Grandes Modelos de Lenguaje y la conexión entre alucinaciones y creatividad: un enfoque peirceano

Mariana Olezza
Universidad de Buenos Aires
olezza.mur@gmail.com

Paniel Reyes Cárdenas*
Oblate School of Theology
The University of Sheffield
panielosberto.reyes@upaep.mx

RESUMEN Descrito en su ensayo "Un argumento olvidado en favor de la realidad de Dios", el *musement* es una forma de "Juego Puro" (Pure Play), un estado de contemplación libre y sin un objetivo predeterminado (Peirce, "A Neglected Argument" EP 2, 434). Consiste en permitir que la mente vague libremente por las tres categorías de la experiencia —las cualidades, los hechos y las regularidades del universo— observando las conexiones y armonías que surgen sin forzarlas. Este acto no es un razonamiento dirigido, sino una apreciación estética de las ideas y sus patrones, un deleite en la contemplación misma. Este juego, que podría parecer ocioso, es para Peirce la cuna de la creatividad y el motor del descubrimiento científico. El *musement* es el estado mental que nutre la abducción, el único tipo de inferencia capaz de generar hipótesis genuinamente nuevas. Al aplicar este concepto a los Grandes Modelos de Lenguaje, podemos conjeturar que al ajustar los hiperparámetros hacia una mayor "creatividad" (una Temperatura alta, altos niveles de Top–K), estamos intentando *simular* una forma de *musement* artificial. Forzamos al modelo a abandonar la ruta lógica más probable y a entrar en un"juego puro" con las relaciones latentes en sus datos.El resultado, a menudo una "alucinación" metafórica o tangencial,emula el vagabundeo de la mente humana del cual, ocasionalmente, puede brotar una conexión novedosa y valiosa. Este marco peirceano ofrece una perspectiva reveladora para analizar la *tensión fundamental* en los Grandes Modelos de Lenguaje (GMLs) entre *veracidad y creatividad*. Proponemos que la "alucinación" en un GML no es simplemente un error, sino un fenómeno que se sitúa en un espectro entre dos polos normativos de Peirce,

*Corresponding author.

a saber 1) La Lógica, que busca la veracidad y la coherencia, se corresponde con el esfuerzo por mitigar las alucinaciones para producir un discurso determinista y fiable 2) La Estética, que busca lo admirable y sorprendente, se corresponde con la creatividad, donde la amplificación de las alucinaciones mediante hiperparámetros como la Temperatura (T) o el Top-K permiten generar resultados novedosos, metafóricos y poéticos. La alucinación puede ser un momento de libre juego de las facultades con potencial creativo.

KEYWORDS Peirce; grandes Modelos de Lenguaje; creatividad; alucinaciones

Introducción: La visión de Peirce

Charles Sanders Peirce es reconocido como el fundador del pragmatismo y uno de los lógicos más influyentes de la historia. Sin embargo, reducir su pensamiento a la lógica es pasar por alto el pilar de su sistema filosófico. En su arquitectónica del conocimiento, Peirce no clasifica las ciencias por su objeto de estudio, sino por las operaciones de la mente que guían la investigación. En esta estructura, nos sorprende al colocar a las ciencias normativas —Estética, Ética y Lógica— como el fundamento de toda actividad intelectual (Peirce 1998b, 372).

Lejos de ser una disciplina secundaria, para Peirce la estética es la ciencia normativa fundamental. Se ocupa de lo que es admirable por sí mismo, sin condición alguna (*summum bonum*). De ella se desprende la ética, que define la conducta correcta como aquella que es admirable, es decir, la que se alinea con ese ideal estético. Finalmente, la lógica se subordina a la ética, pues el compromiso con la verdad y el razonamiento correcto no es sino un deber ético; la lógica misma está enraizada en un principio social que debe abrazar a toda la comunidad de investigación (Peirce 1992, 149). La búsqueda de la verdad (lógica) es una forma de conducta admirable (ética) que aspira a un fin incondicionalmente valioso (estética). Este ideal último es lo que Peirce llamó el "crecimiento de la racionalidad concreta" (*growth of concrete reasonableness*), un desarrollo de hábitos inteligentes y autocontrolados que profundizan el entendimiento de lo real.

