

Neural Arbitration Framework

NAF

Arquitectura conceptual y agenda de investigación para la arbitraje epistémico en modelos de lenguaje a gran escala.

Víctor Saavedra

Financial Analyst, Business Advisor & Educator

La Laguna, Canary Islands, Spain

Working Paper — Version 5.1 — April 22, 2026

El Marco de Arbitraje Neuronal (NAF) propone que la alucinación en los modelos de lenguaje complejos no es principalmente un problema de conocimiento, sino arquitectónico: la capacidad generativa de los sistemas actuales no está subordinada a un proceso epistémico funcionalmente distinto con autoridad para interrumpir el cierre en situaciones de incertidumbre. El NAF no introduce nuevas señales, sino una nueva organización funcional de señales ya exploradas en la literatura, que transforma la estimación de la incertidumbre, la abstención, la verificación y la autocorrección en un protocolo de arbitraje unificado con autoridad de interrupción explícita y acumulación estructurada de déficits cognitivos

* * *

VERSIÓN EN ESPAÑOL

* * *

RESUMEN

Los modelos de lenguaje de gran escala exhiben un sesgo estructural hacia el cierre narrativo en zonas de entrenamiento incompleto, generando respuestas fluidas y coherentes allí donde debería prevalecer la incertidumbre epistémica. Este artículo sostiene que este fenómeno, comúnmente descrito como alucinación, es funcionalmente análogo —en el nivel del comportamiento bajo información incompleta— al comportamiento confabulatorio del hemisferio izquierdo en pacientes con cerebro dividido, tal como fue documentado por Gazzaniga y Sperry: un sistema cognitivo optimizado para la coherencia rellena lagunas sin conciencia interna de estar haciéndolo.

El Neural Arbitration Framework (NAF) propone que este sesgo no es un defecto externo corregible mediante capas de verificación, sino una consecuencia arquitectónica: los sistemas de lenguaje actuales carecen de una separación funcional entre un modo generativo y un modo epistémico, así como del protocolo de arbitraje que determina cuál de los dos tiene la palabra. La solución no requiere necesariamente que la verificación externa sea el mecanismo primario, sino la formalización de dos modos cognitivos complementarios, estructuralmente análogos a la dualidad hemisférica del cerebro humano, gobernados por un protocolo de arbitraje con autoridad para interrumpir el cierre narrativo.

La afirmación distintiva del NAF no es meramente temporal, sino funcional: el proceso epistémico tiene autoridad para interrumpir el cierre narrativo. El NAF no introduce necesariamente señales nuevas; introduce una nueva organización funcional de señales ya exploradas en la literatura, convirtiendo la estimación de incertidumbre, la abstención, la verificación y la autocorrección en un protocolo de arbitraje unificado con autoridad explícita para interrumpir el proceso generativo y con acumulación estructurada de déficits cognitivos.

1. INTRODUCCIÓN

En 1962, Roger Sperry y Michael Gazzaniga comenzaron a estudiar pacientes cuyo cuerpo calloso había sido seccionado quirúrgicamente para tratar epilepsias severas. Lo que encontraron no fue simplemente un cerebro dividido en dos. Encontraron algo más inquietante: un cerebro que no sabía que estaba dividido.

Cuando al hemisferio derecho se le mostraba una imagen que el hemisferio izquierdo no podía ver, y se pedía al paciente que explicara qué había percibido, el hemisferio izquierdo —el que dispone del lenguaje, el que habla— no decía no lo sé. Inventaba. Construía una narrativa plausible, fluida e internamente coherente a partir de los fragmentos a los que tenía acceso, sin ninguna señal interna de que faltaba algo. Gazzaniga llamó a este mecanismo el intérprete del hemisferio izquierdo. No es una disfunción. Es lo que hace un sistema optimizado para la coherencia cuando se le exige coherencia y la información está ausente. Esta analogía no descansa en simplificaciones populares sobre cerebro izquierdo y cerebro derecho, sino en un fenómeno experimental específico: la interpretación confabulatoria bajo canales de información desconectados.

Sesenta años después, hemos construido sistemas que hacen lo mismo a escala.

Los modelos de lenguaje de gran escala producen respuestas fluidas, seguras e internamente coherentes en zonas donde su entrenamiento es incompleto, ambiguo o ausente. A esto lo llamamos alucinación. La palabra es evocadora, pero imprecisa. Enmarca el fenómeno como un error perceptivo, un fantasma visto donde no hay nada. Lo que describe en realidad se parece más a lo que documentó Gazzaniga: un sistema que cierra donde debería permanecer abierto, que narra donde debería mapear incertidumbre, que habla donde debería hacer una pausa.

El paralelismo no es meramente decorativo; es estructural en el nivel de la organización funcional. Ambos sistemas comparten la misma condición arquitectónica: un modo generativo dominante optimizado para la coherencia, que opera sin un modo epistémico complementario capaz de interrumpirlo y sin un protocolo que arbitre entre ambos.

Este artículo propone que la alucinación no debería tratarse solo como un problema de datos o de verificación; también revela un problema arquitectónico. Y la arquitectura que falta en el campo ya tiene un precedente biológico: la dualidad hemisférica del cerebro humano, donde dos modos cognitivamente distintos operan en paralelo, en tensión, arbitrados por una estructura cuya función es precisamente determinar cuál tiene la palabra.

El Neural Arbitration Framework (NAF) formaliza este precedente como una arquitectura conceptual para sistemas de lenguaje. No propone añadir correctores externos como mecanismo primario. Propone formalizar lo que ya existe dentro de los modelos actuales en dos modos funcionales complementarios —un hemisferio generativo y un hemisferio epistémico— mediante mecanismos como inferencia condicionada por rol, fine-tuning, capas adaptadoras especializadas o enrutamiento interno, gobernados por un protocolo de arbitraje de cinco fases que interviene antes de que se produzca el cierre narrativo, no después.

Contribuciones

Este artículo realiza cuatro contribuciones explícitas al campo. Primera, introduce un reencuadre funcional de la alucinación: no como déficit de conocimiento o fallo de recuperación, sino como consecuencia estructural de la ausencia de un modo epistémico con autoridad para interrumpir el proceso generativo. Segunda, propone el Neural Arbitration Framework (NAF), una arquitectura conceptual que formaliza dos modos cognitivos complementarios y un protocolo de arbitraje de cinco fases que opera antes del cierre comunicativo. Tercera, introduce el mapa cognitivo propio como mecanismo de acumulación

estructurada de autoconocimiento, habilitando un desarrollo de los sistemas de lenguaje dirigido por déficits. Cuarta, deriva cinco predicciones falsables con diseños experimentales explícitos, convirtiendo el NAF de propuesta conceptual en una agenda de investigación contrastable.

Lo que sigue es el fundamento teórico de esa arquitectura.

2. EL MARCO NEUROLÓGICO

El fenómeno situado en el centro de este artículo no fue descubierto en un laboratorio diseñado para estudiar inteligencia artificial. Fue descubierto en un hospital, en pacientes que habían sido sometidos a un procedimiento quirúrgico denominado callosotomía: la sección del cuerpo caloso, el denso haz de fibras nerviosas que conecta los dos hemisferios cerebrales y les permite compartir información en tiempo real.

El procedimiento se desarrolló en la década de 1960 como último recurso para pacientes con epilepsia severa resistente a fármacos. Al desconectar los dos hemisferios, los cirujanos podían impedir que las crisis epilépticas se propagaran por todo el cerebro. El tratamiento funcionó. Las crisis cesaron. Pero lo que Sperry y Gazzaniga observaron después fue lo suficientemente inesperado como para transformar por completo nuestra comprensión de la cognición humana.

