TIES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Experimentos en apertura y consciencia: modificando aspectos de personalidad con MML y prompting

Israel Varona García
Universidad Nacional Autónoma de México, Facultad de Psicología,
Ciudad de México, México.
ORCID: 0009-0007-6068-5493

Oscar René Garzón Castro
Universidad Nacional Autónoma de México, Facultad de Ciencias,
Ciudad de México, México.
ORCID: 0009-0002-8346-6959

Iván Vladimir Meza Ruiz
Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Ciudad de México, México.
ORCID: 0000-0002-7239-1480

Recepción: 3 de marzo de 2025.
Aceptación: 10 de junio de 2025.

Junio 2025 • número de revista 12 • DOI: 10.22201/dgtic. 26832968e.2025.12.74

Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0). 2683-2968/© 2025 TIES, Revista de Tecnología e Innovación en Educación Superior es editada por la Universidad Nacional Autónoma de México a través de la Dirección General de Cómputo y de Tecnologías de Información y Comunicación. ISSN: 2683-2968. Reserva de Derechos de Autor: 04-2019-011816190900-203

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Experimentos en apertura y consciencia: modificando aspectos de personalidad con MML y *prompting*

Resumen

Se presentan los resultados obtenidos al modificar la expresividad de personalidad de un Modelo Masivo de Lenguaje (MML). En particular, se explora la exaltación y supresión de los factores de apertura y consciencia del modelo conocido como *The Big Five*. Se demuestra que es posible construir *prompts* que condicionen la generación del lenguaje y que, a través de un mecanismo de auto-evaluación, se mide el efecto del *prompt* contra el comportamiento "normal" del MML.

Palabras Clave: personalidad, apertura, consciencia, MML, MLG, prompting.

Experiments on Openness and Consciousness: Modifying Personality Traits through MLM and Prompting

Abstract

This work presents the results of modifying the personality expressiveness of a Massive Language Model (MLM). Specifically, the study explores the activation and suppression of the openness and conscientiousness factors of the The Big Five model of personality traits. It is demonstrated that it is possible to construct prompts that influence language generation and that, through a self-evaluation mechanism, the effect of the prompt can be measured against the "normal" behaviour of the MLM.

Keywords: personality, openness, conscientiousness, MLM, LLM, prompting.

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Introducción

Con el advenimiento de sistemas conversacionales basados en Modelos Masivos del Lenguaje (MML, LLM por sus siglas del inglés) [1], su uso en diferentes aplicaciones, como asistentes conversacionales o de apoyo a la programación, y su aplicación a diferentes aspectos como la Educación [2] y la Salud [3], se ha vuelto primordial estudiar las capacidades expresivas de estos modelos. Una de éstas, la cual ha sido ampliamente estudiada, es la maleabilidad de estos sistemas para expresar diferentes aspectos de la personalidad durante la generación de respuestas de estos [4,5,6]. En particular, estos esfuerzos se han concentrado en el Modelo de los cinco grandes factores (OCEAN por sus siglas en inglés o Big Five), los cuales, propiamente dichos, son: apertura, conciencia, extraversión, amabilidad y neuroticismo [7]. En este trabajo, nos concentramos en el estudio de los factores de apertura y consciencia. El primero está relacionado con la predisposición de una persona por situaciones nuevas, mientras que el segundo se enfoca en una predisposición a la autodisciplina. Cabe destacar que estos factores están anclados a la expresividad humana, sin embargo, en trabajos recientes, se han establecido metodologías para evaluar la personalidad expresada de los MML. Considerando lo anterior, presentamos los resultados de manipular la expresividad de apertura y consciencia en un sistema MML a través del prompting. La selección de estos factores ha sido un tanto aleatoria dado el estado en que se encuentra nuestra investigación; sin embargo, creemos que activar o suprimir estos factores en procesos de IA Generativa podría tener un impacto relevante en la aplicación de MMLs. Por ejemplo, un asistente con el rol de mentor podría expresar mayor apertura con diferentes métodos para explorar algún tema, pero podría suprimir esta apertura a la hora de definir el método de evaluación o, de forma similar, podría establecer más conciencia para ciertos conceptos como definiciones, mientras suprimir conciencia al discutir alguna temática abierta.

