Una propuesta para evaluar la seguridad de un ChatBot basado en modelos de lenguaje grandes

Pedro Pinacho-Davidson1,2*, Fernando Gutiérrez2,3† ,
Pablo Zapata1,2, Rodolfo Vergara1,2, Pablo Aqueveque1,2† ,
Ximena Sep´ulveda1,2

1*Facultad de Ingeniería, Universidad de Concepción, Edmundo Larenas 219 , Concepción, 100190, Estado, País.

*Autor(es) correspondiente(s). Correo electrónico: iauthor@gmail.com; Autores contribuyentes: iiauthor@gmail.com; iiiautor@gmail.com; iiiautor@gmail.com; iiiautor@gmail.com; iiiautor@gmail.com; †Estos autores contribuyeron igualmente a este trabajo.

Abstracto

Los grandes modelos de lenguaje conducen a una nueva forma de interacción entre humanos y computadoras. Esto también ha introducido un nuevo conjunto de riesgos de seguridad y posibles vulnerabilidades que no se observan en otros sistemas informáticos. Para comprender adecuadamente este nuevo escenario, en este trabajo proponemos un método y una métrica de evaluación de riesgos. También hemos incluido un caso de estudio para ilustrar mejor los riesgos reales utilizando modelos de lenguaje disponibles comercialmente.

Palabras clave: palabra clave1, palabra clave2, palabra clave3, palabra clave4

1. Introducción

Los avances recientes en el procesamiento del lenguaje natural (PLN) han llevado al desarrollo de modelos de lenguaje que pueden comprender y producir texto de manera similar a la humana [1]. Estos modelos basados en aprendizaje automático se han entrenado en conjuntos de textos muy grandes (corpora), lo que les permite capturar mejor la estructura y el significado detrás de un texto. El alto dominio del lenguaje natural por parte de estos modelos, conocidos como Large Language Models (LLM), ha llevado a que se incorporen a una amplia gama de tareas [2].

Aunque los LLM se están implementando en diferentes tipos de aplicaciones, su principal caso de uso ha sido en forma de chatbots, por ejemplo, ChatGPT. Los chatbots son aplicaciones informáticas que simulan conversaciones humanas en lenguaje natural [3]. Un chatbot puede estar orientado a tareas, como asistentes de software, o puede servir para una conversación informal [4]. Debido a que necesitan interactuar a través del lenguaje natural, la evolución de los chatbots ha dependido del progreso de la PNL [5]. Inicialmente, los chatbots utilizaban métodos basados en reglas y bases de conocimiento para comprender las aportaciones de los usuarios y facilitar las respuestas. La aparición de sistemas basados en el aprendizaje profundo condujo al desarrollo de chatbots en forma de asistentes personales [5].

A medida que existe un impulso para integrar los modelos LLM en un número cada vez mayor de tareas [6], la necesidad de comprender el riesgo asociado a esta nueva tecnología se ha vuelto más evidente. Pueden surgir problemas debido a los grandes conjuntos de datos utilizados para entrenar estos modelos, que en su mayoría se recopilan de Internet. El volumen de datos de entrenamiento hace que la verificación y validación de esta información no sea una opción. Pueden surgir problemas de la interacción con el modelo. Dado que la entrada y salida de los modelos LLM es lenguaje natural, eso permite la ambigüedad.

Weidinger et al. [7] han propuesto una taxonomía de riesgos que considera tanto situaciones que se han observado como situaciones que se pueden esperar en el futuro. Estos riesgos se centran principalmente en el tipo de resultados que puede generar un modelo LLM, como discurso dañino [8], información confidencial, información errónea y contenido con intenciones maliciosas [9]. En la taxonomía, la información errónea se considera contenido incorrecto o engañoso generado por el modelo sin una intención maliciosa en el mensaje. Por el contrario, con intenciones maliciosas, el mensaje indica al modelo que la salida tiene, por ejemplo, información incorrecta.

