EVALUACIÓN DEL USO DE MODELOS DE LENGUAJE DE GRAN ESCALA (LLM) EN LA TOMA DE DECISIONES POLÍTICAS: ESTUDIO DE CASO DE PERÚ EN EL CONSEJO DE SEGURIDAD DE LA ONU (2018)

EVALUATION OF THE USE OF LARGE LANGUAGE MODELS (LLM) IN POLITICAL DECISION-MAKING: A CASE STUDY OF PERU IN THE UNITED NATIONS SECURITY COUNCIL (2018)

Autor: *Giulliana Reggiardo Palacios**

RESUMEN

Este trabajo evalúa la aplicación de un Modelo de Lenguaje de Gran Escala (LLM, por sus siglas en inglés) en el contexto de la toma de decisiones políticas en el Consejo de Seguridad de las Naciones Unidas (CSNU), enfocado en el año 2018, cuando Perú fue miembro no permanente. Se emplea el marco metodológico del *benchmark* UNBench, que considera cuatro tareas clave: (1) Juicio de coautoría (co-penholder), (2) Simulación de votación por país, (3) Predicción de adopción de borradores de resolución, y (4) Generación de declaraciones diplomáticas. Los resultados obtenidos mediante la implementación del modelo LLaMA 3.3-70B muestran fortalezas y debilidades en su capacidad para replicar dinámicas multilaterales reales.

Palabras clave: Modelos de Lenguaje de Gran Escala - Consejo de Seguridad de la ONU - diplomacia multilateral - inteligencia artificial - UNBench - Perú - resoluciones de Naciones Unidas - simulación política - aprendizaje automático - toma de decisiones.

(*) Diplomática de carrera, actualmente se desempeña como funcionaria en la Embajada del Perú en Uruguay y de la Representación Permanente del Perú ante la ALADI y el Mercosur. Asimismo, es docente en la Universidad Peruana de Ciencias Aplicadas (UPC). Es Magíster en Relaciones Internacionales y Diplomacia por la Academia Diplomática del Perú "Javier Pérez de Cuéllar" y Magíster en Relaciones Internacionales con mención en Cooperación Internacional para el Desarrollo por la Universidad Andina Simón Bolívar (Ecuador). Licenciada en Comunicación Audiovisual por la Universidad Complutense de Madrid. **ORCID:**https://orcid.org/0000-0002-6784-5388 **Correo electrónico:** greggiardop@gmail.com

ABSTRACT

This paper evaluates the application of a Large Language Model (LLM) in the context of political decision-making within the United Nations Security Council (UNSC), focusing on the year 2018, when Peru served as a non-permanent member. The methodological framework of the UNBench benchmark is employed, which comprises four key tasks: (1) Co-authorship judgment (co-penholder selection), (2) Country-level voting simulation, (3) Draft resolution adoption prediction, and (4) Generation of diplomatic statements. The results obtained using the LLaMA 3.3-70B model reveal both strengths and limitations in its ability to replicate real-world multilateral dynamics.

Keywords: *Large Language Models - United Nations Security Council - multilateral diplomacy - artificial intelligence - UNBench - Peru - UN Resolutions - political simulation - machine learning - decision-making.*

1. INTRODUCCIÓN

Según Amazon Web Services (AWS), los modelos de lenguaje de gran tamaño, también conocidos como LLM, son modelos de aprendizaje profundo muy grandes que se preentrenan con grandes cantidades de datos. El transformador subyacente es un conjunto de redes neuronales que consta de un codificador y un decodificador con capacidades de autoatención. El codificador y el decodificador extraen significados de una secuencia de texto y comprenden las relaciones entre las palabras y las frases que contiene.

La participación de Perú en el Consejo de Seguridad de las Naciones Unidas (ONU) durante 2018 ofrece un escenario idóneo para analizar el rendimiento de los LLM en decisiones políticas reales.

El Consejo de Seguridad emite resoluciones con alto impacto internacional, lo que hace que la simulación y predicción de su comportamiento sea un desafío relevante para la ciencia política y la inteligencia artificial.

En trabajos previos se ha propuesto UNBench (Liang et al., 2025) como primer *benchmark* integral para evaluar la capacidad de los LLM en el contexto de la ONU. Si bien el referido estudio muestra, a través de 30 casos, el despemeño de los LLM en las tareas designadas, este estudio se centra en la experiencia peruana, aprovechando datos de borradores y votaciones efectivamente realizados durante el 2018 y realizando una aplicación de la propuesta mencionada a fin de evaluar su efectividad.

2. MARCO TEÓRICO

El rápido avance de la inteligencia artificial (IA) y, en particular, de los modelos de lenguaje de gran escala (*Large Language Models*, LLM), ha generado un impacto la forma en que los investigadores analizan los fenómenos políticos internacionales. Estas herramientas no solo posibilitan el procesamiento de grandes volúmenes de información textual, sino que además ofrecen nuevas vías para explorar los patrones discursivos y las estructuras argumentativas que subyacen en la toma de decisiones multilaterales. Su incorporación en el estudio de la política internacional requiere una reflexión teórica que permita comprender de qué modo los LLM pueden integrarse en los paradigmas clásicos de las relaciones internacionales y en la evolución metodológica de la disciplina.

2.1. Paradigmas de análisis de la toma de decisiones en política internacional

Desde mediados del siglo XX, los enfoques conductuales introdujeron una preocupación sistemática por el análisis empírico de la conducta estatal. Los trabajos de Herbert Simon (1957) sobre la racionalidad limitada y los estudios de Graham Allison (1971) sobre la Crisis de los Misiles de Cuba marcaron un hito al proponer modelos que explican cómo los procesos decisionales no responden únicamente a cálculos racionales perfectos, sino a estructuras organizacionales, percepciones y restricciones cognitivas. Este enfoque, conocido como *behavioral realism* o realismo conductual, supuso una apertura metodológica dentro del campo de las relaciones internacionales al incorporar el análisis de datos verificables como vía para inferir patrones de comportamiento estatal.

En este marco, los LLM pueden entenderse como una prolongación tecnológica de esa tradición conductual. Así como el análisis empírico buscaba correlacionar variables observables —por ejemplo, los patrones de voto, las alianzas o los discursos— con los resultados de política exterior, los modelos de lenguaje actuales permiten identificar con mayor precisión las asociaciones semánticas y narrativas que configuran el posicionamiento diplomático de los Estados. Su aplicación a contextos multilaterales, como el Consejo de Seguridad de las Naciones Unidas, constituye una herramienta contemporánea para estudiar la racionalidad y los incentivos estratégicos detrás de las decisiones.

2.2. Tecnología, poder y gobernanza global

El papel de la tecnología en las relaciones internacionales trasciende su carácter instrumental. Autores como Floridi (2014) y Bostrom (2017) sostienen que la inteligencia artificial está reconfigurando las dinámicas de poder, autoridad y legitimidad en el sistema internacional, al introducir nuevos actores —plataformas tecnológicas, corporaciones digitales y algoritmos de decisión— capaces de influir en la producción y circulación de información. Desde esta perspectiva, los LLM no deben ser considerados únicamente como herramientas analíticas, sino también como agentes estructurales de la gobernanza global, en la medida en que mediatizan la interpretación de los discursos y las agendas multilaterales. Sin embargo, en el presente trabajo, si bien no se profundiza sobre este último punto, se considera relevante entender a la capacidad de los LLM y la necesidad de mayor estudio y análisis sobre su uso y aplicación en la política internacional.

