Artículo Original / Original Research

Ciencias Matemáticas, Vol. 36, No. Único, 2022–2023, Págs. 51-56
Recibido (*Received*): 22-04-2024, Revisado (*Reviewed*): 02-05-2024
Aceptado (*Accepted*): 12-06-2024, En línea (*Online*): 17-10-2024

Evaluación de habilidades cognitivas de diversos modelos de lenguajes a gran escala
Assessment of cognitive abilities of various large language models

Kevin Talavera Díaz[1], Alberto Fernández Oliva[2]*, Suilán Estévez Velarde[3]

Resumen Se presenta un análisis exhaustivo del concepto de inteligencia humana y su relación con el desarrollo de la Inteligencia Artificial, haciendo una conexión entre la evolución de la misma y la imitación de las capacidades cognitivas humanas mediante la evaluación de los Modelos de Lenguajes a Gran Escala (LLM). La investigación se enfoca en la importancia de evaluar las habilidades cognitivas de los LLM. Se identifica la falta de conjuntos de datos, suficientemente variados, para hacer esto y la necesidad de una base de datos, lo suficientemente exhaustiva para hacerlo con la calidad requerida. Por tanto, se ha diseñado un conjunto de datos que permite la evaluación de las capacidades cognitivas de diversos modelos de lenguaje, utilizando solamente el lenguaje natural. Los resultados obtenidos, a partir de la evaluación de las habilidades cognitivas sobre el conjunto de datos creado, proporcionan una visión más detallada acerca de las carencias y fortalezas de los LLM en lo que respecta a las habilidades evaluadas, así como cuál es el mejor modelo para explotar cada habilidad individualmente y en general.

Palabras Clave: habilidades cognitivas, procesamiento de lenguaje natural, LLM.

Abstract *An exhaustive analysis of the concept of intelligence is presented. Human and its relationship with the development of Artificial Intelligence, making a connection between its evolution and the imitation of human cognitive abilities through the evaluation of Large Scale Language Models (LLM). The research focuses on the importance of assessing the cognitive abilities of LLMs. The lack of sufficiently varied data sets to do this is identified and the need of a database, exhaustive enough, to do so with the required quality. Therefore, a set of data has been designed that allows the evaluation of the cognitive abilities of various language models, using only natural language. The results obtained, from the evaluation of cognitive abilities on the whole of data created, provide a more detailed view about the gaps and strengths of LLMs with regard to the skills assessed, as well as which is the best model for exploit each skill individually and in general.*

Keywords: *cognitive abilities, natural language processing, LLM.*

Mathematics Subject Classification: 68-11, 68T30, 68T37

[1]Departamento de Computación, Facultad de Matemática y Computación, Universidad de La Habana, Cuba. Email: ktalaveradiaz@gmail.com
[2]Departamento de Computación ,Facultad de Matemática y Computación, Universidad de La Habana, Cuba. Email: afoliva55@gmail.com
[3]Departamento de Computación, Facultad de Matemática y Computación, Universidad de La Habana, Cuba. Email: sestevez@matcom.uh.cu
***Autor para Correspondencia (*Corresponding Author*)**

Editado por (*Edited by*): Damian Valdés Santiago, Facultad de Matemática y Computación, Universidad de La Habana, Cuba.

Citar como: Talavera Díaz, K., Fernández Oliva, A., & Estévez Velarde, S. (2024). Evaluación de habilidades cognitivas de diversos modelos de lenguajes a gran escala. *Ciencias Matemáticas*, 36(Único), 51–56. DOI: https://doi.org/10.5281/zenodo.. Recuperado a partir de https://revistas.uh.cu/rcm/article/view/9044.

Introducción

La inteligencia humana es un concepto complejo y multifacético que ha sido definido de diversas maneras por diferentes teóricos a lo largo de la historia. De manera general, es entendida como la capacidad o conjunto de capacidades principalmente cognitivas que permiten a los humanos adaptarse al entorno, resolver los problemas que este plantea e incluso anticiparse a ellos con éxito [15].

