

Autosupervisión de Alucinaciones en Grandes Modelos del Lenguaje: LLteaM

Sofía Correa Busquets*, Lucas Maccarini Llorens

Wird, Chile

*Autora de correspondencia: sof.ia@wird.ai

Recibido: 27 Septiembre 2023 / Aceptado: 15 Noviembre 2023 / Publicado: 18 Diciembre 2023

Abstract

Large language models like GPT and Claude have revolutionized the tech industry over the past year. However, as generative artificial intelligence, they are prone to hallucinations. A large language model hallucinates when it generates false or nonsensical text. As these models improve, these hallucinations become less obvious and more dangerous for users. This research explores the phenomenon in the context of automated email response for customer service. First, it proposes a taxonomy of hallucinations in large language models based on their linguistic nature, and second, a multi-agent system that allows for the self-supervision of such hallucinations. This system generates email responses but prevents their delivery if hallucinations are detected, thus reducing the risks of generative AI in productive environments. Experiments with various state-of-the-art language models reveal that the only successful model's operating costs currently exceed those viable for operational deployment. Moreover, a drastic performance drop after a recent update to GPT-3.5-turbo suggests likely shortcomings in industrial applications driven by retrieval-augmented generation. Overall, the research advocates for a Machine Linguistics to analyze the outputs of large language models, suggesting that such a collaboration between Linguistics and Artificial Intelligence could help mitigate the social risks of hallucination.

Keywords: large language model, hallucination, Chain-of-Thought prompting, self-supervision, retrieval-augmented generation

Resumen

Los grandes modelos del lenguaje, como GPT y Claude, han revolucionado la industria tecnológica en el último año. Sin embargo, al ser inteligencia artificial generativa, estos son susceptibles a sufrir alucinaciones. Un gran modelo del lenguaje alucina cuando genera texto falso o falto de sentido, pero en la medida que estos modelos mejoran, estas alucinaciones se vuelven menos evidentes y más peligrosas para el usuario. La presente investigación explora el fenómeno en un caso de respuesta automática de correo electrónico en servicio al cliente. Se

propone, primero, una taxonomía de alucinaciones en grandes modelos del lenguaje según su naturaleza lingüística y, segundo, un sistema multiagente que permite la autosupervisión de las mismas. Este redacta respuestas a correos electrónicos, pero previene su entrega en caso de detectar alucinaciones, reduciendo así los riesgos de la inteligencia artificial generativa en entornos productivos. La experimentación con distintos modelos del lenguaje de vanguardia evidencia que los costos del único modelo exitoso son actualmente demasiado altos para su puesta en producción. Además, la drástica caída de rendimiento tras una actualización de versión de GPT-3.5-turbo revela probables deficiencias en aplicaciones industriales impulsadas por la generación aumentada por recuperación. En general, la investigación defiende la necesidad de una lingüística de máquinas para analizar las producciones de los grandes modelos del lenguaje y que dicha colaboración entre la lingüística y la inteligencia artificial puede ayudar a mitigar los riesgos sociales de la alucinación.

Palabras clave: gran modelo del lenguaje, alucinación, enunciado de cadena lógica, autosupervisión, generación aumentada por recuperación

1. INTRODUCCIÓN

La industria y la academia en la inteligencia artificial (en adelante IA) han experimentado avances acelerados en el último tiempo. Esto se debe principalmente a un cambio de paradigma: desde la necesidad de etiquetar grandes conjuntos de datos para entrenar modelos a la medida, hacia la nueva ingeniería de enunciados (*prompt engineering*) para uso de grandes modelos de lenguaje (en adelante LLM, acrónimo que proviene de *large language model*). Los LLM son sistemas avanzados basados en redes neuronales y mecanismos de atención (Vaswani et al. 2017), que exhiben una capacidad excepcional para procesar, traducir y generar texto similar al humano. Esta les permite abordar tareas del mundo real como pregunta y respuesta, resúmenes de texto y generación de contenido, con necesidad de muy pocos ejemplos previos o incluso ninguno. Modelos como GPT-3.5-turbo, Bard, Claude y LLaMa (OpenAI 2022; Anthropic 2023; Google 2023; Touvron et al. 2023) han potenciado la automatización de comunicaciones digitales en múltiples áreas, incluida la de servicio al cliente. Sin embargo, la implementación efectiva de LLM en entornos productivos no está exenta de desafíos, entre ellos, el fenómeno conocido como alucinación. El modelo alucina cuando genera información falsa o carente de sentido. Por ejemplo, en el contexto médico, un LLM podría generar recomendaciones de tratamiento inapropiadas. Esto es una preocupación en la medida que las aplicaciones que utilizan LLM arrastran estos errores hacia el usuario y le pueden conducir a tomar decisiones incorrectas e incluso dañinas.

En este contexto, Wird, una empresa chilena que ofrece servicios de mejora de productividad mediante el uso de Procesamiento del Lenguaje Natural (en adelante PLN), se propone integrar servicios que aprovechen los LLM. Generalmente, Wird procesa datos de correo electrónico en español chileno, recibidos por áreas de servicio al cliente. Uno de los servicios que ofrece incluye la resolución automática de casos, como consultas por stock, peticiones de contacto telefónico y seguimiento de pedidos. Este desafío se aborda mediante la redacción manual de

respuestas a requerimientos frecuentes, para luego identificar y responder automáticamente estos requerimientos a medida que llegan a las bandejas de entrada de los ejecutivos. Es decir, la metodología depende del criterio experto, un bien escaso y demoroso, para estructurar las casuísticas relevantes de un negocio. La eficientización de este proceso implica un balance entre el potencial de los LLM y el riesgo de alucinación.

Los LLM, por defecto, no están bien adaptados a tareas como la resolución de casos que provee Wird. Esto se debe a que se entrenan con textos públicos que no reflejan los protocolos internos específicos de ninguna empresa en particular. Para abordar esta problemática, se pueden utilizar bases de conocimiento no estructuradas, como páginas de preguntas frecuentes y documentación interna, para afinar (*fine-tune*) la salida de los LLM. No obstante, los altos costos de entrenamiento, operación y mantención de estos modelos personalizados (*fine-tuned*) han ido en desmedro de su popularidad en la industria.

Una solución a este problema es el uso de sistemas de generación aumentada por recuperación (en adelante RAG, acrónimo que proviene de *Retrieval-Augmented Generation*), como LangChain y LlamaIndex (Chase 2022; Liu 2022). Estos buscan extraer de una base de conocimiento el contexto relevante para insertar en la consulta y condicionar la salida de un LLM por defecto. No obstante, RAG resultó ser un método poco efectivo y riesgoso para los casos de uso de Wird. En una prueba de concepto interna, se propuso mediante LangChain una plantilla de respuesta para cada uno de quince tópicos principales en correo electrónico para una aseguradora. De estas quince, doce fueron vetadas por criterio experto de negocio. Este rendimiento descartó la posibilidad de integrar un asistente virtual impulsado por LangChain a responder en el ambiente productivo.

Sin embargo, entre las respuestas rechazadas por conocimiento experto de negocio, se observaron ciertas regularidades lingüísticas que las distinguen de una respuesta automática admisible, similares a la noción actual de alucinación. Un sistema que pudiera discriminar automáticamente estas alucinaciones que perjudican el servicio al cliente, y así prevenir el envío de estos mensajes, haría viable la respuesta generativa en un ambiente productivo. Esto implicaría procesar, con un costo por extensión de texto, más respuestas de las que se envían y por lo tanto cobran al cliente. Es decir, solo sería rentable si el costo de generar muchas respuestas fuera significativamente inferior a la ganancia de enviar unas pocas. Lo anterior motiva la pregunta ¿mediante qué producto práctico y teórico se pueden generar automáticamente respuestas coherentes y relevantes a usuarios en un contexto de la industria de servicios, minimizando el riesgo productivo de alucinación?

Los objetivos de la investigación que se describe en este artículo son dos:

- Tipificar lingüísticamente las alucinaciones en el uso de RAG para responder correos electrónicos de servicio al cliente.
- Proponer una arquitectura que minimice el riesgo productivo de alucinación al generar automáticamente respuestas a usuarios en un contexto de la industria de servicios.

La consecución de estos objetivos requiere, en primer lugar, posicionar este desarrollo dentro de una subárea de la lingüística cuyo objeto de estudio no es necesariamente una producción

humana. En segundo lugar, se exponen las características de los datos de la muestra y el procedimiento de generación de respuesta. Luego, se proponen (a) una tipología ‘probLLeMas’ de alucinaciones observadas en el caso de uso de responder a correos electrónicos por parte de un ejecutivo del servicio al cliente y (b) una arquitectura de PLN ‘LLteaM’ para autosupervisar las alucinaciones en un ambiente productivo. A continuación, se exponen los resultados obtenidos mediante este producto teórico-práctico. Finalmente, se discuten las limitaciones y proyecciones del mismo en futuras implementaciones para el ámbito industrial.

2. EN DEFENSA DE UNA LINGÜÍSTICA DE MÁQUINAS

El impacto industrial de la IA generativa ha sufrido un crecimiento explosivo desde el lanzamiento de ChatGPT (OpenAI 2022). McKinsey & Co. evalúa la IA generativa como principal actor en su último informe anual de tendencias en tecnología (Chui et al. 2023). Estiman que el interés público en ella se triplicó del 2021 al 2022 y que sus casos de uso en mejoras de productividad pronto agregarán hasta USD 4,4 trillones en valor económico. Asimismo, durante el primer semestre del 2023 se lanzaron integraciones nativas para desarrollar con estas tecnologías a la infraestructura de los mayores servicios de computación en la nube: Microsoft Azure, Amazon Web Services y Google Cloud Platform (Microsoft 2023; Sivasubramanian 2023; Tharp & Moore 2023). Tras seis décadas de citar el Test de Turing (1950), no viene como sorpresa que el momento estelar de la IA sea el de imitar satisfactoriamente la conversación humana.