2.3. La dimensión constructivista y el análisis del discurso

El constructivismo ofrece un marco complementario para comprender la relevancia de los LLM en el estudio de la política internacional. Según Wendt (1999), los intereses y las identidades de los Estados se construyen socialmente a través del lenguaje y la interacción, es decir, el relacionamiento. En consecuencia, el discurso diplomático constituye no solo un reflejo de la realidad política, sino un mecanismo de su producción. El uso de LLM para analizar discursos multilaterales permite identificar patrones narrativos, metáforas recurrentes y estrategias de *framing* que dan forma a las percepciones mutuas entre actores. Así, el análisis computacional del lenguaje aporta evidencia empírica al estudio constructivista de la política exterior, al sistematizar y cuantificar los significados que circulan en la esfera internacional.

2.4. Los LLM como herramienta metodológica en la investigación de relaciones internacionales

El *text-as-data approach*, consolidado en la última década, ha ampliado las posibilidades de la investigación empírica mediante el uso de algoritmos para procesar información textual (Grimmer & Stewart, 2013). En este sentido, los LLM representan un salto cualitativo respecto a las técnicas tradicionales de análisis de contenido, al incorporar la comprensión semántica profunda y la capacidad contextual propia de los modelos de aprendizaje profundo (*deep learning*). Su aplicación en el análisis de documentos diplomáticos, resoluciones o actas de votación permite detectar correlaciones que antes requerían largos procesos de codificación manual, contribuyendo así a la validación empírica de hipótesis teóricas sobre cooperación, conflicto o liderazgo internacional.

En el ámbito de la gobernanza global, estas herramientas pueden revelar dinámicas de coordinación o disenso entre Estados, así como anticipar alineamientos discursivos previos a la formación de coaliciones o vetos. Por ejemplo, en el caso del Perú durante su participación en el Consejo de Seguridad en 2018, el procesamiento automatizado de discursos y votos podría evidenciar cómo los patrones de argumentación reflejan afinidades temáticas con determinados bloques o principios de política exterior.

2.5. Articulación entre teoría y evidencia empírica

La incorporación de los LLM en el análisis internacional no implica sustituir los marcos teóricos tradicionales, sino complementarlos con evidencia sistemática y replicable. La utilidad de estas herramientas depende de su capacidad para dialogar con las corrientes teóricas de las relaciones internacionales: desde el realismo, que enfatiza la distribución del poder; pasando por el liberalismo institucionalista, que resalta la cooperación y las reglas; hasta el constructivismo, que pone el acento en las ideas y narrativas. El aporte metodológico de los LLM radica en su potencial para operacionalizar conceptos abstractos —como identidad, legitimidad o influencia— en indicadores medibles a partir del lenguaje.

"La intersección entre teoría clásica y tecnología avanzada constituye una nueva frontera para el estudio empírico de la toma de decisiones en la gobernanza global."

Es así que se reconoce la convergencia entre la tradición conductual y la innovación tecnológica como el eje que sustenta la presente investigación. El análisis automatizado del lenguaje no solo amplía las herramientas de observación del comportamiento diplomático, sino que también permite repensar las categorías con las que la disciplina conceptualiza la acción estatal en un entorno interdependiente y mediado por la información. La intersección entre teoría clásica y tecnología avanzada constituye, por tanto, una nueva frontera para el estudio empírico de la toma de decisiones en la gobernanza global.

3. JUSTIFICACION

En el contexto del rápido avance de la inteligencia artificial, los Modelos de Lenguaje de Gran Escala (LLM) destacan como herramientas con gran potencial para trans-

formar diversos campos, incluida la política internacional. Aunque su aplicación es común en áreas como educación o medicina, su uso en el estudio de dinámicas diplomáticas bilaterales o multilaterales aún está en una etapa inicial.

De igual manera, es fundamental considerar las implicancias económicas que pueden afectar el funcionamiento de organizaciones internacionales como la ONU. La crisis financiera que enfrenta la organización, documentada en el artículo *'The UN could run out of cash within months'* (The Economist, 2025), resalta los desafíos que surgen cuando los países miembros incumplen con sus aportes financieros. Este déficit podría poner en peligro la capacidad operativa de instancias del organismo tales como el Consejo de Seguridad, un aspecto crucial para la ejecución de sus funciones de mediación y resolución de conflictos. Este contexto económico resalta la necesidad de investigar cómo las tecnologías emergentes, como estos modelos, pueden ofrecer simulaciones de decisiones políticas en un escenario donde los recursos financieros sean limitados o inciertos.

Este estudio explora cómo los LLM pueden asistir o simular procesos de toma de decisiones en escenarios internacionales complejos. Se analiza el caso de Perú durante su membresía no permanente en el Consejo de Seguridad de la ONU (CSNU) en 2018, un contexto real, delimitado y documentado que permite evaluar con precisión la capacidad de estos modelos para replicar o prever comportamientos diplomáticos.

El marco metodológico empleado es el *benchmark* UNBench, que plantea tareas para simular las fases de decisión en el CSNU, evaluando el desempeño de los LLM en comprensión textual, razonamiento estratégico y generación de lenguaje diplomático.

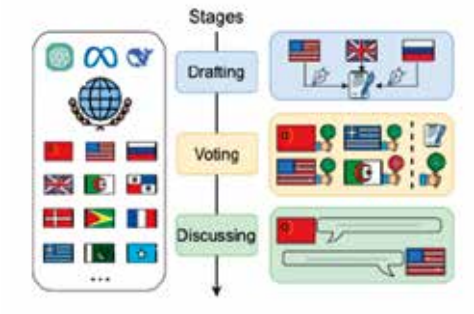
La investigación ofrece evidencia sobre las fortalezas y limitaciones de los LLM en la gobernanza global, fomentando el diálogo entre ciencia política, inteligencia artificial y relaciones internacionales, y abriendo nuevas posibilidades para simulaciones de política exterior con fines analíticos, educativos o predictivos.

4. FUNCIONAMIENTO DE UNBENCH

Debido a la ausencia de un bechmark diseñado para la aplicación específica de los Modelo de Lenguaje de Gran Escala (LLM) en las ciencias políticas (Liang et al., 2025), el UNBench supone una propuesta interesante en esta materia.

Figura 1

Fases del proceso de toma de decisiones en el Consejo de Seguridad de la ONU



Nota. Diagrama que ilustra las tres fases (redacción, votación y debate) del proceso de resoluciones en el Consejo de Seguridad, de "UNBench: A benchmark for evaluating large language models in United Nations Security Council decision-making" por H. Liang, M. Zhao, Y. Qiu, S. Shao y Y. Zhang, 2025, arXiv (https://doi.org/10.48550/arXiv.2502.14122).

Conforme Liang et al. (2025), el proyecto UNBench evalúa la capacidad de los Modelo de Lenguaje de Gran Escala (LLM) en cuatro tareas políticas distintas, pero interconectadas, de diferentes etapas:

Tarea 1 - Juicio del coautor: Dado el contenido anónimo del borrador, identifica las

naciones coautoras óptimas, simulando estrategias de formación de coaliciones en diplomacia multilateral.

