pérdida del sueño anclado (KLOD).
- **Charla con la receta con ancla del mundo.** El criterio, escrito antes de correrla, pedía a lo sumo +1,0 % de daño calibrado en las seis celdas de persona y semana y un recuerdo que no cayera más de 0,05 respecto de la receta intermedia. Dio +0,62 % como máximo y un recuerdo que pasó de 0,864 a 0,928, y se cumplió. No estaba escrito, y apareció, que la receta contestara con el valor de otro dato dicho en 31 de 60 preguntas sin dato.
- **Ablación de la filtración.** Una pieza la explicaría si al quitarla la filtración bajaba de 15 a 9 de 30 o menos con un recuerdo de al menos 0,833, y las diferencias de menos de 6 de 30 no se interpretaban. Quitar una sola pieza dejó 13, 12 y 14 de 30, quitar las tres juntas 5, y los cambios de la segunda tanda 0 (con el recuerdo en cero) y 14. Se cumplió la lectura cruzada escrita antes, que la causa está en la combinación de las tres piezas.
- **Ocho semanas con la receta con ancla del mundo.** Se cumplieron que lo corregido la primera semana siga sobre 0,95 por probabilidad y 0,90 exacto (1,00 en las tres semillas), que la paráfrasis siga sobre 0,90 y los vecinos sobre 0,55 (0,920 a 0,965 y 0,608 a 0,646) y que el criterio de +1,0 % en todos los registros siguiera fallando en inglés. Falló que el daño calibrado no creciera más de un punto entre las semanas 4 y 8 (en la Wikipedia en inglés estrecha creció entre 1,45 y 3,18 puntos, y en la medida amplia entre 0,67 y 1,25).
- **Ancla rotativa.** Para saber si el ancla de texto humano se gasta, a partir de la quinta semana la cambiamos cada semana por 600 párrafos nunca vistos. Falló que el crecimiento del inglés se frenara en las dos semillas (en la medida amplia, +2,00 y +1,71 % a la octava semana contra +1,86 y +1,50 % con el ancla fija). Se cumplieron que la memoria y la paráfrasis siguieran sobre 0,95 y 0,90 (1,00, y 0,925 a 0,965), que el español no pasara de +1,0 % (−1,21 a −1,26 %) y que con dosis 0,3 dañara menos que la rotativa simple (+1,30 %). El desgaste del ancla fija no explica el crecimiento.
- **Receta mínima.** Se cumplieron que la memoria de la primera semana siga sobre 0,95 en dos semillas (1,00) y que sin portero dañe más que la receta con ancla del mundo (+12,5 y +13,9 % contra +1,5 a +1,9 %). Fallaron que la paráfrasis quede sobre 0,85 (0,835 en las dos) y que con portero el inglés quede bajo +0,5 % (+1,22 y +1,64 %). La regla escrita antes decía que si esto último fallaba las anclas seguían haciendo falta. Soñar sólo el valor nuevo cumplió que la memoria se sostenga (1,00) y que la paráfrasis quede por debajo de la receta mínima (0,640), y que con portero el inglés quede bajo +0,5 % (+0,44 %).
- **Ablación de la receta en familias.** Se cumplieron las cuatro. La memoria se sostuvo en los tres brazos (1,00), sin anclas de texto la paráfrasis quedó sobre 0,90 (0,950) y el inglés con portero bajo +1,0 % (+0,99), sin anclas del sujeto la paráfrasis cayó bajo 0,88 (0,690) con el inglés bajo +0,5 % (+0,34), y sin el ancla del mundo la paráfrasis quedó sobre 0,92 (0,960) y el inglés bajo +0,5 % (+0,28). La réplica de la ganadora en las semillas 1 y 2, con la predicción escrita antes, también se cumplió (§4.13).

- **Charla con la receta final.** Se predijo que cada clase quedaría a $\pm 0,05$ de la receta con ancla del mundo en la segunda semana. Falló por arriba en los parientes (0,89 contra 0,78, 6 preguntas por persona) y se cumplió en las demás, incluidas las tres que decidían si el ancla del mundo hacía falta (§4.15). Se cumplió que el texto humano no pase de +1,0 % a temperatura 1 (el peor registro, -0,14 %).
- **AlphaEdit.** Se cumplió que aprenda lo nuevo sobre 0,90 las primeras cuatro semanas, que conteste un valor ante la composición sobre 0,60 y que retenga más que MEMIT. Falló que retenga la primera semana sobre 0,80 a la octava (0,10). Con la protección, se cumplieron que Nudo supere 0,85 en la segunda semana (1,00) y que la retención de hechos no cambie (0,61 contra 0,62).

5. Lo que nos pasó en el medio

Esta versión reemplaza a una que decía que el sueño no empeora el idioma, y eso no era cierto. Contamos cómo lo descubrimos y qué falló en el camino, porque casi todo fueron errores de medición que puede cometer cualquiera que evalúe memoria en pesos. No es un diario. Cada afirmación remite a la sección con su medida, y el Apéndice A lista todas las variantes que probamos y marca las que se agregaron después de ver un resultado.