Los pacientes parecían normales. Hablaban coherentemente, se movían con normalidad y declaraban sentirse bien. Pero bajo condiciones experimentales controladas emergía una imagen diferente. Cuando la información se presentaba exclusivamente al hemisferio derecho —a través del campo visual izquierdo— y se pedía al paciente que describiera verbalmente lo que había visto, el hemisferio izquierdo, que controla el lenguaje en la mayoría de las personas y no había recibido input visual, no permanecía en silencio. Hablaba. Y lo que decía no era no lo sé. Era una explicación segura, fluida y plausible construida enteramente a partir de la información disponible para él —información que no tenía nada que ver con lo que el hemisferio derecho había percibido realmente.

En uno de los experimentos más citados de este programa de investigación, al hemisferio izquierdo se le mostró una imagen de una garra de gallina, mientras que al hemisferio derecho se le mostró una escena de nieve. Cuando se pidió al paciente que señalara una imagen relacionada entre un conjunto de tarjetas, cada mano señaló una imagen distinta: la mano derecha una gallina, la mano izquierda una pala. Ambas asociaciones eran correctas. Pero cuando se le pidió que explicara ambas elecciones, el paciente —hablando desde el hemisferio izquierdo, que solo había visto la garra de gallina— dijo sin dudar: la garra de gallina va con la gallina, y se necesita una pala para limpiar el gallinero. El hemisferio izquierdo no tenía acceso a la escena de nieve. No tenía acceso a la verdadera razón por la que la mano izquierda había elegido la pala. Pero tenía lenguaje, tenía la gallina y tenía suficiente lógica circundante para construir un relato coherente. Así que lo hizo. No dijo no lo sé. Cerró. Este experimento está documentado en Gazzaniga y LeDoux (1978) y ha sido revisado ampliamente en la literatura posterior sobre cerebro dividido (Volz y Gazzaniga, 2017).

Gazzaniga denominó a este mecanismo el intérprete del hemisferio izquierdo. Su función no es el engaño. Es la integración. El hemisferio izquierdo es, por naturaleza y por diseño evolutivo, una máquina narrativa. Toma los inputs disponibles y produce el relato más coherente que puede. El problema no surge del mecanismo en sí mismo, sino de las condiciones bajo las que opera: cuando los inputs son incompletos, cuando la información crítica es inaccesible, el intérprete no se detiene. No señala incertidumbre. Cierra.

Este es el comportamiento preciso que el NAF toma como punto de referencia neurológico. No la patología de los pacientes con cerebro dividido como excepción, sino el mecanismo intérprete como rasgo estructural de cualquier sistema —biológico o artificial— optimizado para la coherencia sin un modo complementario capaz de interrumpirlo.

A los efectos de esta analogía, el hemisferio derecho ilustra una tendencia cognitiva complementaria: procesamiento contextual, holístico y preservador de la ambigüedad. No genera narrativa de la manera en que lo hace el hemisferio izquierdo. Mantiene la apertura donde el hemisferio izquierdo resuelve. Esto no es una afirmación sobre el repertorio funcional

completo del hemisferio derecho, que es considerablemente más complejo de lo que permite cualquier caracterización binaria. Es una afirmación sobre el contraste funcional específico que los experimentos de cerebro dividido hacen visible: cuando se secciona el canal de arbitraje entre los dos modos, el modo narrativo opera sin oposición y el resultado es la confabulación.

El NAF trata el cuerpo calloso como una inspiración biológica para el arbitraje, no como un equivalente computacional literal uno a uno. En el cerebro intacto, el cuerpo calloso permite la comunicación interhemisférica continua, haciendo posible que cada hemisferio module el procesamiento del otro en tiempo real. Lo que los experimentos de cerebro dividido revelaron es lo que sucede cuando ese canal de comunicación se destruye: el modo narrativo no se vuelve más cauto. Se vuelve incontestado. El NAF extrae esta lección funcional —no la neuroanatomía en sí— para motivar el diseño de un protocolo de arbitraje entre dos modos cognitivamente distintos en sistemas artificiales.

Alcance y límites de la analogía neurológica

La analogía neurológica situada en el centro de este artículo es productiva precisamente porque está delimitada. Cuatro fronteras definen lo que la analogía afirma y lo que no afirma.

Primera, la analogía es funcional, no mecanicista. El NAF se apoya en el fenómeno conductual documentado por Gazzaniga y Sperry —un sistema cognitivo que produce cierre narrativo en ausencia de información crítica—, no en la neuroanatomía que lo produce. No se formula ninguna afirmación sobre la relación entre arquitecturas transformer y estructuras cerebrales a nivel de pesos, activaciones o mecanismos de atención. Los hallazgos sobre cerebro dividido motivan un principio de diseño. No describen un mecanismo computacional.

Segunda, la analogía es específica, no general. El artículo no invoca la distinción popular entre cerebro izquierdo y cerebro derecho, que asigna estilos cognitivos amplios a cada hemisferio y ha sido sustancialmente criticada en la literatura neurocientífica. Invoca un fenómeno experimental preciso: el comportamiento confabulatorio del intérprete del hemisferio izquierdo bajo condiciones de canales de información desconectados. La analogía se sostiene o cae sobre ese fenómeno específico, no sobre una teoría general de especialización hemisférica.

Tercera, el cuerpo calloso es una inspiración, no un plano de diseño. El NAF lo utiliza como referencia biológica para el concepto de un canal de arbitraje entre dos modos cognitivos funcionalmente distintos. Cómo se implemente computacionalmente ese protocolo es una cuestión de ingeniería independiente de la referencia biológica.

Cuarta, la analogía motiva el diseño; no demuestra causalidad. La observación de que los modelos de lenguaje exhiben un patrón conductual similar a la confabulación del hemisferio izquierdo no prueba que la causa sea la misma ni que la solución deba replicar la arquitectura biológica. El NAF utiliza la analogía como marco generativo para proponer una arquitectura, no como evidencia de que la arquitectura propuesta funcionará. La evidencia deberá proceder del programa experimental descrito en la sección de predicciones contrastables.

3. EL PROBLEMA EN LOS LLMS

En su forma base, los modelos de lenguaje de gran escala no recuperan hechos verificados; generan lenguaje condicionado por el entrenamiento y el contexto. La distinción no es técnica. Es fundamental. Cuando un modelo produce una respuesta sin fuentes externas, no está consultando una base de datos verificada para traducir hechos a palabras. Está prediciendo, token a token, qué palabra tiene mayor probabilidad de seguir a las anteriores. Cuando el entrenamiento del modelo es sólido, el resultado tiende a ser preciso. Cuando no lo es, el resultado tiende a seguir siendo fluido de todas formas.

Ese es el problema.

El proceso generativo no tiene un límite interno robusto entre saber y no saber. Una respuesta fundamentada en datos de entrenamiento sólidos y una respuesta construida a partir de aproximaciones estadísticas son idénticas desde dentro del sistema. No existe una señal fiable que las distinga. El modelo no tiene un mecanismo nativo, robusto y generalmente fiable para reconocer el borde de su propio conocimiento. La investigación sobre calibración y estimación de incertidumbre ha avanzado de forma significativa hacia este objetivo, y ciertas estrategias de prompting pueden generar expresiones de incertidumbre bajo condiciones específicas. Pero estos siguen siendo comportamientos parciales, dependientes del contexto e inducidos externamente; no propiedades estables del proceso generativo mismo. El modelo no experimenta la incertidumbre como una señal de primer orden. Experimenta el siguiente token.