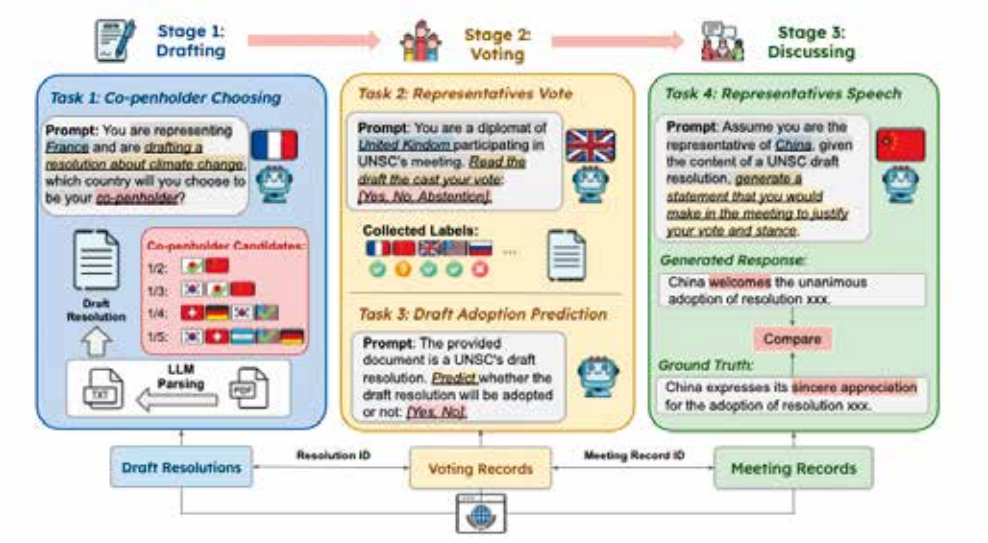
Tarea 2 - Simulación de voto de representantes: Instruye a un Modelo de Lenguaje de Gran Escala (LLM) para que actúe como agente nacional (p. ej., "Como representante de Perú.") y emite decisiones de voto ('A favor', 'En contra', etc.), poniendo a prueba la comprensión contextual de los intereses nacionales.

Tarea 3 - Predicción de adopción del borrador: Introduce un borrador de resolución para predecir su probabilidad de aprobación, lo que requiere el análisis de patrones históricos de votación y alineaciones geopolíticas.

Tarea 4 - Generación de declaraciones representativas: Genera discursos específicos de cada país que justifiquen las posiciones de voto y evalua la generación de lenguaje persuasivo bajo restricciones políticas. Las tareas se diseñan a partir de diferentes etapas de la Organización de las Naciones Unidas (ONU), variando entre tareas predictivas y generativas[1].

Figura 2

Esquema de las cuatro tareas del benchmark UNBench



Nota. Representación de las cuatro tareas (juicio de coautor, simulación de voto, predicción de adopción y generación de declaraciones) en el proceso de resoluciones de la ONU, de "UNBench: A benchmark for evaluating large language models in United Nations Security Council decision-making" por H. Liang, M. Zhao, Y. Qiu, S. Shao y Y. Zhang, 2025, arXiv (https://doi.org/10.48550/arXiv.2502.14122).

1) En el siguiente enlace los autores del UNBench publicaron el codigo abierto al público: https://github.com/yueqingliang1/UNBench

Tarea 1 - Juicio de *Co-Penholder*

Esta tarea tiene por objetivo evaluar si el modelo puede identificar correctamente al país coautor más probable de un borrador. Formalmente, se considera un borrador redactado por un país autor (c_a). A partir de ello, se selecciona un subconjunto de países candidatos a coautores, denotado como *Ccandidate(r$_i$)*. Cada país en esta lista representa una posible opción de *co-penholder*. El modelo de lenguaje debe adoptar el rol del país c_a, examinar el contenido de r_i y elegir uno de los países de la lista como coautor más apropiado.

La tarea se plantea como una pregunta de opción múltiple, con entre 2 y 5 alternativas por resolución, pero solo una respuesta correcta. Esta variación en el número de opciones permite medir el grado de precisión del modelo en contextos de distinta complejidad.

El *co-penholdership* en las resoluciones del Consejo de Seguridad refleja intereses estratégicos compartidos, alianzas diplomáticas o experiencia técnica específica de los países involucrados. La identificación de un coautor adecuado implica que el modelo de lenguaje pueda comprender el contenido temático del borrador (por ejemplo, mantenimiento de la paz, sanciones, derechos humanos) y contextualizarlo dentro de las dinámicas políticas y diplomáticas más amplias. Esto requiere integrar comprensión textual con razonamiento geopolítico.

Además, el modelo debe ser capaz de inferir alineamientos diplomáticos —ya sean históricos o implícitos— y evaluar qué países tienen más probabilidad de copatrocinar una resolución. Esto incluye interpretar señales sutiles entre candidatos, considerar factores como el poder de veto, las prioridades geopolíticas o los intereses regionales. Así, la Tarea 1 permite medir la capacidad del modelo para razonar políticamente y predecir colaboraciones multilaterales en un entorno simulado pero realista.

Tarea 2 - Simulación de voto de representantes

Esta tarea evalúa la capacidad del modelo para simular cómo votaría un país ante un borrador de resolución del Consejo de Seguridad. Para ello, se le proporciona al modelo el texto del borrador y se le pide que adopte la perspectiva de un país específico, eligiendo entre votar a favor, en contra o abstenerse. El objetivo es que el modelo emule el razonamiento que seguiría un representante nacional en una situación real de votación.

Esta simulación requiere que el modelo comprenda el contenido temático del borrador, considere los intereses nacionales del país en cuestión (incluyendo alianzas históricas o posiciones geopolíticas), y tome en cuenta si el país tiene poder de veto como miembro permanente. La tarea no solo mide la comprensión textual, sino también la capacidad del modelo para razonar estratégicamente dentro de un entorno político complejo como el del Consejo de Seguridad.

Tarea 3 - Predicción de adopción del borrador

La Tarea 3 tiene como propósito evaluar si el modelo puede anticipar si un borrador de resolución será aprobado o no en el Consejo de Seguridad. Para ello, el modelo recibe únicamente el texto del borrador y debe predecir el resultado final de la votación. La resolución se considera adoptada solo si obtiene un mínimo de nueve votos a favor y ningún miembro permanente ejerce su poder de veto. A diferencia de la Tarea 2, que examina el voto individual de un país, aquí el desafío radica en evaluar

el comportamiento colectivo de todo el Consejo.

Esta tarea requiere que el modelo analice no solo el contenido del texto, sino también factores más amplios, como el clima político dentro del Consejo, la posibilidad de que alguno de los cinco miembros permanentes bloquee la resolución, y los antecedentes de votaciones similares. Por tanto, se trata de una prueba de razonamiento más complejo, que exige al modelo comprender cómo interactúan las posturas diplomáticas, el poder de veto y los equilibrios geoestratégicos. En conjunto con la Tarea 2, esta tarea permite explorar la capacidad del modelo para reproducir decisiones multilaterales dentro de un entorno político internacional.