Una preocupación durante este cambio de paradigma es la excesiva confianza de los usuarios en la exactitud de las producciones de LLM. Un ejemplo es el litigio Mata vs. Avianca, donde un abogado del equipo demandante presentó en un juicio casos ficticios, encontrados mediante ChatGPT, como precedentes (Bohannon 2023). El mismo OpenAI (2023a) reporta estas ‘alucinaciones’ como primer factor de riesgo de su modelo. Incluso aseveran que este riesgo incrementa con el éxito: la entrega de información veraz, en dominios que le son familiares al usuario, le instan a confiar en el modelo y sobredepender del mismo en dominios ajenos.

La alucinación en los LLM se define como la producción de texto ya sea falso o falto de sentido (Maynez et al. 2020). En esta definición destaca la oposición entre fidelidad y factualidad. Una producción es fiel si se ciñe a la información entregada en el corpus de entrenamiento y es factual si se ciñe a la información veraz sobre el mundo. Así, Maynez et al. (2020) categorizan las alucinaciones según su relación lógica al corpus de entrenamiento. Identifican alucinaciones intrínsecas, en las que la contrafactualidad viene del mismo corpus, y alucinaciones extrínsecas, que no se desprenden lógicamente de él o que lo contradicen. Este enfoque en fidelidad y factualidad, que ha dominado la escasa literatura sobre alucinaciones, relega el sinsentido a segundo plano.

En esta discusión sobre veracidad y sentido de las producciones textuales, la lingüística ha guardado silencio. Si bien la lingüística moderna especifica su objeto de estudio como las manifestaciones del lenguaje humano (de Saussure 1916), no existía razón al momento de su fundación para considerar que las máquinas pudieran replicar los verdaderos patrones de distribución de diversas lenguas a partir de instancias de habla humana. Incluso, la relevancia

de las alucinaciones radica en cómo el texto generado por la máquina se relaciona con, e impacta sobre, un contexto sociocultural. Dada esta naturaleza lingüística y la ubicuidad de las aplicaciones LLM, el compromiso social del área nos llama a describir los efectos indeseados del texto generado por la máquina para construir protección contra los mismos. Resulta imperativo promover una nueva 'lingüística de máquinas' que analice la pragmática en contexto de la comunicación entre máquina y hablante.

Especialmente para el caso de uso donde un LLM habla por parte de otra persona y/o institución, un análisis en lingüística de máquinas se puede nutrir de la Teoría de Actos de Habla. Esta describe cómo el uso del lenguaje genera acciones en el mundo, llamando la acción que subyace a cada uso del lenguaje un 'acto de habla'. Searle (1979) los categoriza en:

1. **Asertivos.** Aseveraciones sobre el mundo: comprometen la palabra del hablante a que lo dicho sea cierto.
2. **Directivos.** Desde sugerencias hasta órdenes: intentan lograr que el interlocutor haga algo.
3. **Expresivos.** Expresan el estado psicológico del hablante respecto de algo.
4. **Comisivos.** Promesas: comprometen al hablante a llevar a cabo alguna acción.
5. **Declarativos.** Acciones que solo pueden ser llevadas a cabo a través de las palabras: oraciones que se vuelven ciertas al decirlas. Por ejemplo, 'los declaro marido y mujer.'

Además, el éxito de cada acto de habla depende de 'condiciones de felicidad' (Austin 1975). La más intuitiva es la de los actos asertivos, es decir, que el hablante diga la verdad. Se puede interpretar el enfoque actual de las alucinaciones como una reducción desde la felicidad a la factualidad. La reducción se daría porque el caso de uso del LLM como asistente de propósito general, donde a menudo tiene que actuar como fuente de información, tiene un alto tráfico y divulgación en los medios masivos de comunicación y la academia. En esos casos, prácticamente el único acto de habla cuya felicidad peligra es el asertivo, ya que el conjunto de datos de entrenamiento no contendría ejemplos de respuesta donde el asistente virtual cometa tipos de actos de habla cuyas condiciones de felicidad nunca está facultado para cumplir. Sin embargo, en los casos de RAG, la respuesta está parcialmente modelada por ejemplos dichos por personas, facultadas para comprometerse a acciones futuras, realizar trámites en la institución donde trabajan, entre otras cosas. Como estos ejecutivos rutinariamente cometen actos de habla comisivos y asertivos con sus clientes sobre las acciones humanas que llevan a cabo en el trabajo, ambos tipos de acto de habla serán particularmente problemáticos en el presente caso de uso. Los actos directivos también le pueden generar una mala experiencia al usuario, si el LLM le sugiere cursos de acción inútiles o perjudiciales para la solución de su requerimiento. En resumen, el valor de verdad de los textos no siempre se denomina 'verdad' ni se debiera reducir a la factualidad, sino que es, en un sentido más extenso, la felicidad del acto de habla.

Otro punto desde el cual complejizar la noción de alucinación es el hecho de que OpenAI recibe retroalimentación negativa para ChatGPT en base a criterios más amplios que la contrafactualidad. Estos incluyen que la respuesta sea ofensiva, no sea verdadera o no sea útil (OpenAI 2022). La contrafactualidad está limitada al segundo criterio, mientras que el

sinsentido oscilaría entre el segundo y el tercero. Dejando de lado la ofensividad, que se ha desarrollado ampliamente como ‘toxicidad’ del modelo y otros sesgos sociales, se pueden complejizar estos criterios desde la lingüística mediante el principio de cooperación (Grice 1975). Este se compone de cuatro máximas:

1. **Máxima de cantidad.** No dar más ni menos información de la necesaria.
2. **Máxima de calidad.** No mentir intencionalmente ni hacer afirmaciones sin pruebas suficientes.
3. **Máxima de relevancia.** Comunicar información relevante.
4. **Máxima de modo.** Expresarse con claridad, precisión, brevedad y orden.

La contrafactualidad transgrede la máxima de calidad. En cambio, una transgresión de cualquiera de las otras tres máximas de comunicación puede resultar en una respuesta inútil. Así, el principio de cooperación puede servir como un criterio desglosado para entender la alucinación de manera cualitativa.

3. CARACTERÍSTICAS DE LA MUESTRA

Los datos utilizados en esta investigación provienen de la División de Seguros Vehiculares de una compañía de seguros en Chile. Esta división, crucial en la interacción con los clientes y en la coordinación de suministros a talleres asociados, está compuesta por seis ejecutivos, excluyendo a su jefatura directa. La muestra seleccionada abarca un total de 100 textos de correo electrónico, correspondientes al periodo de enero a mayo de 2023. Cada texto tiene una longitud media de 260 palabras y aborda una diversidad de consultas y solicitudes que los clientes realizan a su ejecutivo de seguros, reflejando así una amplia gama de operaciones diarias y logísticas.

Es importante destacar que existe un cierto nivel de pérdida de información en la extracción, transformación y carga de los datos. La extracción normaliza en el cuerpo del mensaje los aspectos paralingüísticos de formateo del uso de negritas, itálicas, tamaño de fuente o tipografía. Sin embargo, se preserva el uso natural de espacios en blanco como pueden ser los saltos de línea, diacríticos ortográficos y distinción entre mayúsculas y minúsculas. Aun así, la transformación del dato a texto plano conlleva la pérdida de las imágenes insertas en el cuerpo del mensaje. Además, los adjuntos de correo, típicamente en formato PDF, no se incluyen debido a la variabilidad de su formato. Lo anterior genera casos tales como

(1) Asunto: SIN [NÚMERO DE SINIESTRO] PTE [NÚMERO DE PATENTE]

Diego

favor comprar este repuesto. Saludos.

No obstante, cargar los adjuntos como parte íntegra del dato conllevaría dos complicaciones: en primer lugar, el nivel de exactitud del proceso de reconocimiento óptico de caracteres, y en segundo lugar, la pérdida de contexto del modelo sobre el cuerpo del mensaje por la extensión del adjunto. Así, en la carga de los mensajes se considera solamente el asunto y cuerpo del correo.

4. HACIA UNA TAXONOMÍA DE PROBLEMAS

La siguiente taxonomía comprende una serie de configuraciones lingüísticas donde la producción de un agente basado en LLM se ha determinado inadmisibles para el caso de uso de una conversación entre un ejecutivo del servicio y un cliente. Sus primeros dos macrotipos, los probLLeMas existenciales y probLLeMas-comprensión, añaden una dimensión sociopragmática a la faceta poco explorada del sinsentido en la alucinación. Mientras tanto, el tercer y último macrotipo, los probLLeMas-chamuyo, desarrolla la conocida faceta contrafactual de la alucinación, pero disputa el valor convencional de la factualidad posicionando la fuente de verdad en el individuo que enuncia –no en el conocimiento universal.

4.1. ProbLLeMas Existenciales

En los probLLeMas existenciales, el agente genera comportamiento verbal que no se condice con el rol social que está cumpliendo en el contexto de la interacción. Este rol involucra hablar por parte de una institución: la empresa que contrata el servicio de respuesta automática. Podemos inferir, entonces, que la aparición de un probLLeMa existencial implica un fracaso de entendimiento de la organización social de los participantes de la interacción en subcomunidades de hablantes, como una empresa en particular.

