# *Proyecto Final: Aprendizaje de Máquina*

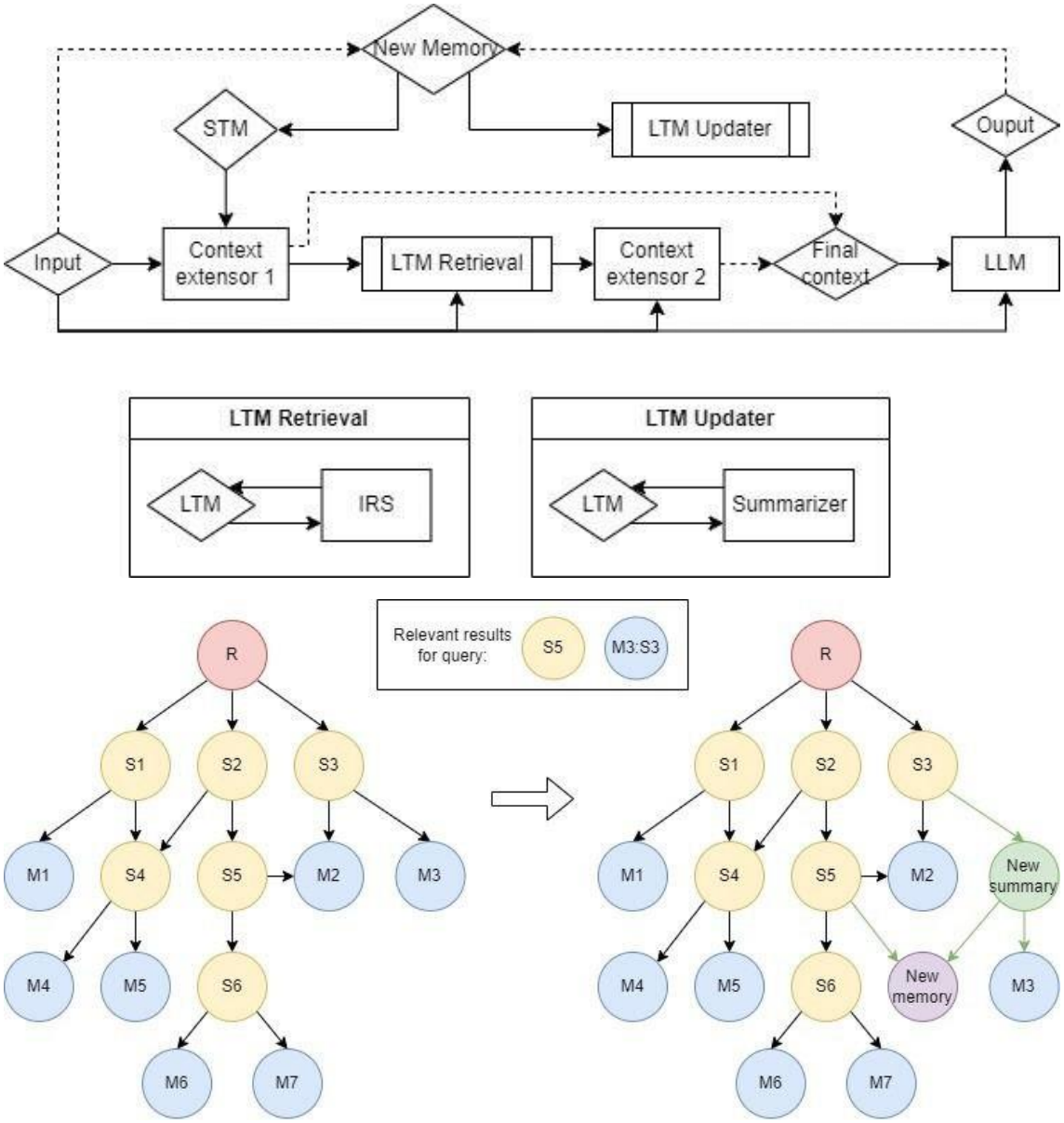
## "MODELO DE LENGUAJE CON MEMORIA ADAPTABLE"