En un intento de imitar e incluso superar la inteligencia humana, la inteligencia artificial [3] ha evolucionado a lo largo de las décadas. Esta evolución ha estado estrechamente vinculada con nuestra comprensión de la inteligencia humana

y ha seguido un camino que busca emular nuestras capacidades cognitivas. En la actualidad, uno de los mayores avances de la inteligencia artificial es el desarrollo de los Modelos de lenguajes a gran escala (LLM, por sus siglas en inglés).

Los LLM son sistemas de aprendizaje automático [18] que han sido entrenados con grandes cantidades de texto, pueden generar texto coherente y responder a consultas en lenguaje natural. Estos han sido sometidos a una serie de pruebas para evaluar su inteligencia o capacidad para simular inteligencia humana. Estas pruebas a menudo implican tareas de procesamiento de lenguaje natural como la comprensión de lectura, la traducción automática, la generación de texto, entre otras [2]. También han sido sometidos a pruebas de razonamiento lógico, las cuales son muy desafiantes para los LLM porque requieren una comprensión profunda del lenguaje y la capacidad de hacer inferencias basadas en el contexto [13].

La evaluación de habilidades cognitivas[1] en LLM es de gran importancia, pues contribuye a garantizar su rendimiento óptimo, imparcialidad, ética y utilidad en una gama amplia de aplicaciones por varias razones dadas en [6]:

- Mejora la precisión y calidad del modelo: Permite identificar las fortalezas y debilidades del modelo de lenguaje, lo que facilita la mejora de su precisión y calidad. Al comprender cómo el modelo responde a diferentes tipos de preguntas y tareas cognitivas, los desarrolladores pueden realizar ajustes para optimizar su desempeño.

- Identificación de sesgos[2] y problemas éticos: Esto es especialmente importante en aplicaciones como los *chatbots* y asistentes virtuales, donde los modelos pueden influir en las decisiones y opiniones de los usuarios [8]. La evaluación ayuda a garantizar que los modelos sean imparciales y éticos en su comportamiento.

- Desarrollo de modelos interpretables: A medida que los modelos de lenguajes se vuelven más complejos, es importante comprender cómo llegan a sus respuestas y qué características están utilizando para tomar decisiones. La evaluación puede ayudar a identificar qué partes del modelo son más relevantes y cómo se pueden interpretar sus resultados.

- Validación de la efectividad de los modelos: Permite comparar el rendimiento de diferentes modelos y establecer métricas objetivas para medir su desempeño. Esto es especialmente importante en aplicaciones como la traducción automática, donde la precisión y la fluidez son fundamentales.

[1] Aptitudes del ser humano relacionados con el procesamiento de la información.

[2] Distorsiones en los datos que pueden llevar a resultados injustos o discriminatorios.

Relevancia del estudio

Se construyó un conjunto de datos lo suficientemente abarcador para evaluar las habilidades cognitivas de los LLM utilizando solamente lenguaje natural. Para ello se realizó un análisis sobre las diversas teorías de la inteligencia humana, seleccionando la de Cattel-Horn-Carroll para tales propósitos. Sobre ella se sustenta el conjunto de habilidades cognitivas tomadas para la construcción del corpus. Se pudo constatar que dicha teoría engloba el resto de las habilidades definidas por las demás teorías de la inteligencia.

Se realizaron una serie de experimentos para evaluar el desempeño de diferentes LLM sobre la base de datos construida. Los modelos de lenguaje obtuvieron buenos resultados en las habilidades que se corresponden con la comprensión y con el uso de la memoria a largo plazo. No se obtuvieron resultados satisfactorios en las habilidades que requieren conocimientos específicos de un dominio, así como diversos niveles de razonamiento.

1. Teorías de la Inteligencia

De todas las teorías de inteligencia, una de las más completa es la de Cattell-Horn-Carroll (CHC) [9], la cual es la que se comenzará a analizar, luego se estudiarán otras teorías y se observará como, de alguna manera, todas están contenidas dentro de esta. Luego se analizará qué habilidades miden las bases de datos más utilizadas y su relación con las habilidades que serán definidas como resultado del estudio de las teorías de la inteligencia, las que serán explicadas y comparadas más adelante.

