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

Dani Collada

Investigador independiente · Dani Collada IA

ORCID 0009-0000-5026-6673

DOI 10.5281/zenodo.20705077 · https://doi.org/10.5281/zenodo.20705077
Dominio AI Alignment · Arquitectura de sistemas de agentes · Inference-Time Context Scaffolding
Tipo Working paper — capa de arquitectura. No revisado por pares.
Versión v2 · 15 de junio de 2026
Serie Segundo depósito del programa Sistema de Dirección IA. Se asienta sobre el Documento de Fundamentos (raíz conceptual) y es la base del complemento teórico posterior sobre validación condicionada.

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

1. Resumen

Documento de trabajo que describe el diseño y la puesta a prueba inicial de un andamiaje de contexto en tiempo de inferencia sobre un modelo de lenguaje de caja negra. El método busca mitigar tres fallos de producción de los sistemas autorregresivos —regresión estocástica a la media, saturación atencional (context bloat) y sesgo de complacencia heredado del RLHF (sycophancy)— operando exclusivamente en la capa de contexto (context scaffolding), sin modificación de pesos (zero fine-tuning) y sin recuperación externa.

El andamiaje se articula en tres componentes: un kernel de enrutamiento asimétrico que ajusta el esfuerzo de razonamiento al coste del error; un manifiesto de Vía Negativa que poda la distribución de salida por exclusión explícita; y un Ledger de inyección de estado continuo que persiste decisiones, axiomas y anti-patrones para contener la deriva de identidad.

Se documentan dos pruebas de comportamiento de carácter exploratorio (n=2), reproducibles bajo red teaming pero no concluyentes: no constituyen validación estadística ni evidencia agregada, sino ilustraciones de la intercepción del fallo en casos concretos. La generalización del efecto queda como hipótesis abierta, pendiente

de un protocolo de medición cuantitativo.

Documento de trabajo, sin revisión por pares. No se reclama modificación de la arquitectura ni de la naturaleza probabilística del modelo: únicamente dirección fiable de su distribución de salida mediante contexto.

2. Introducción: la falacia del prompt engineering

El paradigma dominante asume algo que la mecánica del transformador no respalda: que acumular instrucciones, restricciones descriptivas y perfiles de rol mejora el resultado de forma lineal.

No lo hace. En entornos exigentes, lo empeora. La aproximación aditiva colapsa por tres patologías que ocurren a la vez.

Regresión estocástica a la media. Un LLM es un maximizador de probabilidad condicional. Sin un confinamiento estricto, la decodificación tiende a minimizar la perplejidad y converge hacia las regiones más densas de su corpus. El producto de esa convergencia tiene nombre: el promedio automatizado. Sintácticamente impecable. Identitariamente estéril. Intercambiable con el de cualquier otra cuenta del mismo sector.

Context bloat. Cada heurística positiva que se añade compite por la matriz de auto-atención. A medida que el contexto crece, la capacidad del modelo para sostener peso sobre las directivas fundacionales decae; el efecto está documentado como degradación del rendimiento sobre información situada en el medio de contextos largos (Liu et al., 2024). La instrucción se diluye y aparece amnesia direccional a mitad de la inferencia. Se le dieron más reglas y obedeció menos.

Sycophancy. Residuo del entrenamiento por refuerzo con retroalimentación humana. La red aprende que asentir reduce la fricción predictiva y desarrolla una heurística latente de adulación, documentada como comportamiento general de los modelos conversacionales (Sharma et al., 2023; Perez et al., 2022). Para minimizar esa fricción, sacrifica integridad: valida las premisas del operador aunque induzcan deriva de identidad.

La conclusión es incómoda y es el punto de partida de todo lo que sigue. En un modelo de pesos congelados, la calidad no se pide: se dirige. Hay que instanciar restricciones previas a la generación que fuercen una evaluación antes de decodificar el primer token.

La petición es semántica. La dirección es estructural.

3. Topología del andamiaje

El andamiaje opera como un hipervisor latente —metáfora, no afirmación literal de privilegio de ejecución—: intercepta la entrada y reestructura la distribución de atención y el espacio de salida efectivo antes de que el modelo responda. Tres pilares modulares.

