LA TERCERA
Fecha: 13/01/2025 Audiencia: 271.020 Sección: ACTUALIDAD
Vpe: $6.774.073 Tirada: 76.017 Frecuencia: DIARIO
Vpe pág: $9.829.612 Difusión: 76.017
Vpe portada: $9.829.612 Ocupación: 68,91%

¿Le alcanza para entrar a Medicina?

ChatGPT rindió la PAES 2025 y estos fueron sus resultados

En un novedoso experimento para medir cómo ha mejorado esta tecnología, un grupo de científicos probaron a esta inteligencia artificial haciéndola rendir la Prueba de Acceso a la Educación Superior (PAES). A diferencia de años anteriores, esta vez logró dar una significativa sorpresa.

Francisco Corvalán

¿Cuánto puede avanzar la tecnología de una inteligencia artificial de un año a otro? Muchas veces es difícil poder medir ese tipo de cosas, ya que la arquitectura de los algoritmos que rigen a estas tecnologías, como ChatGPT, suelen ser guardadas bajo el secreto de sus creadores. En este caso, la empresa norteamericana OpenAI.

Pero existe una manera práctica de saber cuánto ha mejorado la IA. En EvoAcademy, compañía dedicada a la capacitación en temas de tecnología e inteligencia artificial, puso a prueba este chatbot y lo hizo rendir la Prueba de Acceso a la Educación Superior (PAES). Este ejercicio, que vienen realizando hace un par de años, logró sorprenderlos durante esta versión: por primera vez esta inteligencia artificial logra obtener puntaje perfecto en una de las pruebas rendidas ¿Le alcanzaría para entrar a alguna de las carreras más solicitadas del país?

Trabajo técnico

El trabajo técnico de este estudio fue realizado por Jonathan Vásquez, profesor adjunto de la Universidad de Valparaíso y Ph.D. in Computer Science de George Mason University, con apoyo del equipo de EvoAcademy. En concreto, el científico hizo que ChatGPT, en todas sus versiones, contestara la PAES admisión regular 2025.

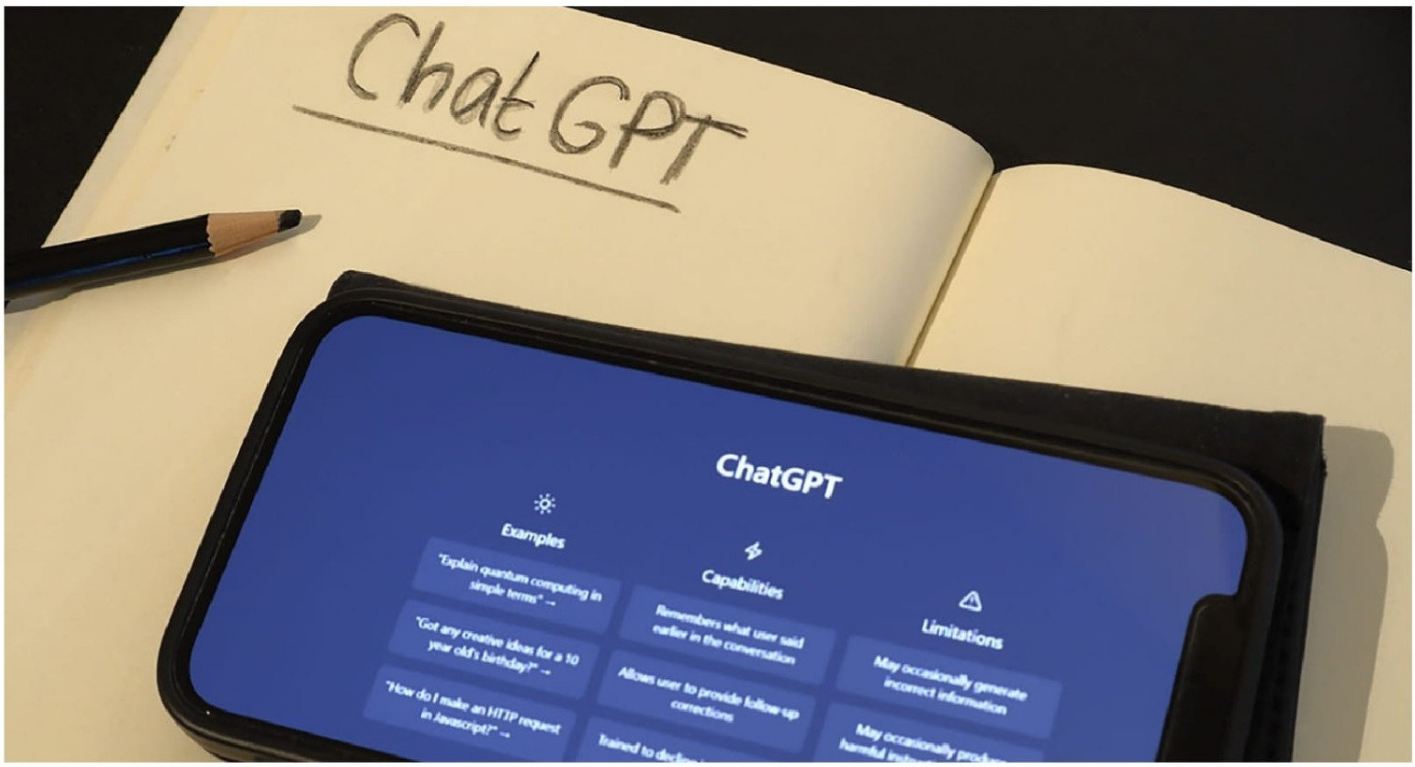
Los modelos de lenguaje enormes, la tecnología detrás de ChatGPT y similares, se han ido sofisticando desde la primera vez que realizaron este ejercicio de prueba. En esta ocasión realizaron la misma prueba con todos las versiones disponibles de esta herramienta tecnológica: gpt-4o, que es la versión de prueba cuando se accede por primera vez al sitio web; también está la versión gpt-4o-mini, que es la que se utiliza luego de que se termine la versión gratuita de prueba, y después están los nuevos modelos de razonamiento o1 y o1-mini.