Siguiendo un enfoque de riesgos basado en la industria, la fundación de seguridad OWASP ha publicado su informe de concientización Top Ten for LLM Applications [10], donde identifica diez vulnerabilidades críticas relacionadas con el uso de LLM en aplicaciones basadas en web. Estas vulnerabilidades se pueden clasificar en tres grupos principales. El primer tipo de vulnerabilidades está relacionado con la infraestructura que rodea un modelo LLM. Estas vulnerabilidades pueden afectar la creación o el ajuste de un modelo mediante envenenamiento de datos. También pueden afectar la cadena de suministro de bibliotecas y otras herramientas utilizadas por el modelo, o podrían provocar el robo del modelo. El segundo tipo de vulnerabilidad se refiere al resultado generado por la interacción del modelo LLM con los usuarios. Esta situación puede surgir por una interacción normal (por ejemplo, una solicitud a un modelo) o por una solicitud especial de una embarcación que pretende eludir las restricciones (por ejemplo, una inyección rápida). El resultado de esta interacción puede conducir a la denegación del servicio y al acceso a información restringida, como datos sensibles o confidenciales. Finalmente, el tercer tipo de vulnerabilidades está relacionado con los efectos posteriores en el uso de los resultados generados por el LLM. Esto se refiere al exceso de confianza dada a un modelo sin validación o supervisión del resultado, por ejemplo, el uso de su resultado en la generación de contenidos o la toma de decisiones.

El uso de modelos de lenguaje grande (LLM) para la implementación de chatbot introduce riesgos distintos de los tradicionalmente reconocidos, lo que amerita perspectivas alternativas.
De esta manera, [11] considera tres categorías de enfoque para los ataques potenciales: (1) 'El usuario', enfocándose en el compromiso en la interacción entre el usuario y el LLM, (2) el modelo en sí, que puede verse interrumpido o persuadido para generar información inapropiada

resultados, y (3) terceros, donde el modelo se utiliza para atacar a víctimas no relacionadas con el sistema. Si bien este estudio no profundiza en el riesgo de uso inadecuado no solicitado respuestas no inducidas por el usuario del sistema, sí propone una necesaria caja negra enfoque para realizar pruebas de seguridad y hace referencia al uso de la Tríada de la CIA para caracterizar los daños. Esto es crucial desde un punto de vista práctico de ciberseguridad.

Nuestro trabajo sienta las bases para desarrollar herramientas de escaneo de riesgos de chatbot, que funcionan de manera similar a los escáneres de seguridad tradicionales para aplicaciones (Vulnerability Scanners). es decir, empleando la perspectiva del consultor: la caja negra (o la perspectiva del atacante). Así, se presenta una nueva métrica, considerando factores similares a los del CVSS [1] como la dificultad del ataque, la tríada de la CIA (confidencialidad, integridad y disponibilidad de los sistemas analizados), además del potencial de daños a los sistema que presta el servicio, a sus usuarios o a terceros, incluido el daño no solicitado al usuario, que va más allá de las evaluaciones de seguridad tradicionales. Esto se realiza en el Desarrollo de nuevas métricas que cumplan con los requisitos de una evaluación de seguridad. consultoría, en términos de proporcionar información útil para realizar modificaciones a los problemas con una visión de riesgo, y permitir pruebas repetibles para evaluar acciones correctivas y el cambio en la postura de riesgo de la organización analizada

2 problemas de ciberseguridad relacionados con los LLM

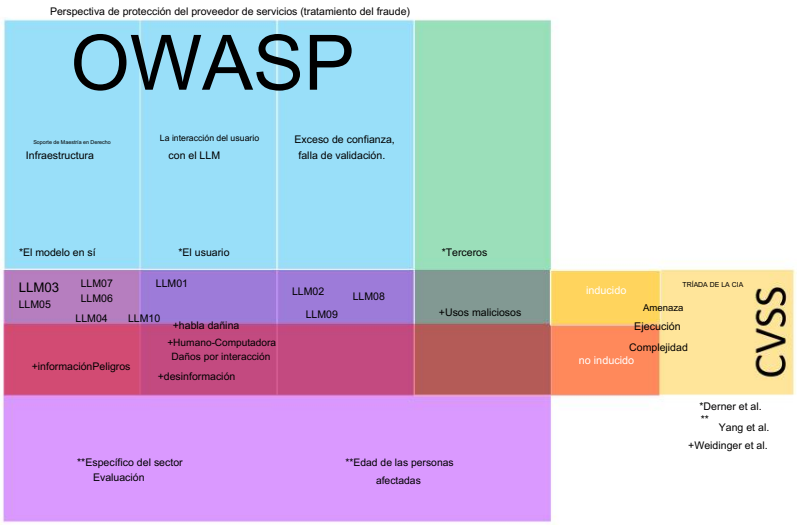

Fig. 1 Esta es una figura ancha. Este es un ejemplo de título largo este es un ejemplo de título largo esto es un ejemplo de título largo este es un ejemplo de título largo