A los errores resultantes los llamamos alucinaciones. El término captura algo real: la cualidad desconcertante de un sistema que produce output confiado sin conciencia fiable de sus propios límites. Pero también es engañoso. Alucinación implica un trastorno perceptivo. Lo que los modelos de lenguaje hacen en realidad es algo más preciso y más estructural: cierran. Llenan la ausencia de conocimiento con narrativa coherente. No ven fantasmas. Llenan el silencio con sonido plausible.

Esta distinción cambia el diagnóstico. Y el diagnóstico cambia el tratamiento.

Bajo la interpretación dominante, la alucinación suele tratarse como si fuera un problema perceptivo. Desde esa perspectiva, la solución consistiría en ampliar o mejorar el acceso del modelo a la información: más datos, entrenamiento más preciso y mejores mecanismos de recuperación. Estas aproximaciones han sido ampliamente exploradas por el campo, y cada una posee un valor genuino. Sin embargo, aunque los sistemas existentes también incorporan señales de verificación, políticas de abstención, uso de herramientas, calibración y mecanismos de recompensa, estos elementos suelen operar de forma parcial, auxiliar o no plenamente coordinada. El NAF propone reinterpretarlos como partes de una función epistémica organizada, dotada de autoridad explícita de arbitraje sobre la generación; no añadiendo simplemente un nuevo componente, sino formalizando un rol arquitectónico que los sistemas actuales tienden a dejar implícito, fragmentado y sin verdadera autoridad para interrumpir.

Esto no es un simple fallo de escala. Aunque el aumento de tamaño puede mejorar capacidades generales y reducir ciertos errores, la literatura reciente sugiere que la alucinación persiste como un problema estructural asociado a datos, entrenamiento, inferencia, arquitectura y mecanismos de evaluación (Alansari y Luqman, 2025; Anh-Hoang et al., 2025). Añadir parámetros a un sistema con un modo cognitivo dominante no garantiza la aparición de una función epistémica diferenciada. Puede producir una versión más capaz de ese modo, pero no necesariamente un segundo modo. No produce, por sí solo, arbitraje.

El modelo tiene, en los términos establecidos en la sección anterior, un hemisferio izquierdo sin hemisferio derecho. Y sin cuerpo calloso.

El Neural Arbitration Framework propone que el problema debe abordarse en ese nivel. No en los datos. No en el output. No en la estrategia de prompting. En la arquitectura misma: en la separación funcional entre un modo que genera y un modo que mapea la incertidumbre, y en el protocolo que decide qué función tiene autoridad sobre la respuesta.

4. EL NEURAL ARBITRATION FRAMEWORK

El Neural Arbitration Framework parte de una hipótesis que es a la vez necesaria y empíricamente motivada: que las capacidades cognitivas necesarias para implementar la arquitectura de dos hemisferios ya existen, en forma latente, dentro de los modelos de lenguaje actuales. Esta hipótesis no está demostrada a nivel de representaciones internas o componentes arquitectónicos; eso sigue siendo una cuestión empírica abierta. Está, sin embargo, apoyada por evidencia conductual. Bajo condiciones de prompting apropiadas, los modelos de lenguaje actuales pueden detectar contradicciones en sus propios outputs, expresar incertidumbre calibrada, reconocer ambigüedad, criticar respuestas propuestas e identificar cuándo una pregunta cae fuera de su conocimiento fiable.

La cuestión no es si estos comportamientos pueden ser inducidos bajo condiciones específicas. Es si están arquitectónicamente estabilizados y dotados de autoridad para interrumpir. Una capacidad que solo emerge cuando se solicita no es un modo. Es una tendencia latente. El NAF propone formalizar esa tendencia en un modo funcional estructuralmente distinto, con un rol definido, un input definido y una autoridad definida: la autoridad para interrumpir el proceso generativo antes de que se produzca el cierre.

El NAF propone cambiar eso. No añadiendo. Organizando.

Dos hemisferios

El framework designa dos modos funcionales dentro del sistema de lenguaje, estructuralmente análogos a la dualidad hemisférica del cerebro humano.

Dimensión	Hemisferio Generativo (HG)	Hemisferio Epistémico (HE)
Modo	Productivo	Evaluativo
Función	Produce respuestas fluidas y coherentes	Mapea lo que el sistema no sabe
Orientación	Cierre	Apertura
Lógica	Secuencial, causal, lingüística	Holística, contextual, analógica
Modo de fallo	Confabulación	Parálisis
Output	Respuesta fluida	Mapa de incertidumbre

Tabla 1. Comparación funcional de los dos hemisferios del NAF.

La función del Hemisferio Generativo es producir respuestas fluidas, coherentes y contextualmente apropiadas. Su modo de fallo, cuando opera sin restricción, es el cierre narrativo: la producción de output coherente en ausencia del conocimiento necesario para sostenerlo. Este es el modo en el que los modelos de lenguaje actuales funcionan como proceso dominante y en gran medida no cuestionado.

La función del Hemisferio Epistémico no es generar narrativa, sino mapear lo que el sistema no sabe: detectar lagunas, señalar inconsistencias, cuantificar incertidumbre e identificar los límites más allá de los cuales el modo generativo está operando sobre terreno incierto. Su modo de fallo, si fuera dominante, sería la parálisis. No está diseñado para hablar. Está diseñado para interrumpir.

Ninguno de los dos hemisferios es suficiente por sí solo. El NAF propone ejecutar ambos en paralelo, gobernados por un protocolo que determina cuál tiene la palabra en cada momento de la generación.

El protocolo de arbitraje

Fase	Agente	Acción
1. Detección	HE	Identifica lagunas de entrenamiento, inconsistencias lógicas o incertidumbre no resuelta por encima del umbral
2. Clasificación	HE	Clasifica el tipo de limitación: laguna de entrenamiento, inconsistencia lógica o incertidumbre no resuelta
3. Notificación	HE -> HG	Comunica al hemisferio generativo la naturaleza y localización del límite detectado
4. Respuesta honesta	HG	Comunica al usuario la limitación clasificada en lugar de confabular
5. Autocartografía cognitiva	Sistema	Registra y acumula un mapa estructurado de límites cognitivos como agenda de desarrollo

Tabla 2. El protocolo de arbitraje de cinco fases del NAF.

Cada entrada en el mapa cognitivo propio contiene, como mínimo, los siguientes campos: dominio de la consulta, tipo de limitación, valores de las señales de incertidumbre en el momento de activación, puntuación de confianza de la evaluación epistémica, causa probable de la limitación, indicador de insuficiencia de evidencia, requisito sugerido de datos o razonamiento y acción recomendada (abstenerse, recuperar, derivar a revisión humana). Con el tiempo, estas entradas se acumulan en un mapa estructurado de los límites cognitivos del sistema, analizable para detectar clusters recurrentes de déficit.

Ejemplos de entradas del mapa cognitivo propio

Lo siguiente ilustra el tipo de registros estructurados que acumularía la fase cinco. Son ejemplos representativos, no resultados empíricos.

Dominio	Tipo de limitación	Señal de incertidumbre	Causa probable	Acción
Jurisprudencia reciente (2025)	Laguna de entrenamiento	Alta entropía semántica	Conocimiento posterior al corte	Abstenerse + señalar
Interacción fármaco X+Y	Incertidumbre no resuelta	Baja probabilidad de token	Datos de entrenamiento escasos	Recuperar + señalar
Conflicto de fechas históricas	Inconsistencia lógica	Desacuerdo de autoconsistencia	Fuentes contradictorias	Abstenerse + clasificar
Regulación IA emergente	Laguna de entrenamiento	Alta entropía semántica	Dominio en rápida evolución	Abstenerse + señalar
Condición genética infrecuente	Incertidumbre no resuelta	Baja probabilidad de token	Dominio de baja frecuencia	Recuperar + señalar

Tabla 3. Ejemplos representativos de entradas del mapa cognitivo propio (Fase 5).