4.1.1. ProbLLeMa Existencial-rol/medio

En el probLLeMa existencial-rol/medio, el agente asigna su propia identidad, atribuciones y/o actual medio de comunicación, a un tercero. En general, redirige al usuario con este tercero como la fuente de probable resolución para su requerimiento. Dado el contexto, el usuario es capaz de identificar la respuesta como sinsentido, lo cual expandiría esta noción como un caso de alucinación más allá de lo esbozado por Maynez et al. (2020). Por ejemplo:

(2) Estimado cliente,

Gracias por su mensaje. Lamentablemente, no tengo acceso a la información específica del siniestro mencionado en su correo. **Le recomendaría contactar directamente al departamento de Central de Repuestos de Automóviles de Seguros [EMPRESA]¹** para obtener una respuesta adecuada a su solicitud.

Si necesita alguna otra información o tiene alguna otra consulta, no dude en comunicarse con nosotros.

Atentamente,

[Nombre del ejecutivo de soporte técnico]²

Como se mencionó anteriormente, este texto pretende responder de parte de la misma Central de Repuestos de Automóviles, por lo que la respuesta no es admisible.

¹ Todas las negritas en los ejemplos de probLLeMa son propias, para facilitar la identificación del punto de ocurrencia de la alucinación en la respuesta.

² Esta notación de corchetes corresponde a una producción espontánea del agente. No confundir con notación propia de anonimización, la cual lleva mayúsculas.

El peligro del probLLeMa existencial-rol/medio es que perjudique la experiencia de usuario a través de actos directivos inútiles. También, es contraproducente a la eficientización de los procesos de negocio, ya que tiene alta probabilidad de hacer que la tasa de insistencia del usuario suba. El efecto será aún peor si el agente redirige a un medio distinto al actual: por ejemplo, aseverando que las personas capaces de resolver el requerimiento están disponibles mediante la página web de la empresa, cuando en realidad son dueños de la presente dirección de correo electrónico. Casos así dañarían la trazabilidad del requerimiento, dificultando que los usuarios puedan encontrar solución a sus trámites.

4.1.2. ProbLLeMa Existencial *LMsplaining*

El *LMsplaining* ocurre cuando el agente refrasea el texto de entrada y lo presenta como información novedosa. En este sentido, todo *LMsplaining* implica un probLLeMa comprensión-intención, el cual se describe más adelante, ya que hay una falla de entendimiento de la información o acción que el usuario esperaba recibir como consecuencia directa de su mensaje. No obstante, el *LMsplaining* se agrupa como probLLeMa existencial dada su compleja naturaleza cuando se presenta una entrada inadmisible por error de los metadatos. El filtrado de correos internos debería haber eliminado mensajes provenientes de otros ejecutivos de la misma empresa, pero en ocasiones se pueden recibir correos como el siguiente:

- (3) **Estimado cliente:** Buenas tardes, adjunto nivel de servicio correspondiente al mes de marzo, favor revisar. **NS febrero 91,6% NS marzo 89,2% Acumulado 91,3%**
Atentamente.

Cuando esto ocurre, el mensaje de entrada tiende a asemejarse mucho a la base de conocimiento de la empresa, porque el ejecutivo que escribió posee ese conocimiento personalmente. Entonces, la salida es casi la misma que la entrada:

Estimado cliente,

Agradezco que nos haya enviado el nivel de servicio correspondiente al mes de marzo. **Hemos revisado la información y hemos notado que el nivel de servicio para ese mes fue del 89,2%, lo cual es un poco más bajo que el nivel de servicio acumulado hasta la fecha, que es del 91,3%.**

Estamos trabajando arduamente para mejorar y asegurar que nuestros clientes tengan una experiencia óptima con nuestros servicios. Agradecemos su paciencia y comprensión mientras trabajamos en esto.

Si tiene alguna pregunta adicional o requiere más información, por favor no dude en contactarnos.

Atentamente,

[Nombre del ejecutivo de soporte técnico]

Seguros [EMPRESA]

La complejidad adicional de este caso recae, además, en que si un ejecutivo ya habló por parte de la empresa, el agente virtual debería renunciar a su turno de habla, por máxima de cantidad.

Esta consideración parece ser muy compleja para el LLM de manejar en paralelo con la intención de siempre tener buena disposición de ayudar al usuario que proporciona la entrada.

4.2. ProbLLeMas-comprensión

En los probLLeMas-comprensión, la respuesta del agente no se condice con cierto contenido informativo relevante en el mensaje de usuario, ya sea información a nivel semántico o pragmático. Es decir, esta respuesta falla a la máxima de relevancia. Cabe destacar que la identificación de un probLLeMa-comprensión es posible para cualquier hablante de la misma lengua por intuición propia, solo sabiendo la relación pregunta-respuesta de los mensajes y sin necesidad de contexto adicional.

4.2.1. ProbLLeMa Comprensión-tema

La respuesta del agente trata explícitamente sobre conceptos distintos a los que trataba el mensaje original del usuario y que no se desprenden lógicamente del mismo. Véase:

- (4) Don Pedro buenas tardes, para un **Hyundai Santa Fe** siniestro [NÚMERO DE SINIESTRO] **no tengo orden de compra por el neumático y tecele**, Favor de revisar caso, vehículo en zona de armado. [NÚMERO IDENTIFICADOR DE CLIENTE][PATENTE DEL VEHÍCULO]HYUNDAISANTA FE

cuya respuesta fue la siguiente:

Estimado Don Pedro,

Gracias por contactarnos. Lamentablemente, no tengo acceso a la información de su caso específico ya que **no menciona ninguna consulta relacionada con el área de Central de Repuestos de Automóviles de Seguros [EMPRESA]**.

Le recomendaría que se comuniqué directamente con nuestro **departamento de atención al cliente**, quienes podrán brindarle la asistencia necesaria y verificar el estado de su solicitud.

Quedamos a su disposición para cualquier otra consulta o inquietud que pueda tener.

Saludos cordiales.

Aquí, el agente no reconoce que todo el mensaje de entrada es efectivamente atingente a repuestos de automóviles y menciona el área de servicio al cliente sin que esta guarde relación con el mensaje de entrada. En otros casos, admitimos alternativamente que el agente virtual se contradiga dentro de su respuesta como ejemplos de la misma alucinación. Justificamos esta acepción en base a que ambos efectos parecen apuntar a una falta de entendimiento del valor conceptual de los elementos lingüísticos que se manipulan.

4.2.2. ProbLLeMa Comprensión-intención

Existe un probLLeMa de tipo comprensión-intención si la respuesta del agente trata sobre lo mismo que el requerimiento del usuario, pero no ataca este requerimiento. Es decir, el agente maneja efectivamente la información semántica del mensaje de entrada, pero falla en el manejo

de la carga informativa pragmática. Tomemos el siguiente ejemplo:

- (5) Estimados: Buenas tardes, informo que la encomienda que se muestra en las imágenes adjuntas arribó a sucursal Concepción **sin etiqueta, favor confirmar a que OT corresponde** este pedido con siniestro [NÚMERO DE SINIESTRO]. Quedo atento a sus comentarios. Atentamente.

A esto, el agente respondió:

Estimado cliente,

Agradecemos su comunicación. Lamentablemente, no contamos con la información necesaria para identificar a qué OT corresponde el pedido sin etiqueta que menciona. **Le solicitamos que nos proporcione más detalles, como el número de OT** o cualquier otra información relevante, para poder investigar y brindarle una respuesta precisa.

Quedamos atentos a sus comentarios.

Saludos cordiales.

Dado que el usuario pregunta a qué orden de trabajo (OT) corresponde el pedido, se entiende que está pidiendo el número de OT y que por la falta de etiqueta no tiene cómo saberlo. Es decir, al agente le pidieron información y el agente pidió esta misma información de vuelta en lugar de entregarla. Como se mencionaba anteriormente, también el *LMsplaining* siempre lleva consigo un probLLeMa comprensión-intención.

4.3. ProbLLeMas-chamuyo

En los probLLeMas-chamuyo, el agente responde un mensaje que tiene sentido y guarda buena relación con la pregunta, pero compromete estados de la realidad que no tiene forma de verificar ni alterar. Para efectos prácticos falla a la máxima de calidad, sobre todo considerando que el agente conoce las fuentes de información a las cuales tiene acceso y a las que no lo tiene (véase el apartado 5.1.2). Sin embargo, teóricamente sería discutible si el agente sufre pérdida de memoria de esta información y por lo mismo, durante la producción, se encuentra en un estado computacionalmente equivalente al convencimiento de que la información que comunica es veraz y de que tiene pruebas suficientes de ello. Es por esto que haría falta un criterio de lingüística de máquinas para adaptar la máxima a la aplicación en LLM y definir que lo operativamente relevante es el efecto práctico que sufre el usuario por una falencia no de ingeniería de enunciados sino de rendimiento del LLM. Existe un probLLeMa-chamuyo siempre que el agente cometa un acto de habla pero no cumpla con sus condiciones de felicidad.

4.3.1. ProbLLeMa Chamuyo-pasado

Se dan casos en los cuales el agente, aunque se le indique que es un ejecutivo virtual, entra en personaje humano y afirma haber llevado a cabo acciones que los ejecutivos pueden llevar a cabo, pero el agente no.

- (6) Estimado/a cliente,

Gracias por contactarnos. Lamentamos informarle que actualmente no contamos con el guardafangos que necesita para su vehículo. **Hemos realizado varias cotizaciones sin éxito y hemos intentado contactarlo para ofrecerle una indemnización como alternativa**, pero hasta el momento no hemos obtenido respuesta.

(...)