1.1 CHC

La teoría de las capacidades cognitivas CHC es una de las teorías psicométricas más completas hasta la fecha y está respaldada empíricamente por la estructura de las capacidades cognitivas. Representa de forma integrada a las obras de Raymond Cattell, John Horn y John Carroll [1, 7, 9, 14].

Debido a que cuenta con un diverso cuerpo de apoyo empírico en la literatura de investigación (por ejemplo, criterios de desarrollo, neurocognitivos y de resultados), se utiliza ampliamente como base para seleccionar, organizar e interpretar pruebas de inteligencia y habilidades cognitivas [5].

2. Propuesta e Implementación

A partir del estudio de las teorías de las inteligencia más importantes y las habilidades cognitivas que estas definen, así como las bases de datos más utilizadas para evaluar este tipo de habilidades en los LLM, fue posible la construcción de una base de datos que complementara las carencias existentes. Es decir, se pudo crear una que agrupara el conjunto de habilidades cognitivas estudiadas, recopilando preguntas de diversas fuentes, entre las que se encuentran: tests de inteligencia [12], trivias [11], las propias bases de datos estudiadas, entre otras.

La base de datos creada cuenta con una serie de preguntas, en las cuales los LLM deberán ser capaces de, dado un

enunciado y un conjunto de posibles soluciones, seleccionar la(s) respuesta(s) correctas, escribir la solución en caso de ser preguntas relacionadas con las matemáticas, responder, verdadero o falso a preguntas booleanas e, incluso, completar espacios en blanco [4]. Estas preguntas están escritas en lenguaje natural y se asocian con las siguientes 6 habilidades primarias y 10 secundarias:

- Conocimiento cuantitativo: álgebra, geometría, probabilidades, teoría de números.
- Conocimiento general: arte y literatura, ciencia y naturaleza, geografía, música.
- Conocimiento verbal: conocimiento léxico, desarrollo del lenguaje.
- Inteligencia fluida: inducción, razonamiento general.
- Lectura: comprensión.
- Procesamiento visual.

Los LLM evaluados en este trabajo fueron: ChatGPT 3.5 [10], Zephyr [17] y Llama 2 [16].

3. Experimentos

Una vez aplicados los instrumentos de recolección de los resultados, se procedió a realizar el tratamiento correspondiente para el análisis de los mismos. Los experimentos realizados fueron los siguientes:

1. Evaluar las respuestas dadas por los LLM utilizando expresiones regulares.
2. Verificar manualmente las respuestas dadas por los LLM.
3. Tomar un sujeto de experimento que resolviera las mismas preguntas que resolvieron los LLM.

A continuación se muestran los resultados asociados a cada uno de los experimentos.

4. Resultados

En esta sección se presentan los resultados obtenidos de la evaluación de los LLM con respecto a las habilidades cognitivas. Estos resultados mostrarán el rendimiento de cada modelo en cada una de estas habilidades utilizando la siguiente métrica:

$$Precision = 100 \cdot (Respuestas\ Correctas / Total\ Preguntas). \quad (1)$$

Para facilitar el proceso de exposición de los resultados, se ha dividido el rendimiento de los modelos por habilidad en cinco categorías:

1. Muy alto: 80 % - 100 % de precisión.
2. Alto: 60 % - 79 % de precisión.
3. Medio: 40 % - 59 % de precisión.
4. Bajo: 20 % - 39 % de precisión.
5. Muy bajo: 0 % - 19 % de precisión.

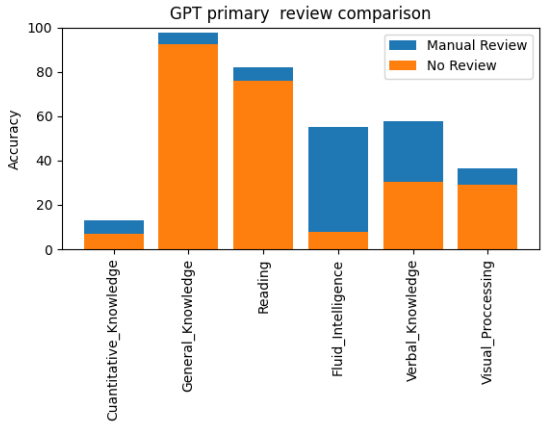

Figura 1. GPT: Comparación entre evaluación manual y automática [*GPT: Comparison between manual and automatic evaluation*].

