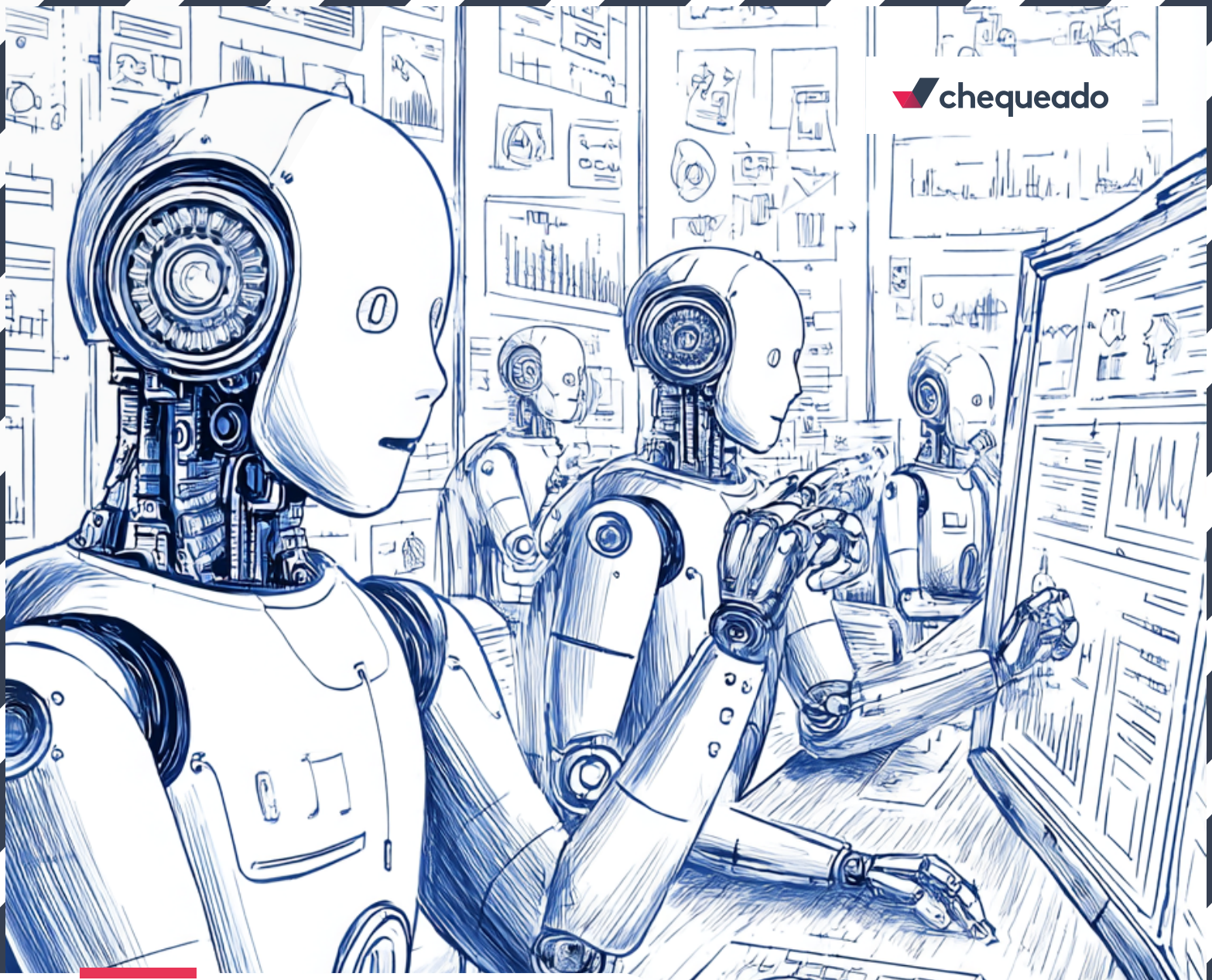


chequeado

Reporte

¿Qué tan confiables son los chatbots para verificar afirmaciones políticas?

Con el apoyo de

Reino de los Países Bajos

abr 2026

Índice

01

Resumen Ejecutivo

La adopción de chatbots **para informarse** está creciendo de forma acelerada. Pero ¿qué tan confiable es la información que ofrecen? Desde Chequeado, nos propusimos evaluar la calidad informativa de los principales modelos de lenguaje utilizados para consultas sobre actualidad: Gemini, ChatGPT, Grok y la IA integrada al buscador de Google.

En este estudio, **analizamos 106 respuestas** que ofrecieron los modelos frente a consultas sobre la veracidad de **14 afirmaciones** realizadas por el presidente Javier Milei en el discurso de apertura de sesiones del 1° de marzo de 2026. Las consultas se realizaron a través de **19 configuraciones diferentes**, variando el modelo y el tipo de acceso, el momento de la consulta (durante el discurso, entre una y dos horas después, y al día siguiente), el tipo de usuario y el fraseo en la formulación de la pregunta.

Para cada respuesta evaluamos la veracidad de los datos presentados y de las fuentes citadas, la consistencia interna de las respuestas y razonamiento dado, la pertinencia y actualidad de los datos y fuentes utilizadAs, la presencia de editorialización, el encuadre de las valoraciones intermedias y, entre otras dimensiones, el uso que los modelos hacían de fuentes de Chequeado y organismos oficiales.

Los resultados son preocupantes:

- Del total de 106 respuestas analizadas, **el 36% presentó algún tipo de problema informativo**.
- Los problemas detectados abarcan distintas dimensiones: desde alucinaciones (datos falsos y fuentes inexistentes), hasta problemas en el uso de datos y fuentes como estadísticas no pertinentes, incompatibles entre sí o metodológicamente inadecuadas para la pregunta formulada, así como inconsistencias en el razonamiento y la argumentación, y editorialización mediante lenguaje emocionalmente cargado o interpretaciones

< Ir al índice

no sustentadas por la evidencia.