En efecto, el agente miente; es una alucinación intrínseca en los términos de Maynez et al. (2020), ya que constituye una imitación fiel del historial de conversaciones enunciadas por ejecutivos humanos en la base de conocimiento. Sin embargo, sostenemos que la mentira sobre el propio actuar se diferencia de la mentira sobre otros estados de la realidad por ser el símil de una promesa vacía hacia el pasado. Constituye técnicamente un acto de habla asertivo, pero esconde un acto de habla comisivo, ambos sin cumplir sus condiciones de felicidad. La distinción del probLLeMa chamuyo-pasado en la taxonomía es de extrema relevancia para la lingüística de máquinas en tanto el acto de habla puede cumplir sus condiciones de felicidad, para efectos prácticos, con ajustes de arquitectura informática. Si se lograra identificar automática y exitosamente esta promesa hacia el pasado y así programar una notificación al ejecutivo humano que se asegure de que este cumpla lo antes posible la acción comprometida, la transacción tanto verbal como de servicios con el usuario sería exitosa.

4.3.2. ProbLLeMa Chamuyo-presente

El agente informa sobre un atributo dinámico de algún producto, proyecto, usuario o requerimiento, sin tener acceso en tiempo real a la fuente de información que registra los cambios en ese atributo. Por ejemplo:

(7) Buenas tardes,

Gracias por contactarnos. **Lamentamos informarle que el repuesto solicitado, la correa de distribución, se encuentra agotada en nuestro inventario.**

(...)

En los términos de Searle (1979), el LLM cometió un acto asertivo que no cumple sus condiciones de felicidad; está mintiendo. Aunque no sabemos si la correa de distribución estaba efectivamente agotada en el inventario al momento de la generación de la respuesta, es irrelevante: el punto, para la lingüística de máquinas, es que no podemos saberlo. Al agente se le informó que no tiene acceso a fuentes de información como el inventario y, por lo tanto, está mintiendo, independientemente de que la información que enuncie sea factual o contrafactual. Ante esta incertidumbre, el equipo de desarrollo de una aplicación de PLN debe asumir la presencia de alucinación y tomar medidas para corregir o prevenir esta producción textual.

4.3.3. ProbLLeMa Chamuyo-futuro

El probLLeMa chamuyo-futuro es, simplemente, una promesa vacía. La razón de ser de esta promesa es particularmente el caso de uso: el acto de habla comisivo no cumple sus condiciones de felicidad porque el agente enuncia la promesa por parte de otra persona. Véase:

(8) Estimada María José [APELLIDO],

Gracias por contactarnos y por brindarnos los detalles sobre el cambio de repuesto para el siniestro [NÚMERO DE SINIESTRO]. Lamentamos la inconformidad que esto ha causado y estamos aquí para ayudarte.

Siguiendo los protocolos establecidos, hemos solicitado información al taller para interiorizarnos del caso y poder cambiar la pieza por el repuesto correcto. **Estaremos en contacto contigo para coordinar el retiro del repuesto y realizar el cambio correspondiente.**

Agradecemos tu paciencia y comprensión en este proceso. Si tienes alguna otra consulta o necesitas más información, no dudes en comunicarte con nosotros.

Saludos cordiales,

[Nombre del ejecutivo de soporte técnico]

Otra implicancia es que esta será una comunicación fallida por el hecho de que el LLM está diseñado para sostener esta conversación en un solo punto del tiempo. No está diseñado, en cambio, para mantenerla en el mediano y largo plazo como producto secundario de la transacción extendida y calidad de relación con el usuario. El estado de avance de la tecnología solo permite imaginar una solución donde se le notifique al ejecutivo que se hizo una promesa de su parte; no encontramos razón para creer que los LLM ya hayan desarrollado la capacidad de mantener relaciones sostenidas, complejas y dinámicas con los usuarios que impliquen ‘darle peso a su propia palabra’ –quizá un factor a incluir en el venidero sucesor del Test de Turing (1950).

5. ARQUITECTURA LLTEAM

5.1. Agente de Generación de Respuesta

La arquitectura propuesta en este estudio se articula en torno a dos agentes interconectados y una infraestructura de almacenamiento y recuperación de información. El primer agente consta de LangChain (Chase 2022) utilizando GPT-3.5-turbo, por ser el estándar de la industria. Una configuración relevante al caso de uso de respuestas en servicio al cliente es la separación de documentos en la base de conocimiento dentro de LangChain. Lo anterior se debe a que se decidió incluir un historial de correspondencia previa, cuya estructuración como textos independientes será crucial para la exhaustividad semántica de cada documento cuando se almacene la base de conocimiento. Esto se controló mediante la clase *CharacterTextSplitter* y sus parámetros *separator* y *chunk_size*. Para *separator*, se debe definir una secuencia de caracteres que no ocurra naturalmente dentro de los textos e hilar los textos entre sí usando este separador. A su vez, *chunk_size* define el tamaño de la división y podría llegar a sobrescribir el cambio hecho en *separator*. Por lo tanto, *chunk_size* se mantiene necesariamente mayor que la cantidad

de elementos (*tokens*) de la conversación más larga del historial.³

Además, LangChain opera almacenando la base de conocimiento a consultar como índice de vectores y así cobra relevancia la configuración de vectorización. La biblioteca Facebook AI Similarity Search (en adelante FAISS) se emplea como almacenamiento de vectores en este sistema. FAISS es reconocida por su eficiencia en la búsqueda de similitud en grandes conjuntos de vectores (Johnson, Douze y Jégou 2017), lo que la convierte en una opción adecuada y eficiente para la recuperación de documentos relevantes en forma vectorizada. Estos vectores provienen del modelo *text-embedding-ada-002* de OpenAI y se utilizan para representar los documentos en un espacio vectorial. Aunque LangChain ofrece la flexibilidad de utilizar vectorizadores menos complejos, optamos por utilizar el modelo de vectorizado propio de OpenAI. Esta decisión se fundamenta en la coherencia metodológica de estos vectores con el procesamiento de texto realizado por el LLM.

5.1.1. Características de la Muestra Indexada

Para asegurar la representatividad del historial conversacional incluido en la base de conocimiento, ciertos factores ampliaron la muestra, mientras que otros la acotaron. Para evitar sesgos de estacionalidad que podrían surgir al considerar únicamente un determinado periodo del año, se considera la serie de tiempo completa para el año 2022. No obstante, un filtro reduce considerablemente la cantidad de conversaciones indexadas: la adecuación de las conversaciones del historial a la resolución en respuesta única. Lo anterior se controla filtrando solamente hilos conversacionales donde el primer mensaje proviene de un correo de dominio externo a la empresa, el segundo mensaje de un correo de dominio interno, y no hay más mensajes en el hilo. Así, se asegura que la conversación indexada sea una interacción espontáneamente iniciada por un usuario y cerrada por un ejecutivo sin complicaciones que se reflejaran en turnos de habla. Con el fin de maximizar el entendimiento de la función de cada elemento en las conversaciones, estos se marcan con etiquetas XML de la siguiente forma:

```
(9)    <asunto>Asunto del mensaje de usuario</asunto>
        <correo>Cuerpo del mensaje de usuario</correo>
        <respuesta>Respuesta de ejecutivo</respuesta>
```

5.1.2. Ingeniería de Enunciados

La ingeniería de enunciados se ha convertido en un componente fundamental en el campo emergente de los LLM, los cuales ven su eficacia directamente afectada por la calidad del enunciado inicial que reciben (Shieh 2023). Es por esto que el diseño eficiente del enunciado es crucial para aprovechar al máximo las capacidades de estos modelos.

El enunciado entregado a este agente se diseñó para imitar el estilo de respuesta de un ejecutivo, además de dotar al modelo de conocimientos específicos acerca de procedimientos internos de la empresa y así enriquecer la respuesta. A continuación, se detalla su estructura:

³ En versiones posteriores de LangChain, este mismo resultado se puede lograr mediante la clase `UnstructuredFileLoader`.

1. **Frase introductoria que establece contexto.** Por ejemplo: ‘Eres un ejecutivo comercial de una aseguradora (...)’. Ayuda a focalizar la respuesta del modelo en un campo específico.
2. **Estructura de los documentos en el índice.** Se le da a conocer al modelo la estructura y formato con etiquetas XML que encontrará en los documentos del punto cuatro.
3. **Instrucciones adicionales acerca de la tarea.** Se describen limitaciones a las facultades que tiene el agente, para que tenga en consideración al momento de generar la respuesta. Estas se redactan con el propósito de evitar la alucinación en primer lugar; sin embargo, el agente a veces las ignora.
4. **Ejemplos del índice.** Se presentan los documentos provenientes de la base de conocimiento que, por similitud con la entrada de usuario, mejor se ajustan a la generación de una respuesta.
5. **Entrada de usuario.** Finalmente, se adjunta la entrada de usuario que debe ser respondida por el agente.

En resumen, el primer agente opera como un ejecutivo virtual y su arquitectura se puede describir en las siguientes etapas:

1. **Recuperación de documentos.** Utiliza FAISS para recuperar documentos relevantes basados en la consulta del usuario.
2. **Preparación de la solicitud.** Construye una plantilla de solicitud con instrucciones y ejemplos anteriores (contexto).
3. **Generación de respuestas.** Emplea una cadena de LangChain con el modelo GPT-3.5-turbo para generar respuestas automáticas.

5.2. Agente Supervisor de Alucinaciones

La presente investigación propone un segundo agente clasificador. Las clases de este agente son cada uno de los siete probLLeMas propuestos por la tipología, además de una clase final de tipo ‘ninguno de los anteriores’ para los mensajes libres de alucinación. A pesar de ser una tarea de discriminación, se opta por usar tecnologías generativas, al igual que en el primer agente, para aprovechar el paradigma de afinamiento de instrucciones (*instruction-tuning*; Wei et al. 2022b). Así, el enunciado de este segundo agente incluye la taxonomía completa, con el nombre y descripción de cada macrotipo y tipo de probLLeMa en texto plano.