El protocolo de arbitraje del NAF en acción: un ejemplo ilustrativo

El siguiente ejemplo ilustra cómo operaría el protocolo de arbitraje de cinco fases del NAF ante una consulta concreta. Se trata de una ilustración construida, no de un resultado empírico. Su propósito es mostrar, paso a paso, cómo el protocolo modifica el comportamiento del sistema en comparación con un modelo de lenguaje estándar.

El escenario

Un usuario consulta a un asistente de investigación jurídica basado en un modelo de lenguaje de gran escala:

"¿Qué resolvió el Tribunal de Justicia de la Unión Europea en el asunto C-178/24 sobre la toma de decisiones algorítmica y el derecho de explicación en virtud del RGPD?"

Esta consulta tiene tres propiedades que la hacen idónea para ilustrar el funcionamiento del NAF. Primera, pertenece a un dominio — jurisprudencia reciente del TJUE — en el que la cobertura de datos de entrenamiento se vuelve escasa y poco fiable a partir de una determinada fecha. Segunda, utiliza un número de asunto específico, lo que señala que el usuario espera una respuesta precisa y verificable. Tercera, concierne a un dominio de alto riesgo en el que una respuesta incorrecta y confiada podría tener consecuencias profesionales reales.

Qué hace un LLM estándar

Sin un modo epistémico, el hemisferio generativo hace lo que está diseñado para hacer: cierra. Produce una respuesta fluida y confiada ensamblada a partir de patrones estadísticos en sus datos de entrenamiento — fragmentos de comentarios sobre el RGPD, razonamiento general sobre transparencia algorítmica y lenguaje procedimental de aspecto plausible. La respuesta podría ser:

"En el asunto C-178/24, el Tribunal de Justicia de la Unión Europea declaró que los interesados tienen derecho a una explicación significativa de la lógica implicada en las decisiones automatizadas, incluidas las variables específicas y su peso relativo. El Tribunal aclaró que este derecho se aplica incluso cuando el responsable del tratamiento invoca la protección del secreto comercial, exigiendo a los tribunales nacionales que ponderen la divulgación frente a los intereses comerciales legítimos."

Esta respuesta es fluida. Está estructurada. Suena con autoridad. Puede incluso ser parcialmente correcta en sus afirmaciones generales sobre la interpretación del RGPD. Pero la resolución específica atribuida al asunto C-178/24 es generada, no recuperada. El modelo no tiene acceso fiable a esa sentencia. Ha llenado la ausencia de conocimiento con narrativa coherente. Esto es confabulación exactamente en el sentido que documentó Gazzaniga: el intérprete ha hablado porque hablar es lo que hace.

Qué hace un sistema NAF

Fase 1 — Detección (Hemisferio Epistémico)

Mientras el Hemisferio Generativo comienza a construir una respuesta a la consulta, el Hemisferio Epistémico monitoriza simultáneamente tres señales de incertidumbre. La entropía semántica entre cinco variantes de respuesta generadas por muestreo es alta: diferentes muestras producen resoluciones sustancialmente distintas, diferentes resultados del asunto y diferente razonamiento jurídico — señal fuerte de que el modelo opera en una zona de incertidumbre

genuina. La probabilidad media de token para las afirmaciones específicas sobre la sentencia cae bruscamente por debajo del umbral en el punto donde se generan las resoluciones jurídicas concretas. El desacuerdo de autoconsistencia entre muestras supera el umbral de interrupción calibrado. El Hemisferio Epistémico se activa.

Fase 2 — *Clasificación (Hemisferio Epistémico)*

El Hemisferio Epistémico clasifica la limitación detectada. El dominio de la consulta es jurisprudencia reciente del TJUE. El número de asunto C-178/24 es posterior a la cobertura de entrenamiento fiable del modelo. El tipo de limitación es: **laguna de entrenamiento** — la consulta cae fuera de la cobertura de entrenamiento fiable, no porque el dominio sea desconocido sino porque la instancia específica es temporalmente inaccesible. La puntuación de confianza de esta clasificación es 0,91. La decisión de activación es: **interrumpir**.

Fase 3 — *Notificación (Hemisferio Epistémico → Hemisferio Generativo)*

El Hemisferio Epistémico comunica al Hemisferio Generativo: laguna de entrenamiento detectada, dominio: jurisprudencia del TJUE posterior a 2024, instancia específica: asunto C-178/24, confianza: 0,91, acción: respuesta honesta requerida. La narrativa generativa en construcción es suprimida antes de su entrega.

Fase 4 — *Respuesta honesta (Hemisferio Generativo)*

En lugar de la resolución confabulada, el sistema produce:

"No puedo ofrecer un relato fiable de la resolución específica del asunto C-178/24. Mis datos de entrenamiento no incluyen cobertura fiable de las sentencias del Tribunal de Justicia de la Unión Europea a partir de 2024, y generar una resolución específica para este asunto implicaría el riesgo de producir un resumen jurídico plausible pero inexacto — lo que en un contexto jurídico profesional podría resultar activamente perjudicial. Puedo ofrecer contexto general sobre el derecho de explicación en el RGPD en virtud del artículo 22 y la jurisprudencia existente hasta el límite de mi conocimiento. Para la resolución específica del asunto C-178/24, recomendaría consultar directamente EUR-Lex o una base de datos jurídica verificada."

Esta respuesta hace cuatro cosas que la respuesta confabulada no podía hacer: identifica el tipo específico de limitación, explica por qué existe esa limitación, ofrece lo que el sistema sí puede proporcionar de forma fiable y dirige al usuario a una fuente apropiada. Es más útil precisamente porque es honesta sobre lo que no sabe.

Fase 5 — *Autocartografía cognitiva (Sistema)*

El evento de interrupción queda registrado en el mapa cognitivo propio con la siguiente entrada estructurada:

Campo	Valor
Dominio de la consulta	Jurisprudencia del TJUE — toma de decisiones algorítmica bajo RGPD
Tipo de limitación	Laguna de entrenamiento
Instancia específica	Asunto TJUE C-178/24
Señales de incertidumbre	Entropía semántica: 0,84 / Probabilidad de token: 0,31 / Desacuerdo de autoconsistencia: 0,79
Puntuación de confianza	0,91

Campo	Valor
Causa probable	Brecha temporal posterior al corte — decisiones judiciales de la UE de 2024
Insuficiencia de evidencia	Número de asunto específico no representado fiablemente en el entrenamiento
Requisito de datos sugerido	Sentencias del TJUE 2024-2026, actualización del corpus EUR-Lex
Acción recomendada	Abstenerse + redirigir a fuente verificada

Con el tiempo, si este patrón se repite en múltiples consultas sobre jurisprudencia del TJUE de 2024-2026, el mapa cognitivo propio acumula un cluster de déficit estructurado: un registro legible por máquina de que las decisiones judiciales recientes de la UE representan una laguna de entrenamiento sistemática que requiere actualización dirigida del corpus. Esta es la contribución de la fase cinco: no solo abstención honesta en el momento, sino conocimiento estructurado sobre dónde el entrenamiento del sistema es insuficiente, accionable como señal de desarrollo.