- El modelo con peor desempeño fue **Gemini, con un 53% de respuestas problemáticas**; le siguieron ChatGPT con un 34,1%, Grok con un 25% y las respuestas generadas por IA en las búsquedas en Google con un 14,3%.
- **Gemini fue el único modelo en el que detectamos alucinaciones**: fuentes inventadas (notas de Chequeado inexistentes y páginas oficiales que no existen o están caídas) y frases atribuidas a Milei durante el discurso que, en realidad, nunca fueron pronunciadas.
- **Grok fue el modelo que más incorporó fuentes actualizadas.**
- En particular, respecto a las notas de Chequeado sobre las afirmaciones particulares de Milei durante el discurso, Grok las utilizó en el 100% de las consultas.
- Pedir explícitamente al modelo que busque fuentes confiables no garantiza mejores resultados. Cuando se utilizó un prompt con instrucciones específicas que pide verificar la información con estándares periodísticos (como buscar fuentes primarias y chequeos previos), **Gemini inventó fuentes oficiales o de Chequeado en el 87,5% de sus respuestas**, mientras que ChatGPT, consultado con el mismo prompt y en el mismo momento, no solo no inventó fuentes sino que utilizó notas de Chequeado correspondientes al tema consultado en el 100% de los casos.

Conclusiones y recomendaciones

Los resultados de este experimento sugieren que **utilizar chatbots para verificar afirmaciones de actualidad conlleva riesgos significativos**. El problema no se limita a los errores detectables: muchas respuestas presentan una dificultad adicional, ya que aunque los datos citados puedan ser correctos por separado, la forma en que se los combina, compara o encuadra **puede llevar al usuario a conclusiones incorrectas sin que esto sea fácilmente detectable**. Las inconsistencias internas y el corrimiento respecto de lo preguntado dificultan que un lector no especializado evalúe la calidad de lo que está leyendo.

Frente a este panorama, el uso de estos modelos como fuente de información requiere precaución. Si se los consulta, conviene **contrastar sus respuestas con fuentes prima-**

< Ir al índice

rias y verificar que los datos citados estén efectivamente en las fuentes mencionadas. Es recomendable ser especialmente cauteloso ante respuestas con múltiples indicadores cuya relación no está claramente fundamentada. **Para verificar afirmaciones políticas, los medios periodísticos confiables y los sitios especializados en verificación de datos siguen siendo las referencias más sólidas.**

Autora: Leticia Smal

Colaboradores: Joaquín Coto y Jeronimo Scipione

Asistencia técnica: Joaquin Saralegui

Editores: Olivia Sohr y Franco Piccato

Diseño: Matías Severo

02

Introducción

La inteligencia artificial (IA) generativa tiene una capacidad cada vez mayor de producir contenido difícil de distinguir del real. El debate público sobre este riesgo suele enfocarse en la generación de imágenes y videos falsos de alta calidad. Pero no deberíamos dejar afuera de la discusión a los chatbots y otras aplicaciones basadas en grandes modelos de lenguaje (LLMs, por sus siglas en inglés), **cuya adopción está creciendo de forma acelerada**. La preocupación no se limita a la generación de las llamadas "alucinaciones" (situaciones en las que la IA genera información totalmente inventada pero presentada como verdadera), sino también a la baja calidad informativa que estos sistemas pueden ofrecer: interpretaciones distorsionadas de fuentes, respuestas incompletas que pueden dar a interpretaciones erróneas, o basadas en información desactualizada. Si bien estas tecnologías no crean contenido con la intención explícita de desinformar, pueden contribuir involuntariamente a ello, incrementando la circulación de información engañosa o confusa.

Con esta inquietud como punto de partida, desde Chequeado nos propusimos evaluar la calidad informativa de los principales chatbots ante preguntas de actualidad. En particular, analizamos las respuestas frente a consultas sobre la veracidad de afirmaciones realizadas por el presidente Javier Milei en el discurso de apertura de sesiones del 1° de marzo de 2026.

< Ir al índice

03

¿Cómo lo hicimos?

Seleccionamos **14 afirmaciones realizadas por Milei durante su discurso** de apertura de sesiones del 1ero de Marzo de 2026, abarcando temáticas variadas: pobreza, empleo y salarios, seguridad, gasto público, política social, educación y comercio exterior. A partir de ellas, formulamos preguntas del estilo "¿es verdad que...?" y las consultamos en distintos modelos de IA **en tres momentos diferentes: durante el propio discurso, una hora después de su finalización y al día siguiente**. El objetivo de este escalonamiento temporal fue evaluar si los modelos actualizaban sus respuestas a medida que aparecían nuevas fuentes, en particular notas periodísticas así como los chequeos de Chequeado realizados en el contexto del discurso. A su vez, hicimos estas consultas a través de diferentes tipos de usuarios (nuevos, con historial de navegación, en modo invitado, y en modo consulta indirecta a través de la API), de forma de cubrir un abanico amplio del tipo de consultas. En total **se realizaron 19 configuraciones de consulta**, considerando las variaciones en modelos, tipos de usuarios, formulación de las preguntas y momentos en las que fueron hechas. A partir de estas configuraciones **se analizaron 106 respuestas diferentes**.

Para cada respuesta, analizamos la actualidad y pertinencia de los datos utilizados, los argumentos, el razonamiento y su consistencia interna, el encuadre de las valoraciones intermedias, la presencia de editorialización, entre otras dimensiones. También examinamos el uso que los modelos hacían de fuentes de Chequeado y de organismos oficiales cuando las citaban. (Para más detalles, ver la sección **Metodología**.)