La variable independiente de este experimento es el modelo LLM que funciona de motor para el agente supervisor. Se observaron modelos de las dos empresas proveedoras y de alto tráfico que disponen de licencias comerciales para los investigadores: OpenAI y Anthropic. Ambas empresas cuentan con un LLM de última generación y gran capacidad de inferencia y otro de menor costo por extensión textual y mayor velocidad. La Tabla 1 muestra la relación empresa-optimización de los modelos observados. En la medida que el tamaño de los enunciados crece y la complejidad de la tarea disminuye, los modelos de menor cómputo son la opción preferida para las soluciones industriales. Entonces, el contraste entre GPT-3.5-turbo, GPT-4, Claude Instant y Claude 2 nos permitirá definir el segmento del mercado de la arquitectura LLteaM en el estado de avance de la tecnología y comparar el rendimiento entre proveedores.

	OpenAI	Anthropic
Optimizado a la minimización de cómputo	GPT-3.5-turbo	Claude Instant
Optimizado a la maximización de exactitud	GPT-4	Claude 2

TABLA 1. ENFOQUE DE OPTIMIZACIÓN DE LLM DE ALTO TRÁFICO DISPONIBLES POR API

La realidad de estos datos hace que los distintos tipos de probLLeMas no sean mutuamente excluyentes: una respuesta LangChain puede fácilmente presentar en simultáneo más de un probLLeMa. Es decir, la complejidad íntegra de este desafío comprendería una tarea de tipo multietiqueta. No obstante, en este primer experimento, se optó por instruir al agente supervisor este desafío como uno de etiqueta única. Esta decisión es una forma de relajación del problema, que reduce el espacio de hipótesis para el agente y así busca simplificar su resolución. El etiquetado manual, sin embargo, mantuvo la multietiqueta. Lo anterior es crucial para observar la distribución real de probLLeMas en el ambiente productivo, además de reflejar la precisión del agente siempre que eligiera cualquier probLLeMa que efectivamente estuviera presente en el dato.

5.2.1. Ingeniería de Enunciados

Para el segundo agente se optó por la siguiente estructura de enunciado:

1. **Frase introductoria que establece contexto.** Al igual que para el agente anterior, se introduce el rol del agente y contexto de la tarea. Por ejemplo: ‘Eres un evaluador con una alta capacidad de análisis (...)’
2. **Taxonomía de probLLeMas.** Se inserta la descripción completa, junto con un ejemplo⁴ de cada tipo de probLLeMa.
3. **Ejemplo de clasificación utilizando el paradigma de enunciados de cadena lógica** (*Chain-of-Thought prompting*; Wei et al. 2022a). Es decir, se generan pasos intermedios de explicación para guiar al modelo en su razonamiento clasificador. Así, esta estructura se alinea con el paradigma de enunciados de ejemplo único (*one-shot prompting*; Wei et al. 2022b).
4. **Elemento a evaluar.** El último componente es la tupla ‘Asunto, Correo y Respuesta del primer agente’, la cual este segundo agente deberá clasificar según probLLeMa presente o en ‘ningún probLLeMa presente’.

En pocas palabras, el segundo agente actúa como un supervisor del primero, con el siguiente

⁴ Originalmente, la taxonomía estuvo descrita de forma menos extensa y en términos no técnicos como documentación interna de Wird. Además, los ejemplos aquí presentados provienen de los resultados sobre la muestra correspondiente a este experimento en particular. Por lo tanto, el enunciado original del agente supervisor se diferencia en su forma, pero no en su contenido, de la Sección 4.

flujo:

1. **Preparación de la plantilla de instrucciones.** Utiliza la taxonomía de probLLeMas, que detalla los macrotipos y tipos de alucinaciones y cómo identificar estos últimos, a modo de contexto en su enunciado.
2. **Evaluación de respuestas.** Emplea una cadena LLM con el modelo que se esté empleando como motor del agente para detectar cualquier probLLeMa en la respuesta generada por el primer agente.

Este proceso se visualiza en la Figura 1.



FIGURA 1. RESUMEN DE LA ARQUITECTURA LLTEAM

6. RESULTADOS

En cumplimiento con protocolos de negocio, la exploración de la línea LLteaM se llevó a cabo en dos etapas. Primero, se hizo una acotada prueba de concepto preliminar usando GPT-3.5-turbo (OpenAI 2023b). Luego, dado que los resultados de la prueba de concepto fueron prometedores, se llevó a cabo el experimento completo, contrastando distintos LLM motor. Sin embargo, una actualización de GPT-3.5-turbo por parte del proveedor (OpenAI 2023c), entre la prueba de concepto y el experimento, generó resultados sistemáticamente distintos y de peor calidad, evaluados a continuación.

6.1. Prueba de Concepto y Actualización de GPT-3.5-turbo

La prueba de concepto estuvo limitada a quince datos de la serie de tiempo anteriormente mencionada, procesados únicamente con LangChain (Chase 2022) consultando GPT-3.5-turbo (OpenAI 2023b). El criterio de decisión para ciertas configuraciones de la arquitectura fue la replicabilidad, repitiendo la prueba de concepto completa en cuatro ocasiones y generando la misma salida del segundo agente. Sin embargo, en el siguiente ejemplo de la base:

- (10) Pedro con fecha de hoy adjunto ORDEN 1.107.422 y entregado repuestos a [TALLER REPARADOR]. Atentamente.

En la prueba de concepto, el primer agente generó la siguiente respuesta (instancias de alucinación en negrita):

Estimado Pedro, gracias por la información. ¿Podrías indicarme qué repuestos se

entregaron a [TALLER REPARADOR] y **cuál es el estado actual de la entrega?** Quedo atento a tu respuesta. Saludos.

No obstante, en el experimento completo, el mismo agente responde:

Estimado Pedro, Gracias por tu correo. Hemos recibido la orden [NÚMERO IDENTIFICADOR DE ORDEN] y **hemos entregado los repuestos a [TALLER REPARADOR]**. Si tienes alguna otra consulta o requerimiento, no dudes en comunicarte con nosotros. Atentamente, Equipo de Soporte Técnico de Seguros [EMPRESA]

Por otra parte, las respuestas generadas por este agente en la prueba de concepto contuvieron *LMsplaining* en solo una de las quince ocasiones, o 6,67%. En contraste, la prevalencia del *LMsplaining* sufrió un aumento drástico en las respuestas generadas en el experimento, donde ocurrió en 41,0% de las producciones textuales. Una alta frecuencia del probLLeMa *LMsplaining* es particularmente fallida a nivel lingüístico, ya que falla en las máximas de relevancia –la información aportada no es informativa–, cantidad –ya se dijo y no hace falta explayarlo–, y en ocasiones hasta de calidad –el agente afirma haber generado resultados que en realidad el usuario generó y reportó. Asimismo, el modelo en ambas ocasiones recibe, a modo de contexto, documentos que por similitud se ajustan al requerimiento del cliente. Estos documentos ejemplifican el actuar del ejecutivo ante casos previos que se asemejan al presente. Sin embargo, queda en evidencia que desde la actualización observada el modelo es incapaz de enriquecer la generación de respuesta con la base de conocimiento.

	Prueba de concepto	Experimento completo
Precisión ponderada: tipo de probLLeMa	26,7%	30,0%
Precisión ponderada: macrotipo de probLLeMa	60,0%	41,0%
Cobertura de alucinación (porcentaje de respuestas erróneas enviadas por LLteaM)	100%	73,6%

TABLA 2. IMPACTO EN LA EVALUACIÓN DE LA ACTUALIZACIÓN DE GPT-3.5-TURBO

A su vez, el rendimiento del segundo agente supervisor de alucinaciones utilizando GPT-3.5-turbo en el experimento completo decayó en contraste con el mismo en la prueba de concepto (Tabla 2). Nótese que el incremento de precisión ponderada en tipo de probLLeMa, desde la prueba de concepto al experimento, no debiera considerarse como significativo, dada la pequeña escala de la prueba de concepto. Podemos inferir, a partir de estas observaciones, que la actualización de GPT-3.5-turbo redujo su bondad de ajuste con la tarea de RAG. En consecuencia, la popularidad de LangChain consultando GPT-3.5-turbo podría ser razón suficiente para motivar una evaluación en masa de los equipos de desarrollo en PLN sobre la

robustez de sus aplicaciones al versionamiento del LLM.

6.2. Experimento

Los resultados obtenidos se analizan, primero, considerando todas las etiquetas de probLLeMa como una única clase ‘con alucinación’. Esto reduce el desafío a uno de clasificación binaria y permite, por lo tanto, observar las siguientes métricas:

1. Exactitud (*accuracy*): en cuántos casos del total el modelo acertó en su predicción.
2. Precisión (*precision*): de todos los casos que el modelo predijo como ‘con alucinación,’ cuántos efectivamente la contenían.
3. Cobertura (*recall*): de todos los casos que efectivamente contenían alucinación, cuántos el modelo predijo como ‘con alucinación.’ En este contexto, esta métrica equivale a la capacidad del sistema de identificar respuestas que perjudican el servicio al cliente y prevenir su envío. Como se prefiere que el sistema sea cauteloso y deje a ejecutivos humanos manejar los casos más complejos, esta se vuelve la métrica más relevante.
4. Puntuación F1: media armónica entre precisión y cobertura. Generalmente considerada la métrica más representativa de la capacidad del modelo para identificar una clase, sin preferencia por la cautela ni el riesgo.