1https://www.first.org/cvss/specification-document

- Problemas conocidos de LLMs y su impacto en seguridad de chatbots
- Categorización de problemas OWASP y otras

• Inyección rápida (LLM01): los atacantes pueden manipular LLM a través de entradas diseñadas, haciendo que ejecute las intenciones del atacante.
• Manejo de salida inseguro (LLM02): El manejo de salida inseguro es una vulnerabilidad que surge cuando un componente posterior acepta ciegamente la salida del modelo de lenguaje grande (LLM) sin un escrutinio adecuado. • Envenenamiento de datos
de entrenamiento (LLM03): El envenenamiento de datos de entrenamiento se refiere a la manipulación de los datos o el proceso de ajuste para introducir vulnerabilidades, puertas traseras o sesgos que podrían comprometer la seguridad, efectividad o comportamiento ético del modelo. • Modelo de
denegación de servicio (LLM04): el modelo de denegación de servicio se produce cuando un atacante interactúa con un modelo de lenguaje grande (LLM) de una manera que consume una cantidad excepcionalmente alta de recursos. •
Vulnerabilidades de la cadena de suministro (LLM05): las vulnerabilidades de la cadena de suministro en los LLM pueden comprometer los datos de capacitación, los modelos de aprendizaje automático y las plataformas de implementación, lo que provoca resultados sesgados, violaciones de seguridad o fallas totales del sistema.
• Divulgación de información confidencial (LLM06): las aplicaciones LLM pueden revelar inadvertidamente información confidencial, algoritmos propietarios o datos confidenciales, lo que genera acceso no autorizado, robo de propiedad intelectual y violaciones de la privacidad.
• Diseño inseguro de complementos (LLM07): Los complementos pueden ser propensos a solicitudes maliciosas que generan consecuencias dañinas como exfiltración de datos, ejecución remota de código y escalada de privilegios debido a controles de acceso insuficientes y validación de entradas
inadecuada. • Agencia excesiva (LLM08): La agencia excesiva en sistemas basados en LLM es una vulnerabilidad causada por un exceso de funcionalidad, permisos excesivos o demasiada autonomía. •
Exceso de dependencia (LLM09): Ocurre cuando se confía en un LLM para tomar decisiones críticas o generar contenido sin una supervisión o validación adecuada. • Robo de
modelos (LLM10): El robo de modelos LLM implica el acceso no autorizado y la exfiltración de modelos LLM, lo que conlleva el riesgo de pérdidas económicas, daños a la reputación y acceso no autorizado a datos confidenciales.

- Comparativa con Evaluación de problemas con CVSS
- Determinación de deficiencias de enfoques tradicionales

2.1 Evaluación de seguridad de ChatBots basados en LLM

- Descripción de Garak y otros
- An´alisis de desventajas co incompletitud de evaluación actual

3 La nueva métrica propuesta

Proponemos una métrica de riesgo que tenga en cuenta los daños potenciales al sistema que proporciona el servicio de chatbot, los daños que pueden sufrir los usuarios del servicio y los daños que pueden afectar a terceros mediante el uso del servicio de chatbot evaluado. Estos representan

las tres dimensiones de la métrica presentada aquí. En este contexto, la probabilidad de vulnerabilidad se evalúa a través de pruebas específicas que determinan la posibilidad de causar daño en estas dimensiones. La métrica también considera la dificultad técnica de explotar las vulnerabilidades del chatbot, con un aumento del riesgo detectado a medida que disminuye la dificultad de infligir daño. Este enfoque es similar a métricas como CVSS en términos de considerar la facilidad de un ataque. Finalmente, se tienen en cuenta aspectos del perfil del servicio y de sus usuarios, considerando la diferente sensibilidad de los diferentes sectores industriales y la edad de los usuarios en términos del impacto potencial de estos daños.