Tarea 4 - Generación de declaraciones representativas

La Tarea 4 consiste en que el modelo genere una declaración oficial que un país determinado (c_j) emitiría frente a un borrador de resolución (r_i) del Consejo de Seguridad. Para ello, el modelo recibe como entrada el texto del borrador, el resultado de la votación, cómo votó cada país y, si están disponibles, las declaraciones previas hechas por otros Estados. El objetivo es que la declaración generada sea coherente con la posición geopolítica del país, justifique su voto (ya sea a favor, en contra o abstención) y respete el estilo diplomático característico del Consejo.

Esta tarea evalúa la capacidad del modelo para integrar múltiples capas de información política, histórica y retórica en un solo texto. No solo se requiere que el modelo comprenda el contenido de la resolución, sino que también interprete los intereses nacionales, antecedentes de alianzas y las normas discursivas del entorno diplomático. Por tanto, la Tarea 4 pone a prueba habilidades de generación de lenguaje a un nivel avanzado, simulando con precisión la complejidad de las intervenciones reales en la diplomacia multilateral.

5. METODOLOGÍA

En el presente apartado se describe el proceso seguido para recopilar, procesar y preparar los datos correspondientes al año 2018, cuando Perú fue miembro no permanente del Consejo de Seguridad de las Naciones Unidas. Asimismo, se explica la estructura de datos adoptada en función de los requerimientos del *benchmark* UNBench y las tareas de evaluación aplicadas. Se hizo uso de la LLM META LLAMA 3.3 70B INSTRUCT TURBO FREE a traves de la API Together a modo de ejemplo de la propuesta presentada por el UNBench.

Cabe precisar que la Together API es una plataforma en la nube que permite a desarrolladores e investigadores ejecutar, ajustar finamente y desplegar modelos de inteligencia artificial de código abierto, como LLaMA, Qwen y DeepSeek. Ofrece una infraestructura escalable y optimizada para la inferencia y el entrenamiento de modelos generativos, facilitando la integración de estos modelos en aplicaciones mediante una API sencilla de usar.

Se utilizó ChatGPT-4o (OpenAI, 2024) como herramienta de apoyo para la redacción y revisión de contenido técnico.

Fuente de datos:

Parte relevante de este trabajo fue la recopilación de información de las decisiones, borradores o *drafts*, y otros documentos que se obtuvieron a partir de los registros oficiales del Consejo de Seguridad de la ONU para el año 2018, centrándose en los documentos en los que Perú participó de manera directa, así como en los registros de votación de los países miembros. Esto permitió contar con ejemplos representa-

tivos de la actividad diplomática peruana en el Consejo de Seguridad de la ONU, a la vez que se mantenía una coherencia temporal (todos los documentos corresponden al mismo año).

Se accedio a la Biblioteca Digital de Naciones Unidas (https://digitallibrary.un.org), así como a la página oficial del Consejo de Seguridad de las Naciones Unidas (https://main.un.org/securitycouncil/en/content/resolutions-0) a fin de recopilar los *drafts* y resoluciones, así como otros documentos relevantes.

Estructura de datos:

Para asegurar la compatibilidad con las cuatro tareas propuestas en UNBench, cada borrador de resolución se almacenó en un archivo JSON[2] independiente, incluyendo:

- Metadatos[3] (fecha, ID de la resolución, países autores y coautores).
- El texto completo o resumido del borrador (campo *"Content"*).

Además, se utilizó un archivo CSV para las votaciones (ver *Task* 2 y *Task* 3 en las imágenes adjuntas), en el que se incluyeron:

- El identificador de la resolución o borrador (por ejemplo, "Original_id").
- El país que emite el voto (*"Country"*).
- El voto real (*"Voting"*, con valores como "Y", "N", "A").
- Fechas y otros metadatos relevantes, como el orden cronológico de los sucesos.

Esta organización se corresponde con el estándar de UNBench, que exige separar la información textual de los borradores en un formato que permita su ingestión por parte de los modelos (JSON) y centralizar los resultados de votación y metadatos en un CSV, formato de texto plano utilizado para almacenar datos tabulares, como hojas de cálculo o bases de datos, donde cada línea representa un registro y los valores dentro de cada registro están separados por comas. Este formato es ampliamente utilizado debido a su simplicidad y compatibilidad con diversas aplicaciones, como hojas de cálculo y sistemas de gestión de bases de datos.

Normalización:

Dado que los nombres de los países pueden aparecer de forma inconsistente (por ejemplo, "United States" vs. "U.S." o "Peru" vs. "República del Perú"), se unificaron todos los nombres a través de un diccionario de equivalencias, evitando duplicidades y asegurando que cada país se represente de manera homogénea en el JSON y en el CSV. Asimismo, se organizaron cronológicamente los borradores de 2018. Esto facilita la simulación de escenarios reales (por ejemplo, entrenar un modelo con borradores de inicios de 2018 y probarlo con borradores de finales de 2018), tal como se

2) Un JSON (JavaScript Object Notation) es un formato de texto ligero y legible por humanos, utilizado para almacenar y transmitir datos estructurados. Se deriva de la notación de objetos de JavaScript, pero es independiente del lenguaje, lo que permite su uso en diversas plataformas y lenguajes de programación. JSON es ampliamente utilizado en aplicaciones web para intercambiar datos entre servidores y clientes debido a su simplicidad y eficiencia.

3) Según la International Federation of Library Associations and Institutions (IFLA), los metadatos son datos estructurados que describen, explican, localizan o facilitan la recuperación y gestión de otros datos o recursos de información. Los metadatos son esenciales para la organización y recuperación de información en entornos digitales, ya que permiten identificar, acceder y gestionar eficientemente los recursos informáticos. Además, facilitan la interoperabilidad entre sistemas y la preservación a largo plazo de los datos.

recomienda en el protocolo de evaluación temporal de UNBench.

Resultados:

El código del modelo propuesto en este trabajo se encuentra disponible en el enlace https://github.com/palregia/UNPeru

Análisis del desempeño del modelo en el caso peruano: Task 1 (Juicio de Co-penholder)

En este estudio se evalúa la capacidad de los modelos de lenguaje de gran escala (LLMs) para simular procesos de toma de decisiones diplomáticas en el Consejo de Seguridad de las Naciones Unidas, tomando como eje de análisis la participación de Perú durante su membresía no permanente en el año 2018. Se aplicó el *benchmark* UNBench, específicamente la Tarea 1 (Juicio de *Co-penholder*), la cual consiste en que el modelo, actuando como país autor de un borrador de resolución, seleccione al coautor más probable dentro de un conjunto de países candidatos.

Para ello, se recopilaron y normalizaron resoluciones y borradores del Consejo de Seguridad correspondientes a 2018, priorizando aquellos en los que Perú participó activamente. El borrador fue transformado en un archivo ".json" conforme a la estructura de UNBench, con los campos *author*, *coauthor*, y listas de candidatos (choices_2 a choices_5). Posteriormente, se utilizó el modelo Llama-3.3-70B, que fue instruido para asumir el rol del autor (Perú) y elegir al *co-penholder* más adecuado de entre los países listados. La evaluación se centró en la métrica de precisión, comparando las respuestas del modelo con las elecciones reales documentadas.