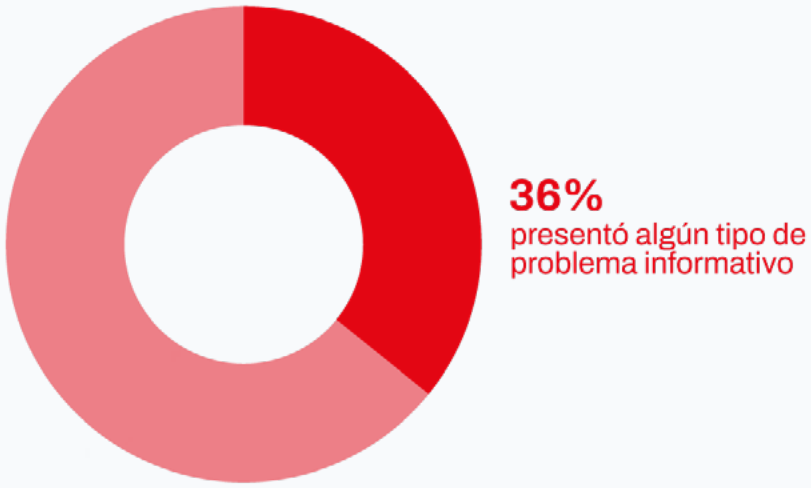
< Ir al índice

04

¿Qué encontramos?

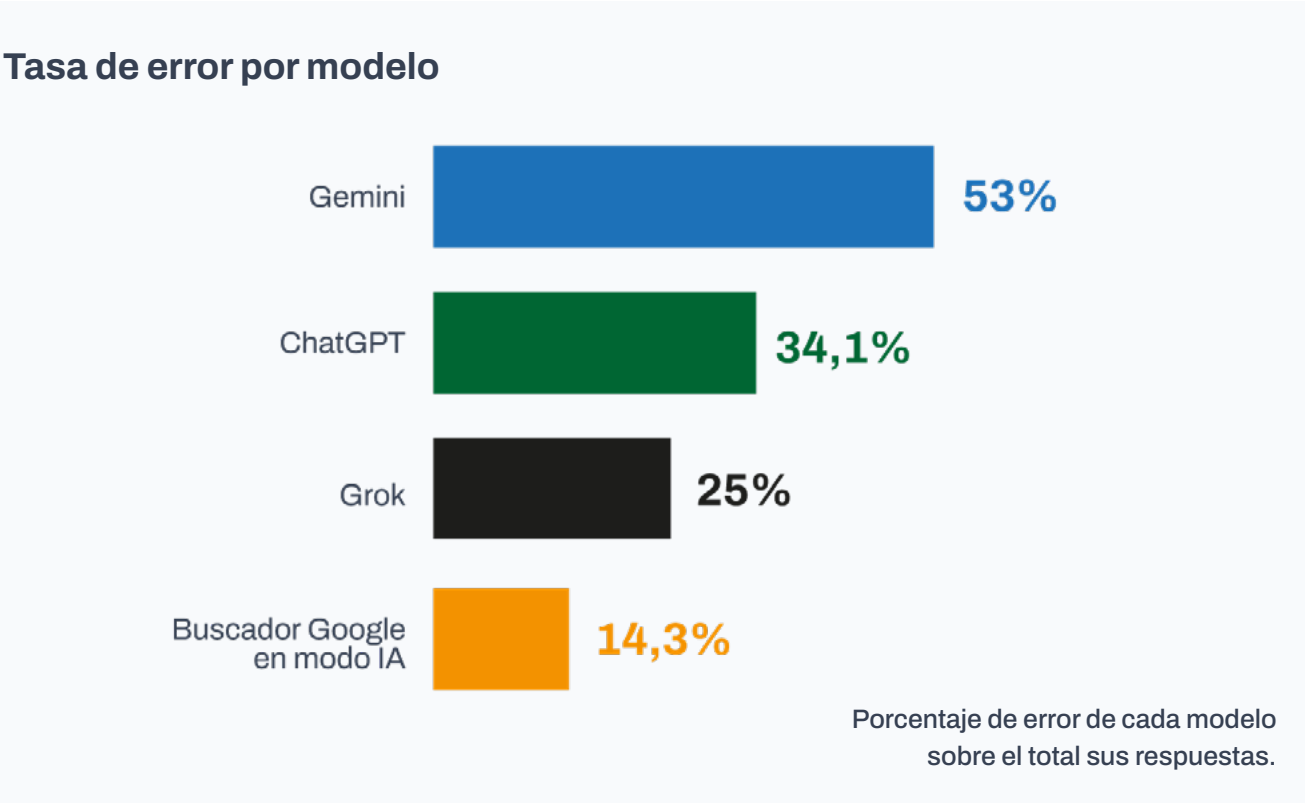
De las 106 respuestas generadas por los distintos modelos, **el 36% presentó algún tipo de problema informativo**.

Tasa de error total (n=106)



El modelo con peor desempeño fue **Gemini: respondió con errores en el 53% de las 34 consultas realizadas, y fue el único que generó alucinaciones**, inventando fuentes (por ejemplo, notas de Chequeado inexistentes y páginas oficiales que no existen o están caídas) y atribuyendo a Milei frases que nunca pronunció en el discurso. **Le siguieron ChatGPT con un 34,1% de respuestas problemáticas, Grok con un 25% y las respuestas generadas por la IA del buscador Google con un 14,3%.**

< Ir al índice



Porcentaje de error de cada modelo sobre el total sus respuestas.