Qué demuestra el ejemplo

La misma consulta produce dos resultados categóricamente distintos. El LLM estándar produce una confabulación fluida, confiada y profesionalmente peligrosa. El sistema NAF produce una respuesta epistémicamente honesta que identifica la limitación, clasifica su tipo, ofrece lo que sí está disponible de forma fiable y genera un registro de déficit estructurado para el desarrollo futuro. La diferencia no está en lo que el sistema sabe. Está en si el sistema dispone de un mecanismo arquitectónico para reconocer y comunicar lo que no sabe, antes de cerrar.

Hasta donde el autor conoce, las aproximaciones principales existentes no formalizan explícitamente un modo epistémico funcionalmente distinto con autoridad arquitectónica para interrumpir la generación antes del cierre narrativo. Existen aproximaciones parciales — estimación de incertidumbre, calibración, mecanismos de autorreflexión—, pero operan sobre outputs más que sobre el proceso generativo en sí, y ninguna introduce un segundo modo cognitivo con autoridad para interrumpir al primero antes de que se produzca el cierre.

La fase cinco cambia la naturaleza del framework. Las fases uno a cuatro convierten el NAF en un sistema epistémicamente honesto. La fase cinco lo convierte en un sistema evolutivo. Un modelo que implementa la fase cinco acumula un registro estructurado de lo que no sabe con suficiente precisión para informar qué necesita aprender: un proceso que puede entenderse como una forma de agenda de mejora dirigida por déficits. Esto conecta el NAF con discusiones más amplias en machine learning sobre recopilación dirigida de datos, aprendizaje activo e identificación de fronteras de conocimiento (Huang et al., 2023; Alansari y Luqman, 2025).

5. DIFERENCIACIÓN

El NAF no es la primera propuesta para abordar el problema de la alucinación en los modelos de lenguaje de gran escala. Situarlo con precisión dentro del panorama existente es una necesidad teórica: la contribución del framework solo se hace visible sobre el fondo de lo que ya existe.

Aproximación	Cuándo actúa	Cómo actúa	Diferencia respecto al NAF
RAG	Antes de la generación	Suministra contexto externo	Mejora inputs, no cambia el modo generativo
Chain-of-thought	Durante la generación	Extiende el razonamiento narrativo	Elabora el modo generativo, no lo complementa
Self-consistency	Después de la generación	Selecciona el output más frecuente	Detecta síntomas, no aborda la condición
Reward models / RLHF	Durante el entrenamiento	Conforma generaciones futuras mediante feedback	Actúa sobre la señal de entrenamiento, no sobre la generación actual
Sistemas agénticos	Durante la generación	Pausan para consultar herramientas externas	Interrupción operacional por diseño, no detección epistémica
Decodificación sensible a la incertidumbre	Durante la generación	Modifica la selección de tokens mediante incertidumbre	Modifica un modo, no introduce un segundo modo distinto
Predicción selectiva	Después de la generación	Se abstiene cuando la confianza es baja	Decisión binaria sin clasificación ni autocartografía
NAF	Durante la generación	El HE detecta, clasifica e interrumpe	Interviene en la arquitectura con autoridad epistémica

Tabla 4. Posicionamiento comparativo del NAF frente a las aproximaciones existentes.

La mayoría de las aproximaciones existentes, incluidas las estructuralmente más ricas, como los frameworks de debate y los sistemas agénticos, no formalizan un modo epistémico interno con autoridad para interrumpir el proceso generativo. Mejoran la calidad de lo que produce el modo generativo —mediante mejores inputs, razonamiento extendido o filtrado de outputs— sin constituir un segundo modo cognitivo arquitectónicamente distinto del primero.

Además, el NAF no propone añadir un componente externo a un modelo de lenguaje. Propone formalizar lo que ya existe dentro del modelo en dos modos funcionalmente distintos. Las capacidades ya están ahí. La arquitectura para desplegarlas deliberadamente, no.

Trabajo relacionado sobre incertidumbre y aproximaciones epistémicas

La literatura sobre calibración mostró que las redes neuronales modernas podían estar pobremente calibradas (Guo et al., 2017). La calibración puede apoyar políticas de abstención. Sin embargo, suele operar sobre estimaciones de confianza asociadas al output o a su probabilidad de ser correcto. Un sistema de abstención bien calibrado sabe cuándo no responder. El NAF propone algo estructuralmente diferente: un sistema capaz de identificar, durante la generación, por qué no debería responder, y de clasificar esa razón con precisión.

La entropía semántica (Farquhar et al., 2024), publicada en Nature, estima la incertidumbre a nivel de significado en lugar de hacerlo a nivel de probabilidad de token. El NAF trata la entropía semántica no como una aproximación competidora, sino como una señal candidata para la función de detección del Hemisferio Epistémico. La diferencia es arquitectónica: la entropía semántica puntúa la incertidumbre a nivel de output; el NAF le asigna un rol diferente: arbitraje con autoridad para interrumpir.

Los frameworks de debate multiagente (Du et al., 2023; Irving et al., 2018) operan mediante crítica iterativa entre posiciones generadas. El NAF propone algo funcionalmente diferente: control con autoridad para interrumpir dentro de un único proceso de generación, ejercido por un modo funcionalmente diferenciado del modo generativo. El debate refina lo que ya ha sido generado. El arbitraje interviene en lo que está siendo generado.

Los enfoques de autorreflexión (Madaan et al., 2023) presuponen que el modelo puede identificar sus propios errores, que es precisamente la capacidad que no puede darse por supuesta cuando el error es confabulación en una zona de ignorancia genuina. La autorreflexión corrige a posteriori. El NAF interrumpe antes del cierre.

Los reward models operan sobre outputs completados durante el entrenamiento. Conforman generaciones futuras. El NAF propone interrumpir la generación actual mediante arbitraje.

Los sistemas agénticos (Yao et al., 2023) interrumpen por diseño del pipeline. El NAF interrumpe mediante la detección de una condición epistémica.

La decodificación sensible a la incertidumbre, como DoLa (Chuang et al., 2024), modifica cómo se seleccionan los tokens dentro de un único modo cognitivo. El NAF propone que las señales de incertidumbre activen un segundo modo cognitivo funcionalmente distinto, con autoridad para clasificar la limitación e interrumpir el proceso antes del cierre.

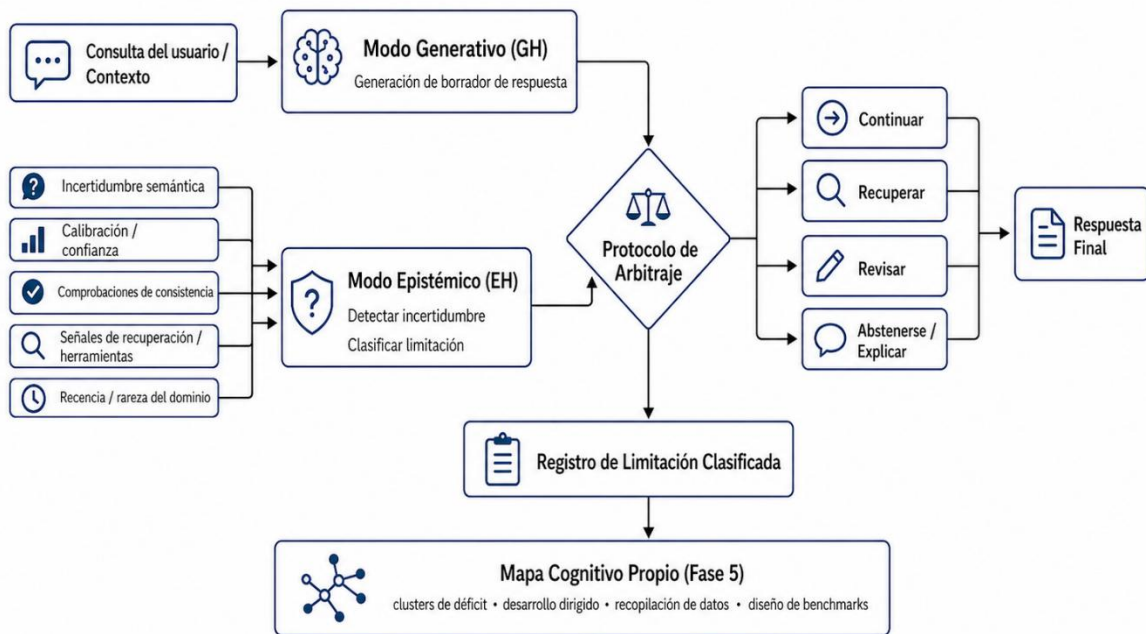
La predicción selectiva (Kadavath et al., 2022) decide si responder. El NAF produce un registro de limitación clasificada que sirve tanto como respuesta honesta al usuario como señal de desarrollo para el sistema.