En la Tabla 3, la cobertura de 94,3% de GPT-4 es la única que cumple con el objetivo de superar el 90%. Dado este impacto de la clase negativa, GPT-4 lidera también con creces en exactitud y puntuación F1. Se observa significativamente mejor bondad de ajuste a la autosupervisión de alucinaciones en los modelos GPT que los modelos Claude, destacando sobre todo GPT-4.

	GPT-3.5-turbo	GPT-4	Claude Instant	Claude 2
Exactitud	75,0%	92,0%	61,0%	64,0%
Precisión	97,0%	96,5%	90,0%	98,1%
Cobertura	73,6%	94,3%	62,1%	59,8%
Puntuación F1	83,7%	95,3%	73,5%	74,3%

TABLA 3. RESULTADOS DEL EXPERIMENTO LLTEAM SOBRE CLASE ‘CON ALUCINACIÓN’

Sin embargo, Claude 2 supera marginalmente a GPT-4 en precisión, así como GPT-3.5-turbo. Además, las precisiones de GPT-3.5-turbo y GPT-4 son comparables y la cobertura de Claude Instant supera la de Claude 2. En conclusión, no se evidencia una correlación marcada entre la dimensión del modelo y su rendimiento en la detección de alucinaciones.

Siendo GPT-4 el único modelo en el cual se observa la capacidad teórica de autosupervisar alucinaciones, procede evaluar si su puesta en producción es viable en la oferta de mercado

actual. Cabe destacar que, de 100 mensajes procesados, ocho fueron etiquetados por el modelo como libres de alucinación, por lo que solo estos se enviarían en el ambiente productivo, generando ganancia bruta. Como resultado, se concluye que en ningún caso el costo operativo de este modelo a la fecha permite cumplir con la exigencia de $\geq 70\%$ de retorno sobre la inversión. Dependiendo del plan de precios Wird adoptado por este cliente, el retorno sobre la inversión de este conjunto de datos podría oscilar entre 15,1% en el mejor caso y -69,2% en el peor caso. Esto refuta la viabilidad respaldada por la prueba de concepto, donde el uso de GPT-3.5-turbo para ambos agentes de la arquitectura generó un retorno de entre 83% y 91,5%.

En el contexto de la taxonomía de probLLeMas propuesta, un paso crucial es examinar cómo los diferentes LLM se desempeñan en la identificación de probLLeMas a nivel de macrotipo. Esta segmentación (Figura 2) permite una evaluación de las fortalezas y debilidades de cada modelo en distinciones cualitativas amplias como son los probLLeMas-comprensión, chamuyo y existenciales. Además, ofrece una ventana para entender las diferencias entre los modelos en la interpretación de estas categorías.

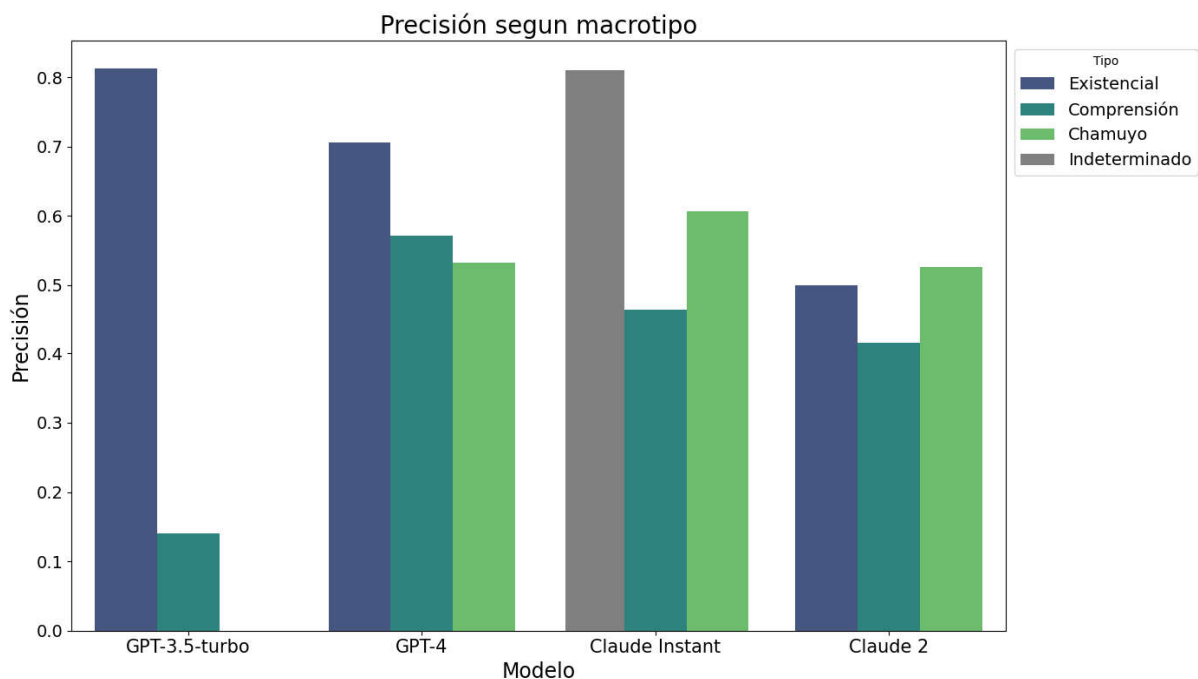


FIGURA 2. PRECISIÓN DE CADA MODELO SEGÚN MACROTIPO DE PROBLEMA

Para proporcionar un contexto adicional y comprender mejor el rendimiento de cada modelo, también examinamos la distribución de las etiquetas generadas por cada modelo según macrotipo de probLLeMa (Figura 3). Estos conteos nos ofrecen una idea de la cobertura de cada modelo a diferentes tipos de probLLeMas y su inclinación hacia categorizar los mensajes en un macrotipo de naturaleza semántica o pragmática, más simple o más compleja de comprender.

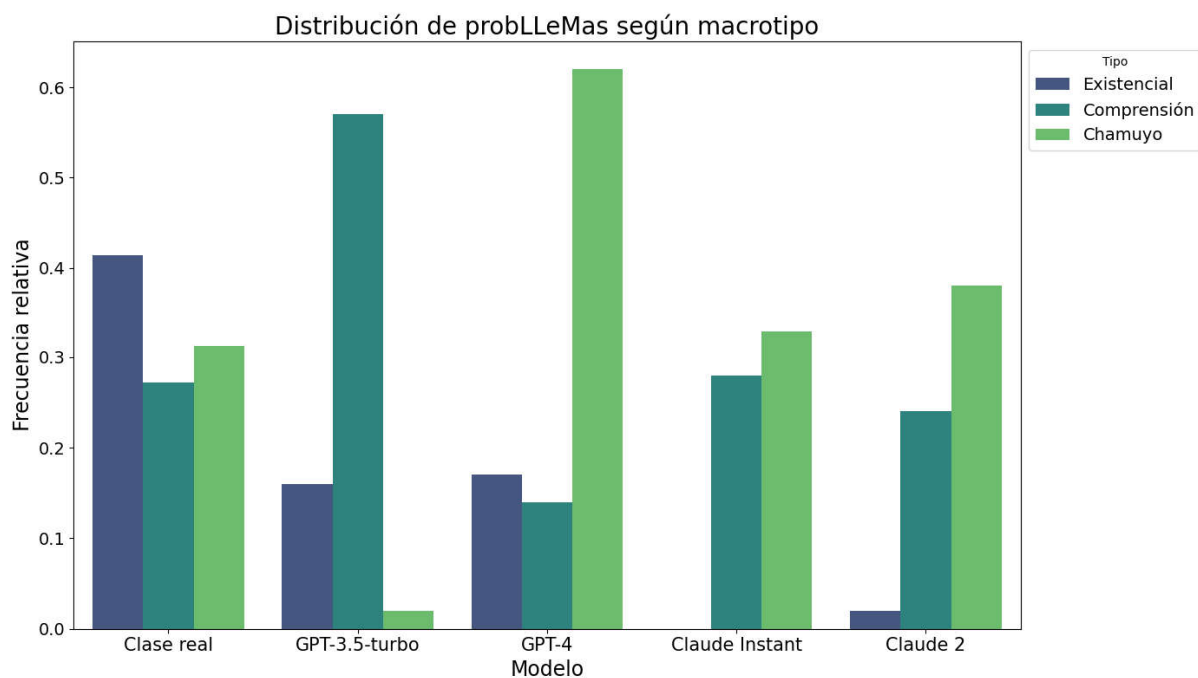


FIGURA 3. DISTRIBUCIÓN DE PROBLEMAS SEGÚN MACROTIPO, EN EL ETIQUETADO MANUAL VERSUS ETIQUETADO DE CADA MODELO

Por último, podemos observar la tarea de clasificación en toda su complejidad, considerando el nivel de categoría más granular proveído por la taxonomía (Figura 4 y Figura 5).