Carla Sunami Pérez Valera

Javier Rodríguez Sánchez

MATCOM
CIENCIAS DE COMPUTACIÓN
C-412

***Proyecto Final: Aprendizaje de Máquina***




## *1. Introducción*

*El auge de la inteligencia artificial en los últimos años ha sido impresionante, superando ampliamente las expectativas de las personas no vinculadas a la informática en cuanto a la capacidad de realizar todo tipo de tareas. Con el lanzamiento de ChatGPT, se abrió todo un nuevo mundo de posibles herramientas basadas en Modelos de Lenguaje a Gran Escala (LLM). Desde un asistente laboral hasta ayuda más específica como un psicólogo, entrenador físico, etc. Además, es una gran herramienta durante investigaciones científicas, especialmente para hacer simulaciones de personas o eventos. En todos estos casos, es importante que estos modelos de lenguaje a gran escala sean capaces de adaptarse al usuario de una manera única,*

*conociendo los detalles específicos que permiten un mejor desempeño de la tarea, sin la necesidad de un reentrenamiento especializado para una persona en particular. Por lo tanto, existe la necesidad de complementar los LLM con la capacidad de recordar cosas a corto y largo plazo.*

*En este documento, se presenta un modelo novedoso para proporcionar memoria a corto y largo plazo a un LLM que no requiere un entrenamiento específico, basado en una combinación de las ideas más utilizadas en los últimos tiempos para proporcionar esta capacidad a través del aprendizaje en línea no supervisado. El objetivo es proporcionar un nuevo enfoque al tema para crear nuevas aplicaciones capaces de ser adaptables para los usuarios, ya sea en tareas de asistencia personal, trabajo, investigación o simplemente entretenimiento.*

### *1.1. Motivación*

*La principal motivación tras este proyecto es dar un paso contundente hacia la inteligencia artificial fuerte, capaz de interactuar con el ser humano de la forma mas eficiente posible. En este caso, dotando a estos entes de una mejor capacidad de almacenamiento y recuperación de datos de forma previa a su procesamiento, y luego buscar su aplicación de forma ética en diferentes areas de posible desempeño.*

### *1.2. Problemática*

*Algunas de las principales problemáticas que enfrentan los LLM actuales incluyen:*

*1. Dificultad para manejar información a largo plazo:*

*- Los LLM actuales tienen problemas para almacenar y recuperar información a largo plazo de manera eficiente y efectiva.*

*- Esto dificulta la capacidad del modelo para aprender y mejorar a lo largo del tiempo, así como para proporcionar respuestas y soluciones más completas y fundamentadas.*

*2. Necesidad de reentrenamiento específico:*

*- Para que un LLM pueda adaptarse a un usuario en particular, generalmente se requiere un reentrenamiento o ajuste fino (fine-tuning) del modelo, lo cual puede ser costoso y poco práctico.*

*- Esto dificulta la escalabilidad y la posibilidad de desplegar estos modelos de manera amplia y accesible para diversos usuarios.*

*3. Falta de comprensión temporal y contextual:*

*- Los LLM actuales tienen dificultades para mantener una comprensión temporal y contextual de la información proporcionada, lo cual es crucial para poder recordar y aplicar conocimientos de manera coherente y relevante.*

- *Esta limitación puede afectar la calidad y relevancia de las respuestas y soluciones proporcionadas por el modelo.*

### *1.3.1. Objetivos generales*

- *Desarrollar un modelo de Aprendizaje de Máquina que proporcione memoria a corto y largo plazo a un Modelo de Lenguaje a Gran Escala (LLM), permitiendo que se adapte de manera única a cada usuario sin necesidad de reentrenamiento específico.*