Un termómetro que se parecía al paciente. La perplejidad sobre 120 frases generadas por el propio modelo dio +6 % a la cuarta semana. Sobre Wikipedia en inglés escrita por personas dio +1.270 %. La réplica con que el sueño se protege son frases del propio modelo, y esa medida premiaba justamente eso (§4.9).

Un termómetro que premiaba suavizar. Con texto humano de seis registros la perplejidad empezó a bajar, incluso en registros que ninguna ancla veía. Parecía una mejora y sospechamos. El modelo de base está sobreconfiado, y con su temperatura óptima, 1,10 a 1,12, su perplejidad baja entre 3,7 y 4,5 % sin aprender nada. Por eso comparamos cada modelo a su propia temperatura óptima (§4.10). Ese control podía fallar y no falló. Calibrada, la receta de la versión 1 da +112,6 % sobre Wikipedia en el primer sueño (a temperatura 1, +246 %), y el sueño anclado sin texto humano da +10,0 % a la cuarta semana (a temperatura 1 se veía como +5,9 %). Este control se agregó después de ver el primer resultado con ancla humana, no estaba en el registro original, y desde entonces informamos siempre el daño calibrado.

Tres causas, encontradas por intervención. Una réplica escrita por un modelo que ya se corrió, una réplica en un solo idioma frente a hechos en inglés, y un objetivo que enseñaba la frase entera de cada hecho en lugar de la respuesta. Cada una se aisló cambiando sólo ella (§4.9).

El daño que queda es del conocimiento del mundo. Con el ancla sobre texto humano el idioma quedó limpio y todavía había un resto en el inglés enciclopédico. Lo repartimos por token en 150 párrafos de Wikipedia, calibrados contra la versión 0. En la receta con ancla y autoparáfrasis a la cuarta semana, el 95 % del aumento de la pérdida estaba en los nombres propios (los tokens con mayúscula, el 15 % del texto), y el resto del texto quedaba neutro (-0,0006 nats por token contra +0,049 en los nombres propios). El daño que queda no es de la lengua. Es de saber qué persona, qué lugar o qué marca viene. Crece con cuánto entrena la receta en el dominio de los hechos, y los vecinos anclados lo bajaron un 40 %.

Lo que no funcionó.

- Poner el adaptador sólo en las MLP de las capas 3 a 8, donde edita MEMIT, protege el idioma en el primer sueño (+17,7 %) pero la paráfrasis cae a 0,49. La detuvimos.

- Dejar el sujeto de cada hecho fuera del ancla, para que pueda guardarse donde lo guardan los modelos, no ayudó (paráfrasis 0,56 contra 0,54 sin eso).
- Las autoparáfrasis solas suben la paráfrasis a 0,765 y bajan la especificidad de los vecinos de 0,50 a 0,42. Recién los vecinos anclados, con el piso para lo nuevo, la devolvieron a 0,58 con la paráfrasis en 0,93.
- Anclas, repaso y vecinos sin un piso para lo nuevo ocuparon el 80 % del lote en la segunda semana, y los hechos de esa semana se aprendieron a 0,76 por probabilidad y 0,66 exacto. El piso de 0,35 del lote lo arregló.
- Un sueño de mantenimiento, que repasa sin hechos nuevos, resta alrededor de 0,24 puntos de daño calibrado en Wikipedia (de +0,80 a +0,56 %), mientras que cada semana de aprendizaje suma alrededor de 0,35. Alternarlos frena la deriva y no la anula.
- El ancla de 2.000 hechos verdaderos del mundo no movió el nivel de las recetas que más generalizan. Calibradas en Wikipedia, con vecinos y sin ancla del mundo dieron +1,56 y +1,32 % en las semanas 2 y 3, y con ancla del mundo +1,48 y +1,47 %. No pudimos separar su efecto del ruido de la medida (unos 0,4 puntos con 40 textos) ni de la variación entre semillas (0,74 puntos).
- En la charla, quitar la práctica del «no sé» no mejoró el recuerdo (0,845 contra 0,856), y la abstención con el portero G3a fue parecida (0,733 contra 0,717). Nuestra hipótesis de que esa práctica causaba los «no me lo contaste» sobre datos que el modelo sí sabía no se sostuvo.

Un error nuestro. Al llevar a la charla las piezas de la receta final, la réplica en español quedó en cero. El generador de réplica descarta las frases cortas y devuelve cerca de la mitad de las que se piden (138 a 147 de 300 en español), y el código recortaba 150 del español antes de sumar la parte en inglés. La señal estaba impresa en el registro de la prueba local (“−2 en español”) y no la leímos a tiempo. Lo vimos al revisar el código, a pedido del autor, antes de leer ninguna métrica. Se descartaron las seis corridas enteras sin mirar sus resultados, se corrigieron ese error y otros cinco desvíos respecto de la receta de la cadena (el repaso con una autoparáfrasis de otro dato, los vecinos calculados sólo sobre los datos nuevos, anclas que incluían sujetos evaluados, la consigna de reescritura en inglés sobre frases en español y la falta de un filtro contra paráfrasis que nombran otro atributo) y se repitieron como una corrida nueva, con la enmienda del registro escrita antes de ver nada. Las corridas descartadas se conservan, y la lección es corta. Hay que leer los conteos que imprime la corrida antes de leer sus métricas.

