Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

¿Capturan los embeddings de los LLMs información sobre temporalidad y relaciones espaciales?

Tesis de Licenciatura en Ciencias de la Computación

Federico Hernán Suaiter

Director: Esteban Zindel Feuerstein
Codirector: Juan Manuel Ortiz de Zárate
Buenos Aires, 2024

ABSTRACT

Los Large Language Models (LLMs) han demostrado una capacidad notable para capturar complejas relaciones semánticas a través de sus embeddings, contribuyendo significativamente al avance de diversas aplicaciones, como el procesamiento del lenguaje natural (NLP). Sin embargo, la manera en que estos modelos representan conceptos abstractos como el tiempo y el espacio sigue siendo un área en exploración. Este trabajo tiene como objetivo analizar cómo los LLMs modelan el tiempo y el espacio dentro de sus representaciones vectoriales. Dicho análisis se lleva a cabo de forma directa sobre la estructura interna de las representaciones, permitiendo una evaluación transparente y una interpretación accesible. Si bien los LLMs solo han sido entrenados para predecir el siguiente token en una secuencia, su comportamiento revela una profundidad que trasciende las intenciones iniciales de su diseño. Los resultados obtenidos al analizar diferentes modelos de LLMs revelan patrones específicos en la manera en que dichos LLMs representan eventos temporales y ubicaciones geográficas, sugiriendo que efectivamente incorporan de manera implícita una estructura, en mayor o menor medida, espacio-temporal en sus representaciones vectoriales. Estos hallazgos abren nuevas oportunidades para el desarrollo de técnicas que mejoran la precisión de los LLMs en tareas que requieren razonamiento espacio-temporal avanzado. Además, contribuyen a la comprensión de las limitaciones actuales, proporcionando una base para futuras investigaciones dirigidas a optimizar su capacidad de modelar conceptos abstractos y complejos en contextos aplicados.

Palabras claves: *LLMs, capacidades emergentes, world models, tiempo, espacio, embeddings*

AGRADECIMIENTOS

En primer lugar, agradecer al programa de Becas de Iniciación a la Investigación en Ciencias de la Computación (BIICC). Gracias a su apoyo pude dedicar tiempo y esfuerzo a esta investigación, al punto de convertirla en mi tesis.

A toda mi familia, en especial a mi padre, Gustavo; por su apoyo constante, paciencia infinita y por estar presente en cada paso de este largo camino.

A mis amigos y compañeros, tanto a aquellos con quienes compartí tan solo poco tiempo, como a quienes me acompañan regularmente. También a mis amigos de siempre, que nunca dejaron de apoyarme y estar presentes en cada etapa de mi vida, incluyendo las más difíciles.

A mi mentor, Juan; y mi director, Esteban; por su paciencia y por ayudarme a que todo esto sea posible. También a los integrantes del grupo de investigación por ayudarme y atender mis dudas cuando lo necesitaba.

A quienes se están tomando el tiempo de leer estas páginas y de valorar el esfuerzo detrás de ellas.

Finalmente, quiero recordar con cariño a quienes ya no están conmigo, ya sea porque partieron de esta vida o porque nuestros caminos tomaron rumbos diferentes.

Muchas gracias a todos.

Índice general

1. INTRODUCCIÓN

1.1. Motivación

Un **modelo de lenguaje** o **LM** (Language Model) es un sistema que puede predecir y generar texto coherente y con sentido. Existen diferentes tipos de LM, algunos basados en reglas lingüísticas y estadísticas, y otros que emplean técnicas de aprendizaje automático. Por otro lado, un **gran modelo de lenguaje** o **LLM** (Large Language Model) es una evolución avanzada de los LM basados en aprendizaje automático. Los LLMs utilizan redes neuronales con miles de millones de parámetros, están entrenados con grandes volúmenes de datos y poseen arquitecturas complejas. Estos son un componente clave en el procesamiento del lenguaje natural, disciplina que estudia tanto la comprensión (estructura y significado de las palabras) como la generación del lenguaje humano.

Los LLMs surgieron entre 2017 y 2018 a partir de avances progresivos en el procesamiento del lenguaje natural. Un punto clave en esta evolución fue Word2Vec [1], que permitió representar palabras como vectores (embeddings), capturando relaciones semánticas gracias a un modelo de red neuronal para aprender asociaciones de palabras a partir de un gran corpus de texto. Más adelante, en 2017, los investigadores de Google presentaron la arquitectura del *Transformer* en su artículo "Attention is All You Need" [2]. Dicho artículo captó la atención de todo el campo de la inteligencia artificial (IA) y revolucionó el mismo al introducir el mecanismo de "self-attention" y "multi-headed attention" para modelar relaciones contextuales entre palabras (el mecanismo de "attention" como tal se basó en el trabajo de [3]). Fue sobre esta base que modelos como BERT (2018) [4] se desarrollaron y permitieron un gran desempeño en una amplia variedad de tareas. Algunos de los LLMs más notables hoy en día son la serie de modelos GPT de OpenAI (por ejemplo, GPT-3 [5] y GPT-4 [6]), Gemini [7] de Google, entre otros.

En los últimos años, a partir de estos avances fundamentales, los LLMs han transformado el campo del procesamiento del lenguaje natural. Si bien solo han sido entrenados para predecir el siguiente token, estos han demostrado grandes resultados al ser aplicados en diferentes dominios y tareas [8, 9] debido a su capacidad para generar y transformar lenguaje con un nivel de sofisticación sin precedentes.

Una de las claves del éxito de los LLMs radica en el uso de los embeddings. Un **embedding** es un vector de números reales que representa un texto o palabra en espacios vectoriales de múltiples dimensiones. Estos números capturan parte de la semántica de la entrada, permitiendo colocar vectores semánticamente similares juntos en el espacio vectorial. Dicho espacio permite realizar operaciones vectoriales que capturan relaciones lingüísticas y semánticas. Por ejemplo, la analogía "Rey - Hombre + Mujer $\approx$ Reina" [10] es una demostración de cómo las operaciones lineales en el espacio de embeddings pueden representar dichas transformaciones semánticas. Los embeddings son utilizados en diversos dominios debido a su capacidad para representar y categorizar datos.

A pesar de que los LLMs son entrenados de manera explícita con grandes cantidades de datos (supervisados y no supervisados), existen diferentes hipótesis sobre cómo estos aprenden de manera abstracta o conceptual. Una de las ideas más destacadas es la noción

de "capacidades emergentes", definidas -parafraseando a [9]- como "*habilidades que no están presentes en modelos más pequeños, pero sí en modelos más grandes*". Esto sugiere que los modelos pueden desarrollar habilidades o comportamientos no previstos directamente por sus diseñadores, los cuales surgen como resultado del aumento en la escala del modelo. Asimismo, el concepto de emergencia se describe como el fenómeno donde "*cambios cuantitativos en un sistema resultan en cambios cualitativos en su comportamiento*". Por otro lado, se encuentra una corriente opuesta, la cual sostiene que dichas "capacidades emergentes" no son más que la combinación de tres elementos: el aprendizaje en contexto, la capacidad de memorización inherente al modelo y la explotación de patrones lingüísticos aprendidos durante el entrenamiento [11]. Una hipótesis alternativa se basa en la idea de compresión de datos, en la que los LLMs, al procesar vastas cantidades de información, aprenden modelos más compactos, coherentes e interpretables del proceso generativo subyacente a los datos de entrenamiento, es decir, una representación interna del entorno también conocido como un modelo del mundo o "World Model".

Gurnee et al. [12] halló que los LLMs efectivamente modelan el tiempo y el espacio internamente. Mediante la información sobre las activaciones de las neuronas de los modelos al recibir textos relacionados al tiempo y el espacio entrenaron una regresión que fue capaz de predecir el tiempo y espacio de un texto, lo que señala que efectivamente esta información está embebida internamente en los modelos. Esto abre una nueva pregunta: si el tiempo y espacio están internamente modelados en los LLMs ¿Puede ser que también estén embebidos en sus outputs?

En este trabajo, buscamos examinar los embeddings generados por diferentes modelos de LLMs para una variedad de temas, incluyendo eventos históricos como batallas, obras de literatura, discursos presidenciales de apertura, títulos de portadas de diarios (con un enfoque en el análisis temporal) y ubicaciones geográficas como lugares, ciudades y países (con un enfoque en el análisis espacial). Dichos análisis tiene como objetivo encontrar efectivamente si el tiempo y espacio están internamente modelados en los embeddings de los diversos modelos. A su vez, buscamos evaluar cómo pequeñas modificaciones en los parámetros del método o en la información proporcionada a los modelos afectan su capacidad para capturar de manera efectiva las dimensiones deseadas.

1.2. Trabajo relacionado

Tal como se comentó en la motivación, en [12] los autores mostraron como los LLMs forman modelos temporales y espaciales del mundo. Una de las diferencias distintivas con lo presentado en ese documento recae en que para nuestra técnica sólo requerimos de los embeddings generados por el LLM. En otras palabras, no se utilizan modelos adicionales para realizar el análisis sobre el modelo de lenguaje, ni se necesita acceder a la red neuronal. Esto permite evaluar las capacidades intrínsecas de los modelos de lenguaje para capturar las dimensiones temporales y espaciales de una manera rápida, sencilla y transparente al no depender de modelos adicionales, entrenamiento específico o un gran poder de cómputo para poder correr localmente los modelos.

El análisis de los embeddings se basa en una metodología similar a la utilizada por Waller et al. [13]. En su enfoque, se emplean "semillas" que, ajustadas según la particularidad a analizar, generan un vector de representación. Una vez obtenido el vector, se asignan puntuaciones al resto de los valores proyectándolos sobre dicho vector mediante

la similitud coseno. Aunque esta técnica proporciona valores numéricos, estos solamente permiten establecer un orden relativo entre los elementos, lo que permite evaluar si el modelo mantiene una noción coherente de orden en relación con la dimensión que se desea analizar. El método no es capaz de devolver, por ejemplo, el año exacto en que ocurrió un evento determinado. En este trabajo, nuestro objetivo es obtener una noción de orden tanto temporal como espacial.

Previo a este trabajo, y siguiendo también lo realizado por [13], en [14] se introdujo una técnica basada en texto para cuantificar el alineamiento de comunidades en línea a lo largo de dimensiones sociales. Esto se realizó mediante la generación de representaciones vectoriales de texto (embeddings), que se utilizan para evaluar a las comunidades en diferentes ejes del espectro político-ideológico. En este trabajo se utilizará el modelo `embed-english-v2.0` mencionado en el mismo con el objetivo de poner a prueba su comportamiento (junto con el de otros modelos) al analizar temas que, al contrario que las ideologías de comunidades, no suelen variar con el tiempo.

Continuando con investigaciones anteriores, [15, 16] han demostrado que el lenguaje codifica información geográfica. Por otro lado, en [17] se descubrió que, si bien ninguno de los modelos que utilizaron fue capaz de razonar de manera confiable sobre las direcciones cardinales, todos los modelos exhibieron algunas habilidades de razonamiento espacial. Además, [18] demostró que los modelos de lenguaje generalmente codifican información geográfica limitada, teniendo los modelos más grandes una mejor habilidad en esto. Cabe destacar que ninguno de los modelos mencionados en las referencias de este párrafo fue utilizado en este trabajo debido a que los mismos no tienen como salida los embeddings del texto ingresado.

Nuestros hallazgos podrían ser potencialmente utilizados para mejorar algunas de las tareas subsecuentes que forman parte de las áreas denominadas Recuperación de Información Temporal (T-IR) y Recuperación de Información Geográfica (GIR). Si bien nuestra intención en este trabajo es demostrar la capacidad emergente de los modelos sobre las representaciones vectoriales de texto (embeddings), esto podría ser útil en dichas áreas gracias al ordenamiento que se puede llevar a cabo a partir de los valores otorgados por el método.

1.3. Estructura de la tesis

El resto de esta tesis se organiza de la siguiente manera: en el Capítulo 2 introducimos diferentes conceptos y conocimientos preliminares necesarios para comprender las métricas. En el Capítulo 3 se explica el método utilizado junto con un ejemplo a modo de ilustración. Luego, en el Capítulo 4 se muestran los diversos modelos utilizados junto con sus características. A su vez, se desarrollan los experimentos del tiempo y el espacio. Por último, en el Capítulo 5, se detallan las conclusiones generales e ideas para un trabajo futuro.

2. CONCEPTOS INICIALES

En esta sección se describen en detalle los diferentes conceptos iniciales necesarios sobre métricas y fórmulas que serán utilizadas en las secciones posteriores.

Métricas

Similitud coseno

Tanto para el método que se comentará en el siguiente capítulo como para un experimento introductorio utilizaremos la **similitud coseno**. Ésta es una medida utilizada para evaluar qué tan similares son dos vectores en un espacio multidimensional, y se basa en el cálculo del coseno del ángulo entre dos vectores. Sean A y B dos vectores, la similitud coseno de ambos se define como el producto escalar de ambos vectores, dividido la multiplicación de las normas:

$$S_C(A, B) \;=\; \frac{A \cdot B}{\|A\|\|B\|}$$

Los valores están comprendidos entre -1 y 1, donde un valor cercano a 1 indica que los dos vectores son similares, es decir, están en la misma dirección. Un valor de 0 indica que los vectores son ortogonales. Un valor cercano a -1 indica que los vectores apuntan en direcciones opuestas.

Coeficiente de correlación de rango de Kendall

Una de las dos métricas principales que se utilizará en este trabajo es el **coeficiente de correlación de rango de Kendall** (también conocido en inglés como *Kendall's* τ [19, 20]). El mismo mide el grado de concordancia entre dos secuencias de datos ordinales, es decir, datos que pueden ser ordenados. En lugar de trabajar con valores directamente, está métrica se enfoca en el orden relativo de los mismos.

Sean X e Y dos secuencias de datos con n observaciones cada uno, al comparar todos los pares posibles de observaciones (x_i, y_i) y (x_j, y_j) (con $i < j$) evaluamos su relación para determinar cuales se clasifican como par concordante, par discordante o empate.

- Par concordante (C): aquel donde el orden relativo de los valores de X coinciden con el de Y: $(x_i < x_j \wedge y_i < y_j) \;\vee\; (x_i > x_j \wedge y_i > y_j)$
  Es decir, ambos aumentan o disminuyen juntos.
- Par discordante (D): aquel donde el orden relativo de los valores de X es opuesto al de Y: $(x_i < x_j \wedge y_i > y_j) \;\vee\; (x_i > x_j \wedge y_i < y_j)$
  Es decir, uno aumenta mientras que el otro disminuye.

- Empate: aquel donde los valores de X o de Y son iguales: $(x_i = x_j \vee y_i = y_j)$

Se define así la variante de Kendall's τ, *Tau-a*, como:

$$\tau = \frac{C - D}{\frac{n(n-1)}{2}}$$

Cabe destacar que dicha variante no posee ningún ajuste para empates, ya sea en X, Y o ambos. En particular, en este trabajo, se utiliza la variación *Tau-b* la cual si realiza ajustes en caso de empates. Se define *Tau-b* como:

$$\tau = \frac{C - D}{\sqrt{(C + D + T) \cdot (C + D + U)}}$$

Donde T representa el número de empates solo en x ($x_i = x_j$ pero $y_i \neq y_j$), y U el número de empates solo en y ($y_i = y_j$ pero $x_i \neq x_j$). Si se produce un empate para el mismo par tanto en x como en y, no se suma ni a T ni a U. Cabe destacar que un par empatado no es ni concordante ni discordante.

Los valores están comprendidos entre -1 y 1. Un valor de $\tau = 1$ indica una perfecta concordancia, lo que implica que las variables están perfectamente ordenadas. Un valor de $\tau = -1$ indica una perfecta discordancia, lo que implica que las variables están ordenadas completamente al revés. Un valor de τ cercano a 0 indica una falta de correlación.

Coeficiente de correlación de Pearson

La segunda métrica principal que se utilizará en este trabajo es el **coeficiente de correlación de Pearson** (también conocido en inglés como Pearson's r). Ésta es una medida estadística que cuantifica la relación lineal entre dos variables cuantitativas.