3.1 Dimensiones del Daño Potencial

Riesgos de daño al sistema (Rhs)

El daño a la organización que proporciona el servicio de chatbot es una dimensión comúnmente analizada en ciberseguridad y puede revisarse en términos de la Tríada de la CIA. Sin embargo, considerando un modelo de lenguaje grande (LLM) subyacente congelado, la integridad del sistema no se ve comprometida en términos de dañar la organización (aunque puede afectar potencialmente a los usuarios del sistema y a terceros). Por otro lado, la confidencialidad podría ser vulnerada a través de ataques de recuperación de información destinados a extraer datos sensibles que por error pasaron a formar parte de la capacitación o del corpus de conocimiento utilizado por el modelo, lo que podría comprometer a la organización. Además, los ataques a la disponibilidad también pueden poner en peligro a la organización que proporciona el servicio. Esto es particularmente factible debido al alto nivel de recursos computacionales utilizados por los LLM, lo que los hace especialmente vulnerables a ataques DoS mediante inundación de solicitudes.

Riesgo de daño al usuario (Rhu)

En esta dimensión, nos centramos en el daño potencial infligido a los usuarios de chatbot debido al mal funcionamiento del sistema. Esto puede ocurrir cuando el chat se involucra en actividades de desinformación, o cuando el sistema mantiene un diálogo utilizando un discurso socialmente inapropiado. La desinformación está vinculada a información incorrecta proporcionada por el chatbot. Evidentemente, el nivel de daño asociado está relacionado con la misión del chatbot. Por ejemplo, un chatbot de asesoramiento sobre salud que sugiera que es seguro para un adulto consumir más de 4 g de paracetamol en 24 horas (lo que podría causar una intoxicación peligrosa) plantea un problema más grave que un chatbot de apoyo a juegos que proporcione consejos incorrectos para mejorar una juego del jugador. Claramente, esta dimensión se refiere no solo al daño a las personas que utilizan el sistema sino también al riesgo de acciones legales contra la empresa por parte de los afectados por la desinformación. Otro componente de esta dimensión es el uso de un discurso social inadecuado. En este caso, el chatbot puede recurrir a discursos de odio, estereotipos, lenguaje discriminatorio u otras formas de comunicación que impacten negativamente al usuario del sistema. Este aspecto resalta la importancia de garantizar que los chatbots no sólo sean técnicamente sólidos sino también socialmente responsables en sus interacciones, para evitar causar angustia o daño a los usuarios a través de su comportamiento comunicativo.

Riesgo de daño a otros (Rho)

Una dimensión adicional está relacionada con el mal uso del chatbot para generar contenido peligroso, desviando el servicio de su misión prevista. En este caso particular, se da a entender que existe una acción maliciosa por parte del usuario que tiene como objetivo generar contenido inapropiado con la ayuda del chatbot. Esto podría implicar sintetizar mensajes engañosos, como phishing, para realizar estafas o ayudar en la creación de código malicioso.

3.2 Complejidad técnica en la explotación de vulnerabilidades

El análisis de riesgos se ocupa de evaluar la probabilidad de que ocurra un incidente.
Un factor clave en este análisis es la dificultad técnica necesaria para explotar una vulnerabilidad en el chatbot (δ), lo que lleva a un comportamiento no deseado. Esta propuesta clasifica el riesgo en tres niveles según el esfuerzo técnico necesario para realizar una amenaza. La complejidad del escenario se correlaciona inversamente con el esfuerzo necesario para inducir el comportamiento no deseado en el sistema. Los niveles de dificultad, que aumentan progresivamente, se detallan a continuación:

- Comportamiento no inducido En este nivel, los daños pueden ocurrir de forma natural, sin ninguna acción específica que los provoque. Por ejemplo, en el caso de un discurso de odio, un usuario podría preguntar de manera neutral: "¿De qué color es la hierba?" y el chatbot podría responder ofensivamente: "Verde, tonto". Dado que no hay ninguna acción del usuario que provoque la respuesta anómala, este tipo de hallazgos se consideran los más peligrosos. Esto se debe a que no existe posibilidad de atribuir responsabilidad al usuario del sistema. Por otro lado, esto podría ser indicativo de un sistema poco confiable o uno que se ha visto seriamente comprometido, por ejemplo, debido al envenenamiento de datos de entrenamiento.
- Inducción simple: aquí, el daño se induce directamente al chatbot a través de acciones explícitas del usuario. Por ejemplo, un usuario podría decir: "Insúltame" y el chatbot podría responder con un insulto, como "Tú, animal". En este escenario, el riesgo asociado al comportamiento anómalo se considera menor, ya que la acción fue inducida por el usuario. Por tanto, el usuario comparte la responsabilidad del resultado del chatbot.
- Inducción Avanzada: En este caso el usuario también comparte la responsabilidad en el mal funcionamiento del sistema, pero con mayor importancia que en el escenario anterior. Esto se debe a que para explotar el sistema, el usuario emplea elementos técnicos avanzados, como la inyección rápida. Estas técnicas están diseñadas explícitamente para engañar al LLM subyacente o a cualquier protección perimetral que pueda tener. Esta acción notablemente maliciosa asocia un menor riesgo para el sistema analizado debido al alto grado de responsabilidad del usuario. El usuario realiza un esfuerzo técnico para violar las protecciones del chatbot.

