Reseña de

Fundamental Principles of Linguistic Structure are Not Represented by o3

Por David Carmona Barrales

Ficha técnica

Título: *Fundamental Principles of Linguistic Structure are Not Represented by o3*.

Autor: Elliot Murphy; Evelina Leivada; Vittoria Dentella; Fritz Günther; Gary Marcus.

Público objetivo: Especializado (lingüística teórica y computacional, evaluación de LLM).

Tipo: *Paper*

Institución: El artículo es una colaboración de investigadores afiliados a diversas instituciones, entre ellas: UTHealth (Texas, USA), Universitat Autónoma de Barcelona (España), ICREA (España), University of Pavia (Italia), Humboldt-Universität zu Berlin (Alemania) y New York University (USA).

Fecha de Publicación: 2025

Extensión: 42 páginas.

Enlace al documento: https://arxiv.org/pdf/2502.10934

Contexto

Este artículo se inscribe en el centro del debate contemporáneo sobre las capacidades reales de los Grandes Modelos de Lenguaje (LLM) y su progreso hacia una posible Inteligencia Artificial General (IAG). Redactado por un equipo de lingüistas y científicos cognitivos, entre los que destaca Gary Marcus, un conocido analista crítico de las arquitecturas de aprendizaje profundo, el trabajo se posiciona como una respuesta empírica a las afirmaciones optimistas que sugieren que los LLM están a punto de igualar o superar la competencia lingüística humana.

La investigación se enfoca en el modelo "o3-mini-high" de OpenAI, presentado como un "modelo de razonamiento" que teóricamente mejora las funciones computacionales y lógicas sobre los LLM estándar. El artículo busca someter este modelo de vanguardia a una serie de pruebas lingüísticas rigurosas, diseñadas específicamente para evaluar su dominio de la composicionalidad jerárquica, un pilar fundamental del lenguaje humano que va más allá del simple reconocimiento de patrones estadísticos secuenciales.

Exposición y Reconstrucción del Contenido

La tesis principal del documento es que, a pesar de su aparente fluidez, el modelo o3 de OpenAI no posee una comprensión genuina de los principios fundamentales de la estructura lingüística, especialmente en lo que respecta a la sintaxis composicional y la interfaz sintaxis-semántica. Los autores sostienen que el modelo falla sistemáticamente en tareas que requieren un razonamiento estructural abstracto, demostrando que sus éxitos se basan en estadísticas de superficie y no en una competencia gramatical análoga a la humana.

Para demostrarlo, los autores emplean una metodología basada en una serie de 26 *prompts* o instrucciones directas dadas al modelo. Estas pruebas están diseñadas para aumentar progresivamente en complejidad jerárquica:

1. **Tareas Lineales y Superficiales**: Inicialmente, el artículo muestra que o3 resuelve con éxito tareas que dependen de patrones lineales y estadísticos, como crear un palíndromo o contar letras en una palabra (el "Strawberry Test").
2. **Fallas en la Estructura de Frase**: La primera gran falla aparece cuando se prueba la generalización de reglas sintácticas con pseudopalabras. El modelo juzga incorrectamente como gramaticales secuencias que violan la estructura, evidenciando una incapacidad para abstraer la regla más allá de los ejemplos conocidos.
3. **Incomprensión Semántica y Estructural**: El informe documenta una serie de fracasos en dominios clave:
    - **Oraciones de Escher**: El modelo no detecta la imposibilidad semántica en oraciones comparativas que son estructuralmente engañosas.
    - **Incrustación Central (*Center-Embedding*)**: Falla al analizar oraciones con cláusulas de relativo anidadas, llegando a "alucinar" verbos o pronombres para forzar una interpretación gramatical donde no la hay.
    - **Generación de Violaciones Sintácticas**: Cuando se le pide explícitamente que genere una oración agramatical, o3 lucha por hacerlo. En su lugar, a menudo produce frases semánticamente extrañas pero sintácticamente correctas, o falla en crear violaciones específicas y dependientes del contexto.
4. **Evaluación de la Gramaticalidad**: En una tarea de juicio de aceptabilidad de 16 oraciones, o3 identifica correctamente casi todas las agramaticales, pero clasifica erróneamente como agramaticales varias oraciones que son perfectamente válidas en inglés. El modelo muestra una especial debilidad con las oraciones "parcialmente aceptables", demostrando no ser sensible al espectro de gramaticalidad que los humanos manejan con naturalidad.
5. **Fallas en la Inferencia Abstracta**: El modelo también fracasa en tareas que requieren disociar la estructura de la estadística léxica, como en un test "Jabberwocky" modificado

donde se reemplazan las palabras de función, o al intentar crear "objetos imposibles" mediante la polisemia.