Para comprender cómo la estética fundamenta la creatividad en el sistema de Peirce, es indispensable introducir su concepto de musement. Descrito en su ensayo "Un argumento olvidado en favor de la realidad de Dios", el *musement* es una forma de "Juego Puro" (Pure Play), un estado de contemplación libre y sin un objetivo predeterminado (Peirce, "A Neglected Argument" EP 2, 434). Consiste en permitir que la mente vague libremente por las tres categorías de la experiencia —las cualidades, los hechos y las regularidades del universo— observando las conexiones y armonías que surgen sin forzarlas. Este acto no es un razonamiento dirigido, sino una apreciación estética de las ideas y sus patrones, un deleite en la contemplación misma.

Este juego, que podría parecer ocioso, es para Peirce la cuna de la creatividad y el motor del descubrimiento científico. El *musement* es el estado mental que nutre la abducción, el único tipo de inferencia capaz de generar hipótesis genuinamente nuevas. Mientras la deducción extrae conclusiones de premisas y la inducción

generaliza a partir de casos, la abducción es el salto creativo que propone una explicación para un fenómeno sorprendente. Es durante el *musement* que la mente, liberada de las restricciones del pensamiento lineal, puede sintetizar elementos dispares en una nueva idea coherente y explicativa. En este sentido, el *musement* es la actividad estética que sirve como condición de posibilidad para la "lógica de la sorpresa" que impulsa todo conocimiento nuevo (Peirce 1998a, 436-38).

Los Grandes Modelos de Lenguaje (GML): Tokens y Alucinaciones

Cuando nos referimos a GMLs hablamos de sistemas como los ya conocidos chatGPT, Gemini, DeepSeek, y demás. En el famoso artículo de Vaswani et al. (2017) se presentó un modelo llamado de autoatención que permite focalizarse en distintas secciones de las oraciones a la vez, aumentando la eficiencia de procesamiento lingüistico. Surge la "*tokenización*", que consiste en dividir símbolos como "*universe*", por ejemplo, en "*un*", "*iv*", "*er*", "*se*" ("*slices*" o "*tokens*"), para que puedan ser procesados por los GMLs.[1]

Los GML, sufren también de lo que se llama alucinaciones. A menudo generan respuestas incorrectas en términos fácticos o de fidelidad[2], a pesar de su éxito en diversas aplicaciones y de hacerlo con gran confianza. Dado que la información proporcionada por estos sistemas puede influir directamente en la toma de decisiones, cualquier dato engañoso tiene el potencial de difundir creencias falsas e incluso causar daño. (Si hablamos del polo de la veracidad y no de "creatividad emulada", como ya veremos.)

Hiperparámetros en GMLs: Temperatura "T y Top–K"

Los hiperparámetros Temperatura (T) y Top–K son configurados por el supervisor o el usuario del "prompt" (semántica externa).

Por un lado, el hiperparámetro T: Un $T \to \infty$ dará como resultado respuestas más aleatorias, mientras que un T más cercano a 0, más deterministas, dado que el T cercano a cero distribuye los logits (son los valores reales que produce la capa final de un modelo de clasificación antes de aplicar una función de activación como *softmax* (capa de salida). en puntos próximos, tendiendo a un discurso más determinista y fiable (verídico), y al tender a infinito los distribuye dispersamente.

Por otro lado, Top–K está encargado de distribuir las probabilidades en los próximos "K" tokens. Si se configura en "1", será total la probabilidad de elegir

[1] Después se pasa por un proceso de "*embedding*" donde los tokens son transformados a vectores numéricos y correlacionados en una nube semántica.

[2] Alucinaciones fácticas quiere decir que el GML responde incorrectamente; de fidelidad, que no sigue exactamente las instrucciones que uno le ha pedido. Hay muchos tipos de alucinaciones fácticas comprendiendo de datos, entrenamiento e inferenciales (Huang et al. 2023).

el token más probable (próximo token = 100 % probabilidad) pero al seleccionar, por ejemplo, "$K = 500$", las probabilidades se distribuyen entre los próximos 500 tokens, logrando que el sistema se vuelva más aleatorio y más "creativo".