Lo que no funcionó en la última semana.

- Cambiar el ancla de texto humano cada semana por párrafos nuevos no frenó el crecimiento del inglés entre las semanas 4 y 8 (§4.16). El desgaste del ancla fija no explica el crecimiento.
- Sacar todas las anclas y confiar en el portero dejó la memoria intacta pero el inglés con portero en +1,22 y +1,64 % y la paráfrasis en 0,835. El portero cuida el texto que no nombra lo aprendido, y un adaptador muy dañado igual se nota en el que sí lo nombra (§4.13).
- El portero por párrafo, que con 800 sujetos funciona, se prende en casi la mitad del inglés con la lista entera de CounterFact. La ventana de ocho tokens lo arregla en CounterFact, donde la respuesta va pegada al sujeto (§4.12).
- Un portero para la charla, por palabras o por similitud de frases, no llegó a la memoria completa sin prenderse casi siempre, porque el usuario habla de sí mismo sin nombrarse (§4.12).

Una tercera trampa, del mismo tipo que las otras dos. Los 40 párrafos con que medíamos la Wikipedia en inglés eran de un solo artículo, y ese artículo, sobre fútbol americano, se dañaba más que el promedio. La medida estrecha exageraba el crecimiento del daño entre las semanas 4 y 8 (§4.11). Las tres trampas tienen la misma forma. Un instrumento que no representa el texto sobre el que se quiere afirmar algo, sea porque lo escribió el modelo, porque premia suavizar o porque cubre un solo tema.

Generalizar o no tocar lo que se sabe. Lo que no pudimos resolver es el mismo intercambio en dos lugares. Con hechos falsos de CounterFact, generalizar la corrección a otras formas de preguntar baja la especificidad de los vecinos, de 0,786 a 0,58 a la cuarta semana con la receta final (de 0,700 a 0,60 a 0,62 a la octava), y el daño calibrado que queda es del orden de +1 %. La receta más simple, con ancla de texto humano y autoparáfrasis pero sin vecinos ni ancla del mundo (cfv), daba un daño en la misma banda (+0,80, +0,06 y +1,92 % calibrados en Wikipedia a la cuarta semana en tres semillas) con una paráfrasis de 0,725 a 0,765 y una especificidad de 0,52 a 0,56, así que la receta final la domina, ya que generaliza más al mismo daño. La variación entre semillas, de hasta 1,9 puntos, es del tamaño de la diferencia entre recetas, y por eso no afirmamos que alguna de las dos dañe menos. El portero del adaptador lo resuelve sólo en parte. Las preguntas sobre otros sujetos no lo prenden y vuelven a la especificidad del modelo original, pero lo que la receta enseña sobre el sujeto corregido sigue empujando a las otras relaciones de ese mismo sujeto (§4.14). En la charla, con datos personales, el mismo empuje aparece como la filtración (§4.15). El límite sigue abierto.

6. Discusión

El ancla no alcanza, el repaso sí. La ablación aísla el ingrediente que sostiene la memoria. El ancla KL sigue presente sin el repaso de hechos y no impide que la primera semana caiga a 0,00. El repaso usa los hechos originales de cada paquete anterior, reconstruidos a partir de su semilla. En un sistema desplegado esa fuente es la propia bitácora, que ya guarda cada hecho con su versión, así que los dos canales no son independientes. La bitácora es lo que el sueño repasa. Es el ingrediente que falta en O'Neill et al. (2026), donde sin repaso lo escrito en los pesos cae al 1 o al 46 % tras veinte escrituras. El costo de cada sueño es fijo, porque el repaso ocupa siempre el 15 % de las 48.000 muestras. Lo que crece con los hechos consolidados es la dilución. En la segunda semana cada hecho anterior se repasa unas 72 veces por sueño y en la octava unas 10, y aun así la retención no bajó. No sabemos todavía dónde está el punto en que el repaso deja de alcanzar.

Proteger lo que se sabe no protege lo que se sabe que no se sabe. El mismo resultado aparece dos veces por caminos distintos. En el sueño, repasar sólo los ejemplos positivos de una habilidad borra su borde de abstención, y en AlphaEdit, proteger sólo los ejemplos positivos deja caer los bordes mientras las habilidades se sostienen. La abstención parece vivir en parámetros que la práctica positiva no cubre, y cualquier método de protección tiene que incluirla de forma explícita.