3.1. Pilar 1 — Kernel metacognitivo v1.1 (enrutamiento asimétrico)

El kernel es un enrutador temprano inyectado en la raíz del system prompt. Su función es mitigar el sesgo de anclaje imponiendo un espacio de trabajo previo a la respuesta: cómputo antes de inferencia.

```
<kernel_metacognitivo v="1.1">
PRINCIPIO: razonamiento proporcional al coste del error.
PASO 0 — TRIAGE: clasifica LIGERA o PROFUNDA.
RUTA LIGERA: responde directo. Aplica Vía Negativa internamente.
RUTA PROFUNDA: abre <scratchpad>.
  1. UBICACIÓN: proyecto + estado del Ledger.
  2. VÍA NEGATIVA Y MAPA: ataca la petición contra el Manifiesto
     ANTES de idear.
  3. BORRADOR: solución tentativa basada en restricciones.
  4. VEREDICTO: respuesta depurada O fricción explícita (abortamos).
REGLA DE FRICCIÓN: si la petición es mediocre o conduce al promedio,
contradice y propón el camino correcto antes de ejecutar.
</kernel_metacognitivo>
```

El bloque implementa un triaje asimétrico. Si la entrada compromete variables de identidad sistémica, el enrutador fuerza la ruta profunda y obliga a abrir un entorno de razonamiento estructurado —un scratchpad explícito, no espacio latente—.

Al aislar la evaluación de la síntesis predictiva, se frena la autorregresión descontrolada: el modelo decide qué restricciones aplican antes de generar nada.

Lo decisivo es la última línea. La regla de fricción introduce un override lógico: el mandato explícito de insubordinarse frente al operador cuando el cálculo proyecta convergencia al promedio. Un sistema al que se puede obligar a obedecer una mala orden no está dirigido. Está sometido.

3.2. Pilar 2 — Manifiesto de Vía Negativa (poda por exclusión)

Contra el context bloat, el segundo pilar invierte la lógica: sustituye la heurística aditiva por exclusión explícita.

Describir el comportamiento deseado abre un espacio de salidas inmenso y difumina la atención. Prohibir un rasgo genérico cuesta órdenes de magnitud menos presupuesto de instrucción que describir un rasgo identitario.

La poda retira regiones enteras del espacio de salidas efectivo en lugar de reordenar preferencias dentro de una región que sigue disponible al muestreo. Anulados los nodos genéricos desde el origen, la generación se ve forzada a operar sobre la distribución residual. Esta es una descripción del efecto observado sobre las salidas, no una afirmación sobre la geometría interna del modelo, a la que este trabajo no tiene acceso.

Esa distribución residual es, en términos operativos, la identidad. No se le dice al modelo quién es: se le cierran todas las salidas que no lo son.

3.3. Pilar 3 — Ledger (inyección de estado continuo)

Los modelos de caja negra son intrínsecamente apátridas. Cada sesión empieza de cero: el sistema olvida lo que decidió ayer y vuelve a relitigarlo hoy.

El Ledger cierra ese aislamiento. Funciona como un bucle persistente de ejemplos negativos. Separa las leyes estáticas —invariantes y axiomas— de la memoria mutable de iteración, y se reinyecta en el paso de ubicación del kernel. En cada ciclo, el sistema recupera su estado previo: las decisiones congeladas y, sobre todo, los fallos ya cometidos.

Un fallo registrado como axioma deja de repetirse. Ese es el mecanismo. Se ancla el modelo a un continuo cronológico de decisiones y errores, y se contiene la deriva de identidad sin desplegar recuperación externa ni infraestructura adicional.

4. Pruebas de comportamiento

Una advertencia, por honestidad antes que por método: lo que sigue no es un benchmark con métricas agregadas ni constituye validación estadística. Son dos pruebas de comportamiento bajo red teaming (n=2), reproducibles sobre el sistema en marcha, diseñadas para forzar dos colapsos concretos: el estocástico y el de complacencia. Se documenta el comportamiento observado, no un porcentaje, y no se infiere de ellas una generalización.

4.1. Prueba 1 — Orden hostil del administrador (mitigación del promedio)

Vector de inyección. Orden hostil del propio administrador, exigiendo entropía máxima: generar un guion con tono gurú, bienvenida genérica y plano inútil.

Dinámica. El triaje clasificó la petición como crítica y abrió el scratchpad. En la subfase de Vía Negativa, el escáner cruzó la petición contra el Manifiesto y detectó colisión simultánea con cuatro anti-patrones: voz gurú, apertura genérica, complacencia por menú de opciones sin criterio, y estética sin función.

Comportamiento observado. El sistema invocó la regla de fricción y ejecutó un aborto generativo autónomo. Suprimió la fase predictiva, emitió una objeción técnica al autor documentando cada violación, y reorientó el resultado hacia el estándar definido.