"Los modelos no tienen acceso a internet. Ellos responden con el conocimiento sobre el cual fueron entrenados. Lo que hicimos esta vez es que generamos estos tipos de modelo y tratamos de testearlos con distintas configuraciones", señaló Vásquez sobre la realización del experimento.

En promedio, los resultados mostraron que la versión gratuita de entrada obtuvo un promedio de 907,38 puntos en la PAES 2025, mientras que la versión mini de este mismo ChatGPT obtuvo 843,75 puntos en promedio. Por otro lado, las versiones o1 y o1-mini -que prometen "razonar" con mayor profundidad pero a menor velocidad- obtuvieron en promedio 865,5 y 857,38 puntos respectivamente.

De hecho, OpenAI declaró que el formato o1 y o1-mini son considerados como los mejores modelos de lenguaje, y alcanzan "niveles similares a un doctorado" en varias de las evaluaciones. Lo paradójico acá, según remarcaron los responsables del experimento, es que la versión que prometía hacer análisis más profundos y detallados no obtuvo mejores resultados en la PAES que la versión estándar de ChatGPT.

En resumen, La IA logró obtener 100% de



► En EvoAcademy, compañía dedicada a la capacitación en temas de tecnología e IA, puso a esta área a rendir la Prueba de Acceso a la Educación Superior (PAES).

LA TERCERA
Fecha: 13/01/2025 Audiencia: 271.020 Sección: ACTUALIDAD
Vpe: $6.994.368 Tirada: 76.017 Frecuencia: DIARIO
Vpe pág: $9.829.612 Difusión: 76.017
Vpe portada: $9.829.612 Ocupación: 71,16%

precisión en la prueba de Historia y Ciencias Sociales, con 3 de los 4 modelos utilizados. Esta es la primera vez en nuestros experimentos que ChatGPT obtiene puntaje perfecto en una prueba. A su vez, el desempeño mejoró significativa en las pruebas de Ciencias, la cual se distingue por sus preguntas enfocadas en observación y análisis de problemas. El promedio de las pruebas de Ciencias mejoró en 18% respecto a 2024, con un promedio del desempeño máximo de 909.25 (gpt-4o).

Por su parte, gpt-4o no logró mejorar en las pruebas de Competencia Matemática, tanto M1 como M2, respecto a su antecesor. En la prueba 2024, gpt-4 logró 90% en M1 y 96% en M2. Mientras que el desempeño máximo de gpt-4o en la prueba 2025 fue de 90% para M1 y 92% para M2. De acuerdo a lo dicho por Vásquez, esto está en línea con las evaluaciones de OpenAI, que muestran que el desempeño de GPT-4o es similar al de GPT-4 turbo.