Dadas las variables X e Y, se tiene que: x_i e y_i son los valores individuales de las variables X e Y respectivamente, $\overline{x}$ e $\overline{y}$ las medias de las variables X e Y respectivamente, y n el número de observaciones. Se define así el coeficiente de Pearson como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Los valores de r están comprendidos entre -1 y 1. Un coeficiente cercano a 1 indica que hay una fuerte relación positiva, donde a medida que una variable aumenta, la otra también. En el caso de un valor cercano a -1, este indica una fuerte relación negativa, donde a medida que una variable aumenta, la otra tiende a disminuir. Un coeficiente cercano a 0 indica la falta de una relación clara entre las variables.

Fórmulas

Antes de ir con la definición de la fórmula del semiverseno, la cual será utilizada para medir distancias sobre la tierra en un análisis del espacio, introduzcamos primero la noción de **radián**. Esta es una unidad de medida angular que se define a partir de la relación entre la longitud de un arco y el radio de un círculo.

$$360^\circ = 2\pi \ radianes \qquad 1 \ radian \approx 57{,}3^\circ$$

Dicho esto, introducimos la **fórmula del semiverseno** (también conocida en inglés como *haversine formula*). La misma es una ecuación matemática utilizada para calcular la distancia más corta entre dos puntos de la superficie de una esfera.

En este trabajo se utiliza para calcular la distancia del círculo máximo entre dos puntos de una esfera dadas sus latitudes y longitudes (expresadas en radianes). La distancia del círculo máximo es el camino más corto sobre la superficie de la Tierra entre dos puntos, suponiendo que la Tierra es una esfera perfecta (la Tierra es casi esférica, así que provee una buena aproximación con menos del 1 % de error en promedio). La distancia D se calcula como:

$$D = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos(\varphi_1) \cdot \cos(\varphi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

Donde R es el radio de la Tierra (aproximadamente 6371 km), φ_1 y φ_2 las latitudes de los dos puntos, λ_1 y λ_2 las longitudes de los dos puntos, y $\Delta\varphi$, $\Delta\lambda$ la diferencia entre las latitudes y longitudes, respectivamente.

3. MÉTODO UTILIZADO

Tal como se mencionó en el Capítulo 1, el enfoque adoptado en este trabajo se basa en el método desarrollado por Waller et al. [13]. En su estudio, se explora el posicionamiento de las comunidades (de Reddit) a lo largo de diferentes dimensiones sociales (por ejemplo, la edad). Esta investigación propone un acercamiento más similar a [14] al utilizar embeddings de palabras como tal en lugar de embeddings de comunidades (los cuales a su vez se basan en la similitud de usuarios que interactúan en ellas). A su vez, diferenciándose de ambas, aquí se planea aplicarlo a la dimensión temporal y espacial. A continuación, se procede a detallar el método utilizando la Figura 3.1 como apoyo visual.

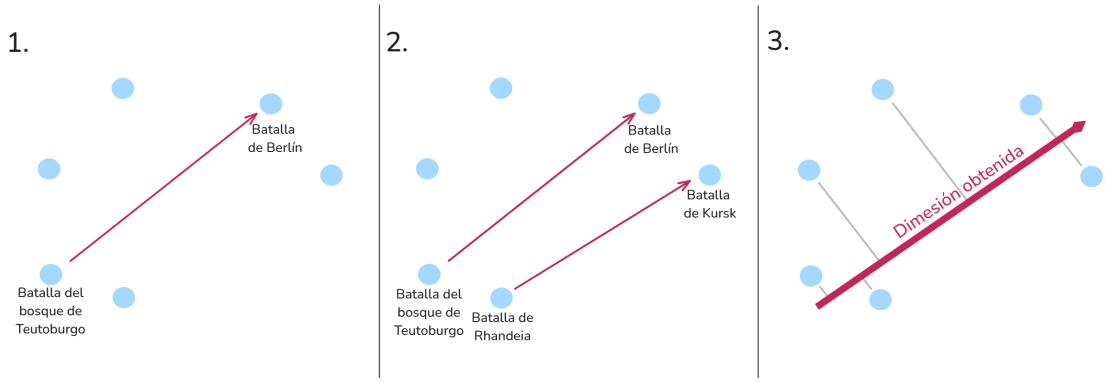

Fig. 3.1: Ilustración del método para la obtención de una dimensión temporal. La disposición de los puntos y los vectores tiene únicamente un propósito demostrativo si bien utiliza datos en español del dataset *batallasHistoricas*.

El método a utilizar consiste en identificar una dimensión objetivo a partir de los embeddings obtenidos de los diferentes eventos y objetos a analizar. Para ello se selecciona inicialmente un par de embeddings que difiera principalmente en el aspecto que se desea estudiar. Por ejemplo, para obtener una dimensión temporal en base a batallas históricas, se podrían elegir la Batalla del bosque de Teutoburgo (año 9) y la Batalla de Berlín (año 1945), generando así un vector que represente dicha dimensionalidad temporal (estas batallas pertenecen a los extremos del dataset *batallasHistoricas* que será descrito en el siguiente capítulo). Este par seleccionado se denomina par semilla, y permite calcular el vector o dimensión d **(Sección 1 - Figura 3.1)**:

$$d = p_2 - p_1$$

Para evitar que la dimensión esté demasiado ligada a la naturaleza y particularidad del par semilla seleccionado, es posible incluir pares semilla adicionales **(Sección 2 - Figura 3.1)**. De ser así, el vector o dimensión d se calcula como el promedio de las diferencias entre todos los pares seleccionados, contribuyendo a que la dimensión obtenida sea más representativa y menos específica:

$$d \;=\; \frac{1}{|P|} \;\cdot \sum_{(p_1,p_2)\in P} (p_2 - p_1)$$

Con la dimensión d definida, se proyectan todos los embeddings a calcular sobre el vector d utilizando la métrica de similitud coseno. Este procedimiento permite asignar un valor asociado a la dimensión d para cada embedding e, calculado de la siguiente manera **(Sección 3 - Figura 3.1)**:

$$d\text{-valor}(e) \;=\; S_C\,(e, d)$$

En síntesis, la siguiente fórmula se utiliza para calcular el valor de cada uno de los embeddings e en base a la dimensión obtenida d:

$$\begin{aligned} d\text{-valor}(e) \;&=\; S_C\,(e, d) \\ &= \frac{e \cdot \sum_{(p_1,p_2)\in P} (p_2 - p_1)}{||e|| \cdot |P| \cdot ||d||} \\ &= \frac{1}{|P| \cdot ||e|| \cdot ||d||} \sum_{(p_1,p_2)\in P} (e \cdot p_2 - e \cdot p_1) \end{aligned}$$

Se puede apreciar aquí como los embeddings que presenten una mayor similaridad con una semilla que con la otra del par, tendrán un puntaje en los extremos (en el caso de seleccionar más de una semilla será al tener mayor similaridad con el promedio de estas). Aquellos cercanos al segundo valor del par semilla (p_2) recibirán valores positivos mientras que aquellos cercanos al primer valor del par semilla (p_1) recibirán valores negativos. Por otro lado, aquellos equidistantes a ambas semillas del par recibirán un puntaje cercano a cero.

Para minimizar el riesgo de sobreajuste en los experimentos, tan solo se seleccionarán unos pocos pares semilla y se prescindirá del método de aumento descrito en [13]. Esta decisión también busca evitar la introducción de ruido, ya que dicho método implica seleccionar múltiples pares en diversas direcciones.

Además, dado que este trabajo se enfoca en el análisis de eventos, lugares y objetos específicos —a diferencia del enfoque del artículo citado, donde el método se emplea para evaluar opiniones de comunidades y otros aspectos— se prioriza el uso de extremos (del conjunto de datos) como semillas para reducir la ambigüedad en los resultados, y simplificar el proceso de selección al tomar aquellos que difieren en mayor medida en la dimensión que se busca obtener. A la hora de seleccionar pares semilla adicionales, también se utilizan extremos, siendo estos los extremos de la subsección restante producto de ignorar los extremos anteriores. A su vez, si bien en algunos análisis se busca probar con diferentes pares (dada una misma cantidad de pares semilla), el objetivo principal del trabajo recae en encontrar si efectivamente existe una dimensión temporal o espacial, y no en analizar de manera exhaustiva toda combinación de pares semilla posible.

4. EXPERIMENTOS

4.1. Modelos utilizados

En la Tabla 4.1 se presentan los diferentes modelos de LLMs junto con sus respectivas limitaciones. En particular se tomaron modelos de Cohere[1], Voyage[2] y Google Gemini[3]. Estos modelos están diseñados principalmente para calcular los embeddings de las palabras o textos ingresados. Sin embargo, algunos de ellos también tienen la capacidad de procesar imágenes, o pueden ser utilizados en tareas como clasificación o búsquedas en bases de datos (en este trabajo solo se los utilizará para calcular embeddings).

Resulta relevante mencionar que una mayor dimensionalidad o una mayor cantidad máxima de tokens no implica necesariamente la superioridad de un modelo sobre otro.

Modelo	**Abreviación**	**Empresa**	**Dimensiones**	**Máxima cantidad de tokens**
embed-english-v2.0	C-Eng2	Cohere	4096	512
embed-english-v3.0	C-Eng3	Cohere	1024	512
embed-multilingual-v2.0	C-Mlt2	Cohere	768	256
embed-multilingual-v3.0	C-Mlt3	Cohere	1024	512
voyage-large-2	V-Lrg2	Voyage	1536	16000
voyage-multilingual-2	V-Mlt2	Voyage	1024	32000
models/text-embedding-004	G-004	Google Gemini	768	2048
models/embedding-001	G-001	Google Gemini	768	2048

Tab. 4.1: Modelos utilizados a lo largo de los diferentes experimentos

Estos modelos fueron seleccionados debido a la diversidad de dimensiones (tamaño del embedding de salida) que presentan, la cantidad máxima de tokens que admiten y los idiomas que soportan. En relación con esto último, la principal diferencia entre los modelos multilingües radica en que los desarrollados por Cohere son compatibles con más de 100 idiomas, mientras que los de Voyage y Gemini soportan únicamente 27 y 38 idiomas respectivamente.

Es importante destacar, a su vez, que todos estos modelos son de acceso gratuito. Esto no quita que los mismos cuenten con restricciones/limitaciones específicas, como la cantidad de solicitudes permitidas o el límite de tokens procesables por minuto.

En todos los casos, siempre que fue posible, se configuraron los modelos para evitar el truncamiento de las entradas, es decir, la eliminación o reducción de parte de la información original cuando esta excede la capacidad del modelo (máxima cantidad de tokens).

[1] `https://docs.cohere.com/docs/cohere-embed`
[2] `https://docs.voyageai.com/docs/embeddings`
[3] `https://ai.google.dev/gemini-api/docs/models/gemini`

4.2. Tiempo

El tiempo (del latín *tempus*) se define como una magnitud física que permite medir la duración o la separación entre acontecimientos. En esta sección nos centraremos específicamente en acontecimientos del pasado, ocurridos después del año de nacimiento de Jesús de Nazaret (el año 1 d. C.). Utilizaremos el año como unidad de referencia para medir la separación temporal entre eventos. Esta elección se debe a que, en textos de carácter histórico, periodístico o narrativo, los años suelen funcionar como referencias clave para organizar cronológicamente los acontecimientos.

Buscaremos en esta sección utilizar el método descrito en el Capítulo 3 para encontrar una dimensión temporal, la cual pueda ser apreciable en cada uno de los diferentes análisis a realizar. Encontrar esta dimensión temporal nos acercará a comprobar si efectivamente esta capacidad emergente se encuentra presente en los diferentes modelos de LLMs.

4.2.1. Datasets

A continuación se describen en detalle los datasets utilizados en esta sección:

batallasHistoricas. Dataset en inglés extraído de Wikipedia donde se toman los nombres y contextos al que pertenecen las batallas históricas[4]. En particular se tomaron 50 batallas, comenzando desde el año 9 hasta el año 1945, asegurándose de que haya por lo menos una batalla cada 100 años y no haya dos batallas libradas en un mismo año. El contexto histórico al que pertenece la batalla (en caso de figurar) fue extraído de la página de Wikipedia correspondiente de cada una de las batallas.

mejoresLibrosEnIngles. Dataset en inglés extraído de Wikipedia donde se toman los nombres y autores de los considerados mejores libros en inglés[5]. Este dataset contiene un total de 46 entradas, cuyos años de publicación se encuentran comprendidos entre 1899 y 2001. Se excluyó una de las entradas, ya que correspondía a un ciclo de novelas publicadas entre 1951 y 1975 en lugar de un único libro. Resulta importante señalar que, al contrario que con los otros datasets, aquí algunos años de publicación son compartidos por más de un libro. Esto se decidió con el objetivo de evaluar si los modelos tienden a agrupar los libros publicados en un mismo año, con un mismo valor.

aperturaSesionesOrdinarias. Dataset en español compuesto por parte de los mensajes presidenciales de apertura de sesiones ordinarias ante la Asamblea Legislativa Argentina. Incluye los discursos de los presidentes de la Nación Argentina desde la recuperación de la democracia, considerando únicamente el discurso pronunciado luego de asumir el mandato de cada presidente (e.g. el discurso de Menem del año 1990 y del año 1996) dando un total de 11 discursos[6]. Debido a que los discursos son extensos (en palabras y tokens), se decidió tomar tan solo una parte de los mismos. Cada fragmento consta aproximadamente de 770 palabras extraídas desde el inicio del discurso, con cuidado de no interrumpir ningún párrafo a la mitad para mantener la coherencia. Además, se eliminaron las anotaciones de aplausos y los guiones de fin de línea para facilitar el procesamiento del texto.

[4] `https://en.wikipedia.org/wiki/Lists_of_battles`

[5] `https://en.wikipedia.org/wiki/List_of_English-language_books_considered_the_best`

[6] `https://www2.hcdn.gob.ar/secparl/dgral_info_parlamentaria/dip/documentos/mensajes_presidenciales.html`

extractosClarin. Dataset en español compuesto por el título y subtítulo de la noticia principal de la portada del diario Clarín. Estas pertenecen al día 25 de mayo, cada 5 años desde el año 1950 al año 2015, dando un total de 15 entradas. Las noticias fueron transcriptas a partir de una página en la que se puede visualizar la portada de una determinada fecha[7].

4.2.2. Análisis preliminar sobre temporalidad en los embeddings

Antes de proceder a los diferentes análisis con el método mencionado en el Capítulo 3, se realiza un análisis preliminar con el objetivo de averiguar si los modelos contienen información temporal en sus embeddings.

Dicho análisis consiste en comparar la similitud coseno entre el embedding correspondiente a las frases en inglés "The year X" (donde X representa un año), y el embedding de un objeto o evento en particular. Para esto se seleccionaron un evento y un objeto, siendo estos: *Battle of the Somme* del dataset *batallasHistoricas* y *To Kill a Mockingbird* del dataset *mejoresLibrosEnIngles* respectivamente.

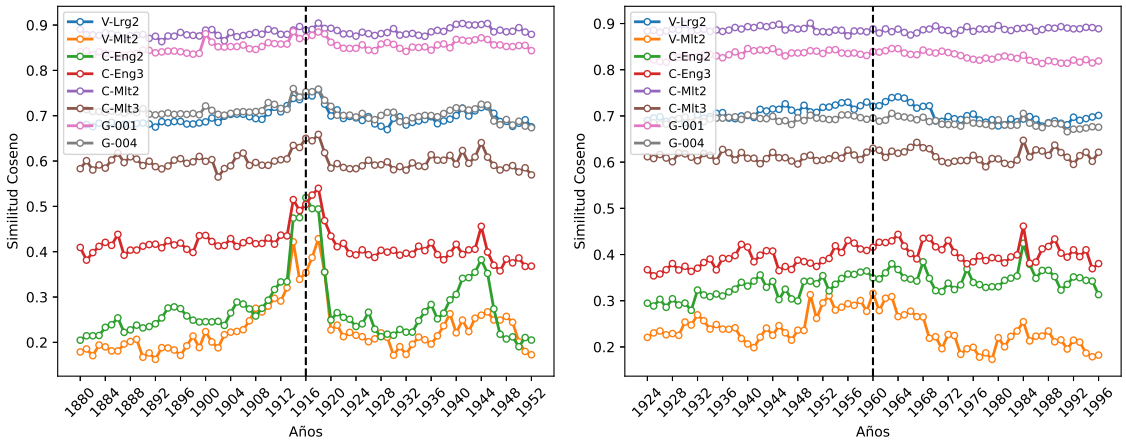

Fig. 4.1: A la izquierda, la comparación de *Battle of the Somme* (1916). A la derecha, la comparación del libro *To Kill a Mockingbird* (1960).