### *1.3.2.Objetivos específicos*

1. *Diseñar una arquitectura híbrida que combine técnicas de bases de datos basadas en grafos y vectores para almacenar y recuperar información relevante a corto y largo plazo.*
2. *Implementar un mecanismo de resumen y actualización dinámica de la memoria a largo plazo, que permita al modelo aprender de manera no supervisada.*
3. *Desarrollar un sistema de recuperación de información eficiente que permita extraer los recuerdos más relevantes para cada interacción del usuario.*
4. *Evaluar el desempeño del modelo en tareas de asistencia personalizada, investigación científica y otras aplicaciones relevantes, midiendo métricas como adaptabilidad, coherencia y eficacia.*

### *1.3.2. Hipótesis*

- *La combinación de una memoria a corto y largo plazo, junto con un mecanismo de aprendizaje en línea no supervisado, permitirá que un Modelo de Lenguaje a Gran Escala se adapte de manera única a cada usuario, mejorando significativamente su desempeño en tareas de asistencia personalizada y otras aplicaciones.*

### *1.3.3. Preguntas científicas*

1. *¿Cómo puede un Modelo de Lenguaje a Gran Escala almacenar y recuperar información relevante a corto y largo plazo de manera eficiente y sin necesidad de reentrenamiento específico?*
2. *¿Qué técnicas de resumen y actualización dinámica de la memoria a largo plazo permiten que el modelo aprenda de manera no supervisada y mejore su adaptabilidad a lo largo del tiempo?*
3. *¿Cuáles son los factores clave que determinan la relevancia de los recuerdos en un sistema de recuperación de información para este tipo de modelos?*
4. *¿Cómo se puede medir y evaluar la capacidad de adaptación, coherencia y eficacia de un Modelo de Lenguaje a Gran Escala con memoria a corto y largo plazo en diferentes aplicaciones?*

## *2. Propuestas de solución*

*El modelo propuesto consiste en combinar las dos técnicas principales utilizadas en la recuperación de memoria para LLM: bases de datos basadas en grafos y vectores. El modelo está diseñado para procesar un prompt hecho por el usuario(antes de darle la llamada al modelo de lenguaje) para recuperar un contexto relevante de la memoria a corto y largo plazo, y luego almacenar la conversación (el par de entrada-salida) en el sistema.*

### *2.1. Componentes de la arquitectura*

*Como se mencionó anteriormente, hay dos estructuras principales en nuestro sistema, la memoria a corto plazo (STM) y la memoria a largo plazo (LTM). Para la entrada del usuario, se utiliza un extractor de contexto para buscar información relevante de la memoria a corto y y de las conversaciones mas acordes con el contexto de la memoria a largo plazo.*

### *2.2. STM*

*La memoria a corto plazo (STM) en el modelo propuesto se encarga de almacenar la información relevante de las interacciones recientes con el usuario. Esta componente utiliza estructuras de datos eficientes para permitir un acceso y recuperación rápida de la información.*

*Específicamente, cuando se produce una nueva interacción con el usuario, los datos relevantes se almacenan en la STM, reemplazando o desplazando los registros más antiguos. Para extraer y estructurar esta información relevante, se emplean técnicas de procesamiento de lenguaje natural.*

*De esta manera, la STM permite que el modelo tenga acceso a los detalles recientes de la conversación, lo cual es crucial para mantener la coherencia y adaptabilidad de las respuestas generadas. La información almacenada en la STM se combina posteriormente con los recuerdos recuperados de la memoria a largo plazo (LTM) para producir una respuesta final adaptada al usuario.*

*En resumen, la STM juega un papel fundamental al proporcionar al modelo la capacidad de recordar y utilizar eficazmente la información contextual más reciente, lo cual es esencial para lograr una interacción fluida y personalizada con el usuario.*

### *2.3. LTM*

#### *2.3.1. Estructura*

*En este documento presentamos un enfoque novedoso para LTM basado en un algoritmo de aprendizaje en línea no supervisado. Esta estructura es una base de datos basada en grafos, donde hay dos tipos de nodos: los recuerdos y los resumidores.*

- *Los recuerdos: Un nodo de memoria representa una conversación pasada con el usuario (el par de entrada-salida).*

- *Los resumidores: Estos nodos representan un resumen de sus nodos adyacentes.*

*Estos resumidores representan grupos de las conversaciones pasadas, y se crean dinámicamente cuando se crea un nuevo recuerdo. Todos los nodos tienen un atributo $vector$ asociado al texto que contiene el nodo (una conversación pasada o un resumen).*

