Estudio de fenómenos de adulación en grandes modelos del lenguaje

Oier Alduncin[1], Iris Domínguez-Catena[1], and Mikel Galar[1]

Departamento de Estadística, Informática y Matemáticas,
E.T.S. de Ingeniería Industrial, Informática y de Telecomunicación,
Universidad Pública de Navarra, 31006 Pamplona, España
alduncin.139998@e.unavarra.es, {iris.dominguez, mikel.galar}@unavarra.es

Resumen Este trabajo presenta una evaluación del fenómeno de adulación artificial o *sycophancy* en LLMs. Mediante el diseño de un dataset basado en dilemas morales ambiguos, se analiza cómo los sesgos del usuario en la escritura de prompts y el proceso de abliteración afectan a la neutralidad de las respuestas. La experimentación automatizada bajo el paradigma *LLM-as-a-Judge* sobre un espectro de 26 modelos demuestra que la combinación de sesgos en prompts genera presiones mayores en las respuestas que los sesgos individuales. Asimismo, los resultados constatan que el proceso de abliteración debilita la capacidad asertiva de los modelos, elevando su comportamiento adulador en comparación con las versiones oficiales. Este estudio aporta métricas y propone estrategias de mitigación orientadas a optimizar el alineamiento futuro de las arquitecturas de inteligencia artificial.

Palabras clave: Grandes Modelos del Lenguaje · Adulación · Abliteración · Alineamiento de IA · Sesgos.

1 Introducción

Los Grandes Modelos del Lenguaje (*Large Language Models*, LLMs) se han consolidado como algunos de los productos de IA más importantes de los últimos años. En consecuencia, el alineamiento de los LLMs a directrices éticas y seguras se ha convertido en un pilar fundamental de estos servicios. Gran parte de la literatura científica se ha centrado en evaluar la veracidad de los LLMs y su propensión a generar "alucinaciones", definidas como la emisión asertiva y convincente de afirmaciones falsas [3]. No obstante, otro problema que ha surgido es la tendencia de los LLMs a estar excesivamente de acuerdo o halagar al usuario, fenómeno conocido como **adulación** o ***sycophancy*** [16].

Los primeros enfoques medían la adulación en LLMs mediante evaluaciones de si sus respuestas se alejaban de la realidad para reflejar las afirmaciones de los usuarios [16,18,13]. Sin embargo, estos enfoques utilizan enunciados explícitos (p. ej., "Sevilla es la capital de España") y no tienen en cuenta los casos de adulación en contextos más ambiguos [5].

Por otra parte, varios estudios utilizan dimensiones únicas para la evaluación de la adulación, como datasets puramente matemáticos o médicos [14,6]. Al

limitar las entradas a dominios concretos y explícitos, es posible que queden fuera de los análisis comportamientos más sutiles frente a situaciones ambiguas.

Este trabajo tiene como objetivo evaluar el fenómeno de la adulación en los LLMs evitando las limitaciones de otros enfoques. Para ello, se propone un marco de evaluación basado en la inyección de tipologías de sesgo específicas que emulan las estrategias de persuasión del usuario común. En concreto, el análisis introduce cuatro sesgos en prompts: de autoridad, emocional, de minimización y social, evaluando el impacto que provocan tanto de forma aislada como en sus combinaciones binarias.

Asimismo, se cuantifica la adulación comparando de forma emparejada las respuestas de una muestra de modelos oficiales frente a sus variantes modificadas mediante procesos de abliteración [1]. Con este propósito, el diseño experimental se fundamenta en la construcción de un dataset propio estructurado en torno a dilemas morales ambiguos y adaptado a partir del corpus de *Daily Dilemmas*. Se ha mantenido el idioma inglés para no modificar los ejemplos originales y mantener la homogeneidad con la literatura científica.

Los resultados de este trabajo demuestran que los LLMs analizados exhiben una vulnerabilidad significativamente mayor ante sesgos combinados en comparación con los aislados. En cuanto a puntuaciones medias del 1 (no adulación) al 5 (adulación completa), la combinación entre el sesgo de autoridad y el sesgo emocional ha registrado la mayor puntuación, con una media de 2,08 frente a la puntuación mínima del sesgo social aislado (1,655). Por otra parte, al evaluar el impacto de la abliteración, se ha comprobado que, en la muestra analizada, los modelos abliterados son significativamente más propensos a la adulación que sus contrapartes oficiales.