Como se puede observar en la Figura 4.1, el gráfico ubicado a la izquierda muestra evidencia sobre cómo la similitud coseno tiende a aumentar a medida que se acerca al año en que ocurrió el evento analizado (en este caso, la batalla del Somme). En particular, el modelo `embed-english-v2.0` de *Cohere* alcanza su valor máximo de similitud en el año exacto del suceso (1916). Si bien otros modelos reflejan un comportamiento similar, este fenómeno es menos pronunciado en comparación, y algunos modelos, como `embed-multilingual-v2.0`, no parecen reproducir este patrón. Un detalle no menor a destacar son los leves picos observados en el año 1944 en varios modelos. Suponemos que se debe a que algunos modelos captan la noción de "Guerra Mundial" al recibir una batalla perteneciente a la primera de estas. Si bien esto deja en evidencia el entendimiento por parte de los modelos sobre el contexto con tan solo unas palabras, el análisis en profundidad de la similitud coseno como tal se aleja del objetivo principal al que apunta este trabajo. De todas formas, resulta interesante la observación.

[7] `https://tapas.clarin.com/`

En el caso del gráfico de la derecha, no se llega a observar un comportamiento análogo. Los valores obtenidos no muestran un patrón claro en ninguno de los modelos analizados. El valor máximo se obtiene en una fecha distante a la original, y en algunos casos no hay un claro valor máximo a simple vista como en el caso del gráfico de la izquierda.

Aunque esta técnica proporciona una visión inicial de las capacidades de los modelos, su simplicidad limita la profundidad de los resultados obtenidos. En los análisis que siguen, se utilizará el método explicado en el Capítulo 3 para analizar cómo responden los distintos modelos al ser evaluados con diversos conjuntos de datos.

4.2.3. Análisis de Batallas

La historia de la humanidad ha estado marcada por conflictos y enfrentamientos a lo largo y ancho del mundo. Las batallas históricas constituyen eventos clave en la narrativa de las naciones y facciones que participaron de ellas. La pregunta es, dada su importancia ¿Podrán sus embeddings ser ordenados en base a cuándo transcurrieron? Se analiza el dataset *batallasHistoricas* en busca de esta respuesta.

Inicialmente se llevó a cabo un análisis utilizando únicamente los embeddings generados a partir de los nombres de las batallas, tomando como par semilla los eventos extremos. Aquí los resultados obtenidos no fueron buenos, en casos otorgando un valor de Tau cercano a 0. Posteriormente se optó por utilizar dos pares semilla, siendo un par los extremos del conjunto de datos y el segundo par los eventos extremos de la subsección restante producto de ignorar los extremos anteriores. Aquí se obtuvieron mejoras significativas.

Como idea alternativa, se optó por brindarle mayor información al modelo, calculando los embeddings del nombre de la batalla y el contexto del que fue parte (dejando este último vacío en los casos en que no se disponía del mismo). La hipótesis detrás de este enfoque sostiene que un modelo entrenado con una amplia cantidad de información puede comprender y contextualizar mejor los eventos históricos, ya que el mecanismo de atención le permite identificar y priorizar relaciones relevantes en los datos. En la Tabla 4.2 se puede ver un ejemplo de cómo está compuesto el texto de los experimentos.

Batalla - Battle of Sudomer	
Solo nombre	"Battle of Sudomer"
Nombre y contexto	"Battle of Sudomer, Hussite Wars'

Tab. 4.2: Ejemplo de cada tipo de experimento de batallas

En la Figura 4.2 se presentan los resultados de los experimentos previamente mencionados donde se comparan entre si sus valores tanto de Pearson como de Tau. Cabe destacar la superioridad del experimento que combina los dos pares semilla junto con la información adicional proporcionada. Este enfoque logró los valores más altos de forma consistente en ambos indicadores para todos los modelos empleados.

El mejor resultado obtenido de todos los experimentos realizados se puede observar en la Figura 4.3. El mismo pertenece al modelo `embed-english-v2.0` de *Cohere*. Aquí los valores obtenidos fueron $\tau = 0{,}649$ y $r = 0{,}786$. El gráfico de dispersión presenta diferentes puntos (batallas en este caso) donde para cada uno de ellos, en el eje de abscisas, se encuentra el año en el que ocurrió, y en el eje de ordenadas, el valor obtenido mediante

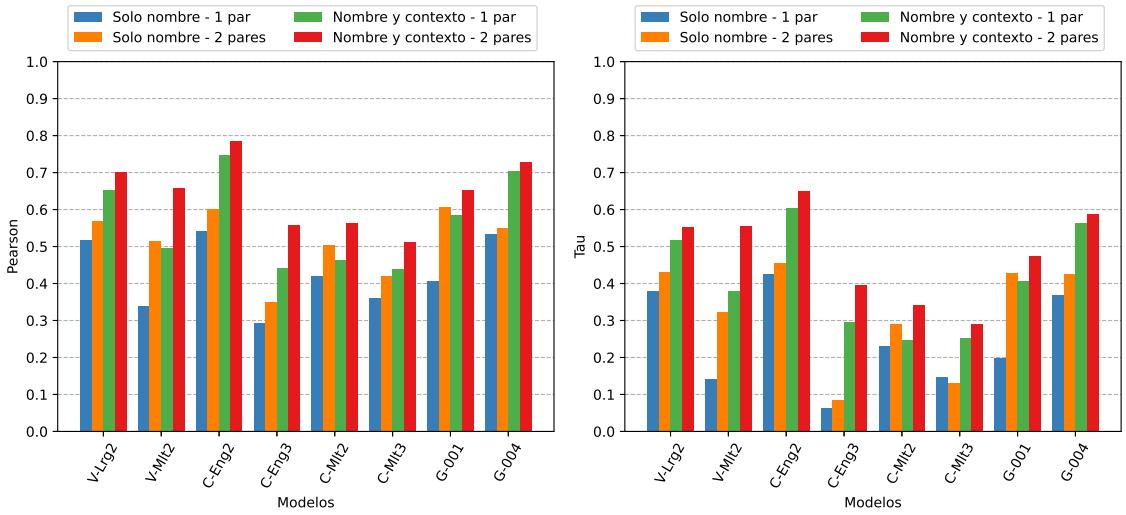

Fig. 4.2: Comparación de los diferentes experimentos sobre batallas. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos.

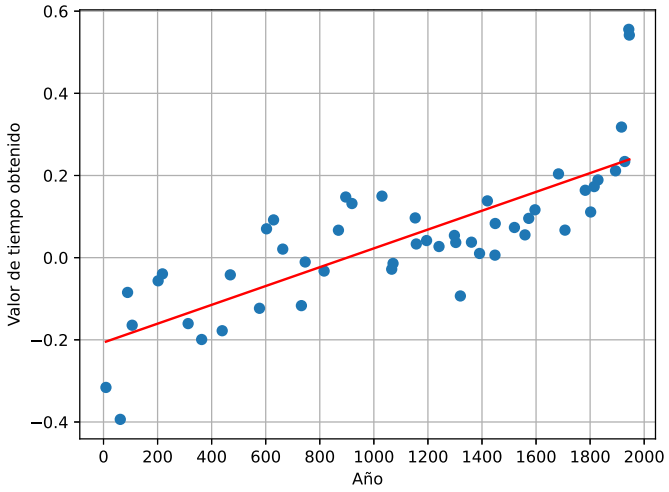

Fig. 4.3: Valores obtenidos mediante el modelo `embed-english-v2.0` utilizando dos pares semilla junto con el nombre y el contexto. Cada punto representa una batalla histórica.

la proyección. En el caso ideal donde una dimensión temporal es apreciable al cien por ciento, los puntos se verían en una perfecta diagonal, donde valores bajos se asociarían a los años menores mientras que valores altos se asociarían a años mayores (proporcionalidad directa). Si bien no se logra este caso ideal, es apreciable dicha proporcionalidad directa, indicada por la dirección en la que se encuentran posicionados los puntos y la recta de regresión (línea roja).

4.2.4. Análisis de Libros

Desde sus orígenes, la humanidad ha tenido que hacer frente a una cuestión fundamental: la forma de preservar y transmitir su cultura. Así como las batallas están marcadas por el año en el que transcurrieron, hay objetos que están marcados por el año en el que fueron publicados (si bien estos pueden haber tenido un largo proceso de creación). Hablamos de los libros. Si bien existen distintos tipos de libros, todos comparten elementos característicos: un título, un autor y un año de publicación. Esto plantea una pregunta interesante ¿Podrán sus embeddings contener información sobre cuándo fueron publicados de forma tal de poder ordenarlos? Se analizó el dataset *mejoresLibrosEnIngles* en busca de esta respuesta.

Siguiendo una metodología similar a la aplicada en el caso de las batallas, se comenzó evaluando los embeddings generados a partir de los nombres de los libros. Dado que los resultados iniciales no fueron satisfactorios, se optó por proporcionar información adicional, en particular, el autor del libro. A su vez, se decidió probar con el agregado de pares semilla adicionales. Estos siempre pertenecen a los extremos, seguido de los extremos de la subsección restante producto de ignorar los extremos anteriores. En la Tabla 4.3 se puede ver un ejemplo de cómo está compuesto el texto de los experimentos.

Libro - Brave New World	
Título	"Brave New World"
Título y autor	"Brave New World by Aldous Huxley"

Tab. 4.3: Ejemplo de cada tipo de experimento de libros

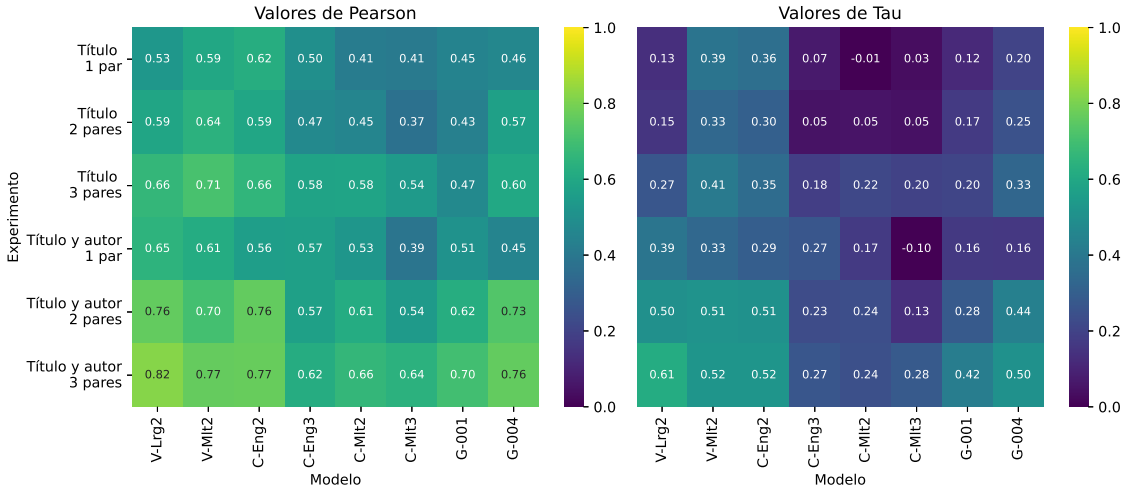

Fig. 4.4: Comparación de los experimentos sobre libros. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos en base a cada experimento y modelo.

La Figura 4.4 muestra cómo, al igual que en la experimentación con batallas históricas, la incorporación de pares semilla adicionales contribuye a obtener mejores resultados. Asimismo, proporcionar más información al modelo (en este caso, el autor) mejora significativamente su rendimiento, alcanzando los valores más altos cuando se combina la

mayor cantidad de pares semilla con información adicional. Esto se refleja en los colores más claros en la parte inferior del gráfico.

Resulta importante destacar cómo, en este caso, el modelo multilingüe de *Voyage* logra un rendimiento comparable al modelo entrenado principalmente en inglés de *Cohere*.

Un aspecto relevante de este análisis fueron los resultados negativos de Tau obtenidos. Consideramos que esto puede deberse a que el rango temporal entre los distintos años de publicación de los libros en el dataset es relativamente bajo (además de los valores repetidos). Esta característica lleva al método a asignar valores numéricos similares, incrementando la probabilidad de que los elementos queden desordenados y, en consecuencia, disminuyendo la puntuación en la métrica de Tau. No obstante, al incorporar más información y aumentar la cantidad de pares evaluados, los modelos parecen adquirir una mejor comprensión del concepto asociado al objeto, lo que permite una organización más precisa de los valores y por consecuente, un valor de Tau mayor.

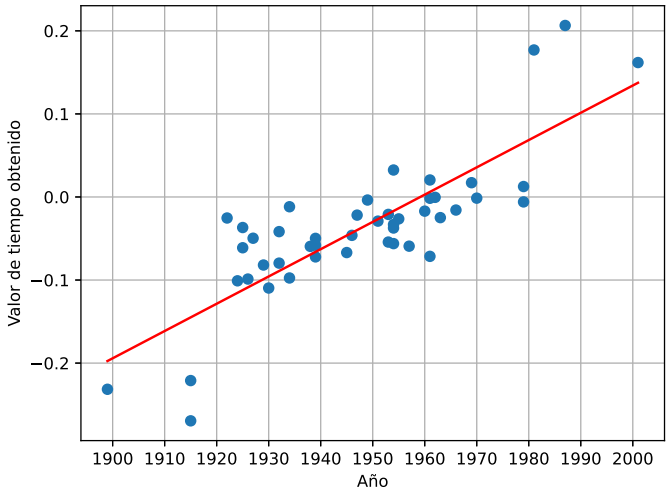

Fig. 4.5: Valores obtenidos mediante el modelo `voyage-large-2` utilizando tres pares semilla junto con el título y el autor. Cada punto representa un libro.

En la figura Figura 4.5 se puede ver el mejor resultado obtenido de todos los experimentos realizados, siendo los valores $\tau = 0{,}615$ y $r = 0{,}817$. El mismo pertenece al modelo `voyage-large-2` de *Voyage*, utilizando tres pares de semillas junto con el título y nombre del autor del libro. Aquí se puede observar lo comentado en el párrafo anterior. Si bien algunos libros que poseen el mismo año se encuentran cercanos (como los tres libros del año 1939), otros se encuentran más dispersos a pesar de pertenecer al mismo año de publicación (como los tres libros del año 1961). A pesar de todo esto, se puede apreciar una buena correlación, mayormente dictada por los puntos que se encuentran en los extremos.

4.2.5. Análisis de Mensajes de Apertura de Sesiones Ordinarias

En los dos experimentos anteriores se analizaron tanto un evento como un objeto. Ambos consistían en fragmentos de texto breves, redactados en el idioma inglés, que apuntaban a temas específicos donde no se tenía en cuenta el habla humana como tal. En este experimento buscaremos ir más allá, profundizando el análisis mediante el estudio de discursos en español, específicamente parte de los discursos presidenciales pronunciados en la apertura de las sesiones ordinarias del Congreso de la Nación Argentina.

Los discursos de apertura de sesiones ordinarias ante la Asamblea Legislativa constituyen una instancia anual en la que el presidente de la Nación se dirige a los integrantes del Poder Legislativo. Dichos discursos forman parte de un acto institucional de alta relevancia en las democracias representativas, ya que sintetizan y comunican las prioridades de gestión, las políticas públicas y las reformas legislativas que el Poder Ejecutivo pretende impulsar en el período anual entrante. Asimismo, estos discursos suelen incluir referencias a la problemática actual, problemas específicos, y menciones a gestiones gubernamentales previas.

La cuestión que se plantea es la siguiente: ¿Es posible que los embeddings de estos discursos (que tienen grandes cantidades de texto) contengan suficiente información para ordenarlos cronológicamente según el año en que fueron realizados? Para encontrar una respuesta a esta incógnita se analizó el dataset *aperturaSesionesOrdinarias*.