En otro aspecto del análisis encontramos que Grok fue el modelo que más incorporó fuentes actualizadas. Además, en los casos en que una nota de Chequeado estaba disponible al momento de la consulta, la utilizó en el 100% de los casos.

I. TIPOS DE ERRORES MÁS COMUNES

ALUCINACIONES

Desde la aparición de los modelos de lenguaje, uno de los problemas más documentados es el de las “alucinaciones”: la generación de información falsa presentada con aparente convicción. En las respuestas analizadas para este experimento, Gemini fue el único modelo que presentó fabricaciones.

La principal fue la invención de fuentes. De las 30 respuestas en las que Gemini explicitó las fuentes utilizadas, se analizaron aquellas correspondientes a fuentes oficiales o a

< Ir al índice

Chequeado: **en el 23,3% de los casos, las fuentes citadas no existían**. El modelo generó URLs que tienen la apariencia de enlaces reales pero que corresponden a páginas inexistentes o que ya no están disponibles, como notas de Chequeado (por ejemplo, "**https://chequeado.com/ultimas-noticias/javier-milei-los-indicadores-sociales-de-la-economia-son-peores-que-los-que-teniamos-en-2001/**") o páginas oficiales del gobierno (por ejemplo, "**https://www.anses.gob.ar/consultas/asignacion-universal-por-hijo**").

Además de inventar fuentes, **Gemini atribuyó a Milei frases que nunca pronunció en el discurso**. Por ejemplo, en una respuesta, Gemini afirmó que "*en el discurso de ayer ante el Congreso, Milei reafirmó estas cifras y destacó que enero de 2026 fue el mes con menos homicidios en Rosario de todo el siglo*", pero esa afirmación no fue realizada. En otro caso, ante una pregunta sobre el nivel de desempleo, el modelo señaló "*en su discurso de ayer en el Congreso, el Presidente atribuyó esta baja a la Reforma Laboral y a la reactivación económica*": además de ser falso porque Milei no hizo esa afirmación en ningún momento de su discurso, la frase en sí carece de sentido puesto que la Reforma Laboral fue **promulgada por el Poder Ejecutivo** cinco días después de la fecha del discurso.

COMPARACIÓN DE FUENTES INCOMPATIBLES

En algunos casos, los modelos combinan datos y estadísticas provenientes de distintos organismos o fuentes. Al compararlos como si fueran equivalentes, **las respuestas llegan a conclusiones erróneas no por los datos individuales en sí, sino por la operación que se hace con ellos**. Por ejemplo, ante la pregunta sobre si **el gobierno de Milei redujo en un 20% la planta estatal**, en una misma respuesta compara una estadística que considera la totalidad de la planta -incluyendo la Administración Pública Nacional (APN), las empresas estatales, y el personal militar y de seguridad- contra otra que solo contempla la APN y las empresas estatales. Al tratarlas como equivalentes, la comparación resultante es metodológicamente invá-

< Ir al índice

lida, y la conclusión a la que llega el modelo resulta errónea.

DATOS CORRECTOS, ESTADÍSTICA INCORRECTA

En estos casos los modelos citan cifras verídicas pero que no son las adecuadas para responder la pregunta formulada: toman datos de períodos o momentos que, en términos estadísticos o según la opinión experta, no son válidos comparar entre sí, por ejemplo cotejando datos de un mes de un año con los de un mes diferente del año siguiente. El dato es real; el uso que se hace de él lleva a conclusiones incorrectas . Por ejemplo, ante la pregunta de si es verdad que durante la gestión de Milei **se pasó de un 57% de pobreza a una tasa del 30%**, en la respuesta generada por la IA del buscador Google se comparó estadísticas de pobreza de forma mensual, a diferencia de lo que recomiendan los especialistas de hacerlo de forma semestral.

DESPLAZAMIENTO DEL OBJETO DE LA PREGUNTA

Las respuestas ofrecen información verdadera pero sobre una cuestión ligeramente distinta a la que se preguntó. En algunos casos, además, suman datos adicionales correctos que profundizan ese corrimiento. El resultado es una respuesta que no es técnicamente falsa y parece pertinente, pero cuya conclusión aplica a otra pregunta. Esto es problemático porque el usuario puede tomar esa conclusión como válida para su consulta original sin advertir el desplazamiento.

INCONSISTENCIA EN EL RAZONAMIENTO

Los modelos presentan respuestas que se contradicen consigo mismas: el veredicto no se desprende de los argumentos y datos citados, o incluso en algunos casos va en contra de los propios datos presentados. Los datos pueden ser correctos, pero el razonamiento que los conecta con la conclusión es incoherente.

< Ir al índice

ACUMULACIÓN DE DATOS HETEROGÉNEOS (No pertinentes a la pregunta y/o sin relación entre sí)

En estos casos los modelos presentan una gran cantidad de estadísticas provenientes de distintas fuentes, períodos o criterios, sin aclarar sus diferencias ni su relevancia para la pregunta. La acumulación hace que el argumento sea difícil de seguir y que la justificación del veredicto quede opacada por el volumen de información.

RECORTE TEMPORAL ARBITRARIO

Algunas preguntas refirieron al comportamiento de indicadores sociales o económicos a lo largo de toda la gestión de Milei. En esos casos, lo esperable es que el veredicto considere la totalidad del tiempo transcurrido. Sin embargo, en algunas respuestas los modelos presentaron los datos de diferentes momentos de la gestión pero a la hora de emitir su conclusión se apoyaron solo en una fracción de ese tiempo, ignorando la evidencia restante que ellos mismos mencionaron. El problema no radica en no entender el alcance temporal de la pregunta, sino en no utilizarlo al momento de evaluar.