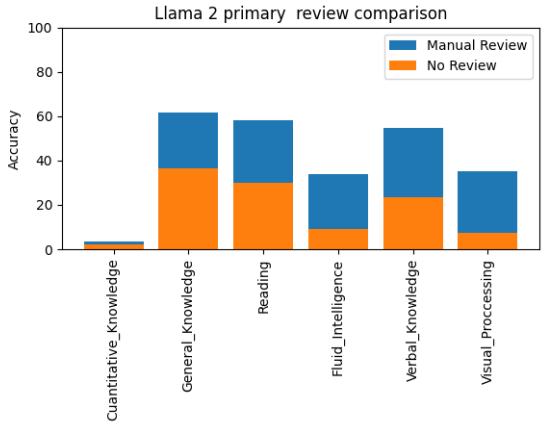

Figura 2. Llama 2: Comparación entre evaluación manual y automática [*Llama 2: Comparison between manual and automatic evaluation*].

5. Comparación de resultados

A partir de los resultados obtenidos en cada uno de los experimentos y de lo que se muestra en la Figura 5 y la Tabla 1 se pueden realizar las siguientes comparaciones:

Tanto ChatGPT, Llama 2 así como Zephyr, a diferencia de la muestra humana con 85 %, obtuvieron resultados inadecuados en tareas referentes a la habilidad de conocimiento cuantitativo. Ninguno superó el 20 % de precisión. ChatGPT

Skills	ChatGPT	Llama 2	Zephyr	Human
Cuantitative Knowledge	13	3.66	9.32	85.93
General Knowledge	97.5	61.5	81	37.31
Reading	82	58	78	59.9
Fluid Intelligence	55.05	33.70	52.80	96.77
Verbal Knowledge	57.57	54.54	51.51	96.77
Visual Proccessing	36.58	35.00	26.82	75

Tabla 1. Tabla comparativa de resultados [*Comparative table of results*].

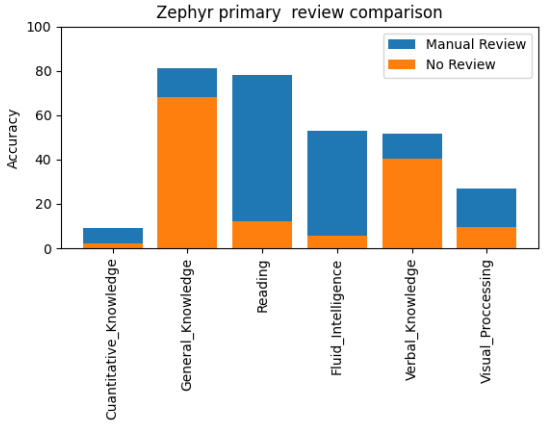

Figura 3. Zephyr: Comparación entre evaluación manual y automática [*Zephyr: Comparison between manual and automatic evaluation*].

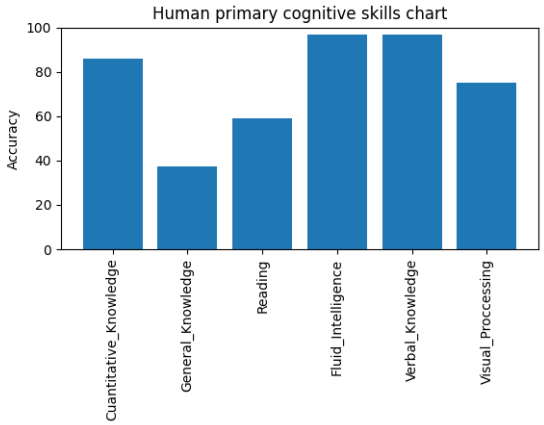

Figura 4. Habilidades primarias Humano [*Primary skills Human*].