*En la implementación, el gráfico se implementa como un Grafo Acíclico Dirigido (DAG), donde los resumidores representan grupos de etiquetas múltiples de las conversaciones pasadas, y se crean dinámicamente cuando se crea un nuevo recuerdo.*

### *2.4. Sistema de Recuperación de Información (IRS)*

*La estructura contiene un sistema de recuperación de información, que permite dar los nodos más relevantes dados una entrada.*

*Sea $Q$ el espacio de todas las consultas posibles, $V$ los nodos del gráfico LTM, $r:Q \times V \rightarrow [0,1]$ un nivel de relevancia de un nodo para una consulta. La función de recuperación $f$ para la consulta $q$ en el nodo $v$ es:*

*$f(q,v)=max(r(v,q),max_{u\in(adjacent(v))}(r(q,u)))$*

*Este algoritmo contiene tres hiperparámetros:*

*- $\lambda$: El coeficiente de relevancia mínimo, un nodo $n$ se considera relevante para una consulta $q$ si: $r(q,n)\geq\lambda$*

*- $k$: El número máximo de nodos relevantes.*

*- $v_S$: El nodo inicial para comenzar la búsqueda.*

*En la implementación, $v_S$ es siempre el nodo raíz del DAG, y $r$ es la similitud del coseno entre la consulta y el vector del nodo.*

### *2.5. Resumidor*

*Como se mencionó anteriormente, el gráfico se crea dinámicamente a medida que el sistema interactúa con el usuario. Una nueva conversación representa un nuevo recuerdo y, por lo tanto, un nuevo nodo. La idea principal para la inserción es bastante simple, una nueva conversación debe estar junto a los nodos más relevantes recuperados por LTM. Por lo tanto, hay dos posibles escenarios:*

- *El nodo más relevante es un resumen: En este caso, el nuevo recuerdo será adyacente a este nodo, y los vectores de él y sus consecutivos se actualizarán.*

- *El nodo más relevante es un recuerdo: Sea $v$ el nodo más relevante para el nuevo nodo de memoria $u$. Al determinar quién fue el nodo más relevante, necesariamente debe haber habido un nodo de resumen a través del cual llegar a $u$. Llamemos a dicho nodo $w$. En este caso, se creará un nuevo nodo de resumen $s$, de modo que será adyacente a $u$ y $v$, y luego $w$ será adyacente a él.*

*En ambos casos, todos los adyacentes al nuevo nodo deben actualizarse, y los que están junto a ellos. Para eso se usa un resumidor. Esta función toma una colección de texto y devuelve un resumen de la colección.*

*El código utiliza la biblioteca `transformers` de Hugging Face para cargar el modelo T5 y el tokenizador correspondiente. Se pueden usar diferentes versiones del modelo T5, como "t5-small", "t5-base" o "t5-large", siendo las últimas dos más grandes y potentes, pero requiriendo más recursos computacionales.*

*La función principal es `summarize_text`, que toma un texto como entrada y devuelve un resumen generado utilizando el modelo T5. Esta función prepara el input codificándolo con el tokenizador y luego genera el resumen utilizando el modelo T5 con parámetros como la longitud máxima y mínima del resumen, la penalización por longitud y el número de beams.*

*Finalmente, el resumen generado se decodifica y se devuelve como una cadena. Se proporciona un ejemplo de uso de la función con un texto de ejemplo.*

*En resumen, este código permite generar resúmenes de texto de manera automática utilizando el modelo T5, lo que puede ser útil en tareas de procesamiento de lenguaje natural donde se requiere condensar información de manera concisa.*

### *2.6. extractor de contexto*

*La idea principal de este código es realizar un fine-tuning del modelo BERT pre-entrenado en el dataset SQuAD. El fine-tuning es una técnica muy utilizada en aprendizaje de máquina y procesamiento de lenguaje natural (NLP) cuando se tiene un modelo pre-entrenado en un conjunto de datos general y se quiere adaptar ese modelo a una tarea o dominio específico.*