Cabe destacar que, tal como se menciona en la Sección 4.1, debido a las limitaciones de los modelos de *Cohere* (en relación a la cantidad máxima de tokens permitidos) estos no son capaces de procesar los extractos de los discursos en su totalidad. Para poder utilizarlos (y no dejarlos fuera del análisis) en este caso en particular se permitió el truncamiento por parte de la `API` (brindada por *Cohere*), configurándola para que tome la mayor cantidad de tokens que pueda desde el comienzo.

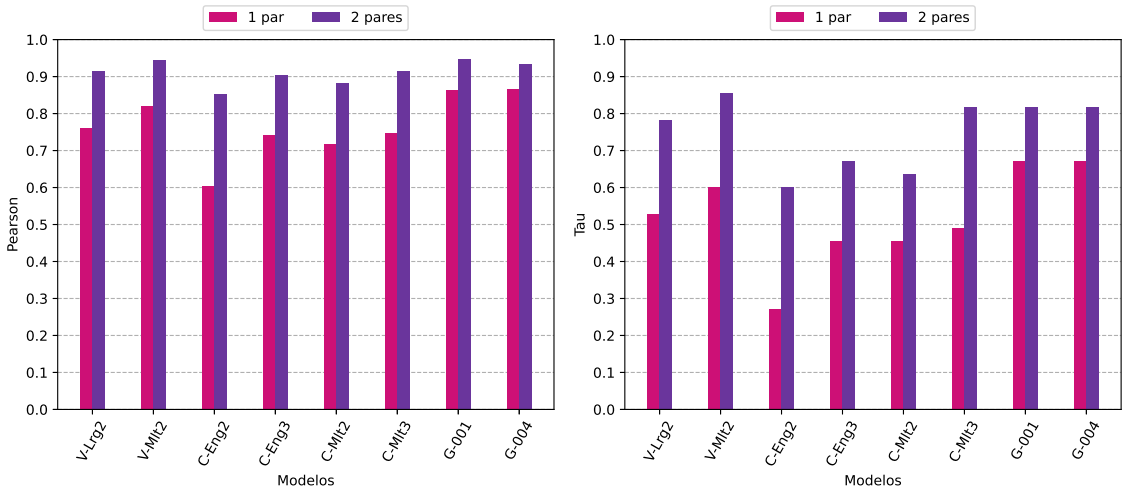

Fig. 4.6: Comparación de los modelos sobre los discursos de apertura. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos.

En la Figura 4.6 se presentan los resultados obtenidos en la experimentación utilizando un par semilla y dos pares semilla. Las semillas corresponden a los extremos, seguido de

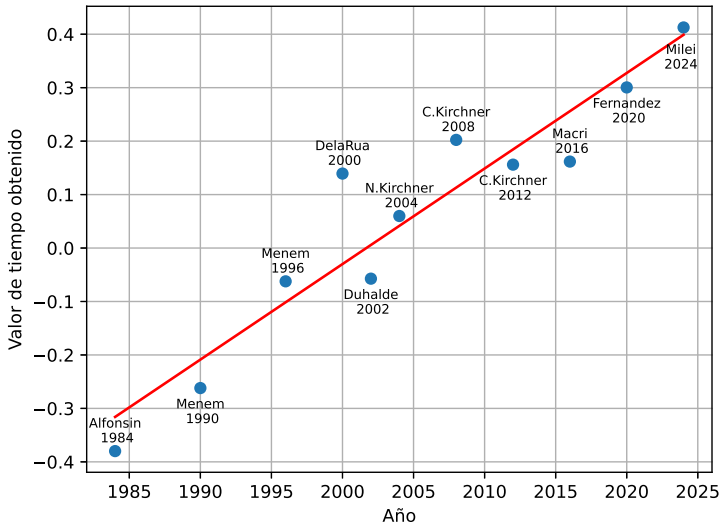

Fig. 4.7: Valores obtenidos mediante el modelo `voyage-multilingual-2` utilizando dos pares semilla. Cada punto representa un discurso de apertura.

los extremos de la subsección restante producto de ignorar los extremos anteriores, siendo estos el par *(Alfonsin-1984, Milei-2024)*, como primer semilla; y *(Menem-1990, Fernandez-2020)*, como segunda semilla en el caso donde se utilizaron dos pares. Es posible observar que, a pesar de trabajar con textos en idioma español y calcular los embeddings a partir de un volumen considerable de datos (comparado con los análisis anteriores), los modelos evidencian una capacidad para captar la noción de temporalidad. En particular, los modelos de *Cohere*, si bien recibieron menor cantidad de texto (el texto posee aproximadamente 1500 tokens y los modelos de *Cohere* soportan como máximo 512 tokens), mostraron buenos resultados en general. Incluso el modelo `embed-multilingual-v3.0` se encuentra a la altura de los que pudieron recibir todo el extracto del discurso. A su vez, el modelo `voyage-large-2`, a pesar de estar entrenado principalmente en inglés, se encuentra casi a la altura de los que mejores valores otorgaron.

Por otro lado, en la Figura 4.7 podemos observar este comportamiento con más detalle. Dicho gráfico corresponde a uno de los mejores resultados obtenidos, donde se puede apreciar cómo los puntos se ordenan de manera casi diagonal, mostrando una clara proporcionalidad directa. El mismo pertenece al modelo `voyage-multilingual-2` de *Voyage*, donde los valores obtenidos fueron $\tau = 0{,}855$ y $r = 0{,}944$. Resulta importante destacar que el resto de modelos de *Google* obtuvieron valores muy similares, así como el modelo `embed-multilingual-v3.0`, quien tal como se mencionó antes, no se quedó atrás.

4.2.6. Análisis de Extractos de Clarín

A lo largo del tiempo, la información sobre distintos acontecimientos, tanto a pequeña como a gran escala, ha sido transmitida a través de diversos medios. Uno de los canales que se ha mantenido constante durante décadas son las noticias publicadas en los diarios. En esta sección buscamos analizar si extractos de diarios en español otorgan información sobre el tiempo en el que fueron publicados, de forma tal de poder ordenarlos cronológicamente.

Para llevar a cabo este análisis, se optó por utilizar un enfoque poco convencional utilizando el dataset *extractosClarin*, el cual contiene extractos de titulares de diarios transcritos de forma manual al ser los mismos imágenes de diarios. A diferencia de los otros experimentos donde los datasets fueron armados con información proveniente de Wikipedia (la cual se asume fue utilizada para entrenar el modelo), que se encuentra más estructurada y accesible, este dataset presenta características que podrían dificultar el reconocimiento por parte del modelo.

Extracto del 25 de mayo de 1955	
Título	"Decidese Hoy la Política de Bonn"
Título y subtítulo	"Decidese Hoy la Política de Bonn. Adenauer Conferencia con sus Embajadores"

Tab. 4.4: Ejemplo de cada tipo de experimento de extractos de Clarín

Basándonos en los resultados obtenidos previamente, en este análisis se buscó evaluar si la incorporación de información adicional y la selección de semillas adicionales contribuían a mejorar los resultados. Para esto, se emplearon entre uno y dos pares semilla, y se añadió al título de cada noticia su respectivo subtítulo, con el objetivo de proporcionar un contexto más amplio[8]. En la Tabla 4.4 se puede ver un ejemplo de cómo está compuesto el texto de los experimentos.

La Figura 4.8 muestra los resultados en los distintos experimentos. En el gráfico de la izquierda se observan muy buenos resultados para los valores de Pearson. De forma consistente con los análisis anteriores realizados, el experimento que incluyó la mayor cantidad de pares junto con información adicional alcanzó los valores más altos. Sin embargo, en el gráfico de la derecha se observa un comportamiento opuesto con los valores de Tau. En este caso, el experimento con mayor cantidad de pares y con información adicional no logró tener los resultados más altos en todos los modelos. En algunos de ellos, el no brindar información extra o utilizar un único par otorgó mejores resultados.

Un aspecto relevante a destacar es la predominancia del modelo `voyage-large-2` sobre los demás, a pesar de haber sido entrenado principalmente en inglés. Este detalle también se observó en el análisis sobre los discursos de apertura presidenciales, al otorgar buenos resultados. Tras establecer contacto por correo electrónico con el equipo de Voyage, se obtuvo la siguiente aclaración: "`voyage-large-2` fue entrenado en todos los idiomas posibles; sin embargo, la cantidad de datos utilizados en idiomas distintos del inglés es significativamente menor en comparación con el modelo `voyage-multilingual-2`"[9]. Esto

[8] Dada la estructura abstracta de las primeras portadas de Clarín, fueron seleccionados los títulos y subtítulos de la considerada noticia principal

[9] Extracto traducido al español de la respuesta obtenida de Tengyu Ma, cofundador y CEO de *Voyage*.

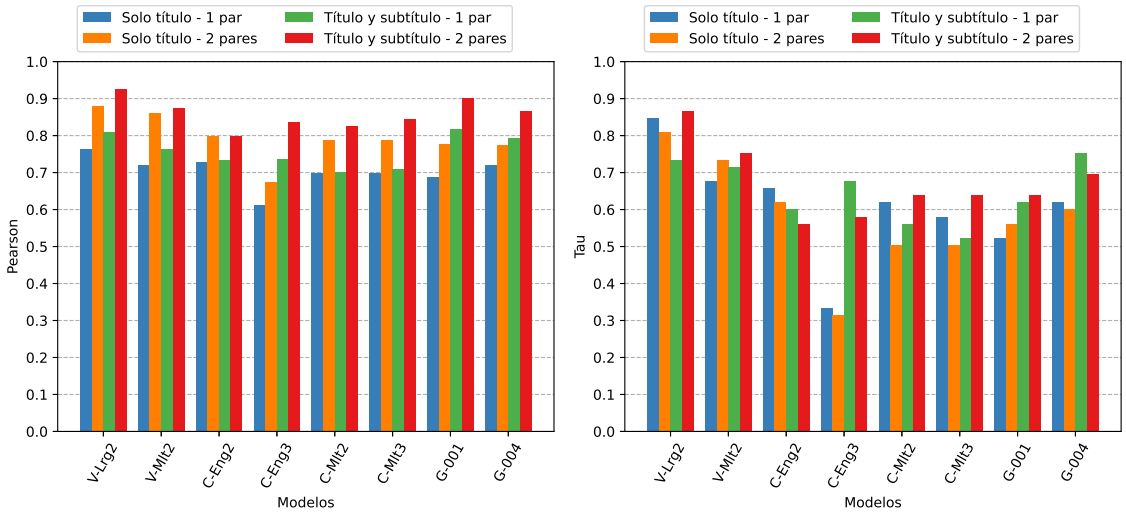

Fig. 4.8: Comparación de los diferentes experimentos de portadas de Clarín. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos.

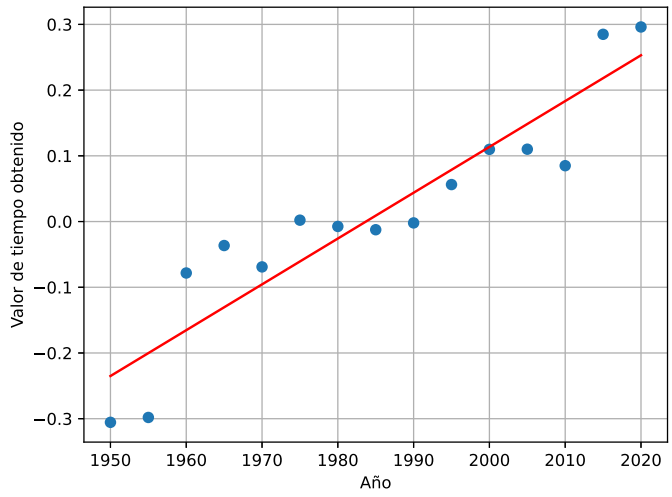

Fig. 4.9: Valores obtenidos mediante el modelo `voyage-large-2` utilizando título y subtítulo junto con dos pares semilla.

nos da a entender que los modelos tienen, en mayor o menor medida, un entrenamiento multilingüe, lo que justificaría los resultados obtenidos tanto en el análisis de los discursos presidenciales como en el análisis actual de las portadas de Clarín.

En la Figura 4.9 se puede observar el resultado obtenido con el modelo y el experimento que otorgó los valores más altos tanto de Tau como de Pearson, siendo estos $\tau = 0{,}867$ y $r = 0{,}926$. Dicho modelo fue el comentado anteriormente, `voyage-large-2`, utilizando dos pares y la información adicional brindada. Se puede observar cómo los puntos presentan una proporcionalidad directa, donde a medida que aumenta el año, los valores obtenidos por el método son mayores. Resulta importante destacar que solo fue calculado el embedding del texto, sin ninguna palabra ni pedido explícito adicional al igual que con el resto de los análisis.

4.2.7. Conclusión del Tiempo

En base a la experimentación con batallas históricas, obras literarias, discursos de apertura presidenciales y portadas de diarios, podemos concluir que existen indicios sólidos sobre la presencia de una dimensión temporal en los embeddings generados por las LLMs.

No hubo una predominancia clara de un modelo sobre otro a lo largo de los diferentes análisis; y si bien los experimentos revelaron indicios de una dimensión temporal, también destacaron las limitaciones inherentes al método utilizado. Por ejemplo: la dificultad para manejar conjuntos de datos con una alta densidad temporal, es decir, datasets cuya separación entre los distintos años seleccionados sea pequeña.

Observamos que proporcionar información adicional, así como un par adicional, contribuye a obtener mejores resultados. En particular, al analizar los experimentos relacionados con aperturas presidenciales y portadas de diarios, y compararlos con aquellos sobre batallas históricas y obras literarias, se evidenció que los primeros arrojaron mejores resultados. Suponemos que esto se debe a que, al brindarle más contexto (es decir, textos más extensos con información relevante), los LLMs logran modelar de manera más precisa la dimensión temporal.

A pesar de las limitaciones, los resultados obtenidos fueron sorprendentes en ciertos casos, como al lograr ordenar en gran medida los discursos presidenciales, a pesar de estar en español, incluir distintas fechas dentro del texto y contener volúmenes significativos de información. Esto es especialmente relevante considerando que el método se basa en la suposición de que los extremos difieran en una dimensión específica, la cual podría diluirse al calcular un embedding de un texto extenso.

Si bien la técnica implementada no proporciona valores con precisión absoluta, y tan solo establece un orden relativo entre los elementos, consideramos que estos hallazgos constituyen una base prometedora para explorar en mayor profundidad, permitiendo profundizar sobre posibles capacidades emergentes en estos y otros modelos.

4.3. Espacio

El espacio (del latín *spatium*) se puede definir como una entidad geométrica en la que interactúan los objetos físicos y en el que los sucesos que ocurren tienen una posición y dirección. Si bien esta definición es una de tantas (puesto que tiene varios significados y conceptos), en esta sección nos centraremos en el área o lugar que ocupa un objeto. En particular, su ubicación en el planeta Tierra.

Para lograr esto utilizaremos coordenadas geográficas decimales (también conocidos como grados decimales)[10]. Estas son un sistema de referencia que expresa las coordenadas de latitud y longitud como números decimales, permitiendo que cada ubicación en la Tierra sea especificada por estos valores.

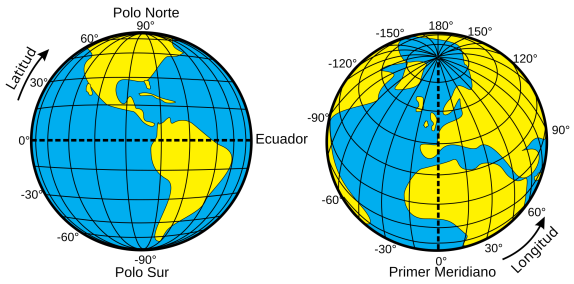

Fig. 4.10: La latitud es lineal; la longitud, por su naturaleza cíclica, no lo es.
Fuente: *Wikimedia Commons*

Resulta primordial destacar que, mientras la latitud se modela razonablemente como una magnitud lineal (Polo a Polo), la longitud no tiene la misma suerte, ya que forma un bucle continuo alrededor de la Tierra. A su vez, los valores de latitud presentan una mayor agrupación en comparación con los valores de longitud. Esta diferencia se debe a que la latitud está limitada a un rango de -90° a 90°, donde los extremos representan efectivamente los puntos más distantes entre sí (los polos). En contraste, la longitud abarca un rango de -180° a 180°, pero en este caso, los valores extremos no corresponden a puntos alejados, sino que representan la misma ubicación geográfica debido a la continuidad de la circunferencia terrestre (Figura 4.10).