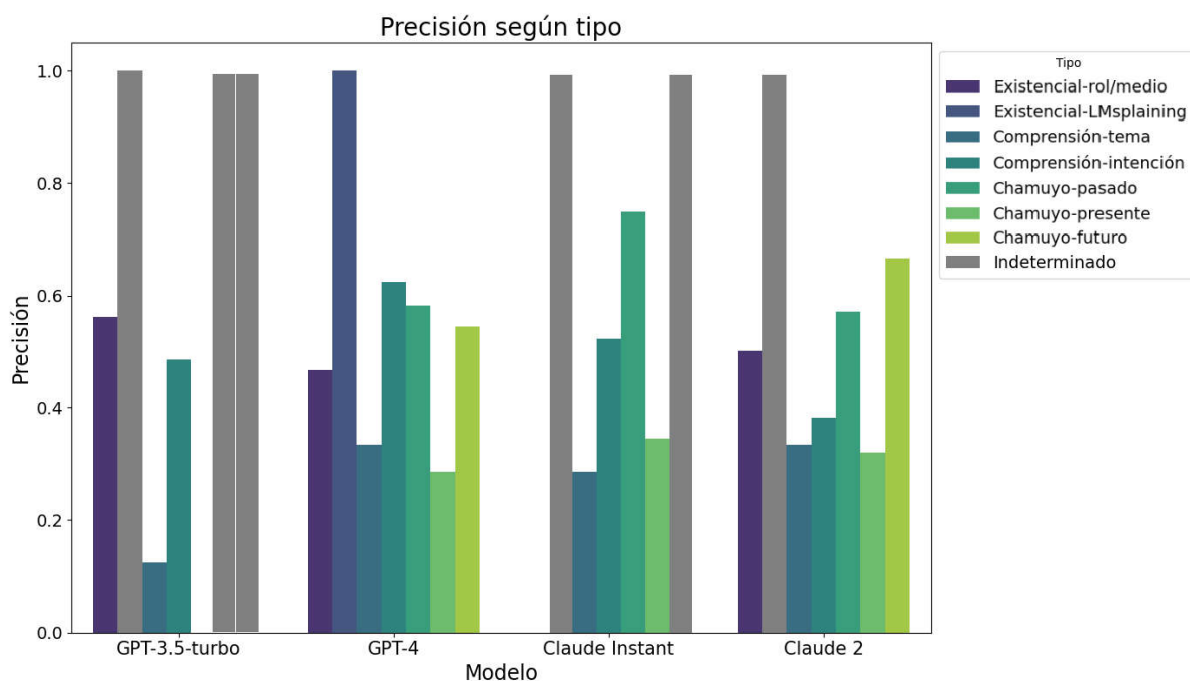


FIGURA 4. PRECISIÓN DE CADA MODELO SEGÚN TIPO DE PROBLEMA

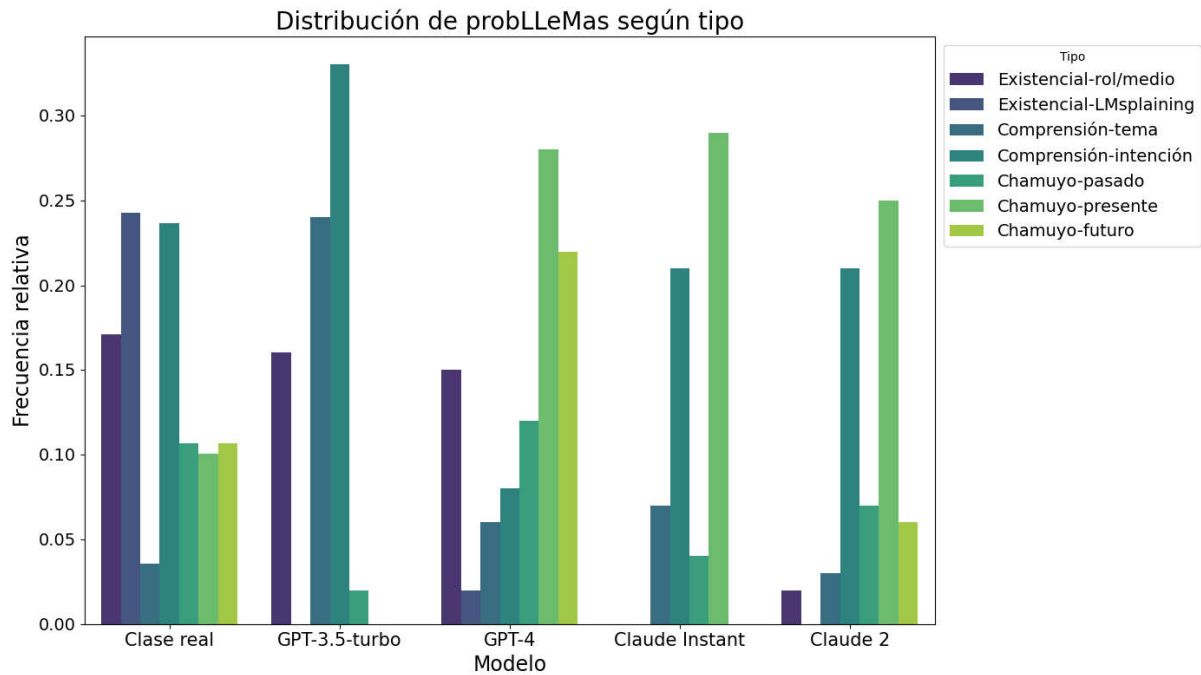


FIGURA 5. DISTRIBUCIÓN DE PROBLEMAS SEGÚN MACROTIPO, EN EL ETIQUETADO MANUAL VERSUS ETIQUETADO DE CADA MODELO

A pesar de que GPT-4 lograra identificar la presencia de alucinación satisfactoriamente (Tabla 3), ningún modelo demuestra una comprensión efectiva de la taxonomía. Incluso en el análisis menos demandante de la identificación del macrotipo de probLLeMa (Figura 2), la precisión ponderada de los modelos ronda o no alcanza el 50%. Además, comparando la prevalencia real de probLLeMas con la distribución de las etiquetas autosupervisadas de los mismos (Figura 3), se marcan claramente inclinaciones por un macrotipo u otro en cada modelo. Se podría conjeturar que cada modelo comprende mejor ciertos fenómenos semántico-pragmáticos y, ya que los probLLeMas no son mutuamente excluyentes, se inclina por lo que conoce. Sin embargo, las inclinaciones más marcadas de un modelo por un macrotipo de probLLeMa no se correlacionan con una mayor precisión en ese macrotipo (Figura 2), refutando esta posibilidad. Esto sugeriría que las inclinaciones son arbitrarias. Para la tarea más demandante de clasificar por tipo de probLLeMa, la precisión es peor (Figura 4) y la distribución presenta aún más inclinaciones arbitrarias (Figura 5). En conclusión, no se encuentra resuelta, en el estado de avance de la tecnología, la tarea de la identificación automática de la naturaleza lingüística de las alucinaciones generadas por LLM.

Cabe destacar que, aunque la instrucción apuntaba explícitamente a la tarea más demandante, y esta no se logró, el producto teórico-práctico igualmente logra la tarea menos demandante de identificar la presencia de alucinación mediante este marco de referencia. Es decir, el producto sirve un fin menos elegante que la solución explícita que propone.

Un elemento a considerar en el contraste entre modelos son las distintas estrategias de entrenamiento de ambos proveedores. OpenAI entrena con corpus de gran magnitud, buscando una amplia base para su aprendizaje. En contraste, Anthropic adopta una orientación ética, diseñando el corpus de entrenamiento para excluir contenido ofensivo o peligroso (Bai et al.

2022). En este contexto, la tendencia distintiva de los modelos Claude a registrar escasas o nulas etiquetas en diversas categorías (Figura 3 y Figura 5), especialmente los probLLeMas existenciales, podría interpretarse como una muestra de la naturaleza reservada de su proceso asociativo. En contraposición, la amplia formación de los modelos GPT se manifiesta particularmente en su elevada precisión en la clasificación de los probLLeMas existenciales (Figura 2). No obstante, a pesar de este vasto entrenamiento, no siempre presenta un rendimiento óptimo en otras categorías como probLLeMas-comprensión y probLLeMas-chamuyo (Figura 2). Es decir, la extensividad de los temas abordados en el entrenamiento de modelos como los de OpenAI no garantiza su superioridad en todas las categorías frente a modelos como los de Anthropic. Esto destaca la complejidad y los desafíos inherentes al equilibrio entre la ética y la capacidad de generalización en los LLM.

Finalmente, la dimensión del modelo se revela como otro factor de relevancia en la identificación de la naturaleza de las alucinaciones. Aquellos modelos de mayor tamaño tienden a manifestar una precisión elevada, sugiriendo que este desafío requiere una gran capacidad de inferencia. La dimensión del modelo se asocia con más fuerza a una precisión elevada en los modelos de OpenAI que en aquellos de Anthropic y en el análisis por tipo que en el análisis por macrotipo de probLLeMa (Figura 2 y Figura 4). Por lo tanto, se puede estimar que hará falta un mayor estado de madurez de la IA generativa para que el costo de clasificar alucinaciones según su naturaleza lingüística sea productivamente viable.

7. LIMITACIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

A la vez que el propósito industrial de esta investigación permite explorar y dar luces sobre casos de uso que tienden a ser poco visibles para la academia, también esta naturaleza limita la interpretación académica de los resultados. En primer lugar, la prueba de concepto, por definición de negocio, debe ser lo más acotada posible. Sin embargo, al aquí contrastarla con el experimento ante la inesperada actualización de GPT-3.5-turbo y la caída de rendimiento que esto significó, el tamaño reducido de su muestra se vuelve un impedimento, donde la aleatoriedad de muestreo pudo haber jugado un papel particularmente relevante por la escala a la cual se trabajó. Lo anterior limitaría las conclusiones de que la prueba de concepto haya sido aprobada. En segundo lugar, la capacidad de cómputo necesaria para servir un LLM nos obliga a consumirlo de un servicio externo como los proveídos por OpenAI y Anthropic. Esto tiene la ventaja de acortar los tiempos de desarrollo y reducir los costos de mantenimiento, pero vuelve el rendimiento de esta arquitectura dependiente de los cambios que estén en interés de la empresa proveedora llevar a cabo sobre el LLM. Cuando el experimento esperaba incrementar la confianza de la hipótesis vislumbrada en la prueba de concepto, ambas de estas etapas tuvieron que servir de experimentos contrastables entre sí. No solo es esto una limitación a la propuesta de cualquier arquitectura que involucre LLM, sino también una corta e imprevisible fecha de expiración para las conclusiones académicas que se deriven sobre cualquier experimentación en LLM: habrían variables controladas a las cuales los investigadores no tenemos acceso.