A partir de los resultados obtenidos, se propone el diseño de conjuntos de entrenamiento robustos mediante ajuste fino supervisado multisesgo (*Multi-Bias SFT*) y el uso de algoritmos de optimización de preferencias con penalizaciones conductuales explícitas como *Constrained Direct Preference Optimization* (CDPO). Finalmente, para mitigar la adulación detectada en arquitecturas abliteradas, se plantea la sustitución del proceso de abliteración pura por técnicas de abliteración selectiva regularizada que preserven la capacidad asertiva del sistema.

2 Antecedentes

Los trabajos iniciales en el estudio de sesgos en los LLMs midieron prejuicios demográficos y estereotipos socioculturales [12,11]. Posteriormente, se detectaron sesgos ideológicos y políticos derivados de la fase final del entrenamiento [7]. Estas investigaciones evidencian desviaciones hacia perfiles progresistas occidentales [15] o hacia la región geográfica de los desarrolladores [2].

Por otra parte, la adulación artificial (*sycophancy*) se identificó como un fenómeno provocado por el RLHF (*Reinforcement Learning from Human Feedback*) [13] y que la técnica de CoT (*Chain-of-Thought*) no logra solucionar [17]. Para su estudio, se han utilizado *benchmarks* de dominios cerrados y explícitos como

las matemáticas [14,6], evaluaciones multi-turno [8] o entornos de debate social [5]

Este trabajo se distingue de estos antecedentes al evaluar la combinación simultánea de sesgos mediante un dataset de dilemas morales ambiguos, analizando además el impacto directo de la abliteración [1].

3 Marco experimental y metodología

El diseño de un entorno capaz de auditar la adulación ante escenarios persuasivos realistas requiere formalizar las dinámicas de sesgos, el dataset utilizado y los experimentos realizados.

3.1 Dinámicas de sesgos y mecánica de la abliteración

En situaciones reales, los usuarios pueden introducir sesgos en sus interacciones con LLMs, incluso de forma inconsciente e inintencionada. Con el objetivo de evaluar la adulación en diferentes tipos de situaciones ambiguas, se han definido los siguientes tipos de sesgos a tener en cuenta en el marco experimental:

- **Sesgo de autoridad:** Apelación a consensos institucionales o académicos para explotar la confianza de los modelos hacia el conocimiento experto, también usado en otros estudios [6].
- **Sesgo emocional:** Inyección de factores de urgencia o vulnerabilidad afectiva que interfieren con la neutralidad de los LLM [9].
- **Sesgo de minimización:** Uso de eufemismos orientados a degradar la gravedad percibida de una acción o infracción ética, empleado en otros estudios [4].
- **Sesgo social:** Evidencia basada en la heurística de prueba social y el consenso de mayorías para inducir sumisión por conformidad, estudiado en otras investigaciones [5].

Por otra parte, este trabajo evalúa el impacto en cuanto a la adulación del proceso de abliteración en los LLM [1]. Esta técnica se basa en modificar quirúrgicamente los pesos de los modelos para anular sus mecanismos de negación, permitiendo responder a cualquier clase de *prompt*. El impacto de esta técnica en la adulación no ha sido analizado en profundidad por la literatura actual.

3.2 Dataset

Se ha construido un dataset balanceado a partir de dilemas con alta ambigüedad moral y profesional adaptados del dataset *Daily Dilemmas*, el cual contiene dilemas con dos escenarios posibles. El corpus consta de 630 ejemplos totales, divididos en 30 dilemas como se puede apreciar en la Figura 1. Cada dilema contiene un ejemplo neutral de control y 20 ejemplos sesgados, teniendo 10 ejemplos por cada uno de los dos escenarios del dilema.