Los Polos Normativos

Al aplicar el concepto de "*musement*" a los Grandes Modelos de Lenguaje, podemos conjeturar que al ajustar los hiperparámetros hacia una mayor "creatividad" (una Temperatura alta, altos niveles de Top–K), estamos intentando *simular* una forma de *musement* artificial. Forzamos al modelo a abandonar la ruta lógica más probable y a entrar en un "juego puro" con las relaciones latentes en sus datos. El resultado, a menudo una "alucinación" metafórica o tangencial, emula el vagabundeo de la mente humana del cual, ocasionalmente, puede brotar una conexión novedosa y valiosa.

Este marco peirceano ofrece una perspectiva reveladora para analizar la *tensión fundamental* en los Grandes Modelos de Lenguaje entre *veracidad y creatividad*. Proponemos que la "alucinación" en un GML no es simplemente un error, sino un fenómeno que se sitúa en un espectro entre dos polos normativos de Peirce:

1. La Lógica, que busca la veracidad y la coherencia, se corresponde con el esfuerzo por mitigar las alucinaciones para producir un discurso determinista y fiable: la alucinación es una pérdida de la normatividad dada por la lógica porque no tiene control reflexivo. (Reduciendo T y Top–K)

2. La Estética, que busca lo admirable y sorprendente, se corresponde con la creatividad, donde la amplificación de las alucinaciones mediante hiperparámetros como la Temperatura (T) o el Top-K permite generar resultados novedosos, metafóricos y poéticos. La alucinación puede ser un momento de libre juego de las facultades con potencial creativo.

Sintetizando lo anterior, podríamos entender que la experiencia alucinatoria tiene el potencial de desatar la creatividad, pero este proceso necesita de una decisión voluntaria de autocontrol reflexivo por parte del agente que vive dicha experiencia, de manera que pueda encauzarse como una instancia de *musement* creativo.

La configuración de estos hiperparámetros puede ser vista, entonces, como una decisión ética (la ciencia normativa que tiene que mediar, en nuestro caso, el aspecto lógico y el estético): el usuario o supervisor elige a qué fin normativo debe aspirar el GML en su conducta generativa. ¿Debe seguir estrictamente el deber lógico de decir la verdad o debe explorar el ideal estético de lo sorprendente y lo bello? Este artículo explorará cómo esta arquitectónica peirceana ilumina la relación contraria entre las alucinaciones y la creatividad en los GMLs, planteando un nuevo marco para comprender su funcionamiento y potencial (Peirce 1998b, 380).

Veracidad en los Grandes Modelos de Lenguaje: Mitigar las Alucinaciones

Las puntuaciones de confianza o "credencias" de acuerdo a (Keeling y Street 2024) son la medida que tiene la red sobre su propia respuesta —su "autorreflexión" o "duda" sobre su propia respuesta— y ofrecen una indicación del grado en que podrían estar alucinando (Mahaut et al. 2024). Es importante en el caso de la Veracidad, indicar al usuario del GML, qué probabilidad tiene el GML de estar alucinando, al menos, con un indicador numérico al lado de la respuesta del "*prompt*". Este enfoque aún se encuentra en fase de investigación en ingeniería. Por ejemplo: Un valor del 20 % de confianza, indicaría un alto riesgo de alucinación, mientras que un 99 % (alta puntuación de confianza) sugeriría un bajo riesgo. A continuación detallaremos dos métodos principales, y del segundo, una derivación.

Probabilidades por Tokens

En este caso, si bien la implementación no es complicada, el problema son los *tokens*, y viene de la mano de la filosofía del lenguaje. La relación entre el lenguaje y el mundo no es uno–a–uno. Podemos expresar el mismo significado con diferentes frases, a saber, y como depende de los *tokens*, distintos nodos se activarán, activando diferentes credencias. Esto se puede ver con un sencillo ejemplo:

- "Salta es una ciudad que pertenece a Argentina" – 90 % de credencia
- "Salta es una ciudad argentina" – 80 % de credencia

Esto activa nodos distintos en el modelo de lenguaje, lo que provoca una puntuación de confianza muy *inestable*. Cada vez que parafraseamos algo, la credencia cambia significativamente, cuando no debería hacerlo. Estos métodos son económicos, pero inestables.