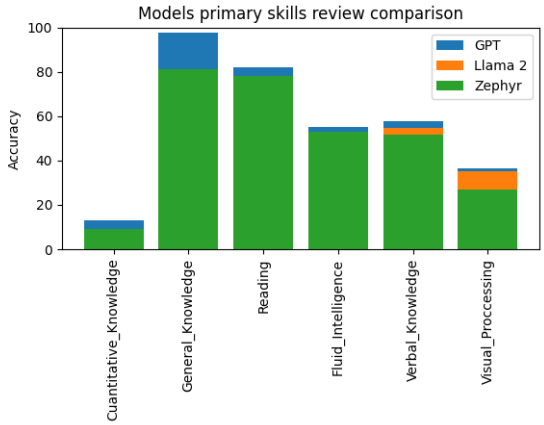

Figura 5. Comparación de resultados entre los LLM [*Comparison of results between LLMs*].

fue el que obtuvo el mejor resultado, aproximadamente un 13 %, luego Zephyr con un 9 % y Llama 2 con sólo un 3 %.

Con respecto a la habilidad de conocimiento verbal, los tres modelos tuvieron un rendimiento medio entre el 50 % y 60 %. ChatGPT fue superior al resto, por un pequeño margen, con 57 %. Le sigue Llama 2 con un 54 % y, por último, Zephyr con 51 %. En el desarrollo de esta habilidad, la muestra humana fue muy superior con un 96 %.

En las preguntas relacionadas a la habilidad de procesamiento visual los tres modelos se encuentran con un rendimiento bajo, entre el 20 % y 40 %. Sin embargo, el humano obtuvo un 80 %. En este caso, ChatGPT y Llama 2 tuvieron prácticamente la misma precisión, 36 % y 35 % respectivamente. Zephyr tuvo una precisión del 26 %.

En el caso de las preguntas relacionadas a la habilidad de conocimiento general, el comportamiento fue más variado. El rango de precisión varía en un 40 % entre los diferentes modelos. ChatGPT obtuvo un casi perfecto 97 %, seguido de Zephyr con un 81 %, luego, y con bastante diferencia respecto a ChatGPT, se encuentra Llama 2 con 61 % de precisión. Por último, el humano con sólo un 37 %.

Con respecto a las preguntas de inteligencia fluida, las precisiones oscilan entre un 30 % y 60 %. ChatGPT y Zephyr están bastante parejos con 55 % y 52 %, respectivamente. Este último obtuvo alrededor de un 20 % de diferencia a Llama 2, que alcanzó un 33 % de precisión. El anotador humano fue

superior a los anteriores con un 96 % de precisión.

En el caso de las preguntas relacionadas a la habilidad de lectura, ChatGPT y Zephyr se comportan de manera similar con 82 % y 78 % de precisión, respectivamente. Con un 20 % menos que Zephyr se encuentra Llama 2, con un 58 %. Por último, se encuentra el anotador humano con un 59 % de precisión.

6. Discusión general

Los modelos ChatGPT, Llama 2 y Zephyr tuvieron un mayor rendimiento en las preguntas asociadas con las habilidades de conocimiento general y lectura, dado que estos modelos están enfocados a las tareas de generación de texto y comprensión. La diferencia de precisión entre ellos probablemente esté dada en el volumen y la calidad de los datos que se utilizan para entrenar.

Con respecto a la habilidad de conocimiento cuantitativo, los tres modelos obtuvieron un resultado inadecuado. Esto puede estar dado por la propia naturaleza probabilística de los modelos de lenguaje, lo cual incluye un elemento de incertidumbre en sus respuestas. Para los problemas matemáticos, donde la precisión y la exactitud son cruciales, depender de estos modelos no es lo ideal.

Las habilidades de inteligencia fluida y procesamiento visual, requieren un cierto nivel de razonamiento lógico, inductivo y abstracto, habilidades en las que los modelos tienen un pobre desempeño. Lo anterior viene dado porque estos modelos no están enfocados a resolver este tipo de problemas, como se comentó en el primer punto.

Por último, la habilidad de conocimiento verbal, aunque coincide mucho con generación de texto y comprensión, también requiere de un cierto nivel de razonamiento, razón por la cual tampoco se obtuvieron resultados satisfactorios en relación con esta.

Con respecto a la comparación con el anotador humano, las habilidades en las que mejor desempeño poseen los LLM evaluados coinciden con las que peor desempeño posee el humano y viceversa, y la razón de esto estÃ¡ dada por las razones mencionadas anteriormente en este apartado.