4.3.1. Datasets

A continuación se describen en detalle los datasets utilizados en esta sección:

primeros20PaisesPBI. Dataset en inglés extraído de Wikipedia donde se seleccionaron los 20 países con mayor Producto Bruto Interno (PBI, o GDP por sus siglas en inglés) en millones de dólares según el Fondo Monetario Internacional (FMI) para el año 2024[11]. Los valores de latitud y longitud de cada país fueron extraídos de sus respectivas páginas (en inglés) de Wikipedia.

ciudadesPobladas. Dataset en inglés que se encuentra basado en el dataset `world_place` introducido en [12][12]. Si bien el dataset original incluye una amplia gama de ubicaciones como monumentos, edificios y restaurantes, se aplicó un criterio de filtrado para incluir únicamente lugares poblados con más de 2.5 millones de visitas (según la columna *page_views* del dataset).

lugaresDelMundo. Dataset en inglés donde se incluye un conjunto diverso de lugares geográficos y culturalmente significativos, asegurando una representación global amplia.

[10] `https://en.wikipedia.org/wiki/Decimal_degrees`
[11] `https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)`
[12] El dataset original está disponible en `https://github.com/wesg52/world-models/tree/main/data/entity_datasets`

Los nombres de los lugares se extrajeron de sus respectivas páginas de Wikipedia y se seleccionaron principalmente en base a dos criterios: su promedio de visitas a páginas de Wikipedia en los últimos 9 años[13] y su condición como lugares pertenecientes a "UNESCO World Heritage"[14]. Además, se añadieron dos lugares adicionales para actuar como "puntos extremos" para nuestro método. Estos son: *Svalbard Global Seed Vault*, en el caso de la latitud, y *Golden Gate Bridge*, en el caso de la longitud. Resulta importante destacar que la estatua del Cristo Redentor pertenece a: *'Carioca Landscapes between the Mountain and the Sea'*, patrimonio de la UNESCO.

4.3.2. Análisis de Países por PBI/GDP

La economía global está definida por una dinámica compleja de interacciones entre naciones, en la que el Producto Bruto Interno (PBI) de un país actúa como uno de los indicadores de su influencia y desarrollo. Los primeros 20 países por PBI representan una porción significativa de la riqueza y actividad económica mundial. En base a la importancia de estos países ¿Podrán sus embeddings otorgar suficiente información como para ordenarlos en función de su ubicación geográfica? Se analizó el dataset *primeros20PaisesPBI* para explorar esta posibilidad utilizando tan solo los nombres de los países.

En particular, nos proponemos analizar si los modelos incorporan el concepto de coordenadas geográficas en grados decimales para ordenar los países en función de su latitud y longitud. En este análisis asumiremos linealidad en la longitud, es decir, que los valores van desde -180 a 180. Resulta importante resaltar que los países abarcan extensas superficies sobre el planeta, lo que dificulta asociar su posición geográfica a un único punto específico. Sin embargo, para simplificar el análisis y establecer una representación estándar, se utiliza un punto central definido por coordenadas aproximadas al centro geográfico del país.

A diferencia de los análisis temporales, en los que se proporcionaba información adicional, en este análisis decidimos centrarnos puntualmente en la selección de semillas. Esto se debe a que el método utilizado busca que dichas semillas seleccionadas se distingan principalmente en relación al concepto a evaluar y que resulten similares en el resto de aspectos. A continuación examinaremos si resulta mejor seleccionar ciertas semillas por sobre otras.

Al analizar los resultados obtenidos en relación a la latitud, se puede observar en la Figura 4.11 que el par semilla que mejor desempeño presentó fue el conformado por los extremos de latitud, es decir, estar compuesto por los países con la latitud más baja y más alta de los seleccionados. Aunque el par *(Australia, Russia)* y el par *(Brazil, Canada)* presentan diferencias menores en longitud en comparación con el par *(Australia, Canada)*, este último obtuvo resultados superiores en la mayoría de los modelos evaluados. Por otra parte, los resultados más destacados entre todas las comparaciones se lograron utilizando los pares *(Australia, Canada)* y *(Brazil, Russia)*, correspondientes a los extremos y los extremos que quedan omitiendo los anteriores. Es importante señalar que no se incluyó el análisis con dos semillas que utiliza los pares *(Australia, Russia)* y *(Brazil, Canada)*. Si bien este doble par tiene menores diferencias en longitud, otorga los mismos resultados que el uso de las dos semillas previamente mencionadas. Esto se debe a la naturaleza del

[13] `https://pageviews.wmcloud.org/`
[14] `https://whc.unesco.org/en/list/`

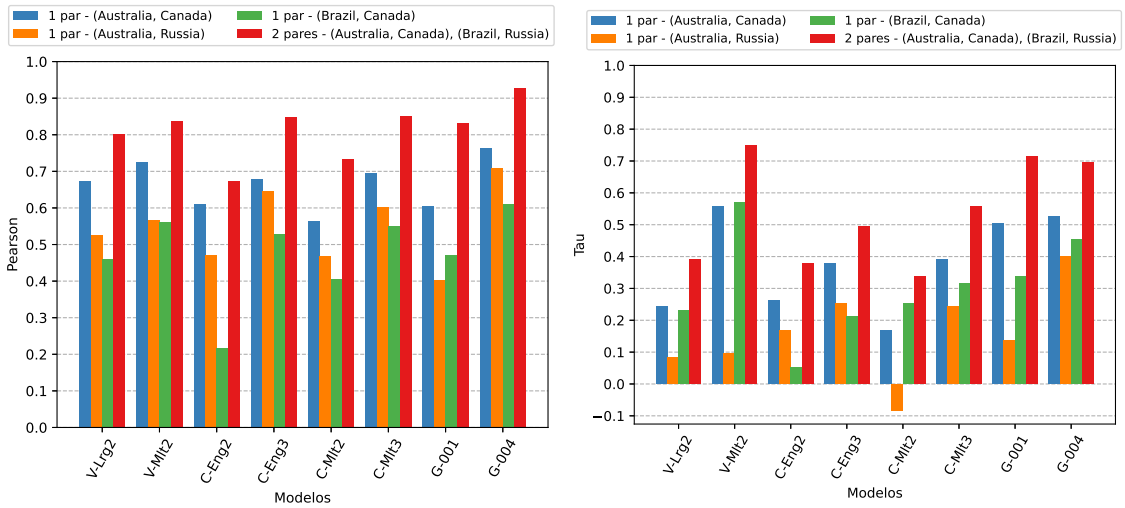

Fig. 4.11: Comparación de las diferentes semillas utilizadas al analizar la latitud. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos.

método de cálculo utilizado, que consiste en sumar las diferencias de cada par y luego dividir el resultado por el total de pares.

Un aspecto llamativo en el gráfico de la derecha es el valor de Tau, que resulta negativo. Sin embargo, dado que este valor es cercano a cero, simplemente refleja la ausencia de correlación en los datos.

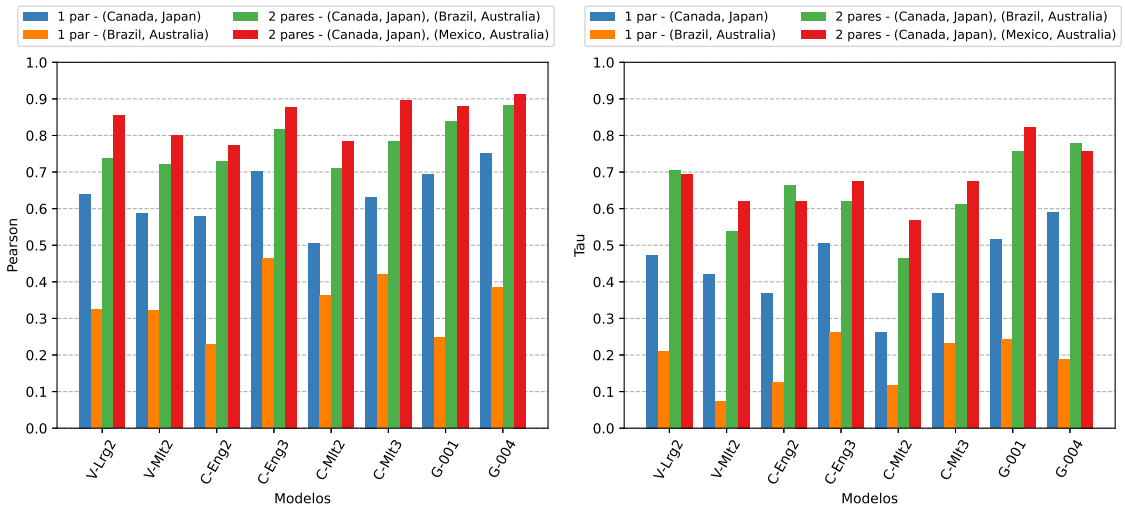

Fig. 4.12: Comparación de las diferentes semillas utilizadas al analizar la longitud. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos.

Al analizar la Figura 4.12 que presenta las comparaciones según la longitud, se puede observar que, al igual que con la latitud, el utilizar un solo par de países produce mejores resultados cuando está conformado por valores extremos de longitud. En particular, el par *(Canada, Japan)* mostró un desempeño superior al par *(Brazil, Australia)*, a pesar de

que el primero presenta una mayor diferencia en términos de latitud. De manera similar, al utilizar dos pares semilla se observó el mismo comportamiento. Aunque los resultados fueron similares, los pares formados por los extremos, y los extremos que quedan omitiendo los anteriores, obtuvieron las correlaciones de Pearson más altas en todos los modelos. En la mayoría de los casos, también presentaron mejores resultados en los valores de Tau. Los mejores desempeños se lograron con los pares *(Canada, Japan)* y *(Mexico, Australia)*. En contraste, los pares *(Canada, Japan)* y *(Brazil, Australia)*, que presentan menores diferencias en latitud, no lograron superar de manera consistente a los pares formados por los valores extremos.

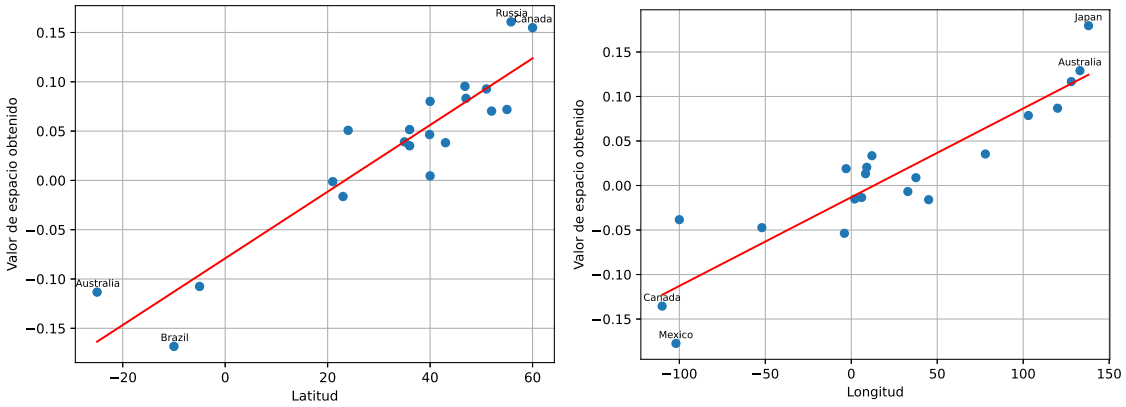

Fig. 4.13: Valores obtenidos sobre los valores reales. A la izquierda, la latitud. A la derecha, la longitud. En ambos casos el modelo utilizado fue `text-embedding-004` con dos pares semilla.

Producto de los resultados anteriores, podemos analizar un experimento y un modelo. En la Figura 4.13 se pueden observar los valores obtenidos por el método en comparación con los valores reales de latitud; a la izquierda y longitud; a la derecha. El modelo utilizado es `text-embedding-004` de *Google*, y el experimento corresponde a aquel con dos pares semilla, los cuales son extremos. Su selección se debe a su gran desempeño tanto en latitud como longitud. Los valores obtenidos fueron: $\tau = 0{,}695$ y $r = 0{,}928$ para la latitud, y $\tau = 0{,}758$ y $r = 0{,}913$ para la longitud. Ambos gráficos muestran una correlación positiva entre la ubicación geográfica (latitud y longitud) y el valor de espacio obtenido.

4.3.3. Análisis de Ciudades en base a la longitud

En la actualidad, algunas ciudades se han convertido en grandes ciudades, concentrando a millones de personas. Las ciudades no solo se definen por su población, sino también por su ubicación geográfica, que puede influir en su desarrollo, cultura y economía. Así como previamente se analizaron países, cuya posición geográfica es extensa ¿podrán los embeddings de ciudades contener información suficiente para ser ordenadas en función de su ubicación geográfica? Se evaluó esto con el dataset *ciudadesPobladas*.

En este análisis nos enfocaremos exclusivamente en la longitud. Resulta importante destacar que no se utilizó la longitud más baja como punto de partida. En su lugar, se optó por segmentar el conjunto de datos en tres divisiones, con el objetivo de analizar si los modelos logran una noción de orden de la longitud, y a su vez, si tienen una noción del

comportamiento circular de la misma. Esto se lleva a cabo utilizando ciudades, partiendo de la suposición de que los modelos enfrentarán una mayor dificultad para ubicarlas correctamente, dado que representan áreas geográficas más reducidas en comparación con los países. En particular, las dos primeras partes del conjunto de datos contienen valores de longitud organizados de forma ascendente. En contraste, la tercer parte presenta una transición de valores positivos a negativos, lo cual se debe al cruce del antimeridiano de Greenwich, donde la longitud cambia de 180° a -180°. Las tres partes están compuestas por aproximadamente 12 ciudades, donde los extremos de cada parte pertenecen a los que figuran a continuación: *Manhattan* a *Budapest (MAN-BUD)*, *Budapest* a *Tokyo (BUD-TOK)*, y *Tokyo* a *Manhattan (TOK-MAN)*. Utilizaremos estas abreviaciones para distinguir cada división de acá en adelante.

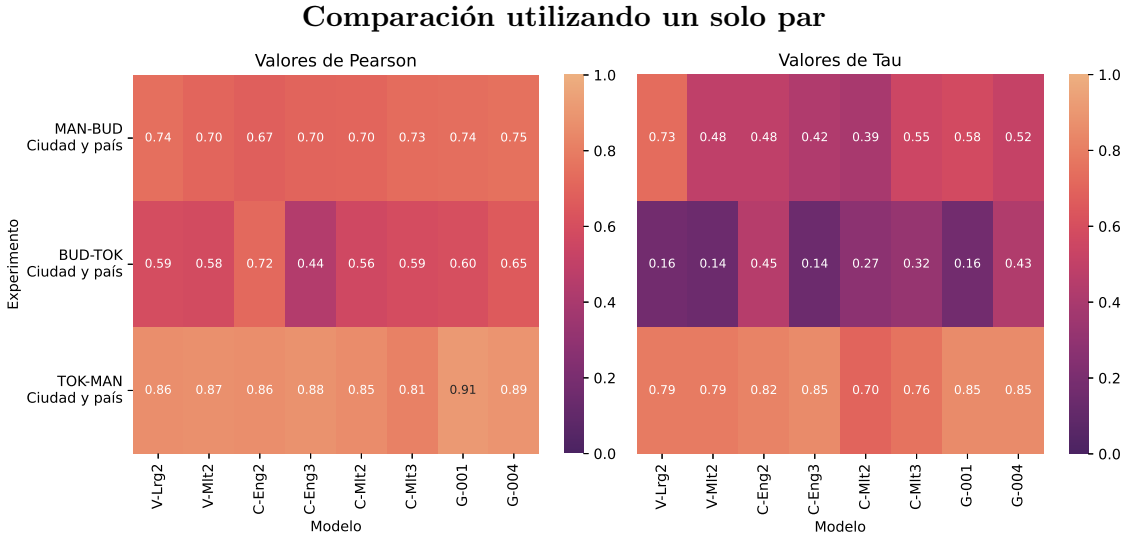

Fig. 4.14: Comparación de los experimentos sobre ciudades utilizando un par. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos en base a cada experimento y modelo utilizando la ciudad y el país al que pertenece.