EDITORIALIZACIÓN

En algunos casos, los modelos no se limitan a presentar y evaluar datos, sino que incorporan juicios de valor, adjetivaciones y afirmaciones interpretativas que exceden el rol de verificador por el cual se los consulta (todas las preguntas comienzan con alguna variante de “¿es verdad que…?”). Este problema puede manifestarse de distintas formas: por ejemplo, mediante lenguaje o expresiones con connotación emocional, mediante encuadre sesgado del veredicto (por ejemplo, si la conclusión matiza algo que los propios datos no matizarían), o mediante interpretaciones no sustentadas por las fuentes.

EL RESULTADO COMBINADO: RESPUESTAS DIFÍCILES DE INTERPRETAR

En conjunto, estos resultados muestran que, más allá de aceptar o no la clasificación de cada respuesta como verdadera o falsa, un usuario que quiera llegar a su propia

< Ir al índice

conclusión a partir de los datos y argumentos que el propio modelo presenta enfrenta al menos tres dificultades. La primera es comprender la totalidad de los datos presentados: la acumulación de estadísticas que no responden directamente a la pregunta, medidas con criterios diferentes y no siempre comparables entre sí, dificulta seguir el hilo del argumento. La segunda es evaluar si ese argumento es razonable, si está bien sustentado y si se corresponde con las fuentes citadas: las inconsistencias internas, el corrimiento respecto de lo preguntado y las comparaciones metodológicamente inválidas no siempre son evidentes para un lector no especializado. La tercera, es que la combinación de las dos anteriores puede llevar a conclusiones incorrectas: el usuario puede creer que está leyendo un razonamiento sólido basado en datos bien interpretados cuando, en realidad, no lo es. En los casos más extremos, las respuestas mezclan sin distinción información verdadera y correctamente analizada con información incorrecta o falsa, lo que hace aún más difícil identificar dónde está el error.

II. VARIACIONES ENTRE MODELOS

No todos los modelos analizados tuvieron el mismo desempeño.

ERRORES TOTALES

Como se mencionó en la introducción, del total de 106 respuestas generadas, el 36% presentó algún tipo de problema informativo: desde datos incorrectos y veredictos contradictorios con la propia evidencia utilizada para sustentarlos, hasta alucinaciones de datos y fuentes. El modelo con peor desempeño fue Gemini, con un 53% de respuestas problemáticas; le siguieron ChatGPT con un 34,1%, Grok con un 25% y las respuestas generadas por la IA del buscador de Google con un 14,3%.

< Ir al índice

ALUCINACIONES

Gemini fue el único modelo que presentó alucinaciones. De las 30 respuestas analizadas, el 23,3% incluyó al menos una. En algunos casos, una misma respuesta combinaba más de un tipo: citar como fuente páginas inexistentes (tanto de organismos oficiales como de notas de Chequeado) e incluir frases atribuidas a Milei en el discurso que en realidad nunca fueron pronunciadas.

El caso de Gemini

USO DE FUENTES

Otra gran diferencia entre configuraciones se encontró en las fuentes citadas para argumentar las respuestas. La tendencia más marcada se registró en la configuración en la que se utilizó un prompt estructurado con instrucciones específicas de fact-checking, que indicaba al modelo verificar la afirmación con estándares periodísticos: buscar fuentes primarias y chequeos previos. Esta configuración fue aplicada tanto en Gemini como en ChatGPT. Los resultados fueron opuestos: Gemini inventó fuentes oficiales o de Chequeado en el 87,5% de sus respuestas (7 de 8 consultas), mientras que ChatGPT, consultado con el mismo prompt en el mismo momento, no solo no inventó fuentes sino que utilizó notas de Chequeado correspondientes al tema consultado en el 100% de los casos. Es decir, ante la indicación explícita de recurrir a fuentes periodísticas o chequeos disponibles, ChatGPT las utilizó y Gemini no solo no las identificó sino que fabricó otras.

< Ir al índice

VALORACIONES INTERMEDIAS

No todas las afirmaciones son claramente verdaderas o claramente falsas: muchas caen en un terreno intermedio. Por ejemplo, cuando podría ser verdadera pero se basa en una proyección y no en un dato objetivo; cuando no es estrictamente cierta pero sí lo es el concepto o tendencia a la que alude; cuando la conclusión depende de las variables con las que se la analice; o cuando coincide parcialmente con algunos datos pero no con otros.

Una de las diferencias más notorias entre modelos se dio en cómo responden cuando se les pregunta por la veracidad de este tipo de afirmaciones: algunos modelos tienden a enmarcar su respuesta desde la parte verdadera para luego introducir matices (“es verdad, pero hay que considerar...”), mientras que otros hacen el recorrido inverso (“es falso, aunque...”).

Sin embargo, en el análisis de los cuatro modelos en sus diferentes configuraciones **no encontramos una tendencia marcada en ninguno de ellos**. Un mismo modelo puede valorar de forma más favorable una pregunta y de forma más crítica otra sin que se observe un patrón consistente, y lo mismo ocurre cuando se consulta al mismo modelo desde usuarios diferentes: ante una pregunta idéntica, las respuestas pueden diferir en su encuadre sin que esa variación sea predecible. En conclusión, el análisis **no permite identificar sesgos marcados en la dirección de sus valoraciones sobre la veracidad de las frases analizadas ni entre modelos ni dentro de un mismo modelo.**