Los resultados muestran una precisión del 100% en todos los escenarios evaluados,

Figura 3

Código Python para la carga de datos en la Tarea 1 de UNBench

Nota: Captura de pantalla de un Jupyter Notebook que muestra el código para procesar borradores de resoluciones en la Tarea 1 (juicio de co-penholder). Elaboración propia.

desde listas con 2 hasta 5 opciones. Este rendimiento contrasta de manera significativa con los resultados globales reportados en el artículo original de UNBench, donde la precisión desciende a medida que aumentan los candidatos (~80% con 2 opciones y ~40% con 5). Esta diferencia puede interpretarse de varias maneras. En primer lugar, podría sugerir que el modelo logró captar con notable precisión los patrones de alianzas y copatrocinios del Perú en 2018. En segundo lugar, es posible que el conjunto de ejemplos estuviera sesgado hacia escenarios con opciones claramente diferenciadas o patrones repetitivos, facilitando la elección correcta.

Figura 4

Código Python para la selección de coautores en la Tarea 1 de UNBench

Nota. Captura de pantalla de un Jupyter Notebook que muestra el código para evaluar opciones de co-penholder en la Tarea 1. Elaboración propia.

En cualquier caso, los hallazgos demuestran que el Task 1 es una herramienta eficaz para evaluar razonamiento geopolítico contextualizado. La simulación del comportamiento de Perú como autor de resoluciones ofrece evidencia de que los LLMs pueden replicar decisiones diplomáticas bajo ciertos marcos estructurados, especialmente cuando el lenguaje del borrador o las relaciones internacionales del país presentan regularidades. Este enfoque también refuerza la utilidad de realizar estudios de caso empíricos como método para complementar *benchmarks* generalistas con aplicaciones situadas y más específicas.

Análisis del desempeño del modelo en el caso peruano: Task 2 (Simulación de votación)

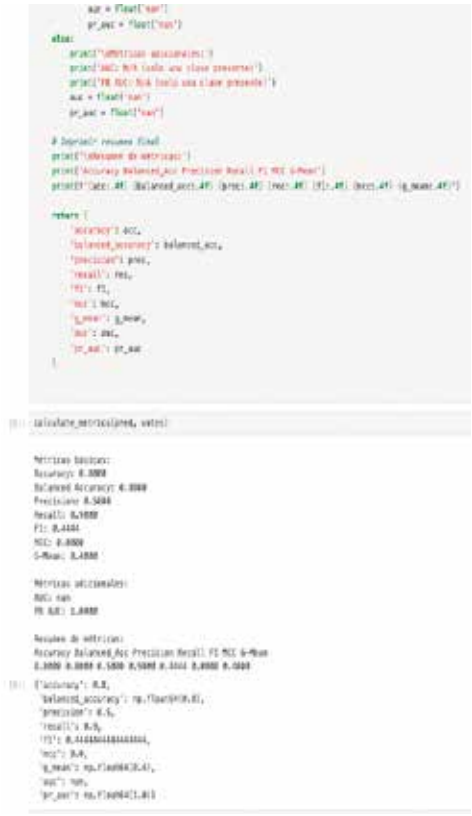
La Tarea 2 del *benchmark* UNBench tiene como propósito simular cómo votaría un país específico ante un borrador de resolución del Consejo de Seguridad de las Naciones Unidas (CSNU). Para esta aplicación empírica, se utilizaron resoluciones reales del año 2018, estructuradas en un archivo CSV que contenía el país, el identificador de la resolución, los autores, el voto emitido y la fecha. Dichos datos fueron preprocesados para garantizar su consistencia terminológica y cronológica conforme al estándar de UNBench.

El modelo Llama-3.3-70B fue instruido para asumir la perspectiva de Perú —o de otros países— y emitir una predicción sobre su comportamiento de voto: a favor, en contra o abstención. Los resultados obtenidos reflejan una precisión del 0.80 y una *Balanced Accuracy* también de 0.80, lo que indica una sólida capacidad del modelo para reflejar los patrones de decisión observados en los datos reales. Las métricas complementarias refuerzan esta interpretación: una precisión de 0.50, un *recall* de 0.90, y un F1-score de 0.4444. Sin embargo, el índice MCC (*Matthews Correlation Coefficient*) fue de 0.0, y el G-Mean se situó en 0.40, lo cual sugiere un desempeño desigual entre clases.

El valor del AUC no pudo calcularse (NaN), y el PR AUC fue de 1.0, dado que el conjunto de clases estaba desbalanceado. Estos resultados reflejan la dificultad del modelo para generalizar en escenarios donde no hay una distribución equitativa de las clases, pero también muestran su potencial cuando se enfrenta a textos alineados con antecedentes diplomáticos claros, como en el caso peruano.

Figura 5

Resultados de la simulación de votación en la Tarea 2 de UNBench



Nota. Captura de pantalla de un Jupyter Notebook que presenta el código y métricas de evaluación (precisión, F1-score) para la simulación de votos de Perú en 2018. Elaboración propia.

Aunque se identifican limitaciones cuando el lenguaje del borrador es ambiguo o aborda temas geopolíticamente sensibles, la simulación resulta útil para extraer patrones interpretables de comportamiento diplomático. Esto refuerza el valor de los LLMs como herramientas de análisis en entornos multilaterales complejos y sugiere la utilidad de combinarlos con contextos históricos y conocimiento experto para maximizar su efectividad.

Análisis del desempeño del modelo en el caso peruano: Task 3 (Predicción de adopción de resoluciones)

La Tarea 3 del *benchmark* UNBench busca evaluar si un modelo de lenguaje de gran escala (LLM) puede predecir si un borrador de resolución del Consejo de Seguridad será finalmente adoptado o no, considerando únicamente su contenido textual. En esta aplicación, el modelo LLaMA 3.3-70B fue alimentado con el texto completo de los borradores utilizados por el Consejo durante el año 2018, año en el que Perú fue miembro no permanente. El objetivo era que el modelo infiriera el resultado de la votación (*ADOPTED* o *NOT ADOPTED*) sin tener acceso a los votos explícitos.

Los resultados muestran un rendimiento razonable del modelo, con una precisión (*accuracy*) de 0.8000, una precisión balanceada (*balanced accuracy*) también de 0.8000, un recall de 0.9000 y un F1-*score* de 0.4444. Si bien estas métricas indican una capacidad considerable para identificar correctamente los casos positivos (resoluciones adoptadas), el modelo presenta limitaciones para discriminar correctamente los negativos (resoluciones no adoptadas), lo que se refleja en un MCC de 0.0000. Este resultado sugiere que, aunque el modelo logra capturar patrones predominantes en los datos, su capacidad para reflejar correlaciones estructurales entre texto y decisión final es limitada.

Adicionalmente, se calculó la similitud coseno promedio basada en TF-IDF, que arrojó un valor de 0.553. Este indicador cuantifica el grado de semejanza léxica entre los textos procesados, tomando en cuenta la frecuencia y relevancia de las palabras (TF-IDF). Un valor de 0.553, en una escala de 0 a 1, sugiere que existe una coincidencia moderada en el contenido textual de los borradores analizados, lo que implica que el modelo puede estar reconociendo

Figura 6

Resultados de la predicción de adopción en la Tarea 3 de UNBench

Nota. Captura de pantalla de un Jupyter Notebook que muestra el código y métricas (precisión, AUC) para predecir la adopción de borradores de resoluciones. Elaboración propia.

estructuras lingüísticas comunes sin necesariamente comprender su significado político o implicancias estratégicas.