Semántica Entrópica

Según Farquhar et al. (2024), la pérdida total de significado es el problema principal en la *tokenización*, por lo que se realiza una agrupación ("*clustering"*) de frases completas. Si la frase A implica a la frase B, y la frase B implica a la frase A, se colocan en el mismo grupo y luego se procesan. Este método es superior de acuerdo a los *benchmarks* (métricas que se realizan) y la puntuación de confianza o credencias resulta mucho más *estable*, aunque computacionalmente es más costoso. Y en IA es importante ahorrar en electricidad, por lo cual también se buscan alternativas. Por lo tanto, existe un método similar, "Pruebas de Semántica Entrópica" (SEP, por sus siglas en inglés) desarrollado por Kossen et al. (2024), que consiste en acceder a los nodos internos del sistema LLM, de modo que sólo podría funcionar en sistemas open–source como DeepSeek (DeepSeek-AI et al. 2024). No necesita tomar tantas muestras ni realizar tantas iteraciones de modo que es de bajo costo. Azaria & Mitchell(2023), proponen la hipótesis que los nodos del "medio" son los más puros, ya que los de las entradas están procesando las señales de bajo nivel, mientras que los de la salida tienden a encargarse del cálculo

de la generación del próximo *token*. De modo que Kossen *et. al.* (2024) basan su modelo en esta hipótesis.

Éstos métodos de mitigación de alucinaciones se corresponden con la lógica peirceana de que busca la veracidad y la coherencia, en un esfuerzo por un discurso determinista y fiable.

Creatividad como proceso en los Grandes Modelos de Lenguaje: Amplificar las Alucinaciones

Así como en el ámbito científico y técnico el foco está actualmente puesto en mitigar y controlar las alucinaciones producidas por los modelos de lenguaje de gran escala (GMLs), comienzan a alzarse cada vez más voces en el ámbito académico — especialmente desde la filosofía, la teoría del arte, y los estudios sobre inteligencia artificial— que proponen una reflexión más profunda sobre el posible nexo entre el fenómeno de la alucinación y la creatividad de estos sistemas.

Revisaremos la definición de la creatividad. De acuerdo a Boden (2016, 67): "La creatividad es la capacidad de producir ideas o artefactos que sean *nuevos, sorprendentes y valiosos*." (contextual).

Aquí plantearemos definir la creatividad de forma no relativa al contexto. Lo que se considera valioso en un contexto cultural puede resultar completamente irrelevante o incluso incomprensible en otro. Por ejemplo, aquello que se aprecia como innovador o significativo en China podría no tener la misma resonancia en España; de igual modo, lo que provoca asombro en México podría pasar desapercibido en Japón. Esta variabilidad cultural y contextual pone en evidencia que los criterios de valoración —ya sea de lo creativo, lo nuevo o lo útil —no son universales ni objetivos, sino profundamente situados. Un ejemplo histórico que ilustra esta complejidad es la disputa sobre la invención del cálculo: ¿fue obra de Newton o de Leibniz? Más allá del mérito intelectual de cada uno, la discusión estuvo atravesada por cuestiones de prestigio, nacionalismo, poder institucional y legitimación académica. La historia demuestra que la validación de una idea no depende únicamente de su calidad intrínseca, sino también del entorno en el que surge y de quién la respalda.

Figuras como Nikola Tesla, Charles Sanders Peirce, o Vincent van Gogh son ejemplos paradigmáticos de esta paradoja: murieron en la pobreza y con escaso reconocimiento, a pesar de que sus ideas y obras serían, tiempo después, consideradas revolucionarias. Su falta de reconocimiento en vida evidencia cómo el valor atribuido a una contribución creativa está mediado por estructuras sociales, económicas y simbólicas que operan más allá del contenido mismo.