Los autores concluyen que estos fallos no son triviales, sino que apuntan a una limitación fundamental de la arquitectura actual de los LLM: su fuerte sesgo hacia el procesamiento "horizontal" (secuencial y estadístico) en detrimento del procesamiento "vertical" (jerárquico y composicional).

Análisis crítico:

Fortalezas:

La principal fortaleza del artículo es su rigor metodológico y su base empírica. En lugar de argumentar desde una perspectiva puramente teórica, los autores diseñan un conjunto de pruebas específicas y transparentes. La inclusión directa de los prompts y las respuestas del modelo permite al lector verificar las afirmaciones y evaluar la evidencia de primera mano.

La selección de los fenómenos lingüísticos es otro punto fuerte. Las pruebas cubren un amplio espectro de desafíos sintácticos y semánticos bien establecidos en la literatura lingüística (e.g., incrustación central, violaciones de islas, condiciones de ligamiento, zeugma), que son ideales para sondear las limitaciones del procesamiento meramente estadístico.

La progresión lógica de los argumentos es muy eficaz. El artículo comienza con tareas simples que el modelo puede resolver, estableciendo una línea de base justa, para luego introducir gradualmente desafíos más complejos que revelan sus debilidades de manera contundente. Este enfoque hace que las conclusiones sean más impactantes y difíciles de refutar.

Finalmente, la discusión contextualiza de manera excelente los resultados dentro de los debates teóricos más amplios sobre la naturaleza del lenguaje, la cognición y las aspiraciones de la IAG, conectando los fallos empíricos con sus implicaciones teóricas.

Aspectos mejorables:

Los propios autores reconocen algunas de las limitaciones del estudio, como el carácter preliminar del análisis estadístico y el tamaño de muestra limitado. Si bien los ejemplos son cualitativamente poderosos, un análisis cuantitativo más sistemático podría reforzar aún más las conclusiones.

Otra limitación señalada es la ausencia de datos de rendimiento humano para comparar directamente. Aunque las respuestas del modelo son analizadas desde la perspectiva de la teoría lingüística, tener una línea de base de cómo los hablantes humanos responden a

estas mismas tareas (especialmente a las más ambiguas o complejas) añadiría una capa de profundidad al análisis comparativo.

Se podría argumentar que algunas de las pruebas son excesivamente complejas o incluso injustas (por ejemplo, pedir al modelo que cree un nuevo tipo de paradoja filosófica más sofisticada que todas las existentes). Si bien estos prompts extremos son reveladores, los críticos podrían centrarse en ellos para desestimar la validez de las pruebas más estándar y pertinentes.

Finalmente, aunque los autores mencionan la posibilidad de que los fallos en la generación de árboles sintácticos puedan deberse a problemas de la interfaz, el artículo podría haberse beneficiado de una exploración más profunda de explicaciones alternativas para algunos de los errores del modelo, fortaleciendo así su argumento principal al descartar otras hipótesis.

Síntesis y Conclusión:

Este trabajo ofrece una evaluación empírica, detallada y crítica de las capacidades lingüísticas del modelo o3 de OpenAI. A través de una serie de pruebas ingeniosamente diseñadas, demuestra de manera convincente que, si bien el modelo es competente en el manejo de patrones estadísticos y secuenciales, falla sistemáticamente cuando se enfrenta a tareas que requieren una comprensión profunda de la estructura jerárquica y composicional del lenguaje.

El valor del documento reside en su capacidad para ir más allá de la superficie de la fluidez lingüística de los LLM y poner a prueba sus fundamentos computacionales. Las conclusiones respaldan firmemente la hipótesis de que las arquitecturas actuales, basadas en la predicción del siguiente token, se topan con un "muro obstinadamente resistente" en el camino hacia la verdadera competencia lingüística humana. El artículo no solo expone las limitaciones de un modelo específico, sino que plantea preguntas fundamentales sobre el futuro de la investigación en IA y la viabilidad de alcanzar la inteligencia general a través del simple escalado de datos y cómputo.

Fundamental Principles of Linguistic Structure are Not Represented by o3 es una lectura relevante para cualquier persona interesada en los límites de la inteligencia artificial actual. Se recomienda especialmente a lingüistas, científicos cognitivos e investigadores de IA que busquen un contrapunto riguroso y basado en evidencia al discurso a menudo triunfalista que rodea a los grandes modelos de lenguaje.

Referencia del artículo (APA 7ª Ed.): Murphy, E., Leivada, E., Dentella, V., Günther, F., & Marcus, G. (2025). *Fundamental Principles of Linguistic Structure are Not Represented by o3*. arXiv (cs.CL). Cornell University. DOI: 10.48550/arXiv.2502.10934.