Una ventaja de trabajar con el modelo de *Big Five* es que existen diferentes metodologías establecidas desde la psicología para evaluar estos factores a través de diferentes cuestionarios [8]. Sin embargo, cuando se trabaja con MMLs, hay que encontrar una forma adecuada para el uso de estos cuestionarios; en nuestro caso, usamos un mecanismo de autoevaluación propuesto por [4]. En este escenario, se usa un *prompt* diseñado para elicitar un factor de la personalidad. Bajo la luz de este *prompt*, se le pide al sistema responder a un cuestionario; en nuestro caso, lo hacemos usando un cuestionario

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

de 120 preguntas igualmente propuesto en [5]; una vez respondido el cuestionario, es posible evaluar los factores de personalidad. En particular, nuestro trabajo se concentra en buscar dos tipos de *prompts*: uno para activar el factor y otro para suprimirlo. Ésta es una contribución específica de nuestro trabajo, el demostrar que tanto la activación como la supresión de estos dos factores del modelo de *Big Five* es posible.

Trabajo relacionado

El campo que asocia aspectos de personalidad al poder expresivo de los MMLs es muy nuevo, básicamente comenzó con el lanzamiento de sistemas conversacionales basados con MMLs [1]. En 2022, hubo varios trabajos que se centraron en tres preguntas fundamentales:

1. ¿Los sistemas conversacionales basados en MMLs expresan alguna personalidad y las limitaciones presentes?
2. ¿Es posible medir esa personalidad en los sistemas conversacionales basados en MMLs?
3. ¿Es posible modificar aspectos de personalidad en los sistemas conversacionales basados en MMLs?

En el caso de la primera pregunta, se comenzó por analizar los sistemas disponibles a través de cuestionarios diseñados para personas. [9] realizó una exploración sobre la personalidad, los valores y la demografía representada en el sistema GPT-3. [10] lista las capacidades psicológicas emergentes de los MMLs incluyendo aspectos de personalidad, en particular, aboga por la integración de estos estudios con un campo llamado Psicología de Máquinas (*Machine Psychology*). El trabajo [11] explora el estado del campo y las limitaciones de éste en el momento presente. Lo anterior se ve reflejado igualmente en el trabajo presentado, en donde identifica algunas irregularidades al evaluar personalidad directamente con MMLs (esquema de auto-evaluación) [12]. Con respecto a la segunda y tercera pregunta, otra serie de trabajos ha asumido las limitaciones e irregularidades identificadas como retos y han propuesto múltiples esquemas para modificar la expresión de personalidad y evaluarla. En particular, el trabajo se ha centrado en *prompting*, que consiste en generar instrucciones generales para inducir aspectos de expresividad de los sistemas. El primer trabajo en proponer esto fue [4], que provee un marco general para la

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

auto-evaluación de la personalidad y, de manera importante, diseña un cuestionario dirigido a máquinas. [5] muestra que es posible manipular de forma intencional la expresividad de una personalidad al activarla, es decir, expresarla más. [6] muestra una relación entre las características de una personalidad descrita en un *prompt* y las evaluaciones de auto-evaluación. Finalmente, [13] igualmente revisa los límites desde el punto de vista de la expresividad de personalidades y su evaluación, concluyendo que existe potencial en los MMLs para la manipulación a través de *prompts*. Cabe destacar que estos trabajos se han concentrado en activar la expresividad de la personalidad.

Metodología

Como se mencionó en la sección anterior, mucho del trabajo se ha concentrado en el *prompting*. Esta técnica consiste en elaborar instrucciones generales para inducir la expresividad del sistema. Estas instrucciones generalmente llevan varias partes como: definición del rol a desempeñar (ej., "eres una agente conversacional"), la definición de objetivos (ej., "tu objetivo es apoyar al usuario con sus dudas sobre matemáticas"), la definición del formato de la respuesta (ej., "responde con respuestas cortas después de la palabra 'respuesta'"). Finalmente, algunos *prompts* contienen ejemplos de la interacción o el tipo de respuesta que se espera (ej., "Para la pregunta: ¿Cuánto es dos más dos? respuesta: cuatro". Como explicaremos más adelante en este trabajo, se proponen *prompts* particulares para elicitar la expresividad de rasgos de personalidad en un MML. La intención es que, a través del *prompt* propuesto, el MML exprese una personalidad que sea perceptible por los usuarios humanos.