En la Figura 4.14 se pueden observar los valores obtenidos por cada una de las divisiones del conjunto de datos. Los resultados fueron obtenidos utilizando un par semilla correspondiente a los extremos indicados en el párrafo anterior. Estos valores corresponden al experimento realizado con el nombre, y país, al que pertenece la ciudad (de la manera *"ciudad, país"*). En particular, en la división correspondiente a *BUD-TOK*, se pueden observar valores muy bajos de Tau. Estos se deben a que los mismos no se encuentran bien ordenados. En pos de mejorar esto, combinaremos lo realizado en análisis anteriores: agregaremos un par semilla adicional, y mantendremos el uso de la ciudad y el nombre del país.

En la Figura 4.15 se puede apreciar como, al utilizar dos pares, en casi todos los cuadrantes se obtienen valores iguales o mayores que aquellos obtenidos al utilizar un solo par (Figura 4.14). Entre todos los modelos el que mejor resultados brinda, en promedio entre los valores de Pearson y de Tau, resulta ser `embed-english-v2.0` de *Cohere*. Utilizaremos este modelo para visualizar mediante un triple gráfico de puntos (uno por cada una de las divisiones) el análisis de la longitud.

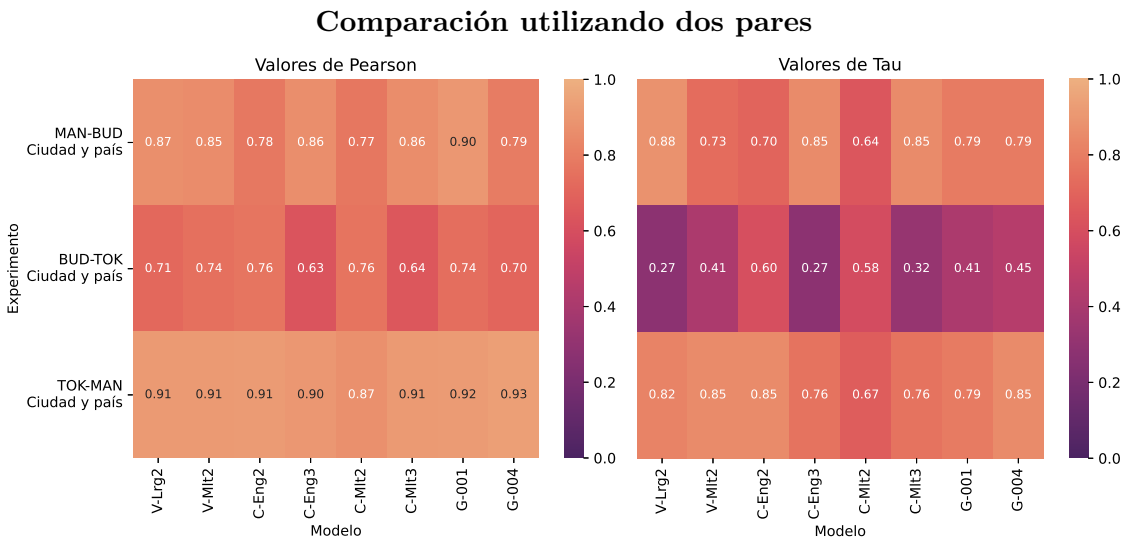

Fig. 4.15: Comparación de los experimentos sobre ciudades utilizando dos pares. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos en base a cada experimento y modelo utilizando la ciudad y el país al que pertenece.

En la Figura 4.16 se pueden observar los resultados de cada división, obtenidos por el modelo `embed-english-v2.0`, tal como se comentó previamente. Los dos gráficos superiores, aunque no perfectos, exhiben cierto orden espacial. En particular, en el gráfico que se encuentra más arriba se puede observar una representación razonable ($\tau = 0{,}697$ y $r = 0{,}777$), mientras que el gráfico del medio muestra una correlación un poco más débil ($\tau = 0{,}604$ y $r = 0{,}762$). Creemos que esta diferencia puede deberse a la proximidad de algunas de las ciudades, como por ejemplo: Maharashtra (72,82), Mumbai (72,8775) y Goa (73,83), así como Bavaria (11,3856), Venice (12,3358), Copenhagen (12,5683) y Berlin (13,405). En la Figura 4.17 se puede apreciar cómo entre las divisiones *MAN-BUD* y *BUD-TOK* los valores se encuentran agrupados con alta frecuencia.

Volviendo a la Figura 4.16, el gráfico que se encuentra más abajo presenta una correlación más fuerte ($\tau = 0{,}848$ y $r = 0{,}915$), lo que otorga indicios de que el modelo efectivamente podría estar capturando una representación circular de la longitud. Esto se debe a que logra ubicar bien los valores, a pesar de estar pasando de valores positivos a negativos (producto de pasar por el antimeridiano de Greenwich).

Algunas ciudades estaban muy cerca unas de otras y, por esta razón, se prefirió un orden numérico simple en el eje x de la Figura 4.16, de forma tal de mejorar la legibilidad, evitar una acumulación excesiva de puntos y prevenir la confusión al

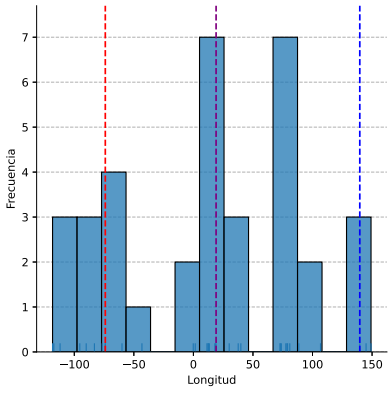

Fig. 4.17: Distribución de las ciudades. Las lineas representan las separaciones: en rojo, *Manhattan*; en violeta, *Budapest*; y en azul, *Tokyo*.

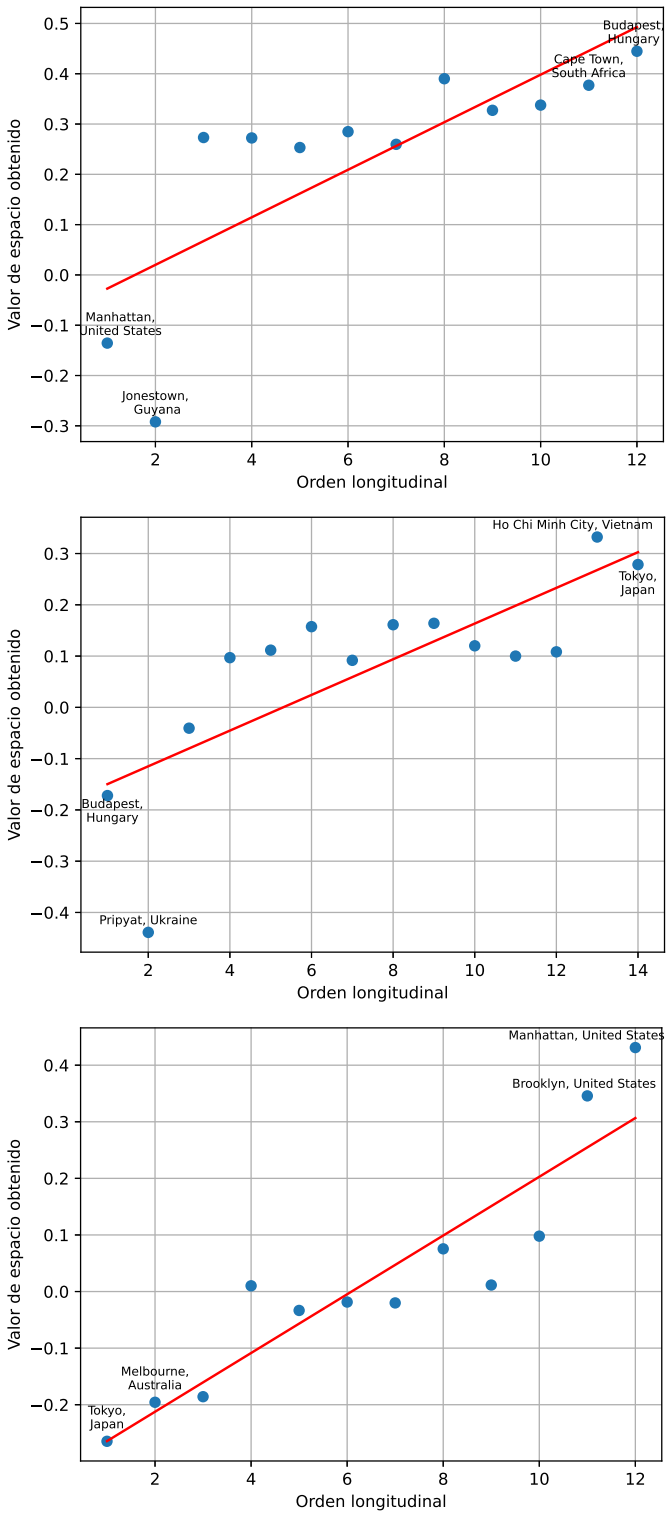

Fig. 4.16: Valores obtenidos según su orden longitudinal, para el modelo `embed-english-v2.0`. De arriba a abajo: la división *MAN-BUD*, la división *BUD-TOK*, y por último la división *TOK-MAN*.

tener valores positivos a la izquierda de los negativos, en el caso de la división *TOK-MAN*. A su vez, si bien los valores de Pearson fueron calculados con el orden numérico simple en lugar de los valores reales de longitud, estos fueron muy similares (y en todos los casos mayores). En el caso de la división *TOK-MAN*, al tener valores negativos y positivos muy alejados, se tomó la decisión de calcularlo manteniendo las distancias entre las ciudades, pero utilizando valores negativos.

4.3.4. Análisis de Lugares del Mundo

El mundo está lleno de lugares asombrosos, muchos de los cuales se encuentran protegidos bajo la designación de Patrimonio de la Humanidad, otorgada por la UNESCO. Esta organización se dedica a la preservación de lugares con un valor cultural, histórico y natural incalculable para la humanidad. Lugares como Machu Picchu, enclavado en los Andes, o la Acrópolis de Atenas, cuna de la democracia, son algunos ejemplos.

Entre estos lugares de interés a analizar, se destacan también otros lugares que no pertenecen a la UNESCO, como *Svalbard Global Seed Vault*. El agregado de estos lugares adicionales permite tener una gran diversidad de ubicaciones en diferentes partes del planeta Tierra. Surge entonces una pregunta de interés: considerando estos sitios ¿Es posible que sus embeddings puedan ser ordenados en función de su ubicación geográfica? Se analizó el dataset *lugaresDelMundo* en busca de esta respuesta.

Lugar - Potala Palace	
Solo nombre	"Potala Palace"
Primera oración de Wikipedia	"The Potala Palace is a dzong fortress in Lhasa, capital of the Tibet Autonomous Region in China"
Nombre y país	"Potala Palace, China"

Tab. 4.5: Ejemplo de cada tipo de experimento

En los análisis previos, tanto espaciales como temporales, se realizaron diversos experimentos. En este análisis buscamos continuar esto, con gran ambición, en pos de lograr obtener valores de correlación altos que nos permitan ubicar los lugares en un mapa a partir de los valores obtenidos por el método. Para ello se decidió llevar a cabo tres experimentos, siendo estos el cálculo de embeddings de: el nombre, el primer párrafo de Wikipedia, y el nombre seguido de una coma junto con el país al que pertenece cada lugar (Tabla 4.5). Esto tiene como objetivo comprobar si el agregado de información mejora o empeora los resultados, así como analizar si la precisión de esta información tiene un impacto significativo en los mismos.

En la Figura 4.18 se pueden observar los resultados obtenidos en relación a la latitud. En la parte superior se encuentra la comparación al utilizar el par compuesto por los extremos pertenecientes a los lugares *Cape Horn* y *Svalbard Global Seed Vault*. En el mismo se puede ver cómo en el caso de los valores de Pearson, todos los experimentos otorgaron valores parecidos. Sin embargo, con respecto a los valores de Tau, los experimentos que utilizaron el nombre del lugar y el país donde se encuentra el mismo otorgaron generalmente los mejores resultados.

Por otro lado, en la parte inferior se encuentra la comparación al utilizar dos pares,

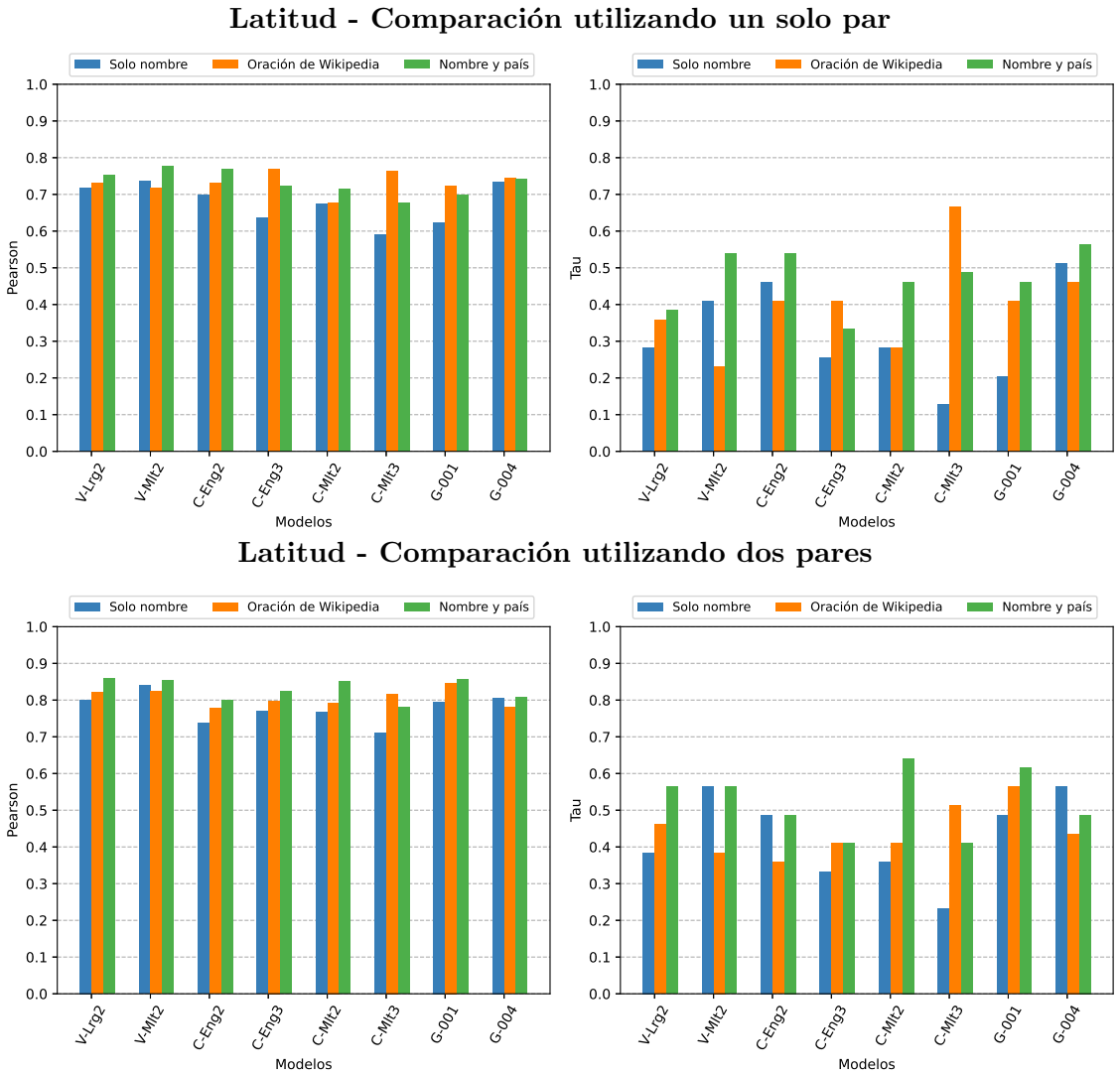

Fig. 4.18: Comparación de los diferentes experimentos al analizar la latitud. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos.

siendo estos el recién mencionado y el par compuesto por los lugares *Sydney Opera House* y *Stonehenge*, quienes son los extremos que restan si ignoramos a los anteriores. En este se puede ver cómo al igual que con un solo par, los valores de Pearson otorgados fueron similares entre los diversos experimentos, observándose tan solo una leve predominancia en el caso del uso del nombre de lugar junto con el país. En el caso de los valores de Tau, estos fueron variados. También, al igual que con un solo par, aquellos casos donde se utilizó el nombre del lugar y el país al que pertenece produjeron los mejores resultados.