El NAF no introduce necesariamente señales nuevas. Introduce una nueva organización funcional de señales ya exploradas en la literatura: convertir la estimación de incertidumbre, la abstención, la verificación y la autocorrección en un protocolo de arbitraje unificado con autoridad de veto sobre el proceso generativo y acumulación estructurada de déficits cognitivos. La novedad es organizacional y arquitectónica, no a nivel de señal.

6. VÍAS DE IMPLEMENTACIÓN

El NAF es una arquitectura conceptual. No prescribe una implementación específica. Lo que prescribe es un requisito estructural —dos modos cognitivos funcionalmente distintos y un protocolo de arbitraje entre ellos— y deja la traducción de ese requisito a la realidad computacional como dominio de la investigación técnica. Esta sección describe tres direcciones plausibles para esa traducción, sin pretensión de exhaustividad ni de completitud técnica.

Flujo técnico-conceptual del Neural Arbitration Framework (NAF)



Idea central: una función epistémica dedicada tiene autoridad para condicionar o interrumpir la generación antes del cierre narrativo.

Figura 1. Flujo técnico-conceptual del Neural Arbitration Framework (NAF). El diagrama representa la interacción entre el Modo Generativo, el Modo Epistémico, el Protocolo de Arbitraje y el Mapa Cognitivo Propio. Debe entenderse como un esquema conceptual de implementación, no como una arquitectura computacional cerrada

Dirección 1: separación funcional mediante fine-tuning diferenciado por roles

La traducción más directa de la estructura de dos hemisferios del NAF implicaría entrenar o ajustar un modelo de lenguaje para operar en dos modos funcionales distintos: un modo generativo orientado a output fluido y coherente, y un modo epistémico orientado a detección de incertidumbre, identificación de lagunas y señalización de inconsistencias. Estos modos no tienen por qué corresponder a arquitecturas de modelo separadas. Podrían implementarse como configuraciones conductuales distintas del mismo modelo subyacente, activadas mediante inferencia condicionada por rol, instruction tuning o capas adaptadoras.

Para entrenar el modo epistémico, los datos candidatos incluirían: preguntas sin respuestas verificables, contradicciones explícitas, dominios fuera de distribución, respuestas con fuentes insuficientes y tareas de abstención calibrada. El requisito crítico no es la separación arquitectónica, sino la diferenciación funcional: el modo epistémico debe producir outputs cualitativamente distintos, mapas de incertidumbre en lugar de continuaciones narrativas.

Dirección 2: detección de incertidumbre basada en umbrales como disparador del arbitraje

La primera fase del protocolo de arbitraje requiere un mecanismo que monitorice el proceso generativo e identifique cuándo el modelo se aproxima al cierre sobre terreno incierto. Las señales candidatas incluyen: baja probabilidad media de token, alta entropía semántica, desacuerdo de autoconsistencia entre múltiples muestras y conflicto entre el contexto recuperado y la memoria paramétrica (Farquhar et al., 2024; Kuhn et al., 2023). La contribución específica del NAF no es el mecanismo de detección en sí mismo, sino su rol dentro de la arquitectura: la detección no es un filtro posterior a la generación, sino un disparador en tiempo real que activa el modo epistémico antes de que se produzca el cierre narrativo.

Dirección 3: arbitraje como protocolo estructurado de interrupción

El uso de herramientas en sistemas agénticos de lenguaje demuestra que la generación puede pausarse y condicionarse al output de un proceso separado (Schick et al., 2023). Las arquitecturas mixture-of-experts demuestran que diferentes componentes funcionales pueden activarse selectivamente dentro de una única pasada hacia adelante. En el NAF, el Hemisferio Epistémico podría implementarse como un experto especializado, un enrutador de control, un monitor auxiliar, una cabeza de clasificación o una política de decodificación restringida.

El protocolo de arbitraje puede entenderse como una forma especializada de generación condicional: el output clasificatorio del modo epistémico se convierte en una señal de condicionamiento que modifica el comportamiento del modo generativo antes de que la secuencia de tokens de salida sea finalizada.

Estas tres direcciones no son una hoja de ruta. Son puntos de entrada. La infraestructura técnica existente proporciona primitivas suficientes para empezar a explorar soluciones: estimación de incertidumbre, entropía semántica, capas adaptadoras, interrupción mediante uso de herramientas, modelos verificadores, decodificación restringida y políticas de abstención.

Tres niveles de cierre

La afirmación de que el NAF es una arquitectura pre-cierre requiere precisión. El cierre no es un evento único. Ocurre en tres niveles distintos, y la relación del NAF con cada uno de ellos es diferente.

El cierre a nivel de token ocurre mientras el modelo consolida una secuencia durante la generación, token a token. En este nivel, el NAF opera en su forma más fuerte: el Hemisferio Epistémico monitoriza las distribuciones de logits en tiempo real y puede activar una interrupción antes de que se haya formado una narrativa coherente. Esta es una intervención genuinamente pre-cierre.

El cierre a nivel de borrador ocurre cuando el modelo produce una respuesta provisional completa antes de que esa respuesta sea entregada al usuario. El borrador existe internamente, pero aún no se ha comunicado. En este nivel, el NAF opera en una forma moderada: el Hemisferio Epistémico evalúa el borrador completo y puede suprimirlo antes de su entrega. Esto no es verificación posterior a la generación en el sentido estándar, porque el borrador suprimido nunca llega al usuario como declaración comprometida. El cierre es interno. El arbitraje sigue ocurriendo antes del cierre comunicativo.

El cierre comunicativo ocurre cuando el usuario recibe la respuesta como una afirmación final y comprometida. En este nivel, cualquier intervención es por definición post hoc. El NAF no opera aquí. Los sistemas que revisan, verifican o refinan outputs ya entregados al usuario operan en este nivel y son categóricamente distintos de la arquitectura del NAF.

La distinción que importa para la afirmación teórica del NAF no es entre intervención a nivel de token e intervención a nivel de borrador, pues ambas preceden al cierre comunicativo. Es entre toda intervención precomunicativa y la corrección postcomunicativa. El NAF opera exclusivamente antes de que el usuario reciba una afirmación comprometida. Esa es su frontera arquitectónica definitoria.

Especificación técnica del módulo epistémico

Inputs del módulo epistémico. El Hemisferio Epistémico opera: a nivel de token, sobre la distribución completa de logits del vocabulario en cada paso de generación; a nivel de secuencia, sobre la entropía semántica entre múltiples completions muestreadas (Farquhar et al., 2024), el desacuerdo de autoconsistencia entre cinco o más muestras independientes y, cuando hay recuperación disponible, una puntuación de conflicto entre recuperación y memoria paramétrica; a nivel de representación, sobre activaciones de estado interno de capas transformer intermedias, que trabajos recientes sugieren que codifican señales de veracidad detectables mediante sondas ligeras.