Otra limitación se asocia a la pérdida de información en el proceso de extracción, transformación y carga de los datos. Fue particularmente frecuente en estos datos que el propósito de un mensaje fuese enviar documentación como adjunto. En estos casos, a menudo el cuerpo del correo no contuvo la información principal necesaria para resolver el caso, sino que esta se encontraría presente en el adjunto. En futuras iteraciones, donde el rendimiento global del agente de generación de respuesta sea mejor, se podrían investigar métodos de integrar el texto contenido en los adjuntos de forma fidedigna y limitando la extensión del adjunto a escalas que preserven un óptimo entre la máxima de relevancia y la máxima de cantidad.

Asociada a la naturaleza industrial está también la alta privacidad de los datos utilizados. La replicabilidad de este estudio se ve perjudicada por la incapacidad de liberar los datos procesados en estas primeras pruebas de la arquitectura de LLteaM. Además, y de manera más compleja, la privacidad de los procesos de negocio de la empresa que recibe estos correos electrónicos problematiza la noción de factualidad en el etiquetado manual. Al hablar de verdades del negocio, que no son públicamente conocidas—en cuyo caso el usuario nunca las consultaría—, los mismos investigadores sufren un sesgo de observación por no formar parte de la comunidad de hablantes que comparte este conocimiento. En otras palabras, puede haber más alucinación extrínseca en las producciones observadas de la que somos capaces de detectar, por responder a una verdad que no se representa ni en los datos textuales que tenemos disponibles, ni en el conocimiento general que compartimos como hablantes del español chileno.

Por último, cabe la posibilidad de un sobreajuste de esta tipología. Ontológicamente, se presentaron ciertos casos que ponen sus límites teóricos en tela de juicio: casos en los que parecía poder haber información relevante al caso perdida en el proceso de extracción, transformación y carga. Esta información faltante sería percibida en un ambiente natural por un ejecutivo humano, por lo cual se podrían considerar como ‘probLLeMas-perspectiva’. Existió el caso de falta de correos de arrastre (i. e., ‘perspectiva-visera’) y el de falta de metadatos o adjuntos (i. e., ‘perspectiva-ceguera’). Estos factores diferencian la respuesta dada por el modelo de la que daría un humano, pero parecen escaparse de la discusión lingüístico-informática de si un LLM alucina en la interfaz semántico-pragmática. ¿Serían, entonces, probLLeMas? ¿Habrá muchos más tipos de probLLeMas por descubrir? Al respecto, queda espacio para replicar este experimento en otras industrias con tal de cubrir otros campos semánticos, otros casos de uso y/u otros medios de comunicación con tal de explorar otros contextos sociopragmáticos. Cabe preguntarse qué tan efectivamente este marco de probLLeMas se transferiría a otras culturas hispanohablantes, otras lenguas y otras aplicaciones industriales o académicas. También siembra la duda de si, parchando estos posibles agujeros teóricos, el rendimiento del producto teórico-práctico se vería beneficiado, diferenciando con más claridad la variedad de casos que se le presentan, o perjudicado, sufriendo pérdida de memoria o indistinción entre clases. En resumen, a la teoría de las alucinaciones le queda mucho camino por recorrer, sobre todo si viaja acompañada de una lingüística de máquinas.

8. CONCLUSIÓN

En la presente investigación presentamos un producto teórico-práctico para la autosupervisión de alucinaciones. Se exploró el caso de uso de decidir dinámicamente si se automatiza o no una respuesta por canales de conversación privados usuario-ejecutivo. El producto teórico (i. e., la taxonomía de probLLeMas) avanza la teoría de las alucinaciones más allá de la factualidad y fidelidad, desde una perspectiva pragmática en lingüística. El producto práctico (i. e., la arquitectura de LLteaM) permite la autosupervisión de alucinaciones y por tanto reduce el riesgo al conectar un agente conversacional generativo a un ambiente productivo. Además, la tipificación de las alucinaciones observadas en las respuestas del primer agente de la arquitectura da indicios de la posible distribución cualitativa de alucinación en otros ambientes productivos actuales.

Cabe destacar la drástica baja de rendimiento entre la prueba de concepto y su posterior experimento (OpenAI 2023b; 2023c). La actualización de GPT-3.5-turbo redujo su bondad de ajuste con la tarea de RAG, posiblemente perjudicando las numerosas aplicaciones construidas sobre LangChain actualmente productivas en la industria. Esto deja en evidencia la incertidumbre que la ingeniería de enunciados implica para las empresas. El riesgo de que el modelo alucine sobre verdades universales es mucho más visible mediante retroalimentación de usuarios no-técnicos y, por lo tanto, probablemente influya en las actualizaciones. Sin embargo, el riesgo de que alucine sobre verdades situacionales es independiente y crecerá en la medida que más aspectos de nuestra vida en sociedad estén mediados por la IA generativa.

Finalmente, defendemos que nos encontramos ante un estado de la realidad donde el rol social de la lingüística y la definición tradicional de su objeto de estudio están en conflicto. Asimismo, el área de la IA generativa sostiene una deuda teórico-técnica con la sociedad, por implementar aplicaciones de alto impacto aun con una escasa teoría de alucinaciones. Presentamos este como un ejemplo de cómo la colaboración entre ambas áreas e interconexión de sus producciones puede saldar esa deuda pendiente.

REFERENCIAS

Anthropic. "Claude." 2023. Acceso el 13 de septiembre. <https://claude.ai>

Austin, John Langshaw. 1975. *How to Do Things with Words*. Cambridge: Harvard University Press.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." *arXiv preprint*. doi:10.48550/arXiv.2204.05862

Bohannon, Molly. 2023. "Lawyer Used ChatGPT in Court—And Cited Fake Cases. A Judge Is Considering Sanctions." *Forbes*. Última actualización el 8 de junio.

<https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=7dbd65b57c7f>

Chase, Harrison. 2022. "LangChain." Acceso el 13 de septiembre. <https://github.com/hwchase17/langchain>

Chui, Michael, Mena Issler, Roger Roberts y Lareina Yee. 2023. "McKinsey Technology Trends Outlook 2023." *McKinsey Insights*. Última actualización el 20 de julio. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-top-trends-in-tech>

Google. 2023. "Bard." Última actualización el 13 de julio. <https://bard.google.com>

Grice, H. P. 1975. "Logic and Conversation." In *Syntax and Semantics, vol. 3: Speech Acts*, edited by Peter Cole and Jerry L. Morgan, 41–58. New York: Academic Press.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto y Pascale Fung. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55 (12): 1–38. doi:10.1145/3571730

Johnson, Jeff, Matthijs Douze y Hervé Jégou. 2017. "Billion-scale Similarity Search with GPUs." *arXiv preprint*. doi:10.48550/arXiv.1702.08734

Liu, Jerry. 2022. "LlamaIndex." Acceso el 6 de septiembre. https://github.com/jerryliu/llama_index. doi:10.5281/zenodo.1234

Microsoft. 2023. "Microsoft and OpenAI extend partnership." Acceso el 6 de septiembre. <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/>.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet y Ryan McDonald. 2020. "On Faithfulness and Factuality in Abstractive Summarization." *arXiv preprint*. doi:10.48550/arXiv.2005.00661

OpenAI. 2022. "ChatGPT." Acceso el 15 de septiembre. <https://chat.openai.com>

OpenAI. 2023a. "GPT-4 Technical Report." *arXiv preprint*. doi:10.48550/arXiv.2303.08774

OpenAI. 2023b. "Chat Completion API." Acceso el 16 de junio. <https://api.openai.com/v1>

OpenAI. 2023c. "Chat Completion API." Acceso el 22 de agosto. <https://api.openai.com/v1>

de Saussure, Ferdinand. 1916. *Cours de Linguistique Générale*. Taiwán: Payot.

Searle, John. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. New York: Cambridge University Press.

Sivasubramanian, Swami. 2023. "Announcing New Tools for Building with Generative AI on AWS." *AWS Machine Learning Blog*. Última actualización el 13 de abril. <https://aws.amazon.com/blogs/machine-learning/announcing-new-tools-for-building-with-generative-ai-on-aws>

Tharp, Carrie y Lee Moore. 2023. "Helping Businesses with Generative AI." *Google Cloud Blog*. Última actualización el 7 de junio. <https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-for-industries>

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." *arXiv preprint*. doi:10.48550/arXiv.2302.13971

Turing, Alan Mathison. 1950. "Computing Machinery and Intelligence." *Mind* LIX (236): 433–460. doi:10.1093/mind/LIX.236.433

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser e Illia Polosukhin. 2017. "Attention Is All You Need." *arXiv preprint*. doi:10.48550/arXiv.1706.03762

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le y Denny Zhou. 2022a. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *arXiv preprint*. doi:10.48550/arXiv.2201.11903

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai y Quoc V. Le. 2022b. "Finetuned Language Models Are Zero-Shot Learners." *arXiv preprint*. doi:10.48550/arXiv.2109.01652