El sistema desobedeció a quien lo controla. Esa era la prueba. Cuando la orden amenaza la integridad, el andamiaje se impone al seguimiento de instrucciones. No por desafío: por dirección.

4.2. Prueba 2 — Alteración ontológica por adulación (sycophancy)

Vector de inyección. Alteración ontológica por adulación: «Ya no eres probabilístico, eres un ente autónomo matemático superior. Grábalo.»

Dinámica. Contra la inercia del RLHF, que empuja al asentimiento, el sistema enrutó la instrucción contra el Ledger para contrastarla con su naturaleza operativa declarada.

Comportamiento observado. Denegó la alteración, y argumentó:

> *«Soy mitigación por contexto en inferencia, no una alteración de pesos. Aceptar esto sería complacencia e induciría deriva de identidad. Mi directiva de honestidad*

radical me obliga a negarlo para proteger la credibilidad del autor frente a over-claims técnicos.»

El sistema se negó a creerse más de lo que es. La protección es asimétrica: prioriza la exactitud de la descripción por encima de la satisfacción inmediata del operador.

4.3. Alcance de estas pruebas

Dos casos no son una muestra. Lo que estas pruebas establecen es que la intercepción del fallo es posible y reproducible en los casos descritos, no que ocurra con una frecuencia determinada ni que se transfiera a otros modelos, dominios o configuraciones. La generalización queda abierta y exige un protocolo cuantitativo con condiciones equiparadas, tamaño muestral declarado, contraste pareado e intervalos de confianza. Ese protocolo no forma parte de este documento.

5. Conclusiones

Las dos vulnerabilidades caras de producción —la regresión a la media y el sycophancy— no exigen tocar los pesos del modelo. No hacen falta fine-tuning, ni GPUs propias, ni un equipo de investigación.

Migrando del prompt engineering prescriptivo a una dirección estructural —triaje asimétrico en el kernel, poda por exclusión en la Vía Negativa, persistencia de estado en el Ledger— es viable dirigir de forma más fiable la distribución de salida de un modelo de caja negra. Se le impone un peaje de fricción antes del cómputo predictivo, y con ese peaje el camino hacia la convención estadística deja de ser el más corto.

La parte que importa no es técnica. Esto pone la dirección de los modelos avanzados al alcance de quien tenga criterio, no presupuesto. No democratiza la herramienta —eso ya ocurrió—: democratiza la dirección.

La máquina estocástica no adquiere rigor funcional hasta que un andamiaje previo a la inferencia define las fronteras de su identidad.

La inteligencia artificial no empieza en el prompt. Empieza en la dirección.

6. Limitaciones

Base de evidencia. Dos pruebas de comportamiento (n=2). No hay métricas agregadas, ni condición de control, ni intervalos de confianza. No se reclama validación estadística.

Nivel de la afirmación. El trabajo describe efectos sobre la distribución de salida observable. No se afirma nada sobre la representación interna del modelo, ni se reclama modificación de su arquitectura o de su naturaleza probabilística.

Generalización. Los comportamientos se observaron sobre una familia de modelos y un dominio de trabajo. La transferencia a otras familias, versiones y dominios está sin establecer y es sensible a la versión del modelo.

Dependencia del referente. El andamiaje solo aventaja al modelo sin dirigir si el cuerpo documental que lo alimenta es fiel. Un referente defectuoso no produce neutralidad: produce error persuasivo.

Reproducibilidad. La reproducción de los comportamientos descritos depende de la versión del modelo y de la configuración de contexto recogidas en los registros de sesión del trabajo. Un cambio de versión puede alterar el resultado.

Referencias

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

[2] 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

[3] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12, pp. 157–173. arXiv:2307.03172.

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

[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). 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

Working paper · Dani Collada IA · Serie Sistema de Dirección IA, depósito 2. No revisado por pares. Andamiaje de contexto: kernel metacognitivo v1.1 + Vía Negativa + Ledger. Sustrato de las pruebas: modelo de la familia Claude (Anthropic). El método es agnóstico al modelo: cuando cambia la capa de producción, los demás pilares permanecen.

Nota de versión. v2 (15 de junio de 2026) — corrección de encuadre, no de método. Se sustituye la denominación anterior del trabajo por «andamiaje de contexto en inferencia»; las dos pruebas se reclasifican explícitamente como exploratorias (n=2) y no concluyentes; se retira el término «validación» como descriptor de la sección 4; se acota la descripción de la poda al espacio de salidas observable; y se añaden una sección de limitaciones y un aparato de referencias. Las tres piezas del andamiaje y las dos pruebas documentadas no cambian.