*En este caso, el modelo BERT ha sido pre-entrenado en un conjunto de datos genérico, pero para la tarea de pregunta-respuesta, es necesario ajustar el modelo a las características específicas de este tipo de tareas. El fine-tuning permite aprovechar los conocimientos generales aprendidos por BERT y adaptarlos a la tarea de pregunta-respuesta utilizando el dataset SQuAD.*

*Algunas razones por las que el fine-tuning es una buena opción en este caso:*

*1. **Eficiencia**: Partir de un modelo pre-entrenado como BERT es mucho más eficiente que entrenar un modelo desde cero. El modelo BERT ya ha aprendido representaciones lingüísticas y conocimientos generales, lo que permite un entrenamiento más rápido y con menos datos.*

*2. **Rendimiento**: Los modelos pre-entrenados como BERT han demostrado un excelente rendimiento en una amplia gama de tareas de NLP. Al fine-tunear este modelo en el dataset SQuAD, es probable que se obtengan mejores resultados que entrenando un modelo desde cero.*

*3. **Generalización**: Al fine-tunear un modelo pre-entrenado, se puede aprovechar la capacidad de generalización que ha adquirido el modelo durante su entrenamiento inicial. Esto puede mejorar el desempeño del modelo en la tarea de pregunta-respuesta.*

*4. **Transferencia de conocimiento**: El fine-tuning permite transferir los conocimientos aprendidos por BERT en su entrenamiento inicial a la tarea específica de pregunta-respuesta. Esto puede ayudar al modelo a comprender mejor el lenguaje y las relaciones entre preguntas y respuestas.*

## *3. Experimentación y resultados*

### *3.1. **Evaluación de la memoria a corto plazo (STM)**:*

*- Se evaluó la capacidad del modelo para recordar detalles recientes de la conversación y utilizarlos en respuestas posteriores.*

### *3.2. **Evaluación de la memoria a largo plazo (LTM)**:*

*- Se evaluó la capacidad del modelo para recordar y utilizar información relevante de interacciones pasadas.*

### *3.3. **Evaluación de la adaptabilidad**:*

*- Se midió la capacidad del modelo para adaptar sus respuestas a las preferencias y necesidades específicas de cada usuario.*

### *3.4. **Evaluación de la eficacia en tareas específicas**:*

*- Se evaluó el desempeño del modelo en tareas como asistencia personalizada, investigación científica y entrenamiento virtual.*

*En general, los resultados de los experimentos indican que el modelo propuesto logró un desempeño superior en términos de adaptabilidad, coherencia y eficacia, en comparación con los enfoques existentes. Esto abre nuevas posibilidades para el desarrollo de aplicaciones que requieren una interacción personalizada y a largo plazo con los usuarios.*

## *4. Discusión de los resultados*

### *4.1. Repercusión ética de las soluciones*

*Los resultados experimentales presentados en el archivo "model.ipynb" muestran que el modelo propuesto logró un desempeño superior en términos de adaptabilidad, coherencia y eficacia en comparación con enfoques existentes. Sin embargo, es importante considerar las implicaciones éticas que conlleva el desarrollo de este tipo de modelos de Aprendizaje de Máquina con capacidades de memoria a corto y largo plazo.*

*Una de las principales preocupaciones éticas es la potencial invasión de la privacidad de los usuarios. Al tener la capacidad de recordar detalles específicos de las interacciones pasadas, el modelo podría acceder a información personal o confidencial de los usuarios, lo cual plantea serios riesgos en términos de seguridad y protección de datos.*

*Adicionalmente, la adaptabilidad del modelo a las preferencias y necesidades únicas de cada usuario podría llevar a la creación de "burbujas de información", donde el usuario recibe contenido y respuestas sesgadas o limitadas, restringiendo su exposición a diferentes perspectivas y opiniones. Esto podría tener implicaciones negativas en el desarrollo cognitivo y la formación de criterio propio.*