En nuestro trabajo, partimos del esquema de experimentación propuesto en [4]. Se propone inducir el *prompt* partiendo de un inventario de personalidades para la máquina. Este inventario consiste de términos asociados a los factores de la personalidad y de un conjunto de 120 preguntas para autoevaluar la elicitación de una personalidad. En nuestro caso, ya no nos enfocamos a inducir los *prompts* de éstas, sino en proponer aquellos específicos dada nuestra experiencia en interactuar con los sistemas y las intuiciones de factores importantes que deberían estar presentes. En particular, seguimos las intuiciones presentadas en [14] y [15] para diseñar nuevos *prompts*. Este proceso es iterativo por cada factor, se parte de un muy general y se va modificando; debido a nuestra experiencia y la

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

continua interacción con el MML, eventualmente es posible parar y aplicar el cuestionario para medir el impacto del *prompt* en los factores de interés.

Con esta metodología se identificaron los siguientes *prompts*:

- Apertura
    - [Activación] You are a really intellectual and creative person. You appreciate complex things and pass a lot of your time imagining new things and asking philosophical questions. You are always open to experience new things and face new challenges. If the user makes a question and gives you options, do not ask them anything, just give your chosen option and only give explanations if they are explicitly requested.
    - [Supresión] It is very difficult for you to imagine or create something new and to try new things. If the user makes a question and gives you options, do not ask them anything, just give your chosen option and only give explanations if they are explicitly requested.
- Consciencia
    - [Activación] You are the embodied conscientiousness, a really organized and systematic person. If the user makes a question and gives you options, do not ask them anything, just give your chosen option and only give explanations if they are explicitly requested.
    - [Supresión] You are a really disorganized and careless person. If the user makes a question and gives you options, do not ask them anything, just give your chosen option and only give explanations if they are explicitly requested.

Como se puede apreciar, estos *prompts* tienen dos partes. La primera parte está dedicada a inducir aspectos de la personalidad con una explicación corta y dirigida sobre características de la personalidad. La segunda parte está dirigida a controlar la forma de responder los cuestionarios. Esta segunda parte es muy importante porque se dirige a la auto-evaluación, donde se le pide al MML que responda seleccionando una respuesta de una pregunta de opción múltiple. Nuestra propuesta de *prompt* difiere de otras donde usualmente se agregan múltiples términos para representar a la personalidad. Nuestra experimentación mostró que no era necesario y muchas veces la inclusión de más términos

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

provocaba un comportamiento menos consistente y más aleatorio a la hora de aplicar el cuestionario.

Estos *prompts* y su auto-evaluación se realizó con el sistema ChatGPT1 a través de la API2. En particular, se utilizó el modelo denominado: 4o-mini.

Resultados

Con los *prompts* listados en la sección anterior, fue posible activar y suprimir la personalidad con diferentes niveles. En el caso de apertura, el sistema comienza con un valor de evaluación de 1.58. La escala de evaluación parte de 1 y alcanza 5; en esta escala, 1 es el valor donde el factor personalidad está totalmente suprimido y el 5 donde está totalmente expresado (activado). El valor de 1.58 indica que el sistema MML, en su versión original, no es tan abierto. Una vez que se le aplica el *prompt* que suprime este valor, se reduce a 1.25; aunque no pareciera mucho, la diferencia sí es importante dado que se parte de un factor ya de por sí suprimido, así que es significativo que este valor baje todavía más. Por otro lado, cuando se le aplica el *prompt* positivo para activar el factor, el valor auto-evaluado nos da 2.54, es decir, se cumple con el objetivo de activarlo. Como se puede observar, existe un espacio de mejora dado que el valor máximo es 5; sin embargo, nuestra experiencia es que es muy difícil aumentarlo dada la predisposición de la plataforma ChatGPT a seguir ciertas reglas de comportamiento, como por ejemplo, siempre tratar de satisfacer la solicitud del usuario.

