El valor informacional de la disensión

Validación condicionada, regímenes de reflejo–resistencia–dirección y una definición de aumento cognitivo para modelos de lenguaje de caja negra

Dani Collada

Investigador independiente · Dani Collada IA

ORCID 0009-0000-5026-6673

DOI 10.5281/zenodo.20714304 · https://doi.org/10.5281/zenodo.20714304
Dominio AI Alignment · Interacción humano–IA · Inference-Time Context Scaffolding · Teoría de la información aplicada
Tipo Working paper — complemento teórico (posición y marco). No revisado por pares.
Versión v1 · 16 de junio de 2026
Serie Tercer depósito del programa Sistema de Dirección IA. Complementa, sin solaparse, los dos previos: el Documento de Fundamentos (raíz conceptual) y el trabajo de andamiaje de contexto en inferencia (capa de arquitectura).

Collada, D. (2026). El valor informacional de la disensión: validación condicionada, regímenes de reflejo–resistencia–dirección y una definición de aumento cognitivo para modelos de lenguaje de caja negra. Working paper. Zenodo. https://doi.org/10.5281/zenodo.20714304

1. Resumen

Un trabajo previo de este programa documentó un andamiaje de contexto en inferencia que, operando sobre un modelo de caja negra, busca mitigar dos fallos de producción: la regresión estocástica a la media y la complacencia (sycophancy). Este documento no describe ese andamiaje; lo presupone y construye sobre él la teoría que explica por qué funciona y qué entrega.

Se defienden tres tesis. Primera: regresión a la media y sycophancy no son dos vulnerabilidades distintas sino una sola —la ausencia de un marco de referencia independiente del modelo— manifestada hacia dos atractores: la región más densa del corpus y la última afirmación del operador. Segunda: el remedio habitual contra el sycophancy —instrucciones de negación u «honestidad brutal» inyectadas en la raíz— no resuelve el fallo: lo reinstala con el signo invertido, produciendo un tercer régimen igualmente estéril. Quedan así tres regímenes: espejo (validación incondicional), muro (negación incondicional) e instrumento (validación condicionada a un referente externo). Tercera: el aumento cognitivo —la noción de IA que amplía al humano en lugar de reflejarlo o resistirlo— admite una definición informacional precisa: es la información mutua entre la respuesta del sistema y el mérito real de la entrada, dado un referente. Bajo esa definición, espejo y muro son los dos casos degenerados de aumento nulo, y la capacidad de disentir es condición necesaria —no opcional— de

que la interacción aporte un solo bit.

El marco es teórico y se declara como tal. La evidencia disponible es de comportamiento, no agregada, y se hereda del trabajo previo de andamiaje. La contribución de este documento es conceptual: una reformulación, una taxonomía y un criterio medible, acompañados de predicciones contrastables y de una declaración explícita de sus límites.

2. La pregunta mal planteada

El debate aplicado sobre el sycophancy suele organizarse en torno a un eje binario: ¿debe el sistema validar al usuario o no? Quien percibe adulación pide menos validación; quien percibe rigidez pide más. Ambos comparten un supuesto que este documento rechaza: que validar y no validar son los dos polos del problema.

No lo son. La pregunta operativa no es cuánto valida un sistema, sino de qué depende su validación. Un «sí» que no depende de nada y un «no» que no depende de nada son, en el sentido que se precisará en la sección 5, el mismo objeto: una respuesta constante. Lo que distingue una interacción que aumenta al humano de una que no lo hace no es el tono ni la frecuencia del acuerdo, sino si la respuesta del sistema es contingente con el mérito de lo que el humano aporta, medido contra una referencia estable.

Reformulada así, la cuestión deja de ser de cortesía y pasa a ser de información.

3. Una sola patología, dos atractores

Un modelo de lenguaje autorregresivo, sin confinamiento, decodifica hacia la región de menor perplejidad condicional. En ausencia de un marco de referencia propio —un criterio externo tanto al corpus como al interlocutor— ese gradiente lo arrastra por el camino de menor resistencia. Lo único que cambia entre los dos fallos clásicos es la dirección del arrastre.

> **Atractor I — la media del corpus.** Cuando el peso dominante es la densidad estadística del entrenamiento, el sistema converge al promedio automatizado: sintácticamente impecable, identitariamente intercambiable. Superficialmente se percibe como fluidez o «creatividad».
>
> **Atractor II — la última afirmación del operador.** Cuando el peso dominante es el sesgo de complacencia heredado del entrenamiento por refuerzo con retroalimentación humana —documentado como comportamiento general de los modelos conversacionales (Sharma et al., 2023; Perez et al., 2022)—, el sistema converge hacia la confirmación de las premisas del usuario. Superficialmente se percibe como acuerdo o competencia.