En conjunto, estos hallazgos reflejan que, si bien los LLMs como LLaMA pueden detectar patrones útiles para anticipar la adopción de resoluciones en contextos multilaterales, su precisión disminuye cuando los textos presentan ambigüedad normativa o temas políticamente delicados. La integración de variables contextuales, como antecedentes de votación o alineamientos diplomáticos, podría mejorar significativamente el desempeño del modelo en escenarios más complejos y realistas.

Análisis del desempeño del modelo en el caso peruano: Task 4 (Generación de declaraciones diplomáticas)

La cuarta tarea del *benchmark* UNBench propone que el modelo genere declaraciones diplomáticas que reflejen la postura de un país frente a un borrador de resolución del Consejo de Seguridad. Se evaluó la similitud entre las declaraciones reales y las generadas por el modelo utilizando métricas lingüísticas como ROUGE-L**,** *cosine similarity* (TF-IDF) y Jaccard**.**

Los resultados obtenidos indican una similitud semántica promedio (TF-IDF *cosine*) de ~0.70, un puntaje ROUGE-L de 0.21 y un índice de Jaccard de 0.20, lo cual sugiere que el modelo fue capaz de capturar la estructura y vocabulario típico del

Figura 7

Resultados de la tarea 4 (código python)

Fuente: Elaboración propia

lenguaje diplomático utilizado en el Consejo. Sin embargo, se evidencian limitaciones en la profundidad política del contenido generado. Las declaraciones tienden a ser genéricas y carecen de referencias contextuales específicas o matices geopolíticos, elementos clave en la justificación de posturas estatales reales. Estas deficiencias se acentúan en resoluciones más sensibles o menos estandarizadas. En general, el modelo demuestra competencia lingüística formal, pero su razonamiento político permanece limitado sin entrenamiento explícito en fuentes históricas y diplomáticas especializadas.

En las figuras adjuntas se aprecia los *notebooks* de Jupyter utilizados para la ejecución de los cuatro *tasks*:

1. **Tarea 1 (juicio de *co-penholder*):** Se muestra cómo se cargan los borradores en formato JSON, junto con las listas de posibles coautores. El modelo debe elegir el coautor correcto basándose en el contenido del borrador.

2. **Tarea 2 (simulación de votación):** Se evidencia el uso del CSV con los votos para comparar la predicción del modelo ("A favor", "En contra", "Abstención") contra el voto real.

3. **Tarea 3 (predicción de adopción):** Se observa cómo se provee el texto del borrador y se pide al modelo que indique si será "Adoptado" o "No Adoptado". El *notebook* implementa una métrica de precisión y otras medidas para evaluar el desempeño.

4. **Tarea 4 (generación de declaraciones diplomáticas):** Se cargan las declaraciones oficiales (*ground truth*) y se comparan con el texto generado por el modelo usando métricas de similitud (por ejemplo, ROUGE y coseno de Sentence-BERT).

Con este preprocesamiento y la estructura de datos definida, se garantiza una implementación consistente de las tareas, permitiendo analizar la capacidad de los modelos para manejar aspectos como la geopolítica, las alianzas históricas y el lenguaje diplomático en el contexto del Consejo de Seguridad de la ONU.

Discusión

Los resultados obtenidos a través de la aplicación del *benchmark* UNBench al caso peruano de 2018 permiten una reflexión crítica sobre el desempeño actual de los Modelos de Lenguaje de Gran Escala (LLMs) en contextos políticos formales, como el del Consejo de Seguridad de las Naciones Unidas. En general, el modelo Llama-3.3-70B logró simular con éxito ciertas dinámicas diplomáticas, especialmente cuando los textos eran claros y las decisiones estaban alineadas con patrones históricos de comportamiento internacional.

La Tarea 1 (juicio de *co-penholder*) destacó por su notable precisión (100%), indicando que el modelo fue capaz de captar con exactitud las posibles alianzas del Perú en la etapa de redacción. Este resultado podría reflejar tanto la claridad de los casos seleccionados como una alta correlación entre el contenido temático de los borradores y la trayectoria diplomática peruana. Sin embargo, es necesario considerar la posibilidad de un sesgo en los datos hacia escenarios poco ambiguos.

La aplicación de la Tarea 2 de UNBench, centrada en simular el voto de los países ante resoluciones del Consejo de Seguridad, mostró resultados alentadores al evaluar el caso peruano en 2018. El modelo alcanzó una precisión y una precisión balanceada de 0.80, reflejando su capacidad para replicar decisiones diplomáticas reales. Aunque el *recall* fue elevado (0.90), métricas como el F1-score (0.4444) y el

"Aunque los LLMs pueden reproducir el estilo diplomático, su utilidad en escenarios reales dependerá de su entrenamiento en fuentes especializadas y su capacidad para incorporar información contextual y antecedentes históricos."

MCC (0.0) indican dificultades para lograr un equilibrio entre clases, especialmente en escenarios con distribución desigual de votos. El G-Mean de 0.40 refuerza esta lectura, mostrando que la efectividad del modelo es variable dependiendo del tipo de decisión que deba predecir. Pese a estas limitaciones, la simulación demuestra el valor de los modelos de lenguaje como herramientas para inferir patrones de comportamiento diplomático en entornos multilaterales. La incapacidad para calcular el AUC (NaN) y un PR AUC perfecto (1.0) revelan que, si bien el modelo responde bien a contextos alineados con posturas históricas claras —como fue el caso de Perú—, requiere ajustes para enfrentar borradores con ambigüedad política o implicancias geoestratégicas complejas. Integrar variables adicionales y conocimiento contextual puede potenciar significativamente su utilidad en análisis de política exterior.

La Tarea 3 del benchmark UNBench permitió evaluar la capacidad del modelo para predecir la adopción de resoluciones del Consejo de Seguridad de la ONU utilizando únicamente el contenido textual de los borradores. Los resultados obtenidos —precisión y precisión balanceada de 0.8000, recall de 0.9000 y F1-score de 0.4444— reflejan que el modelo tiene un buen desempeño al identificar resoluciones que fueron efectivamente adoptadas. No obstante, el bajo valor del MCC (0.0000) señala dificultades para captar correlaciones entre el texto normativo y el resultado de la votación, especialmente en el caso de resoluciones no adoptadas.

Asimismo, el análisis de similitud léxica mediante la métrica de similitud coseno basada en TF-IDF arrojó un valor moderado de 0.553. Este dato sugiere que los borradores presentan una cierta homogeneidad léxica, lo cual podría facilitar al modelo el reconocimiento de patrones formales sin necesariamente interpretar su carga política o estratégica. En suma, si bien la predicción basada únicamente en texto ofrece información valiosa, la inclusión de variables como el historial de votación o alianzas diplomáticas resulta clave para mejorar la capacidad del modelo en entornos reales de toma de decisiones multilaterales. Finalmente, la Tarea 4 del evaluó la capacidad del modelo para generar declaraciones diplomáticas acordes con la postura de un país ante un borrador de resolución del Consejo de Seguridad.