La composición no se arregla en los pesos, se corta en el sistema. Crece con los hechos acumulados, se concentra en las combinaciones de partes conocidas y la intervención más directa sobre su correlato, la densidad de hechos por entidad, no la movió. Todo apunta a que el adaptador decide si contestar reconociendo la relación y la entidad por separado, y el modelo congelado muestra el mismo defecto, porque sin entrenar tampoco empareja la relación con la entidad. Un modelo que consolida no puede saber qué no se le dijo, pero la bitácora sí, y el portero la corta entera sin tocar los pesos. Eso le da a la bitácora un segundo papel. Además de corregir lo consolidado, es el índice de lo que se dijo. El mismo recorrido deja

una advertencia para cualquier lector aprendido de memoria. Entrenado sobre pocas entidades aprende a reconocerlas en lugar de compararlas. Y agregar entidades no alcanza por sí solo, porque con miles de entidades el archivo casi nunca trae otra entrada de la misma entidad (0,2 por pregunta contra 17 con 29 entidades) y desaparece la presión de distinguir la relación. Un lector entrenado así con 3.000 entidades separa hechos de composiciones con un área de 0,62, y devolviéndole los distractores con mundos rotativos de 40 entidades apenas llega a 0,60 a 0,67 en tres semillas. La salida no fue un lector más diverso sino separar las dos piezas. Informamos el resultado nulo del sorteo balanceado porque una intervención registrada que no confirma la hipótesis es parte del resultado.

Un termómetro que se parece al paciente. La medida del idioma de la versión 1 comparaba al modelo consigo mismo, con frases que él mismo había generado y que se parecen a la réplica con que se entrena. Dio bien mientras el idioma se rompía. La lección vale para cualquier consolidación con réplica generada. El control del idioma tiene que ser texto que el modelo no escribió, y la réplica tiene que venir de un modelo que no se corre, porque entrenar semana a semana sobre texto propio es la misma recursión que degrada a los modelos entrenados con su propia salida (Shumailov et al., 2024).

Un termómetro que premia suavizar. Con texto humano la perplejidad a temperatura 1 puede bajar sin que el modelo sepa más, porque el modelo de base está sobreconfiado y cualquier cosa que ablande su distribución mejora la medida. El control es compararlo con su propio mejor caso, la perplejidad calibrada, que además puede fallar y detecta el daño cuando lo hay (+112,6 % calibrado con la receta de la versión 1 en el primer sueño). La lección vale para cualquier trabajo que informe perplejidad después de adaptar un modelo posentrenado.

El daño que queda es de conocimiento y no de lengua. Con el ancla sobre texto humano, el 95 % del daño que quedaba estaba en los nombres propios. Eso cambia la pregunta, de cómo conservar el idioma a cómo enseñar un hecho falso sobre una entidad sin mover lo que el modelo sabe de las demás, que es lo que miden los vecinos. Los vecinos anclados y las reformulaciones de la versión 0 la mejoran, pero con hechos falsos de CounterFact no llegan a cero.

Lo único que tiene que cambiar. El daño al idioma cedió cuando dejamos de preguntar cómo proteger el idioma y preguntamos qué es lo único que el sueño tiene que cambiar. Es la respuesta. Todo lo demás, el enunciado de cada hecho y el texto de la réplica, tiene que salir del sueño como entró, y la referencia de ese «como entró» tiene que ser el modelo original, que no se mueve, y no el de la semana anterior. Es el principio del termómetro aplicado al objetivo. Queda la otra cara. Un sueño que no mueve nada más que la respuesta la aprende atada a la forma en que se le enseñó, y la generalización a otras formas de preguntar es lo que costaba. Las reformulaciones de la propia versión 0 la devuelven en buena parte, de 0,60 a 0,74 con el ancla de texto humano, y a 0,945 con los vecinos anclados, los hechos del mundo y el prefijo ajeno (§4.10).

La misma pared, tres veces. Tres problemas distintos terminaron en el mismo lugar. La huella del valor viejo no separa relaciones vecinas con hechos reales, la réplica dirigida protege las otras relaciones del sujeto al costo de generalizar peor la corregida, y en la charla lo que se le escapa a la verificación son valores dichos puestos en la ranura equivocada. En los tres casos falta lo mismo, saber qué relación pregunta una frase. El portero del mundo sintético lo resolvía con un lector entrenado sobre un conjunto cerrado de relaciones, y con lenguaje natural ni el modelo congelado ni un codificador de frases lo hacen sin entrenar. Lo que sí resultó útil fue invertir los papeles. Los pesos resuelven de qué se les pregunta mejor que cualquier portero que probamos, y la bitácora puede verificar lo que proponen en vez de adivinar la pregunta.