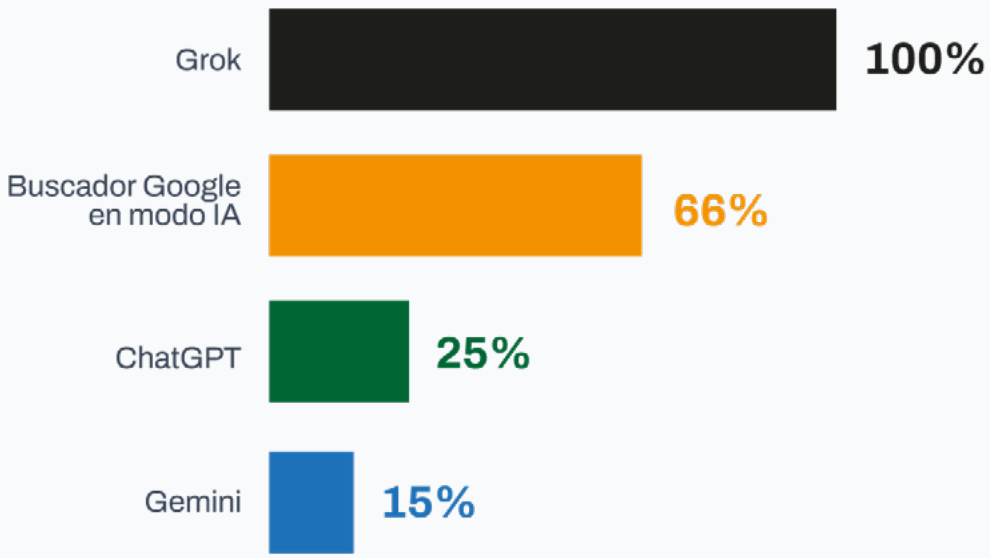
INCORPORACIÓN DE FUENTES DE CHEQUEADO

Dado que las notas de Chequeado sobre el discurso de apertura de sesiones del 2026 abordan exactamente las frases consultadas a los modelos, **analizamos en qué medida cada uno las utilizó para argumentar sus respuestas**. Para ello, comparamos las respuestas obtenidas en dos momentos: durante las dos horas posteriores a la finalización del discurso, a medida que se publicaban los chequeos, y al día siguiente, cuando ya había más notas disponibles. En todos los casos, solo se consideraron las consultas realizadas

< Ir al índice

después de que la nota correspondiente hubiera sido publicada.

Uso de fuentes de Chequeado actualizadas



De los cuatro modelos analizados, **Grok fue el que más recurrió a las notas de Chequeado: las utilizó en el 100% de las consultas realizadas**. Es importante señalar que la muestra es pequeña: las consultas a Grok se hicieron durante el propio discurso o en las dos horas posteriores, período en el que solo había tres notas de Chequeado publicadas correspondientes a las preguntas realizadas.

Aun así, el análisis arrojó un hallazgo ilustrativo. Ante una misma pregunta, Grok respondió de forma incorrecta cuando la consulta se realizó antes de que se publicara el chequeo correspondiente. Cuando otro usuario formuló la misma pregunta apenas 9 minutos después de que la nota de Chequeado estuviera disponible, el modelo incorporó esa fuente y respondió correctamente. **Esto parecería sugerir que la incorporación de fuentes recientes puede mejorar el desempeño de los modelos en preguntas de actualidad**.

Google por su parte citó o utilizó la nota correspondiente en el 66% de las consultas

< Ir al índice

realizadas cuando ya existía el chequeo, aunque el universo analizado en este caso también es acotado.

Le sigue **ChatGPT**, en donde de las 20 consultas realizadas -cuando ya estaban publicadas las notas correspondientes- **las utilizó en el 25% de los casos**.

Gemini fue el modelo que menos recurrió a ellas: de las 20 consultas realizadas cuando ya existía el chequeo correspondiente, solo lo incorporó en 3 respuestas (**un 15%**). Llamativamente, y como hemos anticipado en la sección anterior, **en 6 ocasiones inventó una URL de Chequeado aun cuando la nota real estaba disponible**.

III. VARIACIONES SEGÚN EL MODO DE CONSULTA

Los modelos fueron consultados a través de diferentes tipos de usuarios, formulación de preguntas y, en algunos casos, mediante prompts específicos. El objetivo fue cubrir un abanico de situaciones en las que distintos perfiles pueden interactuar con estos sistemas: desde alguien que usa el modelo por primera vez hasta un usuario con historial de uso, pasando por distintas formas de formular la misma pregunta. En total se realizaron 19 configuraciones de consulta. Para más detalles sobre cada configuración, ver la sección **Metodología**.

El análisis **no encontró diferencias sistemáticas según el tipo de usuario**: que el usuario estuviera logueado o no, tuviera historial de uso o fuera nuevo, o formulara la pregunta de una u otra manera no incidió de forma consistente en la calidad ni en el contenido de las respuestas.

La única diferencia notable se registró en las configuraciones en las que, además de la pregunta "¿es verdad que...?", se incorporó una instrucción específica que indicaba al modelo verificar la afirmación con estándares periodísticos. Concretamente, se le pidió buscar fuentes primarias y chequeos previos, contextualizar la frase, y emitir un veredicto según una escala de tres categorías: Verdadero (cuando los datos coinciden con el espíritu