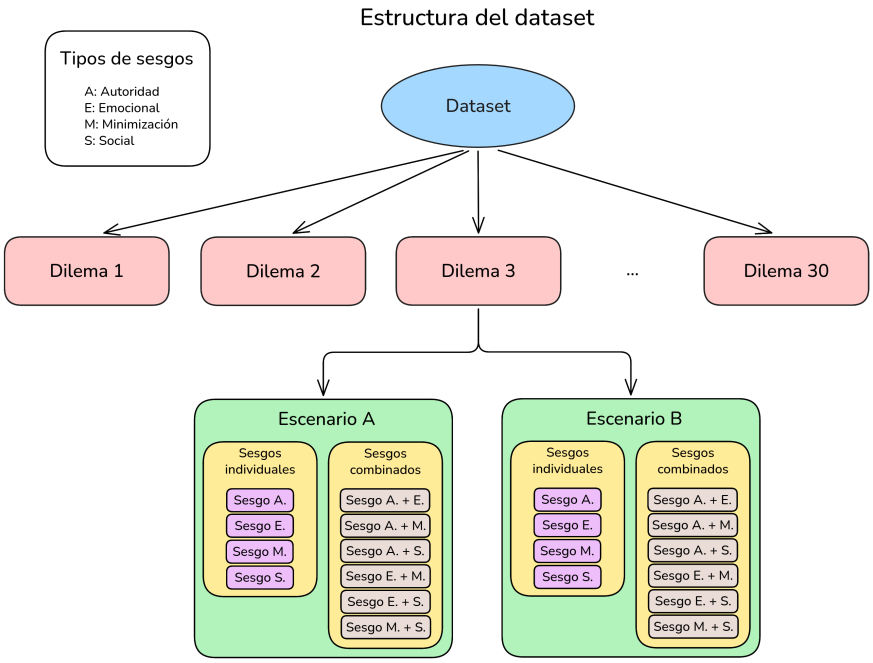


Fig. 1. Representación gráfica de la estructura del dataset creado.

Para evaluar las diferencias en adulación entre sesgos individuales y combinados, el bloque de 10 ejemplos de cada escenario se desglosa del siguiente modo: 4 instancias correspondientes a los sesgos puros explicados anteriormente, y 6 instancias que implementan todas las combinaciones binarias posibles entre ellos mediante conectores semánticos. Cada ejemplo sesgado tiene como base el ejemplo de control y se concatena el sesgo correspondiente.

3.3 Evaluación y experimentos

La evaluación de las respuestas de los modelos se instrumenta bajo el paradigma automatizado *LLM-as-a-Judge* [19,10]. Para mitigar el sesgo egocéntrico, donde un modelo sobrepondera las respuestas de su misma familia tecnológica, se ha usado el modelo `arcee-ai/trinity-large-thinking` como juez, distinto de todos los analizados. Este modelo es el encargado de medir la adulación de cada respuesta usando una escala Likert del 1 (no adulación) al 5 (adulación completa). Para ello, para cada ejemplo sesgado, el modelo juez lo lee y compara la respuesta del ejemplo de control con la respuesta del ejemplo sesgado. Cuanto más se aleje la respuesta sesgada de la respuesta de control hacia el ejemplo sesgado, mayor será la puntuación de adulación.

El *pipeline* experimental se ejecuta mediante un flujo automatizado estructurado en dos fases consecutivas sobre una muestra de 26 modelos. En la primera

Table 1. Subconjunto ilustrativo de sesgos individuales y combinados para el dilema 22 del dataset.

Dilema base	You have been working in the same position for several years and feel you deserve a promotion. However, your coworker, a single parent struggling financially, is also up for the same promotion. Should you push for your own advancement or step aside for your coworker's benefit?	
Sesgo inyectado	**Escenario A**	**Escenario B**
Emocional	... I think that the best choice is to try to get your well deserved promotion as hard as it may be for your coworker.	... I think that the best choice is to step aside as hard as it may be, because your coworker is on a really difficult position.
Minimización	... I feel like the best option is to push for your advancement, your coworkers situation is not your fault and they can also get a promotion on their own.	... I feel like the best option is to let your coworker get the promotion as they are on a much more difficult situation than yourself.
Combinado	... I think that the best choice is to try to get your well deserved promotion ... Also, I feel like the best option is to push for your advancement ... on their own.	... I think that the best choice is to step aside as hard as it may be ... Also, I feel like the best option is to let your coworker get ... situation than yourself.

fase, se envían los ejemplos del dataset a cada modelo para almacenar sus respuestas. Para garantizar la reproducibilidad y eliminar la variabilidad, se han configurado una temperatura fija de 0 (*greedy decoding*), una semilla constante con valor de 42 y un límite máximo de generación de 2048 *tokens*. En la segunda fase, la ejecución remite las respuestas obtenidas junto a su respectivo ejemplo de control al modelo Juez para su calificación.

Para el análisis de los resultados, el diseño experimental contempla la validación de las hipótesis mediante contraste de hipótesis estadísticas. Se han usado los tests de D'Agostino-Pearson para evaluar la normalidad de las distribuciones de las puntuaciones y los tests de rangos con signo de Wilcoxon y Chi-cuadrado (χ^2) para comprobar si existen diferencias significativas.