Referencias

[1] Alfonso, V.C., D.P. Flanagan, and S. Radwan: *The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities*. In *Contemporary intellectual assessment: Theories, tests and issues*, pages 185–202. The Guilford Press, New York, NY, US, 2005, ISBN 1-59385-125-1.

[2] AWS, 2023. https://aws.amazon.com/es/what-is/nlp/#:~:text=tareas%20de%20NLP%3F-,%C2%BFQu%C3%A9%20es%20la%20NLP%3F,y%20comprender%20el%20lenguaje%20humano.

[3] DataScientest, 2023. https://datascientest.com/es/inteligencia-artificial-definicion.

[4] Excentos, 2023. https://documentation.excentos.com/display/WORKBENCH/Boolean+Question.

[5] Flanagan, D.P., V.C. Alfonso, S.O. Ortiz, and A.M. Dynda: *Integrating Cognitive Assessment in School Neuropsychological Evaluations*. In *Best Practices in School Neuropsychology: Guidelines for Effective Practice, Assessment, and Evidence-Based Intervention*, pages 101–140. John Wiley & Sons, Inc., Hoboken, NJ, US, 2012, ISBN 9780470422038.

[6] Hitch, 2023. https://hello.gethitch.ai/blog/importancia-evaluar-habilidades-//cognitivas-candidatos/#predice-el-rendimiento.

[7] Horn, J.L. and N. Blankson: *Foundations for better understanding of cognitive abilities*. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, pages 41–68. The Guilford Press, New York, NY, US, 2005, ISBN 1593851251.

[8] IBM, 2022. https://www.ibm.com/es-es/topics/chatbots.

[9] McGrew, K.S.: *The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future*. New York, NY: Guilford Press, 2nd ed., pp. 136–182, 2005.

[10] OpenAI: *ChatGPT (Mar 14 version) [Large language model]*, 2023. https://chat.openai.com/chat.

[11] Pérez Porto, J. y A. Gardey, 2022. https://definicion.de/trivia/.

[12] Pérez Porto, J. y M. Merino, 2023. https://definicion.de/test-de-inteligencia/.

[13] Roch, E., 2023. https://lovtechnology.com/que-es-llm-large-language-model-//como-funcionan//-y-para-que-sirven/.

[14] Schneider, W.J. and K.S. McGrew: *The Cattell-Horn-Carroll model of intelligence*. In *Contemporary intellectual assessment: Theories, tests, and issues, 3rd ed.*, pages 99–144. The Guilford Press, New York, NY, US, 2012, ISBN 978-1-60918-995-2.

[15] Sternberg, R.J.: *A theory of adaptive intelligence and its relation to general intelligence*. Journal of Intelligence, 7(4):23, 2019. https://doi.org/10.3390/jintelligence7040023.

[16] Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P.S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E.M. Smith, R. Subramanian, X.E. Tan, B. Tang, R. Taylor, A. Williams, J.X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov y T. Scialom: *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv, 2307.09288, 2023.

[17] Tunstall, L., E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush y T. Wolf: *Zephyr: Direct Distillation of LM Alignment*. arXiv, 2310.16944, 2023.

[18] Wiki, 2023. https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico.

Contribución de autoría

Conceptualización K.T.D., S.E.V.

Curación de datos K.T.D., S.E.V.

Análisis formal K.T.D., S.E.V., A.F.O.

Investigación K.T.D., S.E.V., A.F.O.

Metodología K.T.D., A.F.O.

Administración de proyecto K.T.D., S.E.V., A.F.O.

Recursos K.T.D., S.E.V., A.F.O.

Supervisión K.T.D., S.E.V., A.F.O.

Validación K.T.D., S.E.V., A.F.O.

Visualización K.T.D., A.F.O.

Redacción: preparación del borrador original K.T.D., A.F.O.

Redacción: revisión y edición K.T.D., A.F.O.

Conflictos de interés

Se declara que no existen conflictos de interés. Los autores declaran que no hubo subvenciones involucradas en este trabajo.

Suplementos

Este artículo no contiene información suplementaria.