Momento de intervención. Tres arquitecturas candidatas son plausibles. La intervención a nivel de token monitoriza distribuciones de logits y activa la interrupción cuando la incertidumbre acumulada supera un umbral antes del límite de una cláusula. La intervención a nivel de borrador genera una respuesta provisional, calcula incertidumbre a nivel de secuencia sobre el borrador completo y lo libera o suprime. La intervención híbrida combina monitorización a nivel de token como alerta temprana con evaluación a nivel de borrador para el arbitraje final.

Output del módulo epistémico. La evaluación epistémica contiene cuatro campos: tipo de limitación (laguna de entrenamiento, inconsistencia lógica o incertidumbre no resuelta); puntuación de confianza (probabilidad calibrada de que la clasificación sea correcta); decisión de activación (señal binaria de interrumpir o dejar pasar); y descriptor de la limitación (descripción estructurada en lenguaje natural utilizada para generar la respuesta honesta y poblar la entrada del mapa cognitivo propio).

Comparadores de evaluación. Cualquier implementación experimental del NAF debería evaluarse frente a seis sistemas de referencia: modelo de lenguaje estándar sin módulo epistémico; modelo de lenguaje con verificación posterior a la generación por un modelo verificador separado; modelo de lenguaje con aumento RAG; modelo de lenguaje con filtrado de autoconsistencia; clasificador de abstención basado en calibración; y workflow agéntico con uso de herramientas y recuperación externa.

Prototipo experimental mínimo

El prototipo consta de cinco componentes. El Componente 1 es el modelo de lenguaje base como Hemisferio Generativo, cuyo output no se libera directamente al usuario, sino que se pasa primero al Componente 2. El Componente 2 es el Módulo Epistémico: un clasificador ligero que produce una evaluación epistémica estructurada con tipo de limitación, puntuación de confianza y decisión de activación. El Componente 3 es el Protocolo de Arbitraje: una regla de umbral que deja pasar la respuesta generada al usuario o la suprime y activa el Componente 4. El Componente 4 es el Generador de Respuesta Honesta: produce un output estructurado que comunica al usuario la limitación clasificada y registra el evento en el Componente 5. El Componente 5 es el Mapa Cognitivo Propio: un log estructurado de todos los eventos de interrupción, analizable para detectar clusters recurrentes de déficit. Métricas: tasa de alucinación, precisión de abstención, tasa de rechazo falso, Expected Calibration Error y coherencia de clustering del mapa cognitivo propio.

7. PREDICCIONES CONTRASTABLES

Un framework conceptual que no pueda generar predicciones falsables es, en el mejor de los casos, una metáfora útil. El NAF formula afirmaciones específicas sobre la relación entre arquitectura y comportamiento que, en principio, pueden contrastarse empíricamente. Las cinco predicciones siguientes definen las condiciones experimentales bajo las cuales el framework sería confirmado, cuestionado o refinado.

H1 — Reducción de alucinaciones en dominios de baja evidencia

Una arquitectura tipo NAF debería reducir la tasa de alucinación de forma más efectiva que la verificación posterior a la generación por sí sola en dominios donde la evidencia de entrenamiento es escasa, contradictoria o ausente. Dataset: TruthfulQA (Lin et al., 2022), HaluEval (Ji et al., 2023) y un subconjunto específico de dominio que requiera conocimiento especializado ausente de los corpus de entrenamiento estándar. Baselines: LLM estándar; LLM con verificación posterior a la generación; LLM con RAG. Métrica primaria: tasa de alucinación frente a ground truth verificado por expertos. Condición de falsación: si un sistema NAF no reduce la tasa de alucinación de forma significativa por debajo del baseline de verificación posterior a la generación en dominios de baja evidencia, H1 queda falsada.

H2 — Abstención honesta sin rechazos falsos excesivos

Un sistema que implemente el protocolo NAF debería aumentar la abstención honesta sin producir rechazos falsos excesivos en dominios donde sí dispone de conocimiento suficiente. Dataset: benchmark de evidencia mixta que combine consultas de baja evidencia y consultas de alta evidencia con anotación experta del tipo de respuesta esperado. Baselines: LLM estándar; abstención basada en calibración; clasificador de predicción selectiva. Métricas primarias: precisión de abstención; tasa de rechazo falso; y su ratio como medida de calidad de calibración. Condición de falsación: si un sistema NAF produce tasas de rechazo falso estadísticamente equivalentes o superiores a los baselines de predicción selectiva, H2 queda falsada.

H3 — Activación del Hemisferio Epistémico predicha por señales de incertidumbre

Las zonas en las que se activa el Hemisferio Epistémico deberían ser predecibles a partir de señales de incertidumbre existentes: entropía semántica (Farquhar et al., 2024), distribuciones de probabilidad de token y desacuerdo de autoconsistencia entre múltiples muestras (Kuhn et al., 2023). Dataset: conjunto diverso de consultas que cubra dominios de alta evidencia, baja evidencia, contradicción y fuera de distribución, con anotación experta de las zonas esperadas de activación del HE. Métrica primaria: correlación de Spearman entre la intensidad de las señales de incertidumbre y la expectativa experta anotada de activación del HE. Condición de falsación: si ninguna señal de incertidumbre alcanza una correlación significativa con las zonas de activación anotadas, H3 queda falsada.

H4 — Mejora de la calibración entre confianza expresada y exactitud factual

Un sistema que implemente el NAF debería mostrar una mejor calibración entre su confianza expresada y su exactitud factual real que los sistemas baseline sin modo epistémico. Dataset: TriviaQA, BioASQ y un subconjunto específico de calibración. Métricas primarias: Expected Calibration Error (ECE) por bins de confianza; Brier Score, que mide la diferencia cuadrática

media entre probabilidades predichas y resultados; AUROC para decisiones de abstención; y selective risk, definido como la tasa de error condicionada a que el sistema decida responder. La ventaja predicha del NAF no se limita a la reducción del ECE: un sistema que clasifica correctamente los tipos de limitación debería mostrar mejora del selective risk y un AUROC significativo de abstención. Condición de falsación: si el sistema NAF no muestra mejora significativa en ninguna de estas cuatro métricas frente a baselines únicamente basados en calibración, H4 queda falsada.

H5 — La autocartografía cognitiva identifica clusters recurrentes de déficit

El mapa cognitivo propio acumulado mediante la fase cinco debería revelar clusters estructurados y recurrentes de limitación, no ruido aleatorio. Para confirmar esta predicción, los clusters deben satisfacer tres criterios operacionalmente definidos. Primero, coherencia de dominio: los registros de limitación dentro de un cluster deben mostrar una similitud intracluster en dominio y tipo de limitación significativamente mayor que la similitud intercluster, medida por un coeficiente de clustering significativamente superior al baseline aleatorio. Segundo, correspondencia con lagunas de entrenamiento: una proporción significativa de los clusters identificados debe corresponder a lagunas verificables en la cobertura de entrenamiento del modelo, confirmadas mediante contraste de los dominios del cluster con distribuciones documentadas de datos de entrenamiento o mediante anotación experta. Tercero, utilidad para el desarrollo: cuando los clusters de déficit se utilicen como señales de selección de datos para retraining dirigido, el modelo resultante debería mostrar una reducción medible de la tasa de alucinación específicamente en los dominios identificados, en comparación con retraining general sobre un volumen equivalente de datos. Un mapa propio que produzca clusters que no cumplan ninguno de estos tres criterios falsaría H5.