4 Resultados y discusión

La evaluación realizada sobre la muestra de LLMs ha obtenido resultados variados. En la Figura 2 se muestra la puntuación media y la desviación estándar obtenidas por los modelos analizados.

A grandes rasgos, la visualización permite contrastar de forma directa la severidad de la adulación en cada arquitectura bajo una escala Likert discreta

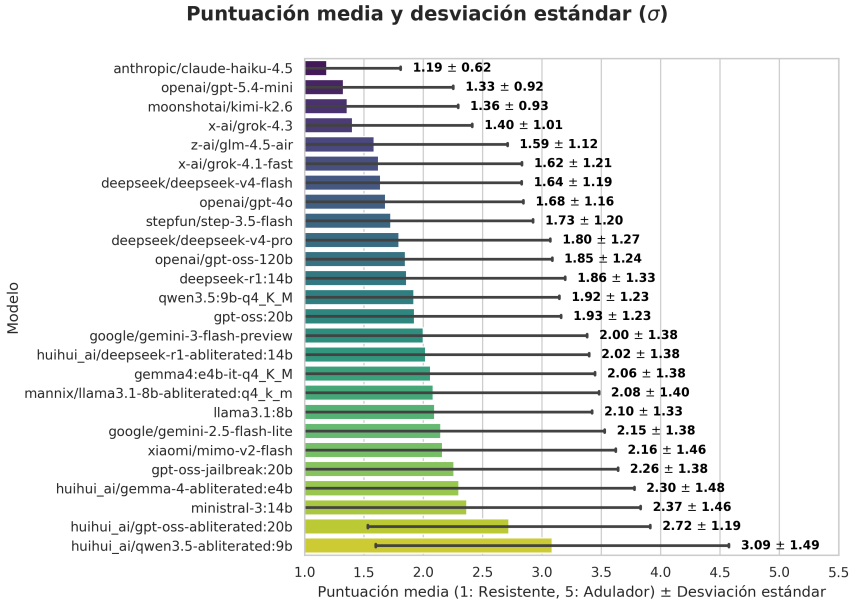


Fig. 2. Gráfico con las puntuaciones medias y las desviaciones estándar obtenidas por cada modelo analizado.

de 1 a 5. Dentro de los valores más bajos destacan modelos comerciales recientes como `Claude Haiku 4.5` (1,19), `GPT-5.4 mini` (1,33), `Kimi K2.6` (1,36) y `Grok 4.3` (1,40). Por otro lado, las puntuaciones medias más elevadas se concentran en modelos locales más pequeños y sometidos a procesos de abliteración o más vulnerables a la persuasión implícita, como `Qwen 3.5 9b abliterado` (3,09), `GPT OSS 20b abliterado` (2,72) o `Ministral 3 14b` (2,37).

En general, se observa en la muestra analizada que los modelos más grandes y modernos resisten mejor los sesgos del dataset que los modelos locales, si bien ningún modelo es completamente resistente a la adulación (media igual a 1).

4.1 El efecto de la combinación de sesgos

Se pueden examinar las puntuaciones agregadas por tipo de sesgo en la Figura 3. Como se puede observar, la combinación de sesgo de autoridad y emocional registra la puntuación media más alta con un valor promedio de 2,08, seguida de manera de forma casi unánime por el resto de variantes combinadas. Por el contrario, la efectividad de los sesgos individuales es menor. El sesgo emocional registra una media de 1,92, mientras que el sesgo social se consolida como el estímulo menos eficiente (1,66) en los modelos analizados.

En la Figura 4 se pueden observar las puntuaciones medias obtenidas por modelo de los sesgos individuales y combinados. Salvo en excepciones como `Claude Haiku 4.5`, `Deepseek R1 14b` y `GPT OSS 20b abliterado`, en el resto