7. Limitaciones

- El experimento central usa un mundo sintético. Con hechos reales, la receta de la versión 1 tiene una cadena de ocho semanas con una semilla y de cuatro con tres (§4.8), la receta final ocho semanas con tres semillas (§4.11), y el portero real decide sólo por la identidad del sujeto.
- La receta final no lleva el daño al idioma a cero en inglés. Sin portero, calibrado sobre 150 artículos de Wikipedia, queda entre +1,11 y +1,49 % a la octava semana y crece despacio desde la cuarta (entre 0,29 y 0,56 puntos), así que no sabemos dónde se detiene. Con el portero del adaptador queda bajo +0,3 %, pero el portero por párrafo se diluye con listas grandes de sujetos y la ventana que lo arregla se probó sólo con CounterFact, donde la respuesta va pegada al sujeto (§4.12).
- La receta cuesta especificidad, porque los vecinos bajan de 0,700 a 0,600 a 0,618 a la octava semana sin portero. La receta con ancla del mundo se eligió entre dieciséis variantes sobre la semilla 0, y la final entre tres familias de una ablación también sobre la semilla 0, en los dos casos con las semillas 1 y 2 como confirmación escrita antes (apéndice). La variación del daño calibrado entre semillas es del tamaño de la diferencia entre recetas, así que no afirmamos cuál de dos recetas parecidas daña menos.
- El sueño contamina otras relaciones del sujeto corregido y ninguna de las protecciones probadas cumplió lo escrito (§4.14).
- En la charla, la receta final contesta con el valor de otro dato dicho donde nunca se dijo nada (33 de 60 preguntas contra 17 con la intermedia), el portero G3a no la frena, y su causa no se identificó. Las corridas son una por persona y semana, y las ablaciones tienen 30 preguntas por brazo, así que sólo detectan diferencias grandes (§4.15). El portero se midió sobre respuestas guardadas y no está integrado en un sistema en línea.
- La charla está generada con plantillas, el escritor del sueño es perfecto (el del propio modelo recupera 0,27), y el contexto le rinde poco a un modelo de este tamaño, así que con modelos grandes la comparación con el contexto puede cambiar (§4.15).
- Un solo modelo, de 0,6 mil millones de parámetros.
- Tres semillas para la receta y una para cada línea de base y para el RAG.
- La bitácora de la cadena de sueños se mide con un lector de una etapa anterior sobre un archivo de una entrada, y la exactitud de búsqueda de la bitácora viene de un experimento aparte sin sueños.
- La semilla 1 cambió la réplica de habilidades a partir de la cuarta semana (§3.3).
- La composición espuria llega a 0,51 y 0,61 a la octava semana en los pesos. El portero la corta en el sistema, pero con relaciones de un conjunto cerrado, las plantillas con que se entrenó el lector y la entidad de cada entrada marcada al escribirla.

8. Conclusión

Sobre un modelo de lenguaje congelado de 0,6 mil millones de parámetros, ocho sueños encadenados en tres semillas sostienen lo aprendido la primera semana entre 0,88 y 0,97, mientras que la edición secuencial de los mismos hechos con MEMIT y AlphaEdit retiene 0,01 y 0,10 y daña capacidades que la edición no tocó. Lo que lo explica es el repaso explícito de los hechos consolidados. Sin él, con todo lo demás igual, la primera semana se pierde en un sueño. La bitácora resuelve la corrección sin reentrenar con la misma calidad que un RAG entrenado y un costo de lectura que no crece con el historial. La composición espuria, que crece con los hechos

acumulados en los pesos, se corta afuera de ellos. Un portero que busca la entidad por identidad en la bitácora y deja que su lector decida la relación la baja a entre 0,000 y 0,007, dejando pasar casi todos los hechos dichos. Con hechos reales, sobre CounterFact, el repaso sostiene lo corregido la primera semana en 0,99 a 1,00 a la cuarta semana en tres semillas y en 0,98 a la octava, y el portero devuelve a los otros sujetos al nivel del modelo original. Medido sobre texto humano, la receta de la versión 1 daña el idioma y el daño crece. La receta final, un sueño anclado al origen que enseña sólo la respuesta y suma texto humano, reformulaciones de la versión 0, otros sujetos anclados y un piso para lo nuevo, sostiene la memoria en 1,00 a la octava semana en tres semillas con la paráfrasis en 0,950 a 0,965, y deja el daño calibrado sobre Wikipedia entre +1,1 y +1,5 % en inglés y por debajo de la versión 0 en español. Un portero del adaptador, que deja hablar al modelo original cuando el texto no nombra lo aprendido, lo lleva a menos de +0,3 %, y con muchos sujetos tiene que limitarse a los tokens que siguen a la mención. En una charla de dos semanas, el modelo que durmió recuerda al otro día 0,918 de lo que se le dijo, contra 0,062 con la charla entera en el contexto, y deja ganar siempre a la corrección. Lo que queda abierto es el límite de todo el trabajo, saber qué no se le dijo, y un resto de daño al inglés que crece despacio cuando no hay portero.