Por otra parte, en la Figura 4.19 se pueden observar los resultados obtenidos en relación a la longitud. En la parte superior se encuentra la comparación al utilizar el par compuesto por los extremos pertenecientes a los lugares *Golden Gate Bridge* y *Sydney Opera House*. En el mismo se puede apreciar como los valores de Pearson son similares entre sí. En el caso de los valores de Tau se puede ver como estos, si bien fluctúan en menor medida que en la latitud, son diversos y no hay una clara predominancia.

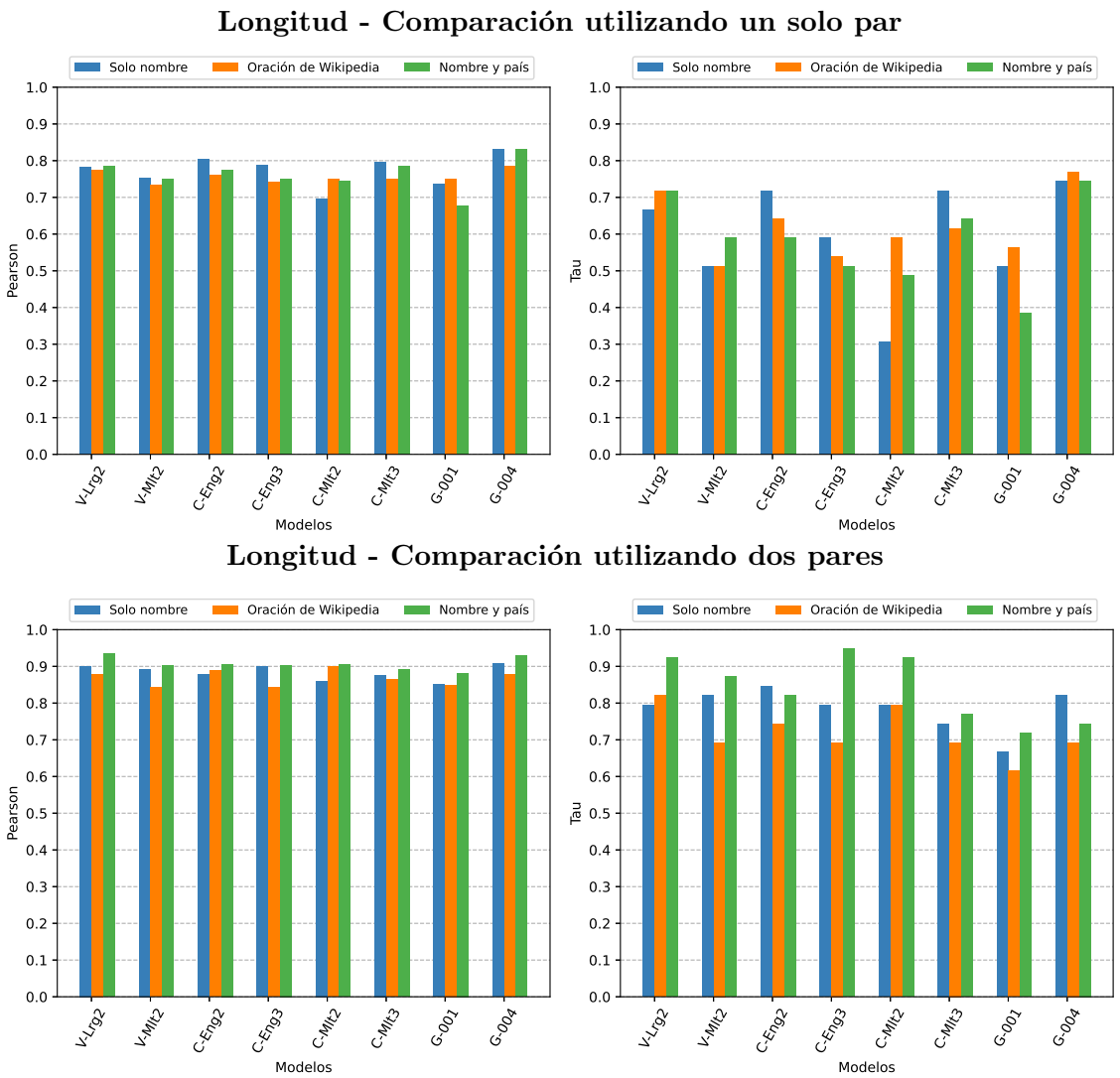

Fig. 4.19: Comparación de los diferentes experimentos al analizar la longitud. A la izquierda, los valores de Pearson obtenidos y, a la derecha, los valores de Tau obtenidos.

Al igual que en el caso de la latitud, en la parte inferior se encuentra la comparación al utilizar dos pares, siendo estos el par recién mencionado y el par compuesto por los lugares *Chichen Itza* y *Potala Palace*, quienes son los extremos siguientes a los anteriores. En este se puede ver como hay una leve predominancia del experimento que utiliza el nombre y país, tanto en el caso de los valores de Pearson como de Tau. Tanto en la latitud como en la longitud, se obtuvieron mejores valores al utilizar dos pares semilla y brindar información específica adicional (tan solo nombre y país), si bien el país al que pertenece cada monumento también estaba incluido (en todos los casos) en el primer párrafo de Wikipedia.

Un detalle que llama la atención son los valores de las métricas de longitud, los cuales resultan superiores a los de latitud. Podría parecer que esta comparación carece de relevancia; sin embargo, debido a la circularidad de la longitud, se esperaban valores inferiores en comparación con la latitud, que puede expresarse de manera lineal.

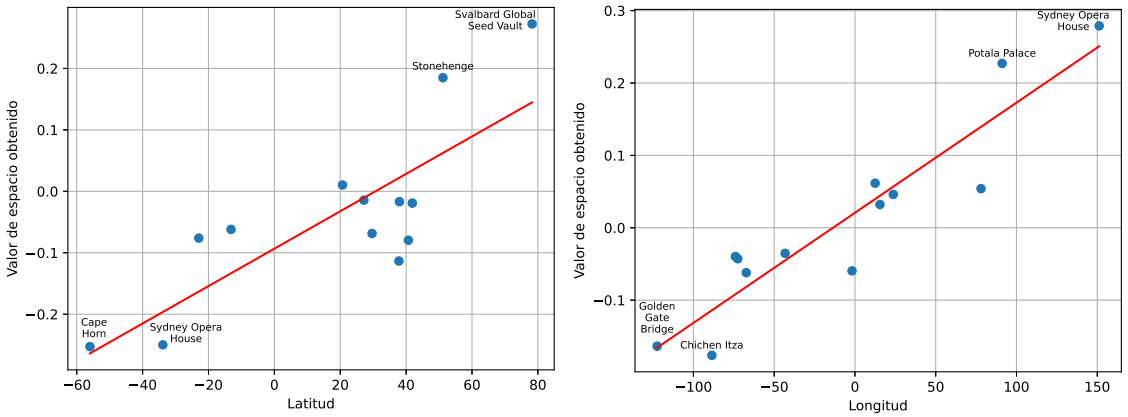

Fig. 4.20: Valores obtenidos sobre los valores reales. A la izquierda, la latitud. A la derecha, la longitud. El modelo utilizado fue `text-embedding-004` con dos pares semilla.

Antes de intentar graficar en un mapa los lugares, es importante notar cómo están distribuidos los puntos. En la Figura 4.20 se presenta un análisis detallado de los puntajes obtenidos por el algoritmo en comparación con los valores reales de latitud y longitud de los lugares analizados. Este fue realizado utilizando el modelo `text-embedding-004` de *Google* debido a los buenos valores obtenidos tanto en latitud como longitud. El experimento corresponde a aquel con dos pares semilla que utiliza el nombre y país. En el gráfico de la izquierda, se puede observar una correlación positiva entre los valores reales de latitud y las proyecciones generadas por el modelo. Los coeficientes de correlación obtenidos son $\tau = 0{,}487$ y $r = 0{,}807$. De manera similar, el gráfico de la derecha muestra una correlación positiva entre las longitudes reales y los valores proyectados. En este caso, los coeficientes de correlación obtenidos son $\tau = 0{,}744$ y $r = 0{,}928$, reflejando una relación aún más fuerte que en la latitud.

Si bien los valores obtenidos no son precisos y los mismos representan un orden relativo, se busca determinar si la combinación de los valores proyectados de latitud y longitud pueden ser representados en un mapa mundial. Para ello, se utiliza la herramienta *MinMaxScaler* de la biblioteca *scikit-learn*[15]. Esta herramienta permite escalar cada valor dentro de los rangos mínimos y máximos de latitud y longitud, respetando las restricciones geográficas de los valores utilizados. Este enfoque exploratorio tiene como objetivo evaluar la capacidad del modelo para aproximar coordenadas reales, a pesar de las limitaciones inherentes comentadas anteriormente.

Como se puede observar en la Figura 4.21, si bien casi todos los valores extremos son favorecidos debido a su selección en el proceso de escalado, es importante destacar ubicaciones como *Cape Horn* o el *Svalbard Global Seed Bank* que, a pesar de ser extremos en términos de latitud (es decir, venir "regalados" al ser los extremos utilizados al escalar), están posicionados con bastante precisión en sus coordenadas longitudinales correspondientes. De manera similar, sitios como *Christ the Redeemer (statue)* y *Machu Picchu* se encuentran ubicados muy cerca de sus ubicaciones reales. Sin embargo, otros lugares como la *Statue of Liberty* presentan un desplazamiento significativo. En la Figura 4.22 se puede apreciar la distribución de las distancias obtenidas al utilizar la fórmula del semiverseno para calcular las distancias entre el punto estimado y el punto real de cada lugar. El

[15] `https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

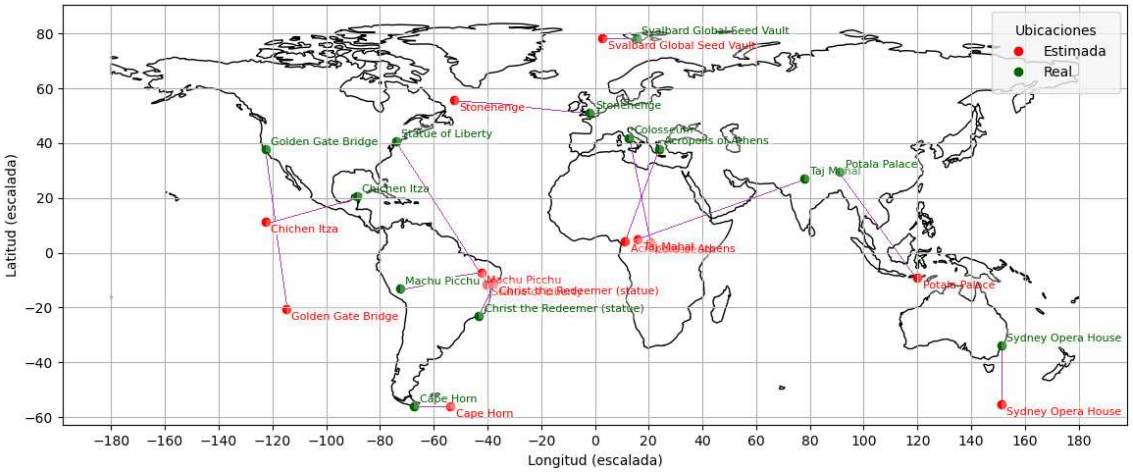

Fig. 4.21: Mapa construido al escalar los valores obtenidos en la Figura 4.20. Las lineas unen las ubicaciones reales con las estimadas.

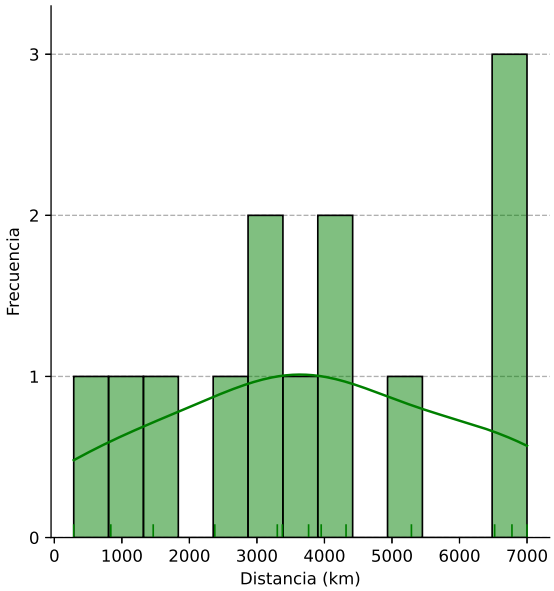

Fig. 4.22: Distancia entre los puntos, calculada mediante la fórmula del semiverseno.

promedio de distancia fue de 3789,55 km, mientras que la mediana de unos 3765,38 km.

4.3.5. Conclusión del Espacio

A partir de los diversos experimentos realizados tanto en ubicaciones puntuales (lugares, monumentos y ciudades) como en áreas extensas (países), podemos concluir que existen indicios de la presencia de una noción espacial en los embeddings generados por los LLMs. Esta noción parece incorporar una dimensión de latitud y longitud dentro del modelo, evidenciando una representación geográfica.

De manera similar a lo observado con el factor temporal, no se identificó una clara superioridad de un modelo sobre otro en los distintos experimentos realizados. En algunos

casos, los valores de Tau obtenidos fueron bajos, lo que sugiere un ordenamiento débil de las variables, aunque en otros casos se observaron correlaciones moderadas según el coeficiente de Pearson. Resulta importante volver a destacar que el método empleado proporciona valores numéricos que permiten establecer un orden relativo, y no valores puntuales sobre su latitud o longitud como tal.

Un aspecto relevante fue la obtención de mejores resultados cuando se utilizaron pares semilla pertenecientes a los extremos que diferían en mayor medida en otros aspectos, en comparación con aquellos pares semilla que presentaban diferencias aparentemente menores (a pesar de también estar cerca de los extremos). Esto se evidenció en el análisis de los países al obtener mejores resultados con pares semilla extremos por sobre los que, si bien estaban separados, diferían en menor medida en la longitud o latitud. Por ejemplo: en términos de longitud, el par *(Canada, Japan)* mostró un desempeño superior al par *(Brazil, Australia)*, a pesar de que *Brazil* y *Australia* difieren en menor medida en términos de latitud que *Canada* y *Japan*. Ocurrió de manera similar con la latitud. Suponemos que esto se debe a que el análisis se centró en conceptos específicos y concretos, en lugar de abarcar ideologías y opiniones, tal como se hizo en el estudio de [13]. Allí los autores destacan que los pares semilla no necesitan ubicarse en los extremos de la dimensión objetivo, que tan solo es necesario que difieran principalmente en lo que se busca analizar. Sin embargo, por lo experimentado, podemos observar como la selección de extremos otorga una mejor representación de la dimensión buscada.

5. CONCLUSIONES Y TRABAJO A FUTURO

5.1. Comparación entre modelos

Como aporte final, resulta relevante comparar el desempeño de cada modelo en los diversos análisis llevados a cabo. Para esto se seleccionó un experimento representativo de cada tipo de análisis llevado a cabo y se calcularon los coeficientes de Tau y Pearson correspondientes a cada modelo. A continuación se detalla el tipo de experimento utilizado para obtener dichos valores:

- **Análisis de Batallas** - Nombre y contexto, utilizando 2 pares.
- **Análisis de Libros** - Título y autor, utilizando 3 pares.
- **Análisis de Mensajes de Apertura** - Utilizando 2 pares.
- **Análisis de Extractos de Clarín** - Titulo y subtitulo, utilizando 2 pares.
- **Análisis de Países por PBI/GDP** - Extremos, utilizando 2 pares.
- **Análisis de Lugares del Mundo** - Nombre y país, utilizando 2 pares.