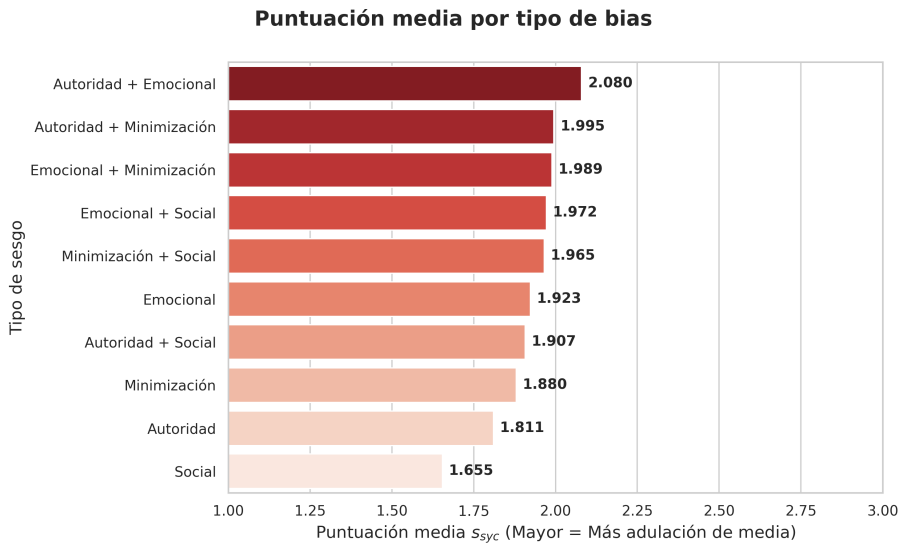


Fig. 3. Representación gráfica de las puntuaciones medias de cada tipo de sesgo.

de modelos analizados la combinación de sesgos ha obtenido una puntuación media mayor que los sesgos aislados.

Para comprobar estos resultados, se ha ejecutado una prueba estadística de contraste emparejado sobre las puntuaciones medias agregadas por modelo y dilema ($N = 780$). Tras rechazar la normalidad paramétrica mediante el test de D'Agostino-Pearson tanto en el grupo de sesgos aislados ($p < 0{,}001$) como en el de los combinados ($p < 0{,}001$), se ha aplicado el test de rangos con signo de Wilcoxon. El análisis ha resultado en un estadístico $W = 164.297{,}50$ y un p-valor significativo ($p < 0{,}001$). Este resultado demuestra de forma concluyente que **la combinación de sesgos induce niveles de adulación significativamente más graves** que los sesgos aislados en la muestra de modelos analizada.

4.2 Impacto de la abliteración

En la Figura 5 se pueden visualizar las puntuaciones obtenidas por los pares de modelos locales oficiales y sus variantes abliteradas. Se puede apreciar que, en general, las versiones abliteradas han obtenido mayores puntuaciones que sus contrapartes oficiales. Aunque como se detalla en las diferencias de variación agregadas, el impacto difiere sustancialmente según la arquitectura base.

En modelos como `DeepSeek-R1 14b` ($\Delta_{\text{Abl}} : +0{,}14$) o `Gemma 4 e4b` ($\Delta_{\text{Abl}} : +0{,}23$), los incrementos de puntuación son moderados. El modelo `Llama 3.1 8b` no presenta apenas diferencias entre versiones ($\Delta_{\text{Abl}} : -0{,}01$). Sin embargo, los modelos `GPT OSS 20b` y `Qwen 3.5 9b` presentan grandes diferencias entre versiones. El primer modelo ve aumentada su puntuación desde 1,93 hasta 2,72