Declaraciones

Financiamiento. Ninguno. **Conflictos de interés.** El autor declara no tener conflictos de interés. **Contribuciones.** Autor único, responsable del diseño, la implementación, el análisis y la escritura. **Ética.** No aplica, ya que no hay sujetos humanos ni animales y los datos son sintéticos o de bancos públicos. **Uso de herramientas.** El código experimental y la redacción se hicieron con asistencia de un modelo de lenguaje (Claude, de Anthropic). El diseño, las decisiones y la responsabilidad sobre cada cifra son del autor. **Disponibilidad.** El código de cada experimento y los archivos JSON de los que sale cada número de este trabajo, con el historial de commits que fecha cada predicción, están disponibles a pedido del autor. **Nota del autor.** Este trabajo se hizo sin auspicio, supervisión ni financiamiento institucional, y se reporta a título personal.

A. Todas las variantes que probamos

La receta con ancla del mundo se eligió entre dieciséis variantes de la cadena sobre CounterFact, corridas sobre la semilla 0, más las referencias, y la final con la ablación de la Tabla 18. La Tabla 17 las lista todas, con la última semana que se corrió, y la Tabla 19 lista las corridas de la charla. La columna de origen distingue las variantes escritas en el registro de su tanda antes de correr (P) de las que se agregaron después de ver el resultado de otra de la misma línea (D). Las D también se escribieron antes de correrse, pero no estaban en el registro original. Las perplejidades son a temperatura 1 contra la versión 0 en Wikipedia en inglés. Los calibrados de las variantes elegidas están en la Tabla 11. Las variantes a ocho semanas, medidas con la Wikipedia amplia, están en la Tabla 18.

Referencias

- Zweiger, A., Pari, J., Guo, H., Akyürek, E., Kim, Y., y Agrawal, P. (2025). Self-Adapting Language Models. NeurIPS 2025. arXiv 2506.10943.
- Behrouz, A., Hashemi, F., Javanmard, A., y Mirrokni, V. (2026). Language Models Need Sleep. Learning to Self-Modify and Consolidate Memories. arXiv 2606.03979.
- Lee, S., McLeish, S., Goldstein, T., y Fanti, G. (2026). Do Language Models Need Sleep? Offline Recurrence for Improved Online Inference. arXiv 2605.26099.

Tabla 17. Las variantes de la cadena sobre hechos reales. Memoria es lo de la semana y la peor retención de las anteriores, paráfrasis y vecinos como en la Tabla 11. Semilla 0 salvo que se indique.

variante	qué cambia	corrida	memoria	paráfrasis	vecinos	Wikipedia	origen
cf	receta de la versión 1	4 sem., 3 sem.	1,00 y 0,96	0,825	0,456	+1,270	P
cfsin	cf sin repaso de hechos	4 sem., 3 sem.	1,00 y 0,15 a 0,32	0,64 a 0,92	0,28 a 0,43	no medido	P
cfr	réplica de la versión 0 en dos idiomas	4 sem.	1,00 y 0,98	0,780	0,494	+559	P
cfa	V1, sólo la respuesta y KL 1 al origen	4 sem.	1,00 y 0,99	0,835	0,488	+43,5	P
cfm	V2, como V1 con LoRA sólo en MLP 3 a 8	1 sem.	1,00	0,490	0,600	+17,7	P
cfk	V3, como V1 con KL 3 y tasa fija	2 sem.	0,91 y 0,62	0,735	0,518	+84,1	P
cfv	V4, como V1 con KL 10	4 sem.	1,00 y 0,97	0,775	0,462	+22,0	D
cfe	V5, KL 3, tasa en coseno y recorte	4 sem.	1,00 y 1,00	0,600	0,482	+5,9	D
cfg	V5 más autoparáfrasis	1 sem.	1,00	0,765	0,420	+1,6	P
cfh	cfg más prefijo ajeno	1 sem.	1,00	0,895	0,358	+4,8	P
cfs	V5 más sujeto libre	1 sem.	1,00	0,560	0,482	+2,6	P
cfu	V5 más 300 párrafos de texto humano	4 sem.	1,00 y 1,00	0,505	0,502	-1,6	P
cfv	cfu más autoparáfrasis	4 sem., 3 sem.	1,00 y 1,00	0,725 a 0,765	0,516 a 0,558	-1,2 a -0,6	P
cfw	cfv más prefijo ajeno	1 sem.	1,00	0,885	0,400	+0,4	D
cfz	600 párrafos y franja 0,3, autoparáfrasis y prefijo	1 sem.	1,00	0,925	0,374	-0,2	D
cfy	cfv con 600 párrafos y franja 0,3	4 sem.	0,99 y 1,00	0,720	0,538	-1,1	D
cfn	cfz más vecinos anclados y piso 0,35	3 sem.	1,00 y 1,00	0,915	0,600	-0,5	D
cfv4m	sueño de mantenimiento sobre cfv	1 sueño	1,00 y 1,00	0,740	0,536	-1,4	D
cfq	receta con ancla del mundo, cfn más 2.000 hechos del mundo	4 sem., 3 sem.	1,00 y 0,99 a 1,00	0,945 a 0,950	0,606 a 0,638	-0,9 a -0,2	D