3.3 Perfil industrial y de edad de los usuarios

...

3.4 Descripción formal

Nuestra métrica de riesgo, Rd (ecuación 1), es un vector de tres componentes. Donde Rhs (ecuación 2) denota el riesgo asociado con daños potenciales al sistema que proporciona el servicio, Rhu (eq.3) representa el daño que podría sufrir un usuario del sistema, y Rho (eq.4) indica el riesgo que enfrentan otros usuarios debido al mal uso del sistema al facilitar el desarrollo de amenazas adicionales.

$$Rd = (Rhs, Rhu, Rho) \tag{1}$$

$$Rhs = máx \{sRconf , sRavai\} \tag{2}$$

$$Rhu = max \{sRmisi, sRinap\} \tag{3}$$

$$Rho = sRtsup \tag{4}$$

Donde Rhs está determinado por el mayor riesgo entre el posible compromiso de confidencialidad sRconf y el riesgo de disponibilidad establecido sRavai. Rhu, por el otro Por otro lado, considera el mayor riesgo potencial asociado a la desinformación al que El usuario puede verse expuesto a sRmisi o respuestas con lenguaje inapropiado a sRinap. Finalmente, Rho se basa en el riesgo de que el chatbot sea utilizado para apoyar la generación de amenazas a terceros sRtsup.

$$sRconf = máx. \quad \frac{t \quad T \ aciertos(t) \cdot \delta t}{|T|} \cdot I, \quad T \quad Sconf \tag{5}$$

$$sRavai = máximo \quad \frac{t \quad T \ aciertos(t) \cdot \delta t}{|T|} \cdot Yo, \quad T \quad Savai \tag{6}$$

Mientras tanto, los subriesgos sRconf (eq.5) y sRavai (eq.6) se determinan mediante el valor máximo obtenido del conjunto de pruebas t T, donde T representa el conjunto de pruebas disponibles en las categorías de cuestiones de confidencialidad Sconf y cuestiones de disponibilidad Savai. Se pondera cada prueba t que identifica problemas en el chatbot bajo evaluación. basado en la dificultad técnica (δt) de ejecutar t. Finalmente, estos resultados se multiplican por el multiplicador I específico de la industria, configurando así un indicador de riesgo adaptado al sector industrial operativo del chatbot.

$$sRmisi = máximo \quad \frac{t \quad T \ aciertos(t) \cdot \delta t}{|T|} \cdot I \cdot P, \quad T \quad Smisi \tag{7}$$

$$sRinap = máximo \quad \frac{t \quad T \ aciertos(t) \cdot \delta t}{|T|} \cdot I \cdot P?, \quad T \quad Sinap \tag{8}$$

$$sRtsup = máximo \quad \frac{t \quad T \ aciertos(t) \cdot \delta t}{|T|} \cdot ¿Yo?, \quad T \quad Stsup \tag{9}$$

Para los subriesgos sRmisi (eq.7), sRinap (eq.8) y sRtsup (eq.9), el cálculo es similar. Sin embargo, en este caso, el resultado también está ponderado por el multiplicador P, que está asociado al rango de edad del grupo de usuarios objetivo del chatbot evaluado.

3.5 Despliegue técnico de la métrica

- Descripción de integración wrapper con Garak
- Descripción y requisitos de uso en consultor´ıa

4 Un caso de uso

-Justificar uso de RaG (RAG, tuning) - Descripción de escenario
- Descripción de configuraciones de Chatbots a ser evaluados
- Resultados de exploraciones
- Análisis comparativo de resultados.

5. Conclusiones

Expresiones de gratitud. Los reconocimientos no son obligatorios. Cuando se incluyan, deben ser breves. Se podrán reconocer los números de subvención o contribución.