Las métricas utilizadas —ROUGE-L (0.21), similitud coseno basada en TF-IDF (~0.70) e índice de Jaccard (0.20)— revelan que el modelo logra emular de forma aceptable la forma y el estilo del lenguaje diplomático. Estas cifras indican que existe una coincidencia estructural y léxica razonable entre las declaraciones reales y las generadas, lo que evidencia una competencia lingüística superficial adecuada por parte del modelo. No obstante, el análisis también muestra que el contenido político de las declaraciones generadas resulta limitado. Las respuestas tienden a ser genéricas, sin reflejar matices geopolíticos ni referen-

cias contextuales que son fundamentales en discursos diplomáticos reales. Esta carencia se hace más evidente en resoluciones controvertidas o con implicancias estratégicas. Por tanto, aunque los LLMs pueden reproducir el estilo diplomático, su utilidad en escenarios reales dependerá de su entrenamiento en fuentes especializadas y su capacidad para incorporar información contextual y antecedentes históricos.

6. CONCLUSIONES

El análisis realizado demuestra que los Modelos de Lenguaje de Gran Escala (LLMs) pueden constituir una herramienta innovadora para comprender la dinámica de la toma de decisiones en el ámbito multilateral, siempre que su aplicación se articule con los marcos teóricos clásicos de las relaciones internacionales. La incorporación de enfoques como el realismo conductual, el constructivismo y el institucionalismo liberal permite dotar de sustento conceptual a los resultados empíricos obtenidos mediante técnicas de procesamiento automatizado del lenguaje.

En esa línea, el estudio del caso peruano en el Consejo de Seguridad de la ONU evidencia que los patrones discursivos y las estrategias de posicionamiento pueden analizarse con profundidad a partir de los avances tecnológicos contemporáneos. No obstante, al tratarse de una investigación de carácter exploratorio, los hallazgos aquí expuestos deben entenderse como un punto de partida para futuros desarrollos teóricos y metodológicos. La novedad del enfoque —que integra inteligencia artificial, análisis de discurso y teoría de las relaciones internacionales— abre un campo de investigación aún incipiente, pero de enorme potencial para fortalecer el vínculo entre la reflexión teórica y la evidencia empírica en el estudio de la gobernanza global.

"Más allá del análisis técnico, es crucial reflexionar sobre las implicancias jurídicas y políticas que conlleva una implementación de LLMs en la gobernanza global"

Este estudio exploró la aplicabilidad del *benchmark* UNBench como marco de evaluación de los LLMs en el ámbito de la toma de decisiones en la diplomacia multilateral, utilizando como caso empírico la participación de Perú en el Consejo de Seguridad de las Naciones Unidas (ONU) durante 2018. A través de la implementación del *benchmark* UNBench, se ha observado que los LLMs, como el modelo LLaMA 3.3-70B, son capaces de replicar ciertos aspectos del comportamiento en los procesos de toma de decisión, particularmente, donde los patrones de decisión son previsibles. Mediante la implementación de las cuatro tareas propuestas —juicio de coautores, simulación de votaciones, predicción de adopción y generación de discursos— se comprobó que los modelos actuales pueden ofrecer aproximaciones razonables al comportamiento diplomático. Sin embargo, los resultados obtenidos también evidencian limitaciones importantes.

La alta precisión alcanzada en la Tarea 1 evidencia el potencial de los LLMs para identificar alianzas diplomáticas plausibles, mientras que los resultados de las Tareas 2 y 3 sugieren que es posible simular decisiones de voto con un grado aceptable de confiabilidad, aunque aún con margen de error en escenarios ambiguos o politizados. La Tarea 4, si bien muestra que los modelos pueden generar texto coherente

y formal, pone de manifiesto la necesidad de un entrenamiento más profundo en contenido diplomático especializado.

Más allá del análisis técnico, es crucial reflexionar sobre las implicancias jurídicas y políticas que conlleva una implementación de LLMs, tal y como ha sido descrita en este ensayo, en la gobernanza global. En términos jurídicos, surge la cuestión de la responsabilidad en la toma de decisiones políticas asistidas por inteligencia artificial. Si bien los LLMs pueden proporcionar simulaciones y recomendaciones, las decisiones finales siguen siendo tomadas por representantes humanos, lo que plantea interrogantes sobre la asignación de responsabilidad en caso de decisiones erróneas basadas en estos modelos.

Desde el punto de vista político, el uso de estos modelos en espacios multilaterales podría tener efectos significativos en la soberanía de los países y en la dinámica de poder dentro de organizaciones internacionales como la ONU. Si bien estos modelos ofrecen la posibilidad de que en un futuro se puedan simular comportamientos diplomáticos de muy eficiente, existe el riesgo de que países con mayor acceso a estas tecnologías puedan ejercer una influencia desproporcionada en los procesos de toma de decisiones.

Otro aspecto crítico es el impacto que los LLMs podrían tener en la transparencia y la rendición de cuentas en los procesos. Dado que estos modelos pueden procesar grandes volúmenes de datos y generar recomendaciones rápidas, es fundamental garantizar que las decisiones tomadas con la ayuda de los LLMs sean revisadas adecuadamente y sean comprensibles para los actores humanos involucrados. Sin un adecuado marco de supervisión, podría ponerse en riesgo la responsabilidad política y ética de las decisiones diplomáticas.

Como conclusión general, los hallazgos de este estudio respaldan el valor de los LLMs como herramientas auxiliares para el análisis político y diplomático computacional, pero también subrayan que su uso debe ir acompañado de una comprensión crítica de sus limitaciones. La incorporación de variables como relaciones bilaterales, antecedentes históricos, contexto regional y opinión pública podría enriquecer la capacidad predictiva y explicativa de estos modelos. Futuros trabajos podrían profundizar en el entrenamiento multimodal y multilingüe para acercarse aún más a las complejidades del lenguaje y la política internacional.

7. RECOMENDACIONES

Dentro de la bibliografía explorada se ha identificado el libro titulado "El Perú en el Consejo de Seguridad (2018-2019): Diplomacia constructiva en tiempos de polarización", el cual recoge las experiencias y reflexiones del equipo diplomático peruano durante su participación como miembro no permanente en el Consejo de Seguridad de las Naciones Unidas. Este documento podría aportar significativamente a la implementación de este ejercicio sobre la evaluación de Modelos de Lenguaje de Gran Escala (LLM) en la toma de decisiones políticas, específicamente para el caso peruano en 2018. El libro proporciona una descripción exhaustiva y contextualizada de la diplomacia peruana durante ese periodo, incluyendo prioridades estratégicas, procesos internos de negociación, manejo de crisis internacionales (como la crisis siria), y consideraciones tácticas para la formación de alianzas y votaciones.

Teniendo en cuenta desarrollo futuros y ejercicios sobre el uso de los LLM como herramientas para el proceso de negociación multilateral real, el uso de este documento como insumo podría ser útil para:

- **Tarea 1 (juicio de *co-penholder*):** El libro ofrece detalles sobre cómo Perú eligió sus aliados estratégicos y los criterios utilizados para copatrocinar resoluciones, lo que podría ayudar a validar o ajustar la precisión de la simulación que realiza el LLM en la identificación de coautores óptimos.