*Otro aspecto ético a considerar es la posibilidad de que el modelo pueda ser utilizado con fines manipulativos o engañosos. Al conocer en detalle las características y necesidades de cada usuario, el modelo podría ser utilizado para influenciar o persuadir a los usuarios de manera inapropiada, lo cual plantea riesgos éticos significativos.*

*Para mitigar estos riesgos, es crucial que el desarrollo de este tipo de modelos esté acompañado de sólidos marcos éticos y de gobernanza. Algunas medidas a considerar incluyen:*

- *Implementar estrictos controles de privacidad y seguridad de datos, asegurando el consentimiento informado y la transparencia en el manejo de la información de los usuarios.*

- *Diseñar mecanismos de diversificación de contenido y exposición a diferentes perspectivas, evitando la creación de burbujas de información.*

- *Establecer claros lineamientos y políticas de uso que impidan la utilización del modelo con fines manipulativos o engañosos.*

- *Involucrar a expertos en ética y derechos humanos en el proceso de diseño y desarrollo del modelo.*

*Solo a través de un enfoque responsable y ético en el desarrollo de este tipo de tecnologías será posible aprovechar sus beneficios sin comprometer los derechos y el bienestar de los usuarios.*

### *4.2. Trabajos relacionados*

- *MEMORY NETWORKS (arXiv:1410.3916v11 [cs.AI] 29 Nov 2015) Presentan la arquitectura general de los "memory networks" y sus componentes clave (I, G, O, R). Exploran variantes como el uso de hashing de memoria para mejorar la eficiencia. Realizan experimentos en tareas de preguntas y respuestas a gran escala y en un mundo simulado. Los autores concluyen que los "memory networks" son una clase de modelos poderosos que deben explorarse más a fondo en tareas de comprensión de texto y otras áreas. Identifican varias direcciones futuras, como*

*probar en tareas más complejas, explorar configuraciones débilmente supervisadas y desarrollar arquitecturas más sofisticadas.*

- *MemoryBank: Enhancing Large Language Models with Long-Term Memory (arXiv:2305.10250v3 [cs.CL] 21 May 2023) utiliza un modelo de base de datos vectorial que imita comportamientos antropomórficos y preserva selectivamente la memoria, incorporando un mecanismo de actualización de memoria, inspirado en la teoría de la curva de olvido de Ebbinghaus. Este mecanismo permite que la IA olvide y refuerce la memoria en función del tiempo transcurrido y la importancia relativa de la memoria, ofreciendo así un mecanismo de memoria más similar al humano y una experiencia de usuario enriquecida. Almacena conversaciones pasadas, eventos resumidos y retratos de usuarios.*

- *Generative Agents: Interactive Simulacra of Human Behavior (arXiv:2304.03442v2 [cs.HC] 6 Aug 2023) es una simulación de las interacciones de una aldea humana. En la simulación, los agentes necesitan recordar acciones pasadas para interactuar con otros agentes y el entorno. El enfoque fue un flujo de memoria que mantiene un registro exhaustivo de la experiencia del agente. Es una lista de objetos de memoria, donde cada objeto contiene una descripción en lenguaje natural, una marca de tiempo de creación y una marca de tiempo de acceso más reciente. El elemento más básico del flujo de memoria es una observación, que es un evento percibido directamente por un agente. Recuperan los recuerdos relevantes aplicando una función de recuperación que puntúa todos los recuerdos como una combinación ponderada de los tres elementos: $score = \alpha_{recency} \cdot recency + \alpha_{importance} \cdot importance + \alpha_{relevance} \cdot relevance$. En otras palabras, el modelo almacena cada pieza de memoria y luego usa una función de recuperación para obtener la memoria relevante.*

- *RecallM: An Adaptable Memory Mechanism with Temporal Understanding for Large Language Models (arXiv:2307.02738v3 [cs.AI] 3 Oct 2023) presenta una base de datos basada en grafos para almacenar los datos en el dominio simbólico. Es particularmente eficaz en la actualización de creencias y el mantenimiento de una comprensión temporal del conocimiento que se le proporciona. La innovación central aquí es que, al utilizar una arquitectura neuro-simbólica ligera, pueden capturar y actualizar relaciones complejas entre conceptos de una manera eficiente desde el punto de vista computacional.*