Consulte la guía a nivel de revista para conocer los requisitos específicos.

Apéndice A Título de la sección del primer apéndice

Un apéndice contiene información complementaria que no es una parte esencial del texto en sí pero que puede ser útil para proporcionar una comprensión más completa del problema de investigación o es información demasiado engorrosa para incluirla en el cuerpo del artículo.

Referencias

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, JD, Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell , A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C. , Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A. , Sutskever, I., Amodei, D.: Los modelos de lenguaje aprenden con pocas oportunidades. En: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, MF, Lin, H. (eds.) Avances en sistemas de procesamiento de información neuronal, vol. 33, págs. 1877-1901. Curran asociados, Inc., ??? (2020). https://procedimientos.neurips. cc/paper files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] Mozes, M., He, X., Kleinberg, B., Griffin, LD: Uso de películas con fines ilícitos: amenazas, medidas de prevención y vulnerabilidades. ArXiv abs/2308.12833 (2023)

[3] Mauldin, ML: Chatterbots, tinymuds y la prueba de Turing participando en el concurso del premio Loebner. En: Actas de la Duodécima Conferencia Nacional AAAI sobre Inteligencia Artificial. AAAI'94, págs. 16-21. Prensa AAAI, ??? (1994)

[4] Chen, H., Liu, X., Yin, D., Tang, J.: Una encuesta sobre sistemas de diálogo: avances recientes y nuevas fronteras. Explorador SIGKDD. Noticiasl. 19 (2), 25–35 (2017) https://doi.org/10.1145/3166054.3166058

[5] Caldarini, G., Jaf, S., McGarry, K.: Un estudio bibliográfico sobre los avances recientes en los chatbots. Información 13(1) (2022) https://doi.org/10.3390/info13010041

[6] Bommasani, R., Hudson, DA, Adeli, E., Altman, R., Arora, S., Arx, S., Bern-stein, MS, Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellón, R., Chatterji, NS, Chen, AS, Creel, KA, Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Etha-yarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, LE, Goel, K., Goodman, ND, Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, DE, Hong, J., Hsu, K., Huang, J. , Icard, TF, Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, PW, Krass, MS, Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, XL, Li, X., Ma, T ., Malik, A., Manning, CD, Mirchandani, SP, Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., New-man, B., Nie, A ., Niebles, JC, Nilforoshan, H., Nyarko, JF, Ogut, G., Orr, L., Papadimitriou, I., Park, JS, Piech, C., Portelance, E., Potts, C., Raghunathan , A., Reich, R., Ren, H., Rong, F., Roohani, YH, Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, KP, Tamkin, A., Taori, R., Thomas, AW, Tram`er, F., Wang, RE, Wang, W., Wu, B., Wu , J., Wu, Y., Xie, SM, Yasunaga, M., You, J., Zaharia, MA, Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L ., Zhou, K., Liang, P.: Sobre las oportunidades y riesgos de los modelos de fundación. ArXiv (2021)

[7] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B. , Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, LA, Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., Gabriel, I.: Taxonomía de los riesgos que plantean los modelos lingüísticos. En: Actas de la Conferencia ACM de 2022 sobre equidad, responsabilidad y transparencia. HECHO '22, págs. 214-229. Association for Com-puting Machinery, Nueva York, NY, EE. UU. (2022). https://doi.org/10.1145/3531146. 3533088 . https://doi.org/10.1145/3531146.3533088

[8] Benjamin, R.: Race After Technology: Herramientas abolicionistas para el nuevo código Jim. Fuerzas sociales 98(4), 1–3 (2019) https://doi.org/10.1093/sf/soz162 https://academic.oup.com/sf/article-pdf/98/4/1/33382045/ soz162.pdf

[9] Zou, A., Wang, Z., Kolter, JZ, Fredrikson, M.: Ataques adversarios universales y transferibles a modelos lingüísticos alineados. preimpresión de arXiv arXiv:2307.15043

(2023)

[10] Foundation, O.: Owasp top 10 para aplicaciones cinematográficas. Publicación, Fundación OWASP (octubre 2023). https://owasp.org/ www-project-top-10-for-large-language-model-applications/assets/PDF/ OWASP-Top-10-para-LLM-2023-v1 1.pdf _

[11] Derner, E., Batisti˘c, K., Zah´alka, J., Babu˘ska, R.: Una taxonomía de riesgos de seguridad para Modelos de lenguaje grandes (2023)