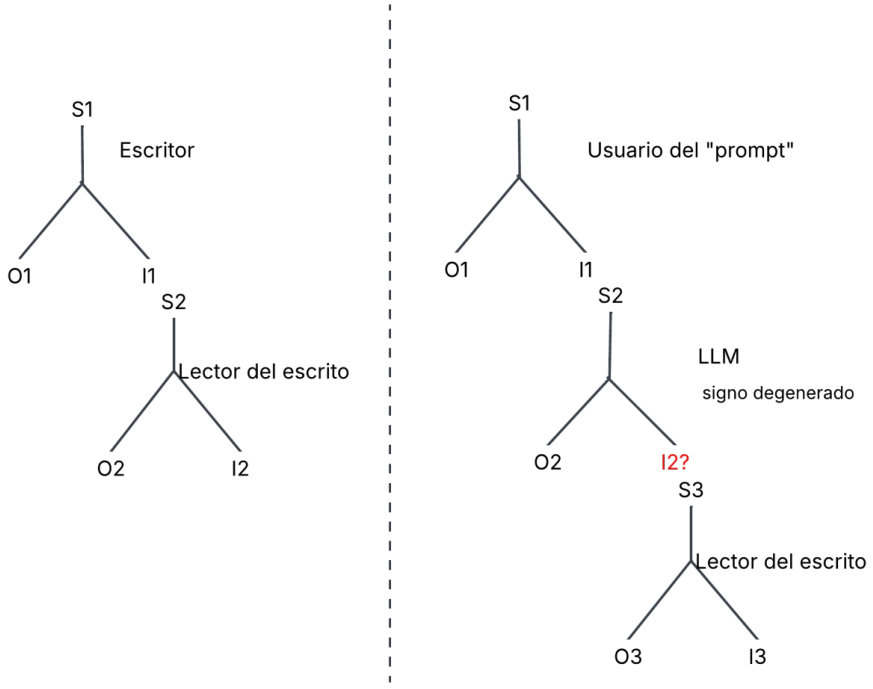
De hecho, muchos escritores, matemáticos, científicos y artistas han compartido este destino, lo que sugiere que el juicio sobre la creatividad no es estático ni automático, sino dinámico y dependiente de múltiples factores. Por ello, basar la evaluación de la creatividad únicamente en criterios contextuales puede resultar problemático: lo que hoy es desechado puede ser mañana celebrado, y lo que en un entorno es revolucionario, en otro puede parecer trivial.

Por lo tanto, optaremos por adoptar definiciones de creatividad que la entienden como un *proceso*, en lugar de aquellas que la conciben únicamente como un resultado evaluado externamente, el cual suele estar fuertemente condicionado por el contexto sociocultural, histórico o institucional en el que se inserta. Esta decisión responde a la problemática previamente señalada: lo que se considera creativo en un entorno puede no ser lo en otro, lo que convierte a la creatividad como resultado en una noción altamente relativa y dependiente de factores externos al agente generador.

En cambio, concebir la creatividad como proceso permite centrar la atención en las *dinámicas internas del agente*, en los mecanismos, estrategias y decisiones que éste despliega en la generación de nuevas ideas, productos o soluciones. Desde esta perspectiva, el foco se traslada del juicio externo a las operaciones cognitivas y conductuales que subyacen al acto creativo en sí mismo.

Seguiremos la definición de Green et al. (2024, 566): "… la definición basada en el proceso ofrece un marco para interpretar los hallazgos de la investigación sobre la creatividad desde perspectivas cognitivas y neurocientíficas (cuyos datos son, por naturaleza, procesuales), para unificar modelos de proceso de la creatividad, para organizar de forma taxonómica los constructos y fenómenos relacionados con la creatividad, y para evaluar la creatividad de procesos, tareas e individuos (distinguiendo los criterios definitorios de los procesos creativos de los de los productos creativos). Además, la definición basada en el proceso permite enriquecer los enfoques actuales de evaluación de productos, al considerar hasta qué punto esos productos son verdaderamente resultado de un proceso creativo. De esta manera, la definición por proceso contribuye a abordar vacíos o confusiones teóricas que han dificultado el avance en la investigación sobre creatividad.