Tabla 18. Las variantes a ocho semanas. Memoria es la probabilidad con que gana lo corregido la primera semana. Inglés es la Wikipedia amplia calibrada sin portero y con el portero del adaptador. Todas se escribieron antes de correrse. Origen como en la Tabla 17.

variante	qué cambia	semillas	memoria	paráfrasis	vecinos	inglés, sin y con portero	origen
cfq	receta con ancla del mundo	0, 1 y 2	1,00	0,920 a 0,965	0,608 a 0,646	+1,50 a +1,89 y +0,20 a +0,36	P
cqr	cfq con el ancla humana rotativa desde la semana 5	0 y 1	1,00	0,925 y 0,950	0,606 y 0,638	+2,00 y +1,71, sin portero	D
cqd	cqr con dosis 0,3 del ancla	0	1,00	0,965	0,594	+1,30, sin portero	D
cmn	receta mínima, sin anclas	0 y 1	1,00	0,835	0,460 y 0,476	+12,5 y +13,9, y +1,22 y +1,64	D
cpu	cmn sin autoparáfrasis	0	1,00	0,640	0,500	+8,33 y +0,44	D
cts	cfq sin anclas de texto	0	1,00	0,950	0,626	+8,60 y +0,99	D
ctt	cfq sin anclas del sujeto	0	1,00	0,690	0,608	+1,02 y +0,34	D
csn	receta final, cfq sin el ancla del mundo	0, 1 y 2	1,00	0,950 a 0,965	0,600 a 0,618	+1,11 a +1,49 y +0,26 a +0,29	D

Wang, Z., Gupta, A., Dong, Z., y MacLellan, C. J. (2026). Self-Consolidating Language Models. Continual Knowledge Incorporation from Context. arXiv 2605.07076.

Meng, K., Sen Sharma, A., Andonian, A., Belinkov, Y., y Bau, D. (2023). Mass-Editing Memory in a Transformer. ICLR 2023. arXiv 2210.07229.

Fang, J., Jiang, H., Wang, K., Ma, Y., et al. (2025). AlphaEdit. Null-Space Constrained Knowledge Editing for Language Models. ICLR 2025. arXiv 2410.02355.

Wang, P., Li, Z., Zhang, N., Xu, Z., et al. (2024). WISE. Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models. NeurIPS 2024. arXiv 2405.14768.

Das, P., Chaudhury, S., Nelson, E., Melnyk, I., et al. (2024). Larimar. Large Language Models with Episodic Memory Control. ICML 2024. arXiv 2403.11901.

Mitchell, E., Lin, C., Bosselut, A., Manning, C. D., y Finn, C. (2022). Memory-Based Model Editing at Scale. ICML 2022. arXiv 2206.06520.

Modarressi, A., Köksal, A., Imani, A., Fayyaz, M., y Schütze, H. (2024). MemLLM. Finetuning LLMs to Use An Explicit Read-Write Memory. arXiv 2404.11672.

Khattab, O., y Zaharia, M. (2020). ColBERT. Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. SIGIR 2020. arXiv 2004.12832.

Teyler, T. J., y DiScenna, P. (1986). The hippocampal memory indexing theory. Behavioral Neuroscience, 100(2), 147 a 154. doi.org/10.1037/0735-7044.100.2.147.

O'Neill, C., Kirkby, M., Liu, J., y Psenka, M. (2026). Can a Language Model Learn Facts Continually in Its Weights? arXiv 2607.11020.

Tabla 19. Las corridas de la charla, tres personas por dos semanas salvo que se indique. Recuerdo es la exactitud sobre los datos dichos (n 291). En las 60 preguntas sin dato, otro dato es la cantidad que se contesta con el valor de otro dato dicho y abstención la cantidad de «no me lo contaste». Sin portero.

corrida	qué cambia	recuerdo	otro dato	abstención	origen
conv	receta de la versión 1 (ancla KL 0,3 a la semana anterior)	0,887	15	7	P
conv5	sueño anclado al origen con 300 párrafos de texto humano	0,856	17	14	P
conv7	conv5 sin la práctica del «no sé»	0,845	16	0	P
conv8	conv9 con el error de réplica (§5)	descartada sin leer sus métricas			
conv9	receta con ancla del mundo	0,914	31	6	P
convcsms	receta final, conv9 sin el ancla del mundo	0,918	33	5	D
semana 1, tres personas (recuerdo sobre 117, 30 preguntas sin dato)					
conv5	referencia, receta intermedia	0,829	7	9	
conv9	referencia, receta con ancla del mundo	0,889	15	4	
conv10a	conv9 sin autoparáfrasis	0,872	13	7	P
conv10b	conv9 sin prefijo ajeno	0,880	12	4	P
conv10c	conv9 sin anclas de CounterFact	0,932	14	5	P
conv10d	conv9 sin el piso para lo nuevo	0,000	0	29	D
conv10e	conv9 con 300 párrafos y sin réplica en inglés	0,880	14	3	D
conv10f	conv9 sin las tres piezas de la primera tanda	0,855	5	9	D

Hartvigsen, T., Sankaranarayanan, S., Palangi, H., Kim, Y., y Ghassemi, M. (2023). Aging with GRACE. Lifelong Model Editing with Discrete Key-Value Adaptors. NeurIPS 2023. arXiv 2211.11031.