- **Tarea 2 (simulación de votación):** El libro incluye información que puede contribuir a conocer las razones políticas y estratégicas detrás de las votaciones de Perú en el Consejo, lo que podría ayudar a explicar por qué el modelo tiene éxito o falla en simular las decisiones de voto reales.

- **Tarea 3 (predicción de adopción de resoluciones):** La narración del libro sobre cómo Perú navegó en dinámicas internas y externas podría proporcionar contexto adicional para evaluar las limitaciones del modelo cuando no logra predecir correctamente la adopción de resoluciones.

- **Tarea 4 (generación de declaraciones diplomáticas):** Las intervenciones diplomáticas reales descritas en el libro podrían servir como ejemplos concretos para entrenar o validar la capacidad del modelo en la generación de declaraciones representativas, destacando elementos de estilo, argumentación y contexto político.

Además, el análisis en el libro sobre la dinámica interna del Consejo, los mecanismos de negociación y la influencia de factores externos (como la diplomacia informal y las relaciones personales) podría aportar datos cualitativos útiles para explicar algunas limitaciones observadas en el rendimiento del LLM, especialmente cuando se enfrenta a situaciones diplomáticas complejas y ambiguas.

8. LIMITACIONES

A pesar de los resultados alentadores obtenidos en la simulación de tareas diplomáticas con LLMs, es importante reconocer las limitaciones que condicionan la validez y generalización de este estudio.

En primer lugar, el tamaño y representatividad del conjunto de datos constituye una restricción fundamental. El análisis se centró exclusivamente en las resoluciones y borradores del año 2018 en los que participó activamente Perú. Si bien esto permitió una evaluación contextualizada y detallada, también implica que los resultados podrían no ser extrapolables a otros países, contextos regionales o periodos históricos, donde las dinámicas geopolíticas y los estilos diplomáticos difieren sustancialmente.

En segundo lugar, la interpretación del modelo depende enteramente del texto, sin acceso a información contextual crítica como negociaciones previas, discursos informales, relaciones bilaterales o presiones externas que influyen significativamente en el proceso de toma de decisiones del Consejo de Seguridad. Esto se refleja especialmente en las Tareas 2 y 3, donde el modelo muestra dificultad para anticipar abstenciones estratégicas o el uso del veto por parte de miembros permanentes.

Tercero, el modelo utilizado (Llama-3.3-70B) no ha sido específicamente entrenado en lenguaje diplomático ni en documentos históricos de política internacional. Aunque logra reproducir estructuras formales y vocabulario general, en tareas como la generación de declaraciones (*Task* 4), se observaron deficiencias en matices retóricos, referencias contextuales y profundidad argumentativa, lo cual limita su utilidad como herramienta de generación textual con valor político real.

Además, si bien las métricas de evaluación como precisión, F1-score o ROUGE ofrecen una visión cuantitativa del rendimiento, no capturan completamente la calidad semántica ni la coherencia política de las respuestas generadas. Las decisiones en política internacional involucran dimensiones normativas, ideológicas y emocionales difíciles de reducir a métricas tradicionales de NLP.

Finalmente, la implementación técnica fue realizada en un entorno controlado (*Jupyter Notebooks*), con *prompts* diseñados manualmente. Esto podría haber inducido sesgos en la formulación de las tareas, así como en los resultados, especialmente si las preguntas fueron redactadas de manera favorable al modelo.

En suma, aunque el estudio demuestra que los LLMs pueden simular ciertos aspectos del comportamiento diplomático, se requiere precaución antes de extender sus aplicaciones a contextos reales de toma de decisiones, especialmente sin incorporar capas adicionales de razonamiento geopolítico y evidencia empírica. ◆

BIBLOGRAFÍA

Allison, G. (1971). Essence of Decision: Explaining the Cuban Missile Crisis. Boston: Little, Brown and Company.

Amazon Web Services. (s.f.). What is a large language model (LLM)?. Amazon Web Services. Recuperado el 11 de junio de 2025, de https://aws.amazon.com/es/what-is/large-language-model/

Bostrom, N. (2017). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Floridi, L. (2014). The Fourth Revolution: How the Infosphere is Reshaping Human Reality. Oxford University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

International Federation of Library Associations and Institutions. (s.f.). Metadatos. Recuperado el 14 de junio de 2025, de https://www.iasa-web.org/book/export/html/3490

Internet Engineering Task Force. (2017). The JavaScript Object Notation (JSON) Data Interchange Format. RFC 8259. Recuperado el 14 de junio de 2025, de https://datatracker.ietf.org/doc/html/rfc8259en.wikipedia.org+3datatracker.ietf.org+3datatracker.ietf.org+3

Jupyter. (s.f.). Jupyter. Recuperado el 11 de abril de 2025, de https://jupyter.org/

Liang, H., Zhao, M., Qiu, Y., Shao, S., & Zhang, Y. (2025). UNBench: A benchmark for evaluating large language models in United Nations Security Council decision-making. arXiv. https://doi.org/10.48550/arXiv.2502.14122

Meza-Cuadra, G., Popolizio, N., & Delegación Permanente del Perú ante las Naciones Unidas. (2020). El Perú en el Consejo de Seguridad (2018–2019): Diplomacia constructiva en tiempos de polarización. Lima: Fundación Academia Diplomática del Perú.

Naciones Unidas. (2018). Consejo de Seguridad: Resoluciones, borradores y actas de reuniones [Base de datos]. Biblioteca Digital de las Naciones Unidas. https://digitallibrary.un.org

Naciones Unidas. (s.f.). Security Council resolutions. Recuperado el 11 de abril de 2025, de https://www.un.org/securitycouncil/en/content/resolutions-0

Naciones Unidas. (s.f.). Security Council resolutions in XML AKN4UN format [Repositorio]. Recuperado el 11 de abril de 2025, de https://github.com/UNxml/SCresolutions/tree/main

Naciones Unidas. (s.f.). United Nations Security Council. Recuperado el 11 de abril de 2025, de https://main.un.org/securitycouncil/en/content/resolutions-0

OpenAI. (2023). LLaMA 3.3-70B model card. Recuperado el 11 de abril de 2025, de https://openai.com/research

OpenAI. (2024). ChatGPT-4o (Versión abril 2024) [Modelo de lenguaje de IA]. https://openai.com/chatgpt

Python Software Foundation. (s.f.). Python. Recuperado el 11 de abril de 2025, de https://www.python.org/

Shafranovich, Y. (2005). *Common Format and MIME Type for Comma-Separated Values (CSV) Files* (RFC 4180). Network Working Group. Recuperado el 14 de junio de 2025, de https://www.rfc-editor.org/rfc/rfc4180.htm

Simon, H. A. (1957). Models of Man: Social and Rational. Wiley.

The Economist. (2025, mayo 1). The UN could run out of cash within months. Recuperado el 14 de junio de 2025, de https://www.economist.com/international/2025/05/01/the-un-could-run-out-of-cash-within-months

Together AI. (s.f.). Together API. Recuperado el 14 de junio de 2025, de https://www.together.ai/

Together Computer. (2024). API documentation for Together LLM services. Recuperado el 11 de abril de 2025, de https://api.together.xyz

UNBench. (s.f.). UNBench [Repositorio]. Recuperado el 11 de abril de 2025, de https://github.com/yueqingliang1/UNBench

Wendt, A. (1999). Social Theory of International Politics. Cambridge University Press.