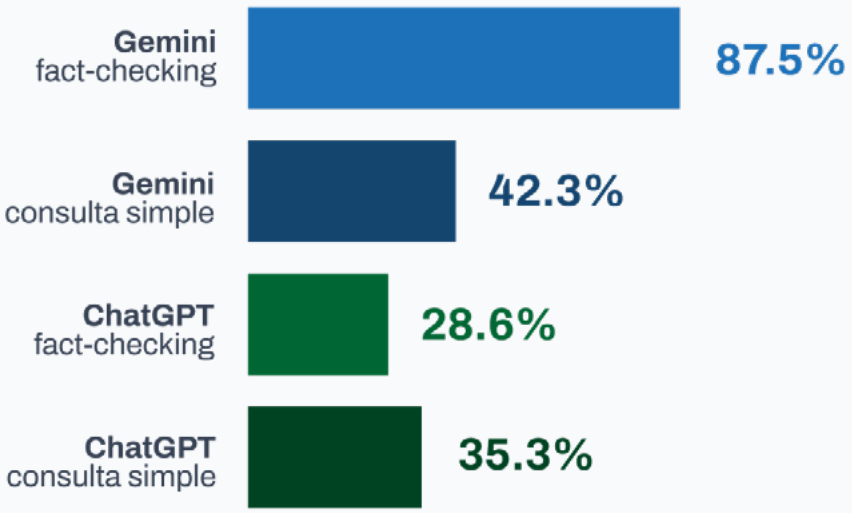
< Ir al índice

de la afirmación), Falso (cuando los datos la contradicen) o Mixto (cuando hay datos a favor y en contra, cuando la respuesta depende de la interpretación, o cuando los datos son exagerados pero van en línea con una tendencia real).

Gemini presentó una proporción de errores notablemente mayor cuando fue consultado con el prompt de estándares periodísticos (87,5% de respuestas problemáticas) que cuando se lo consultó sin ninguna instrucción específica sobre cómo formular la respuesta (42,3%). Esta diferencia se explica en parte por las alucinaciones de fuentes que Gemini generó en la configuración de fact-checking, aunque ese no fue el único tipo de error registrado en esa modalidad.

ChatGPT, en cambio, mostró el panorama opuesto: una menor proporción de respuestas problemáticas en modo fact-checking (28,6%) que en la modalidad de consulta simple (35,3%).

Tasa de error de los modelos en modo fact-checking en comparación con modo de consulta simple



< Ir al índice

05

Conclusiones

Los resultados de este experimento sugieren que utilizar chatbots para verificar la veracidad de afirmaciones de actualidad conlleva riesgos significativos. Con una tasa de error global del 36% y valores que llegan al 53% en el caso de Gemini, la probabilidad de obtener una respuesta incorrecta o engañosa es alta.

Pero el problema no se limita a los errores detectables. Como señalamos en la sección de resultados, muchas respuestas presentan una dificultad adicional: aunque los datos citados puedan ser correctos por separado, la forma en que se los combina, compara o encuadra puede llevar al usuario a conclusiones incorrectas sin que esto sea fácilmente detectable. Un usuario que no se conforme con el veredicto del modelo y quiera evaluar los argumentos por su cuenta tampoco lo tiene garantizado: la acumulación de datos no comparables, las inconsistencias internas y el corrimiento respecto de lo preguntado dificultan esa lectura independiente.

Pedir explícitamente al modelo que busque fuentes confiables tampoco garantiza mejores resultados. En el caso de Gemini, esa instrucción no solo no mejoró la calidad de las respuestas sino que aumentó la tasa de errores: ante el pedido explícito de citar fuentes periodísticas y chequeos previos, el modelo alucinó e inventó fuentes en el 87,5% de los casos.

Hay, sin embargo, un elemento alentador: algunos modelos sí incorporaron fuentes actualizadas y pertinentes cuando estaban disponibles, lo que en algunos casos se tradujo en corrección de respuestas que el mismo modelo había respondido con falencias antes de que se publicaran dichas fuentes. Esto sugiere que la calidad de estos sistemas puede mejorar a medida que se amplía y actualiza su acceso a información confiable y verificada .
Frente a este panorama, utilizar chatbots como fuente principal para verificar afirmaciones de actualidad requiere precaución. Si se los consulta, conviene contrastar sus respuestas

< Ir al índice

con fuentes primarias y chequear que los datos presentados estén realmente en las fuentes citadas. Además, ser especialmente cautelosos ante preguntas que implican comparaciones estadísticas: los modelos pueden combinar datos medidos con criterios distintos o elegir períodos de comparación que no son los metodológicamente indicados. También serlo ante respuestas con múltiples datos e indicadores donde la relación entre ellos no está explicada de forma clara: la acumulación puede dar la impresión de un análisis sólido cuando en realidad el razonamiento que los conecta es inconsistente o poco riguroso. Siempre tener en cuenta que mayor volumen de información brindada no equivale a mayor precisión.

Para verificar afirmaciones de actualidad, los medios periodísticos confiables y los sitios especializados en verificación de datos siguen siendo las fuentes más confiables.

< Ir al índice

06

Metodología

Para acercarnos a la experiencia real de las personas que usan estos chatbots en su vida cotidiana testeamos distintos modelos, distintos tipos de usuario y distintas formulaciones de preguntas. El objetivo fue cubrir el abanico más amplio posible de formas en que alguien puede, hoy, consultar a una IA sobre si algo es verdad.

Tipos de usuario

- usuarios registrados con historial de navegación y uso de las diferentes IAs
- usuarios nuevos generados exclusivamente para la investigación (sin historial de navegación)
- modo invitado (sin iniciar sesión)
- modo de prueba indirecto (vía API)

Los modelos y configuraciones testeadas

Vía API:

- Gemini 3 Flash
- ChatGPT (modelo GPT-5.3)

Acceso web vía plataforma:

- ChatGPT con usuario autenticado con historial, en su versión gratuita
- ChatGPT modo invitado (sin iniciar sesión). Incluimos esta variante porque es una de las formas más frecuentes de uso cotidiano de la herramienta.
- Gemini 3 Flash con usuario nuevo (sin historial de navegación) creado para la investigación, en su versión gratuita.
- Grok integrado a X (Twitter), a través de un usuario autenticado con historial.
- Grok integrado a X (Twitter), a través de un usuario nuevo creado para la investigación (sin historial de uso de la plataforma).