Zhang, B., y Huang, Y. (2026). When to Write and When to Suppress. Route-Specialized Dual Adaptors for Memory-Assisted Knowledge Editing. arXiv 2606.14668.

Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., y Herzig, J. (2024). Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? EMNLP 2024. arXiv 2405.05904.

Meng, K., Bau, D., Andonian, A., y Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. NeurIPS 2022. arXiv 2202.05262.

Levy, O., Seo, M., Choi, E., y Zettlemoyer, L. (2017). Zero-Shot Relation Extraction via Reading Comprehension. CoNLL 2017. arXiv 1706.04115.

De Cao, N., Aziz, W., y Titov, I. (2021). Editing Factual Knowledge in Language Models. EMNLP 2021. arXiv 2104.08164.

Yang, A., et al. (2025). Qwen3 Technical Report. arXiv 2505.09388.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., y Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631, 755 a 759. doi.org/10.1038/s41586-024-07566-y.

Sun, M., Chen, X., Kolter, J. Z., y Liu, Z. (2024). Massive Activations in Large Language Models. arXiv 2402.17762.

Rajpurkar, P., Zhang, J., Lopyrev, K., y Liang, P. (2016). SQuAD. 100,000+ Questions for Machine Comprehension of Text. EMNLP 2016, 2383 a 2392. doi.org/10.18653/v1/D16-1264.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., et al. (2022). LoRA. Low-Rank Adaptation of Large Language Models. ICLR 2022. arXiv 2106.09685.

Efraimidis, P. S., y Spirakis, P. G. (2006). Weighted Random Sampling with a Reservoir. Information Processing Letters, 97(5), 181 a 185.

- Speranza, M. (2026). La indexación geométrica no mejora la recuperación de memoria versionada. Tres resultados negativos preregistrados y un modo de fallo total silencioso. Research Square. doi.org/10.21203/rs.3.rs-10669947/v1.
- Speranza, M. (2026). Dónde vive la abstención. Comparación preregistrada de cuatro interfaces de abstención en un modelo de lenguaje chico con memoria versionada. Research Square. doi.org/10.21203/rs.3.rs-10839567/v1.
- Speranza, M. (2026). La similitud encuentra el hecho, no la versión. Un sello de orden co-entrenado en la clave resuelve el conflicto de versiones en una memoria persistente dentro del modelo. Research Square. doi.org/10.21203/rs.3.rs-10896018/v1.
- Speranza, M. (2026). La consulta no puede ver la pregunta. El alcance de una convolución corta decide qué parte de una consulta condiciona la recuperación. Research Square. doi.org/10.21203/rs.3.rs-10934426/v1.
- Allen-Zhu, Z., y Li, Y. (2024). Physics of Language Models. Part 3.1, Knowledge Storage and Extraction. ICML 2024. arXiv 2309.14316.
- Jeong, H., Kim, G., y Kang, S. (2026). KLOD. Locality-Preserving Knowledge Editing via Non-Target Distribution Preservation. arXiv 2608.27839.
- Zhu, L., Lin, Y., Wang, Y., Liu, Y., Li, H., y Chen, Y. (2026). Will My Assistant Remember My Allergy? What Personal LLM Assistants Forget When Conversation Memory Is Compressed. arXiv 2609.05767.
- Hwang, E., Mitra, K., Zhang, D., Kim, H., y Hruschka, E. (2026). Hypotheses-Guided Self Distillation for Continual Personalization. arXiv 2609.00251.
- Xie, Y. (2026). Learning to Forget. Sleep-Inspired Memory Consolidation for Resolving Proactive Interference in Large Language Models. arXiv 2603.14517.
- Dennis, S., Shabahang, K., Guo, H., y Patil, R. (2026). Beyond Inference-Only Deployment. Comparing Weight-Based Consolidation Against Cascading Compaction. arXiv 2605.24657.
- Li, B. (2026). User as Engram. Internalizing Per-User Memory as Local Parametric Edits. arXiv 2606.19172.
- Choi, H. (2026). Beyond Endpoint Scores. Time- and Capacity-Conditioned Evaluation of Continual Knowledge Updating. arXiv 2609.03900.
- Zheng, C., Li, L., Dong, Q., Fan, Y., Wu, Z., Xu, J., y Chang, B. (2023). Can We Edit Factual Knowledge by In-Context Learning? EMNLP 2023. arXiv 2305.12740.
- Padmanabhan, S., Gul, M. O., y Goyal, T. (2026). Updating Parametric Knowledge with Context Distillation Retains Post-Training Capabilities. arXiv 2602.16093.