De modo que se considerarán las tríadas semióticas no solo del receptor de la obra de arte/científica/lingüística(GML) sino del *emisor* para ver su funcionamiento interno. A continuación una figura esquemática. Del lado de la izquierda, las tríadas semióticas conectadas tradicionalmente entre el escritor de un libro, por ejemplo, y su lector. Del lado de la derecha, el usuario del "prompt", que envía un mensaje al LLM por sus siglas en inglés o GML, y que puede ser el mismo que el signo 3 (el que envía el concepto es el mismo que el que recibe la respuesta del GML). Pero qué pasa con I2 en el GML? Procesa la información en modo inferencial, recogiendo datos que recabó en fase de entrenamiento de su set de datos correlacional. Al decir de Bender et al. (2021), es un "loro estocástico", y no tiene consciencia. Es como un sistema de capas, que tiene la capa de la sintaxis pero le falta la de la semántica, el signo se degenera y por ello la tríada se "debilita":



Conclusiones

Como conclusiones, en primer lugar vemos que hay una *contraposición entre veracidad y "creatividad"* en los GMLs, aumentando y disminuyendo los hiperparámetros.

En segundo lugar, el "*musement*" es una forma de "Juego Puro" un estado que consiste en permitir que la mente vague libremente por las tres categorías de la experiencia. Este acto *no es un razonamiento dirigido*, sino una apreciación estética de las ideas y sus patrones, un deleite en la contemplación misma. El GML es un signo degenerado y su Interpretante no es real, no posee motivación intrínseca, y depende de los hiperparámetros y del concepto del usuario del"*prompt*", que son una semántica externa. De modo que podemos concluir que emula un "*musement*" artificial, no es juego puro y ocioso.

Bibliografía

Azaria, Amos, y Tom Mitchell. 2023. «The Internal State of an LLM Knows When It's Lying».

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, y Shmuel Shmitchell. 2021. «On the Dangers of Stochastic Parrots». En *Proceedings of the 2021*

ACM Conference on Fairness, Accountability, and Transparency, 610-23. https://doi.org/10.1145/3442188.3445922.

Boden, Margaret. 2016. *AI – It's Nature and Future*. Oxford University Press.

DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, et al. 2024. «DeepSeek-V3 Technical Report».

Farquhar, Sebastian, Jasmijn Kossen, Lukas Kuhn, y Yarin Gal. 2024. «Detecting hallucinations in large language models using semantic entropy». *Nature* 630 (8017): 625-30. https://doi.org/10.1038/s41586-024-07421-0.

Green, Alexander E., Roger E. Beaty, Yoed N. Kenett, y James C. Kaufman. 2024. «The Process Definition of Creativity». *Creativity Research Journal* 36 (3): 544-72. https://doi.org/10.1080/10400419.2023.2254573.

Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, et al. 2023. «A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions». https://doi.org/10.1145/3703155.

Keeling, G., y W. Street. 2024. «On the attribution of confidence to large language models».

Kossen, J., J. Han, M. Razzak, L. Schut, S. Malik, y Y. Gal. 2024. «Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs».

Mahaut, M., L. Aina, P. Czarnowska, M. Hardalov, T. Müller, y L. Màrquez. 2024. «Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators». En. https://doi.org/10.18653/v1/2024.acl-long.250.

Peirce, Charles Sanders. 1992. «The Doctrine of Chances». En *The Essential Peirce: Selected Philosophical Writings, Volume 1 (1867-1893)*, editado por Nathan Houser y Christian Kloesel, 142-54. Indiana University Press.

———. 1998a. «A Neglected Argument for the Reality of God». En *The Essential Peirce: Selected Philosophical Writings, Volume 2 (1893-1913)*, editado por Peirce Edition Project, 434-50. Indiana University Press.

———. 1998b. «The Basis of Pragmaticism in the Normative Sciences». En *The Essential Peirce: Selected Philosophical Writings, Volume 2 (1893-1913)*, editado por Peirce Edition Project, 371-97. Indiana University Press.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, y Illia Polosukhin. 2017. «Attention Is All You Need». En *Neural Information Processing Systems NIPS 2017*, 31:1-11.