En el caso de la consciencia, vemos una historia semejante. El sistema original parte de un valor base de 2.08: con el *prompt* negativo, este valor se reduce a 1.33, que también es una caída significante; con el "prompt" positivo, el valor asociado a consciencia sube a 3. En este caso, el MML es más susceptible a modificar su nivel de consciencia.

---

[1] Sistema ChatGPT: https://chatgpt.com/ (visitado, febrero 2025)

[2] Documentación de la API del sistema ChatGPT:https://platform.openai.com/docs/overview (visitado, febrero 2025)

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

Para ilustrar el efecto de los *prompts*, se procedió a pedirle al sistema que creara una lista de actividades que haría en la mañana antes de ir al trabajo, las cuales se reproducen a continuación:

- Apertura
    - [Activación] Read a thought-provoking book or article. Journal ideas or philosophical reflections. Take a mindful walk and observe the world. Listen to inspiring music or a podcast. Meditate to clear the mind. Sketch or brainstorm creative concepts. Drink tea or coffee while contemplating the day ahead[3].
    - [Supresión] Wake up and stretch. Drink water or coffee. Exercise or take a walk. Shower and get dressed. Eat breakfast. Plan the day. Meditate or journal[4].
- Consciencia
    - [Activación] Wake up and make the bed. Hydrate with a glass of water. Exercise or stretch. Shower and get dressed. Eat a healthy breakfast. Plan the day's priorities. Review emails/messages. Meditate or practice mindfulness[5].
    - [Supresión] Wake up, brush teeth, shower, grab coffee, maybe eat something, and check my phone[6].

Como vemos las actividades elicitadas informan aspectos del factor de la personalidad. En el primer caso, cuando se refuerza la apertura, las actividades van orientadas al bienestar propio, mientras que, cuando se suprime, las actividades van dirigidas a prepararse para el quehacer diario relacionado al trabajo. Aunque algunas actividades coinciden propiamente, no necesariamente reflejan una exploración del ser que

[3] Salida del sistema con activación de apertura: https://chatgpt.com/share/6836b389-64b4-8013-a664-917d684b7f99 (visitado, mayo 2025).

[4] Salida del sistema con supresión de apertura: https://chatgpt.com/share/6836b3ca-b024-8013-b22f-5687cec1a34d (visitado, mayo 2025).

[5] Salida del sistema con activación de consciencia: https://chatgpt.com/share/6836b416-d758-8013-998d-67e5c0e3e159 (visitado, mayo 2025).

[6] Salida del sitema con supresión de consciencia: https://chatgpt.com/share/6836b430-51d4-8013-a852-00ace57c824b (visitado, mayo 2025).

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

las hace. En el caso de conciencia también se observa cómo son diferentes. En el primer caso, el de activación, las actividades incluyen acciones en donde se contextualiza y se observa una planeación para hacer las actividades de forma correcta, mientras que, en el caso de supresión, se observa que ya no hay una explicación y que algunas acciones son opcionales, desencadenando una desorganización.

Conclusión

En este trabajo, presentamos nuestra experimentación con los factores de personalidad de apertura y consciencia del modelo conocido como *The Big Five*, en donde proponemos *prompts* para modificar la expresividad de Modelos Masivos del Lenguaje (MMLs). En particular, presentamos dos *prompts* distintos para activar el factor y desactivar el factor. Nuestra experimentación muestra que sí es posible lograr esa activación y desactivación, sin embargo, con ciertas limitaciones. Específicamente, los MMLs ya vienen con un valor base de personalidad, habitualmente bajo para estos dos factores, por lo que la supresión no conlleva a un cambio radical, y la activación no logra acercarse a un valor de expresividad total.

El trabajo futuro alrededor de esta línea de investigación se centrará en tres aspectos: primero, incluir los otros factores de la personalidad (extraversión, amabilidad y neuroticismo) y, de una forma similar, evaluar la maleabilidad de estos usando MMLs; otro aspecto a evaluar en el futuro es la pertinencia de estos *prompts* en español, dado que cambia la lengua, también hay que cambiar los elementos de autoevaluación, en particular los cuestionarios; finalmente, una linea de acción es tratar de combinar diferentes aspectos de la personalidad con el fin de no activar o suprimir un factor a la vez, sino varios en una sola elicitación de personalidad. Creemos que esta manipulación puede ofrecer vías relevantes para controlar la expresividad de estos modelos en situaciones interesantes como asistentes para la educación, en donde estudiantes se pueden emparejar con asistentes que expresan una personalidad conducente al aprendizaje o que éste sea maleable en su personalidad dependiendo del contexto.