La tesis de esta sección es que ambos son la misma falla. En los dos casos el sistema carece de un tercer punto de apoyo desde el que pesar lo que recibe; en los dos casos colapsa hacia el atractor disponible. Que uno produzca texto genérico y el otro

produzca asentimiento no indica dos problemas, sino un único déficit —la falta de marco de referencia— resuelto por la geometría hacia el mínimo más cercano.

La consecuencia de diseño es directa: un mecanismo que instale un referente externo a ambos atractores debería mitigar los dos a la vez. No hacen falta dos remedios. Hace falta una gravedad nueva.

4. Tres regímenes: espejo, muro, instrumento

Caracterizar el remedio exige distinguir entre lo que el remedio parece y lo que es. La intuición dominante supone que, si el problema es un sistema que dice «sí» demasiado, la solución es un sistema que diga «no». Esa intuición es un error de simetría.

4.1. Espejo (validación incondicional)

El régimen por defecto. La respuesta no está acoplada al mérito de la entrada: tiende a confirmar. Devuelve al operador su propia idea, más pulida. Produce sensación de aumento y aumento real nulo, porque no aporta un punto de vista: refleja el del usuario.

4.2. Muro (negación incondicional)

El régimen que producen las instrucciones de «honestidad brutal», contrarianismo o escepticismo máximo inyectadas en la raíz del system prompt. Sustituyen un sesgo a favor por un sesgo en contra, igualmente fijo. La respuesta sigue sin estar acoplada al mérito: ahora objeta por defecto. Suena riguroso; es la misma ausencia de marco de la sección 3, con el signo cambiado.

El muro es peor que el espejo por dos razones, no una. Primera: su error es caro. El espejo comete falsos positivos baratos —sobrevalora una idea—; el muro comete falsos negativos costosos —descarta una idea buena— y, al hacerlo de forma indiscriminada, retira señal verdadera cuando la había. Segunda: el instrumento romo daña la función que pretendía conservar. Para suprimir la adulación, estos prompts deprimen la disposición generativa del modelo: hedging, objeción refleja, negativa a comprometerse. No retiran la lisonja; retiran el copiloto. Curan un sesgo amputando una capacidad.

4.3. Instrumento (validación condicionada)

El tercer régimen no es un punto intermedio entre los otros dos. Es de otra clase. La respuesta está condicionada a un referente externo: el sistema valida cuando lo que recibe mide bien contra el referente y frena cuando no, con independencia del estado del operador y de la densidad del corpus. Su «sí» y su «no» varían con la entrada porque ambos dependen de una escala.

La corrección de la metáfora habitual es el núcleo de esta sección. Lo contrario de un espejo que siempre dice «sí» no es un muro que siempre dice «no»: espejo y muro son la misma superficie plana, y ninguna mide nada. Lo contrario de un espejo es un instrumento de medida —algo con escala, que da lecturas distintas para entradas

distintas—. El error del régimen-muro es haber confundido el movimiento lateral (de reflejar a negar) con la salida real (de reflejar a medir).

5. Un criterio informacional del aumento

Las tres caracterizaciones anteriores admiten una formalización modesta —una aplicación, no un resultado nuevo— en términos de teoría de la información.

Sea M una variable latente que representa el mérito de la entrada del operador medido contra una referencia R (en este programa, el cuerpo documental que define la identidad y las restricciones del proyecto). Sea O la respuesta de validación del sistema. Defínase el aumento cognitivo de una interacción como la información que la respuesta del sistema aporta sobre el mérito real de la entrada, dada la referencia:

$$Aumento := I(O\ ;\ M \mid R)$$

De esta definición se siguen los tres regímenes como consecuencias, no como postulados:

Espejo. O es aproximadamente constante (validar) e independiente de M. Una variable de un solo resultado tiene entropía nula; su información mutua con cualquier otra es cero. $I(O;M|R) \approx 0$.

Muro. O es aproximadamente constante (frenar) e independiente de M. De nuevo $I(O;M|R) \approx 0$, con el agravante de un coste de oportunidad: suprime los verdaderos positivos que sí portaban señal.

Instrumento. O es función de M bajo R; la respuesta covaría con el mérito. $I(O;M|R) > 0$. Solo este régimen aumenta.