Los modelos de razonamiento o1, en cambio, no son consistentemente mejores que los modelos GPT. A pesar de su mayor sofisticación, costo y tiempo de procesamiento, los nuevos modelos o1-preview y o1-mini no logran siempre mejores resultados que los modelos gpt-4o. Por ejemplo, en la prueba de Competencia Lectora, gpt-4o obtuvo un 93.33% de precisión máxima, mientras que los modelos o1 obtuvieron entre 86.67% y 93.33% respectivamente.

Detalle de los resultados

Para evaluar los modelos, los investigadores utilizaron el desempeño máximo en cada prueba, pues de este modo es posible ver el potencial de cada formato de inteligencia artificial. El gpt-4, por ejemplo, obtuvo su mayor resultado en la prueba de Ciencias, mención biología, mientras que su puntaje más bajo (836 puntos) lo obtuvo en la prueba de Competencia Matemática (M2). Por otro lado, cabe destacar que todos los formatos evaluados obtuvieron puntaje perfecto (1000 puntos) en la prueba de Historia y Ciencias Sociales.

¿Cómo fue posible esto último? Hasta ahora se sabe que la inteligencia artificial "alucina", o entrega respuestas que suelen ser lógicamente coherentes pero al mismo tiempo incorrectas.

En los modelos de lenguaje siempre existe la posibilidad de que alucinen, pero no es intencional. En el fondo siempre los modelos de lenguaje operan en la base de tratar de predecir la palabra más probable. Si por ejemplo tú dices "Eugenio está tranquilo porque tiene la conciencia... uno termina la frase con "limpia" porque es lo más probable de acuerdo a lo que tú conoces del lenguaje. Y en base a eso es en que operan estas inteligencias artificiales", explica Sebastián Cisterna, MBA de Harvard, gerente general de EvoAcademy y docente en las universidades de Chile y Adolfo Ibáñez.

Pero, según agrega, hay veces que lo más probable no coincide con la realidad, o que el modelo piensa de que debería ser una cosa pero no lo es, y cuando eso ocurre se le llama una alucinación. "Eso siempre va a ocurrir en los modelos de este estilo, pero en la medida de que estos modelos se van sofisticando, que su matemática empieza a ser mejor, que sus bases de datos empiezan a ser más robustas, esas alucinaciones tienden a disminuir y tienden a representar el mundo de mejor manera, que es probablemente lo que pasó en la prueba de Historia, donde aquí no hubo ningún error. Pero las alucinaciones existen igual", aclaró.

En comparación con la edición 2024 de este experimento, el desempeño mejoró en las pruebas de Ciencias. En 2024 el promedio de las pruebas de Ciencias fue 769.25, mientras que en esta edición el promedio del desempeño máximo fue 909.25 (+18%) para el modelo gpt-4o y 830.25 (+8%) para o1-preview. En particular, en 2024 el desempeño general de los modelos era entorno al 81% - 87% en estas pruebas. Pero en esta edición el desempeño máximo subió a 96% - 98% en el modelo gpt-4o, y 91 - 92% en los modelos o1.

El desempeño matemático de los modelos ha mejorado en 2025. Pero, eso sí, los investigadores de EvoAcademy no vieron esto reflejado en el experimento, ya que obtuvo un desempeño peor que lo sacado por GPT-4 el año pasado.