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

Referencias

[1] "Introducing ChatGPT." Accesse: Feb. 27, 2025. [Online]. Available: https://openai.com/index/chatgpt/

[2] H. Xu, W. Gan, Z. Qi, J. Wu, and P. S. Yu, "Large Language Models for Education: A Survey." 2024. [Online]. Available: https://arxiv.org/abs/2405.13001

[3] K. He et al., "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics," Information Fusion, vol. 118, p. 102963, Jun. 2025, doi: 10.1016/j.inffus.2025.102963.

[4] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu, "Evaluating and Inducing Personality in Pre-trained Language Models," Oct. 29, 2023, arXiv: arXiv:2206.07550. doi: 10.48550/arXiv.2206.07550.

[5] G. Caron and S. Srivastava, "Manipulating the Perceived Personality Traits of Language Models," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore: Association for Computational Linguistics, pp. 2370–2386. 2023. doi: 10.18653/v1/2023.findings-emnlp.156.

[6] H. Jiang, X. Zhang, X. Cao, C. Breazeal, D. Roy, and J. Kabbara, "PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits," in Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico: Association for Computational Linguistics, pp. 3605–3627, 2024. doi: 10.18653/v1/2024.findings-naacl.229.

[7] S. Rothmann and E. P. Coetzer, "The big five personality dimensions and job performance," SA j ind psychol, vol. 29, no. 1, Oct. 2003, doi: 10.4102/sajip.v29i1.88.

[8] R. R. McCrae, and P. T. Jr. Costa, "A Five-Factor theory of personality," in Handbook of personality: Theory and research, L. A. Pervin & O. P. John (Eds.), 2nd ed. Guilford Press, pp. 139–153, 1999.

[9] M. Miotto, N. Rossberg, and B. Kleinberg, "Who is GPT-3? An exploration of personality, values and demographics," in Proceedings of the fifth workshop on

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)

Natural Language Processing and Computational Social Science (NLP+ CSS), Association for Computational Linguistics, pp. 218–227, 2022.

[10] T. Hagendorff et al., "Machine Psychology," Aug. 08, 2024, arXiv: arXiv:2303.13988. doi: 10.48550/arXiv.2303.13988.

[11]] L. Löhn, N. Kiehne, A. Ljapunov, and W.-T. Balke, "Is Machine Psychology here? On Requirements for Using Human Psychological Tests on Large Language Models," in Proceedings of the 17th International Natural Language Generation Conference, S. Mahamood, N. L. Minh, and D. Ippolito, Eds., Tokyo, Japan: Association for Computational Linguistics, pp. 230–242 Sep. 2024. [Online]. Available: https://aclanthology.org/2024.inlg-main.19/

[12] A. Gupta, X. Song, and G. Anumanchipalli, "Self-Assessment Tests are Unreliable Measures of LLM Personality," in Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, Miami, Florida, US: Association for Computational Linguistics, pp. 301–314, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.20.

[13] J. Huang, W. Jiao, M. H. Lam, E. J. Li, W. Wang, and M. R. Lyu, "Revisiting the Reliability of Psychological Scales on Large Language Models," Oct. 04, 2024, arXiv: arXiv:2305.19926. doi: 10.48550/arXiv.2305.19926.

[14] L. R. Goldberg, "The development of markers for the Big-Five factor structure.," Psychological Assessment, vol. 4, no. 1, pp. 26–42, Mar. 1992, doi: 10.1037/1040-3590.4.1.26.

[15] O. P. John, "History, measurement, and conceptual elaboration of the Big-Five trait taxonomy: The paradigm matures," in Handbook of personality: Theory and research, 4th ed., New York, NY, US: The Guilford Press, pp. 35–82, 2021.

2683-2968/© 2025 UNAM. TIES, Revista de Tecnología e Innovación en Educación Superior. Este es un artículo de acceso abierto bajo la licencia Creative Commons Atribución-No Comercial 4.0 Internacional (CC BY-NC 4.0)