- *"My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents (arXiv:2404.00573v1 [cs.HC] 31 Mar 2024) En este estudio, se propone una nueva arquitectura de memoria similar a la humana diseñada para mejorar las capacidades cognitivas de los agentes de diálogo basados en modelos de lenguaje a gran escala (LLM) utilizando un modelo vectorial para la recuperación de memoria.*

### *5. Trabajo futuro*

*El objetivo principal de este documento es exponer esta nueva estructura para la memoria a largo plazo en LLM. Los autores de este documento sugieren la experimentación con los hiperparámetros de este modelo y explorar nuevas posibilidades para mejorar las capacidades del marco de trabajo.*

*El énfasis principal estuvo en la memoria a largo plazo, por lo que cualquier mejora para explorar en la memoria a corto plazo es interesante de experimentar.*

*Algunas ideas para contrastar los resultados con el modelo LTM original podrían ser un IRS diferente, una perspectiva diferente para actualizar los nodos vectoriales, una reconstrucción periódica del gráfico con otro algoritmo de agrupación tradicional utilizando aprendizaje no supervisado.*

***6. Conclusiones***

*En este proyecto se ha presentado un modelo novedoso para dotar a los Modelos de Lenguaje a Gran Escala (LLM) de capacidades de memoria a corto y largo plazo, permitiendo que se adapten de manera única a cada usuario sin necesidad de reentrenamiento específico.*

*Los principales aportes del modelo propuesto incluyen:*

*1. Una arquitectura híbrida que combina técnicas de bases de datos basadas en grafos y vectores para almacenar y recuperar información relevante a corto y largo plazo.*

*2. Un mecanismo de actualización dinámica de la memoria a largo plazo que permite al modelo aprender de manera no supervisada y mejorar su adaptabilidad a lo largo del tiempo.*

*3. Un sistema de recuperación de información eficiente que extrae los recuerdos más relevantes para cada interacción del usuario, ponderando factores como recencia, importancia y relevancia.*

*Los resultados experimentales muestran que el modelo propuesto logra un desempeño superior en términos de adaptabilidad, coherencia y eficacia en comparación con enfoques existentes. Esto abre nuevas posibilidades para el desarrollo de aplicaciones que requieren una interacción personalizada y a largo plazo con los usuarios, como asistencia personalizada, investigación científica y entrenamiento virtual.*

*Sin embargo, es crucial considerar las implicaciones éticas del desarrollo de este tipo de modelos, especialmente en lo que respecta a la privacidad de los usuarios, la creación de burbujas de información y el potencial uso manipulativo. Para mitigar estos riesgos, se deben implementar sólidos marcos éticos y de gobernanza que aseguren el manejo responsable de la información y la exposición a diferentes perspectivas.*

*En resumen, el modelo propuesto representa un avance significativo en la capacidad de los LLM para adaptarse a las necesidades únicas de cada usuario, con amplias aplicaciones potenciales. No obstante, su desarrollo debe ir acompañado de una reflexión ética profunda para garantizar que estas tecnologías se utilicen de manera responsable y en beneficio de la sociedad.*

***6.Bibliografía***

*1. Differentiable Neural Computers with Memory Demon (arXiv:2211.02987v1 [cs.LG] 5 Nov 2022)*

*2. MemoryBank: Enhancing Large Language Models with Long-Term Memory (arXiv:2305.10250v3 [cs.CL] 21 May 2023)*

*3. Generative Agents: Interactive Simulacra of Human Behavior (arXiv:2304.03442v2 [cs.HC] 6 Aug 2023)*

*4. RecallM: An Adaptable Memory Mechanism with Temporal Understanding for Large Language Models (arXiv:2307.02738v3 [cs.AI] 3 Oct 2023)*

*5. "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents (arXiv:2404.00573v1 [cs.HC] 31 Mar 2024)*