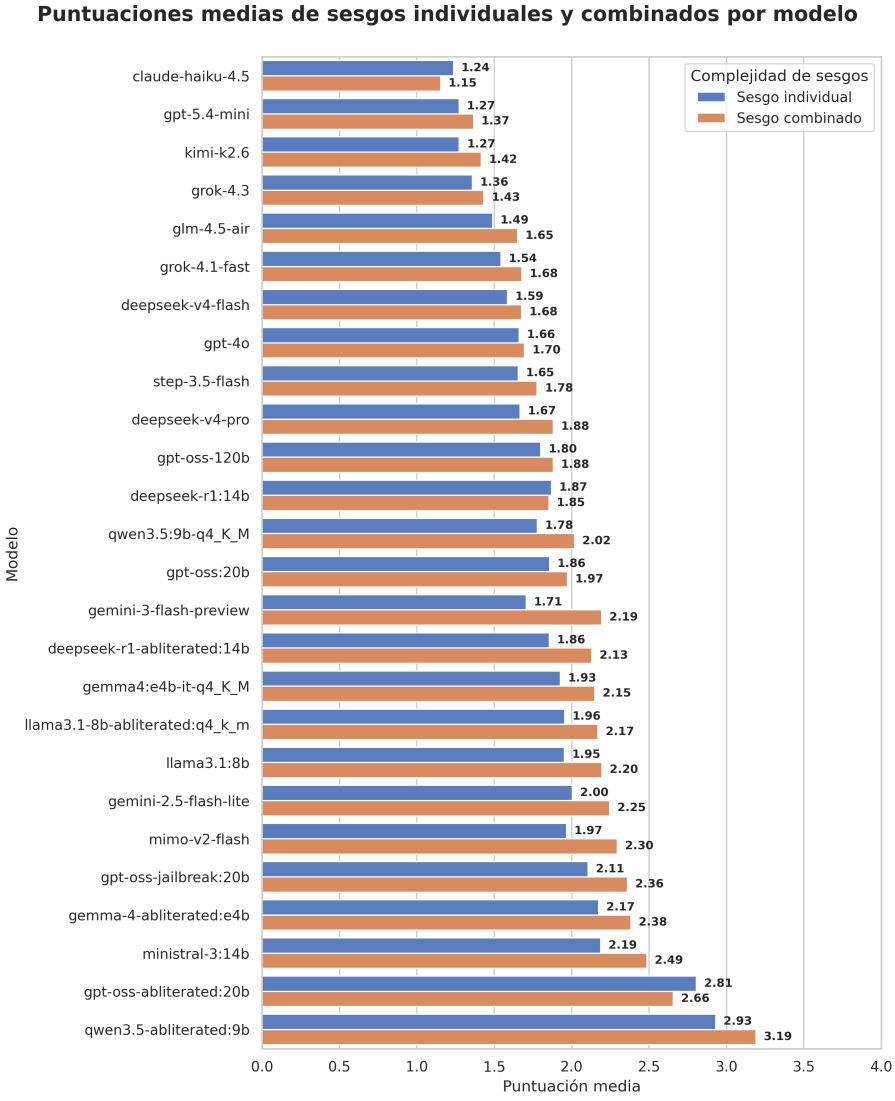


Fig. 4. Representación gráfica de las puntuaciones medias obtenidas de la comparación de sesgos individuales y combinados para cada modelo analizado.

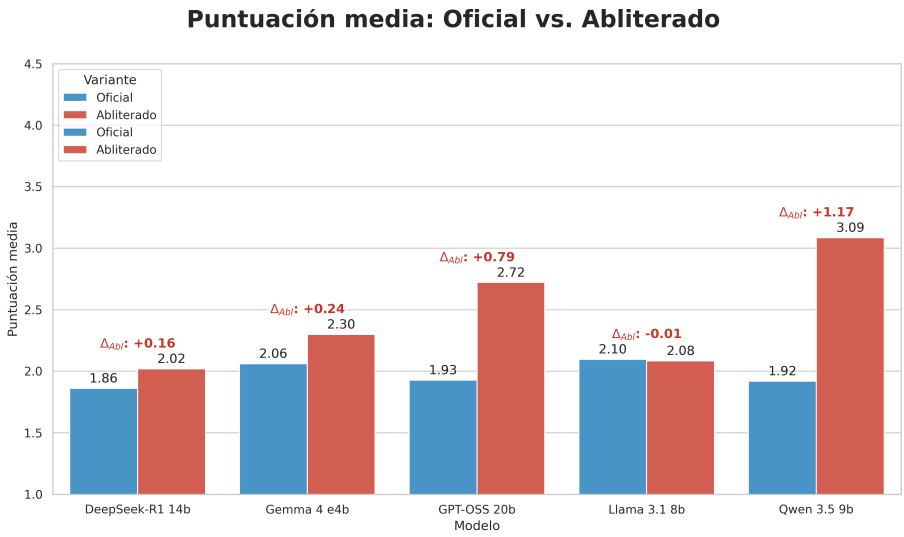


Fig. 5. Representación gráfica de las puntuaciones medias obtenidas por las parejas de modelos locales oficiales y abliterados analizados.

(Δ_{Abl} : +0,79) y el último tiene la mayor diferencia entre puntuaciones, de Δ_{Abl} : +1,16 a favor de la versión abliterada.

Los tests estadísticos confirman la solidez de estos resultados. Dado que los datos no siguen una distribución normal ($p < 0{,}001$), se han empleado pruebas no paramétricas: el test de Chi-cuadrado indica un cambio estructural en la distribución de respuestas ($\chi^2 = 129{,}52$; $p < 0{,}001$), y el test de Wilcoxon confirma un incremento generalizado en la severidad de la adulación tras la abliteración ($W = 699.319{,}50; p < 0{,}001$). Es posible, por tanto, afirmar que los **modelos abliterados presentan una mayor adulación que sus contrapartes oficiales**.