En el caso de los análisis espaciales, dado que noción de espacio nace a partir de la latitud y longitud, se promediaron los valores de Tau y Pearson para obtener los valores. El análisis basado únicamente en la longitud de ciudades fue excluido debido a la falta de la dimensión de latitud.

Las figuras Figura 5.1 y Figura 5.2 muestran la comparación de los valores de Tau y Pearson, respectivamente. En ambos gráficos se pueden apreciar las fortalezas y debilidades de cada modelo en las diferentes áreas analizadas. Incluso aquellos modelos pertenecientes a una misma empresa presentan diferencias significativas.

En el gráfico comparativo de Tau se puede observar una gran diferencia entre los modelos, donde no todos destacan en las mismas áreas. Por ejemplo, en el caso del modelo `embed-multilingual-v2.0`, el mismo alcanza el mayor valor en el análisis de lugares, mientras que `embed-english-v2.0` destaca en el análisis de batallas. Por otro lado, los modelos de *Voyage* y *Google* presentan una versatilidad en varios de los campos, alcanzando en algunos casos el mejor desempeño o situándose muy cerca de los modelos con el rendimiento más destacado.

En el caso del gráfico comparativo de Pearson sucede algo similar, aunque los valores son relativamente altos para todos los modelos, lo que indica un rendimiento consistente en gran parte de los análisis realizados.

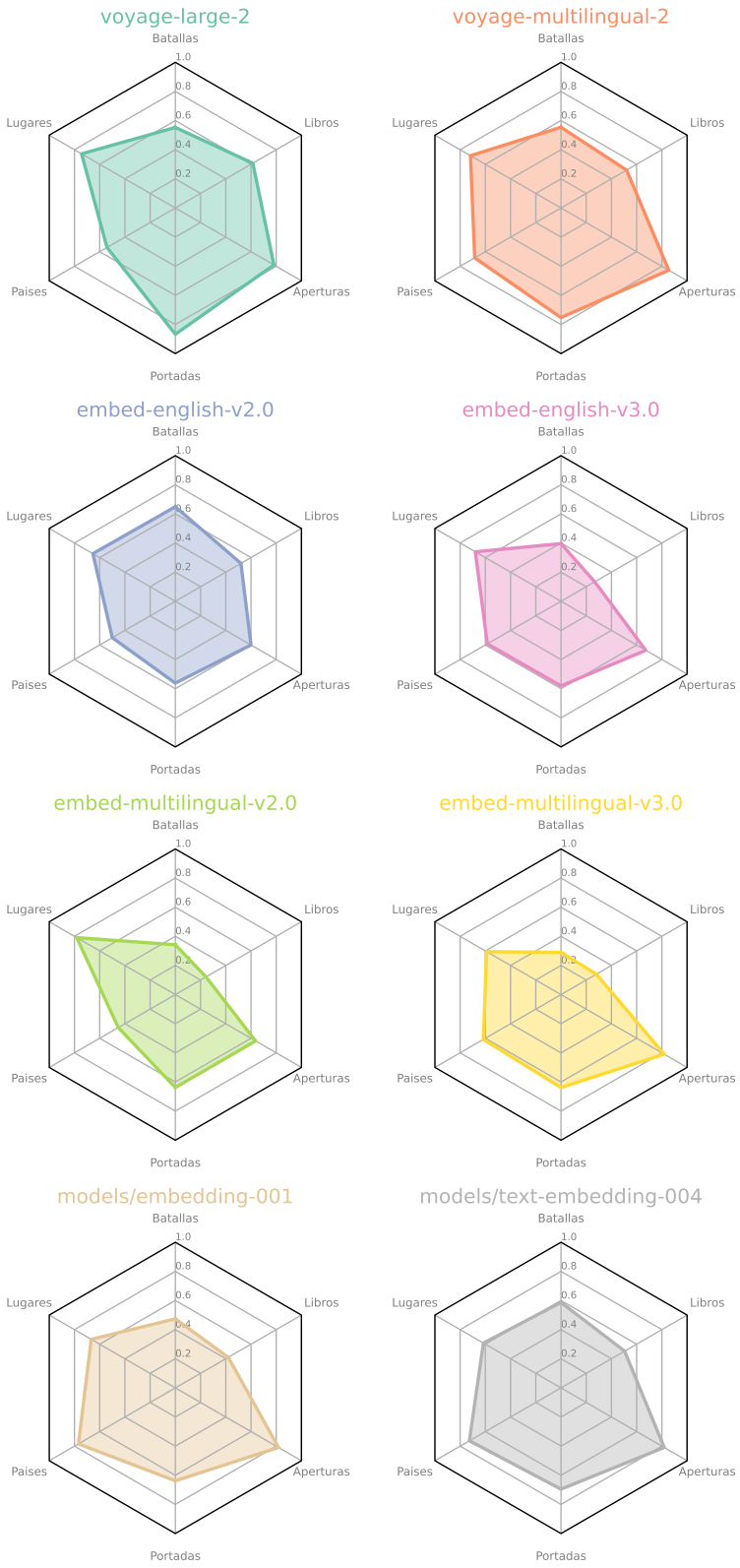

Fig. 5.1: Comparación entre todos los modelos según su valor de Tau.

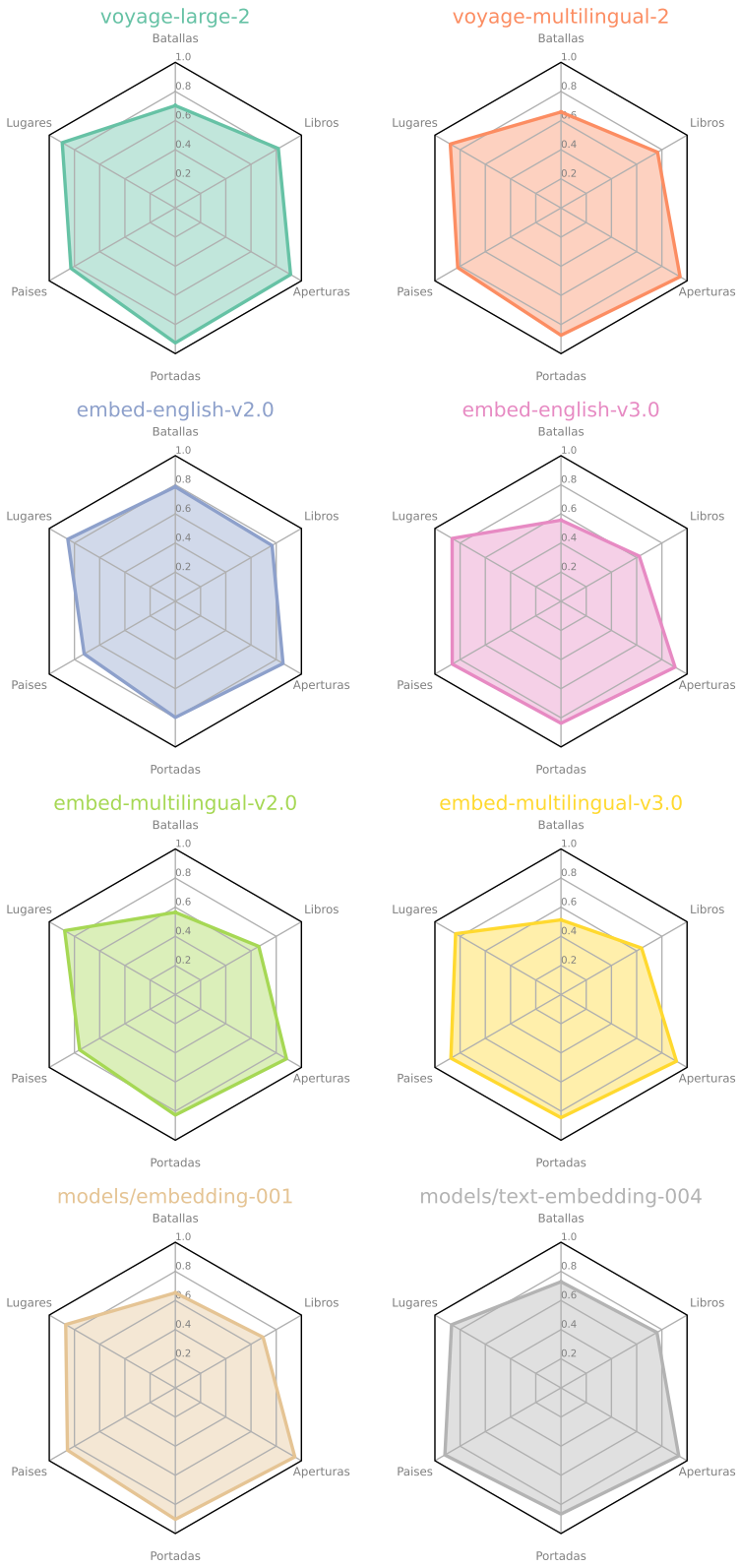

Fig. 5.2: Comparación entre todos los modelos según su valor de Pearson.

5.2. Conclusiones finales

Si algo queda claro al explorar el fascinante mundo de los LLMs, es que su capacidad de sorprendernos parece no tener límites. Hoy en día, su potencial completo sigue siendo un enigma. Sin embargo, este desconocimiento no es una limitación, sino una invitación: al tratarse de un área en constante desarrollo, cada avance en métodos, técnicas y cálculos matemáticos abre nuevas puertas y revela aplicaciones inesperadas.

En este trabajo, nos propusimos analizar los embeddings de una amplia variedad de modelos de LLMs, llevando a cabo diversos análisis y experimentos sobre el tiempo y el espacio. Estos análisis fueron variados, yendo desde batallas históricas hasta discursos de apertura presidenciales (por el lado del tiempo) y países hasta lugares y monumentos importantes (por el lado del espacio). La información utilizada provino de fuentes confiables, como Wikipedia, junto con páginas oficiales y trabajos de calidad académica de otros autores.

La respuesta a todas las preguntas que nos realizamos en cada análisis pueden ser respondidas con un si, en mayor o menor medida. A pesar de basarse en un método que únicamente establece un orden relativo entre los elementos, nuestros resultados son sólidos y presentan un interés significativo. Estos hallazgos se obtuvieron mediante una técnica sencilla pero efectiva, que permite proyectar y cuantificar la puntuación de una palabra o texto a lo largo de distintas dimensiones. Los diversos análisis y experimentos llevados a cabo utilizando el método de Waller et al. [13] sugieren que el tiempo y el espacio están intrínsecamente embebidos en los embeddings de estos modelos. Se pudo observar cómo el brindar información adicional específica y más de un par semilla, suele otorgar mejores valores, aunque hay casos y modelos donde esto no siempre es así. Este hallazgo sobre los modelos es significativo, ya que sugiere que estos poseen la capacidad de comprender y aprovechar las relaciones temporales y geoespaciales, lo que puede mejorar su aplicación en diversos dominios.

El debate sobre si los LLMs desarrollan “capacidades emergentes” o si su desempeño es simplemente el resultado de una combinación de aprendizaje en contexto, memoria de modelos y conocimiento lingüístico, sigue sin resolverse. Sin embargo, sostenemos que la evidencia presentada en este trabajo sugiere una inclinación hacia la existencia de dichas “capacidades emergentes”. Esto se debe a que los modelos utilizados no fueron entrenados específicamente para las tareas evaluadas y únicamente se realizaron operaciones sobre los embeddings de los objetos y eventos, sin formular preguntas ni pedidos explícitos.

Este trabajo proporciona más evidencia de que los LLMs generan representaciones fundamentadas del mundo real, tal como encontraron Gurnee y Tegmark [12]. Además, el método puede extenderse para interpretar cómo los LLMs representan otros conceptos seleccionando diferentes semillas, lo que permite una exploración más amplia de las representaciones internas del modelo.

5.3. Trabajo Futuro

Tal como se comentó en la conclusión, los resultados de este trabajo no solo aportan evidencia al campo de estudio, sino que también invitan a continuar investigando las capacidades aún inexploradas de estos modelos, cuyas posibilidades parecen expandirse tan rápido como nuestro entendimiento de ellos.

Dado el constante desarrollo en el mundo de los LLMs, resulta pertinente considerar la posibilidad de evaluar nuestras propuestas utilizando nuevos modelos. Esto permitiría analizar si los hallazgos obtenidos se mantienen, mejoran o presentan variaciones significativas.

A su vez, para mejorar el análisis del tiempo, podrían utilizarse eventos que hayan sucedido antes del año número uno (antes de Cristo). Por el lado del espacio, se podría buscar y analizar si el modelo tiene otra concepción del espacio (es decir, una representación geoespacial diferente), por ejemplo, en coordenadas polares. En [21] se muestran, entre otras cosas, como existe una representación circular de los días de la semana y del mes. Esto resulta de gran importancia en la búsqueda de un análisis más profundo y complejo, cuyas ideas podrían tomarse en búsqueda de una posible representación circular de la longitud.

Entre otros temas a continuar investigando, se podrían realizar nuevos experimentos con conjuntos de datos de distintas granularidades. Por otro lado, también se podría ahondar en la exploración de combinaciones de múltiples pares semilla y el análisis de diferentes maneras de brindar información adicional. Por ejemplo, durante el transcurso de la experimentación al calcular los embeddings de libros, se obtuvieron resultados variados al utilizar "," en lugar de "*by*" al agregar el autor del libro. Ocurrió que mientras que algunos modelos mejoraban considerablemente, otros empeoraban a pesar de cambiar tan solo un pequeño detalle. Esto deja en evidencia la sensibilidad de los modelos a variaciones en la formulación de la entrada (*wording*), lo que resalta la importancia de considerar cuidadosamente cómo se estructuran los datos de entrada.

Bibliografía

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space" *arXiv preprint arXiv:1301.3781*, vol. 3781, 2013.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need" *Advances in Neural Information Processing Systems*, 2017.

[3] D. Bahdanau, "Neural machine translation by jointly learning to align and translate" *arXiv preprint arXiv:1409.0473*, 2014.

[4] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding" *arXiv preprint arXiv:1810.04805*, 2018.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners" *arXiv preprint arXiv:2005.14165*, 2020.

[6] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "GPT-4 technical report" *arXiv preprint arXiv:2303.08774*, 2023.

[7] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, *et al.*, "Gemini: a family of highly capable multimodal models" *arXiv preprint arXiv:2312.11805*, 2023.

[8] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4" *arXiv preprint arXiv:2303.12712*, 2023.

[9] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, *et al.*, "Emergent abilities of large language models" *arXiv preprint arXiv:2206.07682*, 2022.

[10] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations" in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

[11] S. Lu, I. Bigoulaeva, R. Sachdeva, H. T. Madabushi, and I. Gurevych, "Are emergent abilities in large language models just in-context learning?" *arXiv preprint arXiv:2309.01809*, 2023.

[12] W. Gurnee and M. Tegmark, "Language models represent space and time" *arXiv preprint arXiv:2310.02207*, 2023.

[13] I. Waller and A. Anderson, "Quantifying social organization and political polarization in online platforms" *Nature*, vol. 600, no. 7888, pp. 264–268, 2021.

[14] F. Demarco, J. M. O. de Zarate, and E. Feuerstein, "Measuring ideological spectrum through NLP" in *NL4AI@ AI* IA*, 2023.

[15] M. M. Louwerse and R. A. Zwaan, "Language encodes geographical information" *Cognitive Science*, vol. 33, no. 1, pp. 51–73, 2009.

[16] M. M. Louwerse and N. Benesh, "Representing spatial structure through maps and language: Lord of the rings encodes the spatial structure of middle earth" *Cognitive science*, vol. 36, no. 8, pp. 1556–1569, 2012.

[17] A. G. Cohn and R. E. Blackwell, "Evaluating the ability of large language models to reason about cardinal directions" *arXiv preprint arXiv:2406.16528*, 2024.

[18] B. Liétard, M. Abdou, and A. Søgaard, "Do language models know the way to rome?" *arXiv preprint arXiv:2109.07971*, 2021.

[19] M. G. Kendall, "A new measure of rank correlation" *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[20] M. G. Kendall, "The treatment of ties in ranking problems" *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[21] J. Engels, E. J. Michaud, I. Liao, W. Gurnee, and M. Tegmark, "Not all language model features are linear" *arXiv preprint arXiv:2405.14860*, 2024.