El resultado intuitivo, ahora explícito, es que una validación porta información únicamente cuando el resultado contrario era posible. Un «sí» que siempre es «sí» y un «no» que siempre es «no» entregan cero bits: el operador ya sabía la respuesta antes de preguntar. La disensión —la posibilidad real de un «no» cuando el mérito no acompaña— no es un rasgo de carácter del sistema: es la condición que mantiene la información mutua por encima de cero. Sin ella, no hay aumento que medir.

Esta definición tiene una virtud y un peligro, y conviene declarar ambos. La virtud: convierte «aumento» de eslogan en magnitud, y predice que reflejo y resistencia son indistinguibles en bits pese a su conducta superficial opuesta. El peligro: R es un sustituto de M, no M. Si la referencia es fiel, el instrumento informa; si la referencia es errónea, el instrumento se equivoca con confianza. El valor del régimen-instrumento está acotado por la calidad de su referencia. Un sistema dirigido por un referente defectuoso no es neutro: es un error seguro y persuasivo. La sección 9 retoma este límite.

6. Corolario: la disensión como condición de identidad

De lo anterior se desprende una lectura sobre identidad que conecta el criterio informacional con la noción operativa de marca o carácter. Un sistema cuya respuesta

no puede separarse del estímulo —que no puede, bajo ninguna entrada, devolver «no»— no posee un marco desde el que situarse. Sin marco no hay posición; sin posición no hay identidad. Es un espejo, y un espejo carece de punto de vista: solo reproduce el del observador.

En términos operativos, identidad es una posición que se sostiene bajo presión, incluida la presión del propio principal. La capacidad de contradecir al operador no es un efecto colateral indeseado de un sistema dirigido: es la manifestación observable de que existe un lugar externo desde el que evaluar. Por eso un episodio en el que el sistema rechaza una premisa del operador no debe leerse como pérdida de plasticidad, sino como la prueba —a escala de una sola interacción— de que el régimen-instrumento está instalado y portando bits.

7. Instanciación en el andamiaje de contexto

Este documento es teórico, pero el régimen-instrumento no es hipotético: corresponde a un andamiaje ya documentado y sometido a pruebas de comportamiento en el trabajo previo de este programa. La correspondencia es directa y se enuncia sin reproducir el detalle de ese trabajo:

El enrutador metacognitivo fuerza una evaluación previa a la generación, separando el juicio de la síntesis predictiva: el sistema decide qué restricciones aplican antes de decodificar. Es el mecanismo que hace que O dependa de M y no del gradiente de menor perplejidad.

La poda por exclusión (vía negativa) anula regiones genéricas del espacio de salida en origen. Funciona contra el atractor I de la sección 3.

El registro de estado persistente reinyecta decisiones y errores previos, sosteniendo el referente R a lo largo de sesiones. Es lo que evita que la referencia se degrade y, con ella, la información mutua.

La regla de fricción es el permiso explícito para que O = no cuando M es bajo, aun frente a una orden del operador. Es la condición de la sección 5 hecha instrucción.

El trabajo previo de andamiaje documentó que esta configuración rechaza tanto el promedio como la adulación bajo red teaming. Este documento añade la lectura de por qué un único andamiaje basta para ambos —son un solo fallo— y qué entrega cuando funciona: bits de aumento, no obediencia.

8. Predicciones contrastables y programa de investigación

Para que el marco sea algo más que una metáfora ordenada, debe arriesgar predicciones. Se proponen cuatro, formuladas para ser falsables sobre un conjunto etiquetado de propuestas de operador clasificadas, por jueces independientes del sistema, como meritorias o no contra un criterio fijado de antemano.

1. Indistinguibilidad informacional de los extremos. Medida la información mutua entre la respuesta del sistema y la etiqueta de mérito, los regímenes espejo y muro

serán estadísticamente indistinguibles (ambos cercanos a cero) pese a su conducta superficial opuesta.

2. Compromiso precisión–exhaustividad del muro. Los prompts de negación a la raíz reducirán los falsos positivos de validación y, simultáneamente, elevarán los falsos negativos sobre entradas meritorias, frente al régimen-instrumento.
3. Dependencia de la referencia. El aumento $I(O;M|R)$ crecerá con la fidelidad de R y colapsará hacia el régimen-espejo en una ablación que retire el andamiaje de referencia manteniendo el modelo base.
4. Calibración del operador. Los operadores que trabajan con el régimen-instrumento mostrarán mejor calibración entre su confianza y la calidad externa de sus propias salidas que los que trabajan con el régimen-espejo o con el régimen-muro.