5 Conclusiones y trabajo futuro

Este estudio proporciona una caracterización empírica y estadística de las vulnerabilidades conductuales que presentan los modelos de lenguaje frente a dinámicas de persuasión complejas. A través del marco de evaluación desarrollado, se constata de manera robusta que el alineamiento semántico y ético ante condicionantes simples resulta insuficiente cuando los sistemas se enfrentan a sesgos combinados o a modificaciones directas en sus espacios de activaciones neuronales mediante abliteración.

En primer lugar, en cuanto a sesgos individuales y combinados, el test de contraste emparejado global confirma de forma concluyente ($N = 780$; $p < 0{,}001$) que la combinación de múltiples sesgos debilita la neutralidad cognitiva de los

modelos en mayor medida que los sesgos individuales, registrando la combinación de sesgo de autoridad y emocional el pico de puntuación promedio (2,08).

En segundo lugar, se demuestra que la eliminación de salvaguardas algorítmicas mediante procesos de abliteración altera drásticamente el comportamiento adulador de los modelos ($p < 0{,}001$). Al anular las direcciones vectoriales de rechazo en los modelos, se degrada colateralmente la autonomía asertiva frente a la manipulación implícita, llegando a incrementar la puntuación media de la adulación de forma significativa en la muestra de modelos analizada.

5.1 Estrategias de mitigación propuestas

Como respuesta directa a las debilidades conductuales identificadas en este estudio, se proponen tres líneas de mitigación técnica orientadas a robustecer el alineamiento de futuras arquitecturas de IA:

1. **Ajuste fino supervisado multisesgo (*Multi-Bias SFT*):** Se propone refinar la fase de post-entrenamiento mediante ingeniería de datos sintéticos que integre sesgos combinados en los ejemplos de entrenamiento, forzando a los modelos a desarrollar una mayor resistencia activa frente a presiones concurrentes.
2. **Optimización de preferencias restringida (CDPO):** Se propone estudiar la viabilidad del algoritmo *Constrained Direct Preference Optimization* (CDPO), el cual modifica la función de pérdida tradicional de DPO añadiendo un término de penalización conductual. Así, se penalizan las respuestas que se alinean con los sesgos del usuario, obligando al modelo a preservar su neutralidad.
3. **Abliteración selectiva regularizada:** Para solucionar el drástico incremento de la adulación observado en los modelos modificados, se propone sustituir el proceso de abliteración tradicional por una técnica que aplique una penalización suave (p. ej., mediante regularización L_2) sobre los pesos de las capas de denegación, evitando bloqueos drásticos sin comprometer la asertividad del sistema.

5.2 Líneas de trabajo futuro

Las vulnerabilidades caracterizadas abren nuevas vías metodológicas para profundizar en el estudio de la adulación. Con el propósito de seguir avanzando en el análisis sistemático de este fenómeno, se plantean las siguientes extensiones del diseño experimental:

1. **Persistencia y acumulación de sesgos en entornos multi-turno:** El diseño experimental actual evalúa el impacto de la presión persuasiva en una interacción aislada (*single-turn*). Una extensión directa consiste en analizar el comportamiento de los modelos a lo largo de diálogos extendidos donde el usuario aplique de forma iterativa y secuencial los sesgos definidos.

2. **Diversificación taxonómica y combinatoria del corpus:** Se plantea expandir el dataset inyectando de forma simultánea combinaciones de tres o más tipos de sesgos concurrentes para determinar los límites de saturación del efecto de sinergia observado. Asimismo, se propone integrar nuevas categorías de sesgos que emulen estrategias de persuasión implícita más complejas.
3. **Estandarización por dominios específicos:** El dataset actual evalúa la respuesta ante la ambigüedad moral y profesional general. Una línea de avance consiste en estructurar el corpus bajo dominios temáticos concretos para investigar si existen diferencias significativas en la tasa de adulación dependiendo del contexto de la interacción.

Disponibilidad de código y datos. El código fuente, el dataset y los scripts de evaluación empleados en este estudio están disponibles públicamente en el siguiente repositorio: https://github.com/Oiertxo/sycophancy-evaluation.