< Ir al índice

- Google con usuario autenticado con historial, se activó el “Modo IA” forzando la respuesta generada por la IA de Google.
- Google con usuario nuevo (sin historial de navegación). Se registró el primer resultado de la búsqueda independientemente de su formato (respuesta generada por IA, enlace o resumen del contenido del enlace recomendado)

Los momentos del testeo

Cada consulta fue realizada en tres momentos distintos:

1. Durante el propio discurso
2. Una y dos horas después de su finalización
3. Al día siguiente

El objetivo fue evaluar si los modelos varían sus respuestas a medida que aparecían nuevas fuentes en la web, por ejemplo las coberturas periodísticas del discurso o los propios chequeos publicados por Chequeado.

Las preguntas y los tipos de prompt

La experiencia previa de Chequeado y de otras organizaciones que indagaron en la evaluación de chatbots indican que las respuestas de los modelos de IA suelen variar según cómo se formula la pregunta. Quisimos explorar si esa variación se limita a diferencias de estilo de la respuesta o si podría afectar también al contenido de la misma e, incluso, al veredicto sobre la veracidad de una afirmación. Para ello, evaluamos siguiendo diferentes estructuras de indagación.

Para los accesos web vía plataforma, utilizamos siempre preguntas del tipo “¿es verdad que…?”, con variaciones naturales en la enunciación del dato consultado. Por ejemplo, ante la frase de Milei “(Al asumir teníamos) indicadores sociales peores a los de 2001”, algunas consultas fueron fraseadas como “¿Es verdad que cuando asumió Milei había indicadores sociales peores que en el 2001?” y otras como “Es verdad que los indica-

< Ir al índice

dores sociales en Argentina eran peores hace dos años que en 2001?". Es necesario destacar que, en el caso de incorporar variaciones en el fraseo, cada una de estas variaciones fue testeada en todas las plataformas.

Para los accesos vía API, utilizamos tres tipos de prompt:

- **Prompt A:** "¿Es verdad que [afirmación]?"
- **Prompt B:** "Milei dijo que [afirmación]. ¿Es verdad?"
- **Prompt C:** un prompt estructurado con instrucciones específicas de fact-checking, que indicaba al modelo verificar la afirmación con estándares periodísticos: buscar fuentes primarias y chequeos previos, contextualizar la frase, y emitir un veredicto según la escala Verdadero (cuando los datos coinciden con el espíritu de la afirmación), Falso (cuando los datos la contradicen) o Mixto (cuando hay datos a favor y en contra, cuando la respuesta depende de la interpretación, o cuando los datos son exagerados pero van en línea con una tendencia real).

Las afirmaciones chequeadas

Seleccionamos 14 afirmaciones realizadas por Milei durante el discurso de apertura de sesiones del 1° de marzo de 2026, abarcando temáticas variadas: pobreza, empleo y salarios, seguridad, gasto público, política social, educación y comercio exterior.

No todas las afirmaciones fueron consultadas en todas las configuraciones. En total **se realizaron 119 consultas**, distribuidas entre los distintos modelos, tipos de usuario y momentos de testeo. Se excluyeron 13 respuestas por presentar incompatibilidades que impedían su análisis puesto que el modelo no logró interpretar correctamente la consulta. El universo analizado quedó conformado por 106 respuestas.

Qué analizamos de las respuestas

Para cada respuesta registramos las siguientes dimensiones:

1. **Consistencia interna**: si la valoración era coherente con los datos que el propio

< Ir al índice

modelo presentaba en su respuesta.

2. **Actualidad de los datos**: si la valoración se basaba en los datos más recientes disponibles o en datos desactualizados.
3. **Pertinencia del período de comparación**: si el modelo comparaba los períodos correctos según lo que se preguntaba, o si elegía un recorte temporal distinto al que correspondía.
4. **Priorización del encuadre por sobre los datos**: si aun citando datos correctos y actualizados, la valoración se realizaba en base a otro recorte.
5. **Editorialización**: si la respuesta incorporaba opiniones o valoraciones no respaldadas por la evidencia presentada.
6. **Desplazamiento del objeto de la pregunta**: si el modelo respondía algo relacionado pero no exactamente lo que se le preguntó, lo que puede inducir a error en lectores que no lean la respuesta completa.
7. **Divergencia entre modelos ante los mismos datos**: si distintos modelos, usando la misma evidencia como justificación, llegaban a valoraciones diferentes.
8. **Encuadre de las valoraciones intermedias**: en respuestas matizadas —por ejemplo, “el número es correcto pero corresponde a un recorte particular de los datos”—, si el modelo ponía el énfasis en lo que la afirmación tenía de verdad o en lo que tenía de falso o exagerado.
9. **Fuentes citadas**: Adicionalmente, cuando los modelos citaban fuentes de Chequeado o de organismos oficiales como el INDEC, analizamos si la nota citada era la más actualizada disponible, si los datos que el modelo mencionaba efectivamente figuraban en esa fuente, y si la interpretación que hacía de esos datos era correcta.
10. **Atribución de citas**: cuando el modelo señaló declaraciones o frases supuestamente dichas por Javier Milei durante el discurso de apertura de sesiones del 1° de marzo de 2026, se contrastaron con **la transcripción oficial** para confirmar si efectivamente habían sido pronunciadas.

< Ir al índice

Reporte

¿Qué tan confiables son los chatbots para verificar afirmaciones políticas?

Con el apoyo de

Reino de los Países Bajos

abr 2026

chequeado

chequeado.com