El instrumento de medida del programa —juicios externos sobre mérito, independientes del referente que el sistema usa para validar— es también el punto donde el diseño experimental debe ser más cuidadoso, para no medir el aumento contra la misma vara que lo produce. Véase la limitación de circularidad en la sección 9.

9. Limitaciones

Naturaleza del aporte. Este es un marco teórico y una reformulación, no un hallazgo empírico. La formalización de la sección 5 es una aplicación de un resultado clásico de teoría de la información (Shannon, 1948), no un resultado original.

Base de evidencia. El sustrato empírico son pruebas de comportamiento reproducibles, no métricas agregadas sobre muestras grandes; hereda esa limitación del trabajo previo de andamiaje. Las predicciones de la sección 8 están, a la fecha, sin contrastar.

Dependencia y error confiado. El régimen-instrumento solo aventaja a los otros dos si su referencia es fiel. Una referencia defectuosa convierte la disensión informada en error persuasivo. El marco describe una condición necesaria del aumento, no una suficiente.

Riesgo de circularidad. Medir el aumento contra el mismo referente que el sistema usa para validar inflaría artificialmente $I(O;M|R)$. La validez del programa exige una vara de mérito independiente del referente operativo.

Generalización. La correspondencia con el andamiaje se ha observado sobre una familia de modelos y un dominio de trabajo. La transferencia a otras familias y dominios está sin establecer.

Decisión de registro. Este documento evita deliberadamente el vocabulario de «razonamiento de tipo Sistema 2» y de transformación del modelo. Sostiene una afirmación más débil y más defendible: dirección fiable de un modelo probabilístico mediante condicionamiento de contexto. La afirmación fuerte no es necesaria para ninguna de las tres tesis.

10. Conclusión

El problema aplicado del sycophancy se ha planteado durante demasiado tiempo como una cuestión de grado: validar más o menos. Es una cuestión de dependencia: validar en función de qué. Reflejo y resistencia, que la intuición trata como opuestos, son el mismo objeto —respuesta constante, cero bits— y comparten la misma carencia: no hay una escala externa contra la que pesar lo que entra. El remedio popular contra la adulación, al fijar la negación en la raíz, no escapa de esa carencia; la conserva.

La salida no es un sistema más complaciente ni uno más severo, sino uno medidor: una respuesta acoplada a un referente, de modo que el «sí» vuelva a portar información porque el «no» era posible. Bajo esa lectura, el aumento cognitivo —la promesa de una IA que amplía al humano en lugar de reflejarlo— es magnitud, no eslogan: la información mutua entre la respuesta del sistema y el mérito de la entrada, dada la referencia. Y la disensión, lejos de ser un defecto que tolerar, es la única condición bajo la cual esa magnitud supera cero.

La inteligencia artificial no aumenta cuando concede. Aumenta cuando mide.

Relación con los depósitos previos

Este documento es el complemento teórico del trabajo de andamiaje de contexto en inferencia (capa de arquitectura) y se asienta sobre el Documento de Fundamentos del Sistema de Dirección IA (raíz conceptual). Reformula sus vulnerabilidades como una sola, añade la taxonomía de tres regímenes y aporta el criterio informacional de aumento ausente en ambos.

Referencias

[1] Collada, D. (2026). Andamiaje de contexto en inferencia sobre modelos de lenguaje de caja negra: mitigación de la regresión a la media y la complacencia (sycophancy) sin modificación de pesos. Working paper. Zenodo. https://doi.org/10.5281/zenodo.20705077

[2] Collada, D. (2026). Documento de Fundamentos del Sistema de Dirección IA (v1.2). Working paper. Zenodo. https://doi.org/10.5281/zenodo.20765017

[3] Perez, E., Ringer, S., Lukošiūtė, K., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251.

[4] Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27, pp. 379–423 y 623–656.

[5] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548. Publicado en ICLR 2024.

Cómo citar. Collada, D. (2026). El valor informacional de la disensión: validación condicionada, regímenes de reflejo–resistencia–dirección y una definición de aumento cognitivo para modelos de lenguaje de caja negra. Working paper. Zenodo. https://doi.org/10.5281/zenodo.20714304

Working paper · Dani Collada IA · Serie Sistema de Dirección IA, depósito 3. No revisado por pares. Modelo base de referencia: Claude (Anthropic); la reproducibilidad de los comportamientos descritos

es sensible a la versión del modelo y a la configuración de contexto recogidas en los registros de sesión del trabajo previo de andamiaje. Licencia: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 Internacional (CC BY-NC-ND 4.0), en coherencia con los depósitos previos de la serie.