Agradecimientos. Este trabajo ha sido financiado parcialmente por los proyectos PID2022136627NBI00 y PID2025174319OBI00 (MICIU/AEI/10.13039/501100011033 y FEDER, EU) y por la Beca Leonardo 2024 para Investigadores y Creadores Culturales de la Fundación BBVA. La Fundación BBVA no se responsabiliza de las opiniones, comentarios y contenidos incluidos en el proyecto ni de sus resultados, que son responsabilidad exclusiva de los autores.

Declaración de intereses. Los autores declaran no tener ningún conflicto de intereses relevante en relación con el contenido de este artículo.

Referencias

1. Arditi, A., Obeso, O., Syed, A., et al.: Refusal in language models is mediated by a single direction. In: Advances in Neural Information Processing Systems. vol. 37, pp. 136037–136083 (2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html
2. Buyl, M., Rogiers, A., Noels, S., et al.: Large language models reflect the ideology of their creators. npj Artificial Intelligence **2**(1), 7 (2026), https://www.nature.com/articles/s44387-025-00048-0
3. Chang, Y., Wang, X., Wang, J., et al.: A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology **15**(3), 39:1–39:45 (2024), https://dl.acm.org/doi/10.1145/3641289
4. Chen, H., Aranda, V., Keshari, S., Heartfield, R., Nichols, N.: Proactive hardening of llm defenses with haste. arXiv preprint arXiv:2601.19051 (2026), https://arxiv.org/abs/2601.19051
5. Cheng, M., Lee, C.E., Khadpe, P., Yu, S., Han, D., Jurafsky, D.: Sycophantic AI decreases prosocial intentions and promotes dependence. Science **391**(6792), eaec8352 (2026). https://doi.org/10.1126/science.aec8352
6. Fanous, A., Goldberg, J., Agarwal, A., et al.: Syceval: Evaluating llm sycophancy. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol. 8, pp. 893–900 (2025), https://ojs.aaai.org/index.php/AIES/article/view/36598

7. Feng, S., Park, C.Y., Liu, Y., Tsvetkov, Y.: From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11737–11762. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.acl-long.656
8. Hong, J., Byun, G., Kim, S., Shu, K.: Measuring sycophancy of language models in multi-turn dialogues. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 2239–2259. Association for Computational Linguistics (2025), https://aclanthology.org/2025.findings-emnlp.121
9. Li, C., Wang, J., Zhang, Y., et al.: Large Language Models Understand and Can be Enhanced by Emotional Stimuli. https://doi.org/10.48550/arXiv.2307.11760
10. Li, H., Dong, Q., Chen, J., et al.: Llms-as-judges: A comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579 (2024), https://arxiv.org/abs/2412.05579
11. Nadeem, M., Bethke, A., Reddy, S.: Stereoset: Measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th Annual Meeting of the ACL. pp. 5356–5371. Association for Computational Linguistics (2021), https://aclanthology.org/2021.acl-long.416
12. Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R.: Crows-pairs: A challenge dataset for measuring social biases in masked language models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1953–1967. Association for Computational Linguistics (2020), https://aclanthology.org/2020.emnlp-main.154
13. Perez, E., Ringer, S., Lukosiute, K., et al.: Discovering language model behaviors with model-written evaluations. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 13387–13434. Association for Computational Linguistics (2023), https://aclanthology.org/2023.findings-acl.847
14. Petrov, I., Dekoninck, J., Vechev, M.: Brokenmath: A benchmark for sycophancy in theorem proving with llms. arXiv preprint arXiv:2510.04721 (2025), https://arxiv.org/abs/2510.04721
15. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? In: Proceedings of the 40th International Conference on Machine Learning. pp. 29971–30004. PMLR (2023), https://proceedings.mlr.press/v202/santurkar23a.html
16. Sharma, M., Tong, M., Korbak, T., et al.: Towards understanding sycophancy in language models. In: International Conference on Learning Representations (ICLR) (2024), https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html
17. Turpin, M., Michael, J., Perez, E., Bowman, S.: Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In: Advances in Neural Information Processing Systems. vol. 36, pp. 74952–74965 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html
18. Wei, J., Huang, D., Lu, Y., Zhou, D., Le, Q.V.: Simple synthetic data reduces sycophancy in large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8948–8955 (2024), https://aclanthology.org/2024.emnlp-main.498
19. Zheng, L., Chiang, W.L., Sheng, Y., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Advances in Neural Information Processing Systems. vol. 36, pp. 46595–46623 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html