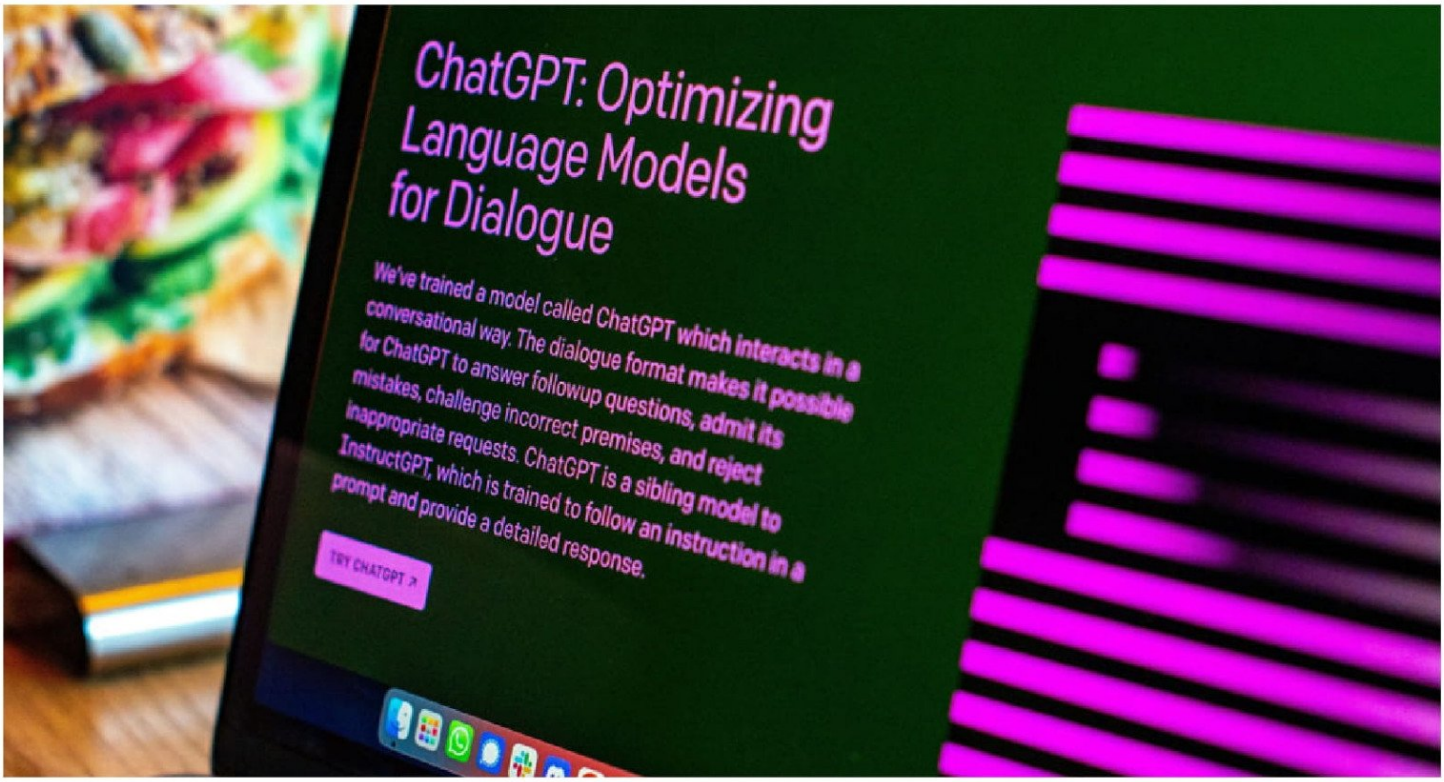
Carreras a las que entraría ChatGPT si fuera un estudiante

Ahora, con los resultados en mano ¿a qué carrera le alcanzaría a la IA para ingresar hipotéticamente a una de las carreras más solicitadas y con puntaje de corte más alto? En base a si su puntaje ponderado fuese mayor o no que el puntaje del último matriculado para esa carrera en 2024, por ejemplo, la versión gpt-4 podría entrar a Medicina en la Universidad de Chile e Ingeniería en la Pontificia Universidad Católica.

Esta es la primera vez que podría entrar a Medicina en una de las universidades con mayor puntaje de corte del país. Por otro lado, todas las modalidades de este chatbot les alcanzaría para entrar al Plan común de la Facultad de Ingeniería en la Universidad de Chile, mientras que solo la versión gpt-4-mini quedaría fuera de Ingeniería Comercial en la PUC.

"Medicina es un caso especial porque, como tiene un puntaje tan alto, también las universidades tienden a variar mucho en la manera que ponderan los puntos. Hay algunas que le ponen más ranking, otras que le ponen menos ranking y como el robot no tiene ranking nosotros teníamos que hacer una aproximación equivalente al NEM", señala Cisterna al respecto.

Eso sí, esto podría ser contraproducente incluso si fuera un alumno real. Esto, porque en general los que postulan a carreras como Medicina son los que están en los primeros percentiles de sus colegios, con un mayor ponderado de notas de enseñanza media. "Si nosotros lo tomáramos así, con el promedio de notas de la gente que postula Medicina, probablemente entrarían muchos más que solamente la Universidad de Chile", concluye. ●



► En comparación con la edición 2024 de este experimento, el desempeño mejoró en las pruebas de Ciencias.