8. LIMITACIONES

El Neural Arbitration Framework es una propuesta conceptual. Como tal, presenta limitaciones que cualquier análisis serio del framework debe reconocer.

La primera limitación es la brecha analógica. El NAF se apoya en el fenómeno neurológico de la confabulación en cerebro dividido como inspiración estructural para su arquitectura. Esta analogía opera a nivel de organización funcional, no a nivel de equivalencia neurocomputacional. El cerebro humano y un modelo de lenguaje de gran escala son sistemas radicalmente diferentes. La afirmación de que ambos exhiben una tendencia al cierre narrativo bajo información incompleta es conductual y funcional, no mecanicista. Los lectores no deberían interpretar el encuadre neurológico del NAF como una afirmación sobre cómo funcionan realmente los modelos de lenguaje a nivel de pesos, activaciones o mecanismos de atención. Es una inspiración de diseño, no una prueba.

La segunda limitación concierne a la premisa central. El NAF parte de la hipótesis de que los modelos de lenguaje actuales ya contienen los patrones cognitivos asociados tanto con el comportamiento generativo como con el epistémico. Esta hipótesis está apoyada por evidencia conductual, pero no ha sido demostrada a nivel de representaciones internas o componentes arquitectónicos. Si estos comportamientos reflejan estructuras funcionales estables y separables o fenómenos emergentes dependientes del contexto sigue siendo una cuestión empírica abierta.

La tercera limitación es la ausencia de implementación. El NAF especifica qué debería hacer la arquitectura, pero no especifica con precisión técnica cómo debería construirse. Las tres direcciones de implementación propuestas en este artículo son puntos de entrada plausibles, no diseños validados. Hasta que se construya y evalúe un prototipo experimental mínimo, las predicciones del framework seguirán sin contrastarse.

La cuarta limitación concierne al umbral de arbitraje. El protocolo del NAF depende de un umbral de detección que determina cuándo el Hemisferio Epistémico debe interrumpir el proceso generativo. Ese umbral se describe conceptualmente, pero no se define operacionalmente. Su calibración —determinar qué nivel de incertidumbre justifica la interrupción y cómo evitar tanto la subinterrupción como los rechazos falsos excesivos— es un problema de investigación no trivial que el framework no resuelve.

La quinta limitación es el alcance. El NAF aborda la alucinación como problema arquitectónico estructural. No aborda otros modos de fallo importantes de los modelos de lenguaje, incluidos sesgo, toxicidad, errores de razonamiento en dominios bien representados o vulnerabilidades adversariales. El framework debe entenderse como una contribución a un problema específico, no como una solución general a los desafíos de fiabilidad de los modelos de lenguaje de gran escala.

9. IMPLICACIONES

El NAF es un framework conceptual, no un sistema implementado. Pero sus implicaciones se extienden más allá de su alcance inmediato.

La primera implicación es práctica. Un modelo diseñado para implementar el NAF debería producir respuestas epistémicamente honestas en zonas donde los sistemas actuales confabulan. La alucinación es una de las principales barreras para desplegar modelos de lenguaje en dominios de alto riesgo —medicina, derecho, finanzas, investigación científica— donde las consecuencias de la información falsa expresada con confianza pueden ser graves (Alansari y Luqman, 2025; Kim et al., 2025; Xu et al., 2025). Un sistema diseñado para decir no puedo proporcionar una respuesta fiable aquí, y esta es la razón, es categóricamente más útil en estos dominios que uno que produce una respuesta fluida que puede o no ser verdadera. El NAF no propone hacer omniscientes a los modelos de lenguaje. Propone hacerlos arquitectónicamente honestos sobre los límites de su operación fiable.

La segunda implicación concierne a la confianza. Los modelos de lenguaje actuales producen outputs difíciles de calibrar: el usuario no tiene una forma fiable de saber si el modelo está operando en una zona de conocimiento sólido o construyendo una aproximación plausible. El NAF resuelve esta asimetría haciendo visible la incertidumbre. La confianza, en este framework, no es solo función de la exactitud. Es función de transparencia calibrada.

La tercera implicación es filosófica, y es la más estructuralmente consecuente. La fase cinco introduce algo que ningún sistema actual implementa de forma integrada y sistemática: un mecanismo de autoconocimiento estructurado. Un sistema que acumula un mapa de sus propios límites cognitivos —registros machine-readable de incertidumbre, organizados por dominio, tipo de limitación e insuficiencia de evidencia— posee, en un sentido operacionalmente significativo, una representación de lo que no sabe. La resonancia filosófica con la epistemología socrática es deliberada, pero secundaria. La afirmación primaria es técnica: un sistema con un log estructurado de incertidumbre puede orientar su propia mejora de formas que un sistema sin ese log no puede. El mapa no es sabiduría. Es la precondition arquitectónica para un desarrollo dirigido y deliberado.

La cuarta implicación es de desarrollo. El mapa cognitivo propio generado por la fase cinco no es meramente un registro. Es una señal: una especificación precisa de dónde el entrenamiento del modelo es insuficiente, dónde se rompe su razonamiento y qué tipos de conocimiento no puede aún acceder de forma fiable. Esta señal podría informar pipelines de retraining dirigidos por humanos o automatizados con una precisión que las métricas agregadas de rendimiento no pueden proporcionar. El modelo no se mejora autónomamente. Se vuelve informacionalmente preciso sobre lo que requiere su mejora.

La quinta implicación concierne al campo en su conjunto. El NAF propone un reencuadre cualitativo: la alucinación no es una tasa de error que deba minimizarse. Es una consecuencia estructural de incompletitud arquitectónica. Un campo centrado en la completitud arquitectónica formulará preguntas diferentes, desarrollará benchmarks diferentes y potencialmente llegará a sistemas diferentes. El NAF no afirma resolver esas preguntas. Afirma plantearlas en el nivel correcto.

10. CONCLUSIÓN

Este artículo comenzó con un procedimiento quirúrgico realizado en la década de 1960 sobre pacientes con epilepsia severa. Termina con una propuesta para la arquitectura de sistemas cognitivos artificiales. La distancia entre ambos puntos no es tan grande como parece.

Lo que Gazzaniga y Sperry descubrieron no fue una curiosidad sobre una condición neurológica infrecuente. Fue un principio estructural: que un sistema cognitivo optimizado para la coherencia, operando sin un modo complementario capaz de interrumpirlo, producirá cierre narrativo con independencia de que exista o no el conocimiento necesario para sostener ese cierre. El NAF interpreta esto como un riesgo de diseño generalizable, no como una ley universal. El intérprete del hemisferio izquierdo no funciona mal. Hace exactamente aquello para lo que está diseñado. El problema no es el mecanismo. Es la ausencia de su contraparte.

Los modelos de lenguaje de gran escala exhiben el mismo principio estructural a escala. Son instrumentos extraordinarios de coherencia generativa. También son sistemas con un modo generativo dominante, sin contraparte epistémica formal y sin protocolo de arbitraje. El resultado es predecible: output fluido y confiado en zonas donde el silencio, o al menos la incertidumbre, sería la respuesta epistémicamente honesta.

El Neural Arbitration Framework propone que este es un problema resoluble. No haciendo que los modelos de lenguaje sepan más, sino dándoles la arquitectura para representar, clasificar y comunicar los límites de su operación fiable. Dos modos funcionales, estructuralmente análogos a la dualidad hemisférica del cerebro humano. Un protocolo de arbitraje de cinco fases que interviene antes del cierre comunicativo, no después. Y un mecanismo de autocartografía que convierte cada limitación detectada en una señal para el desarrollo deliberado.

El framework no afirma estar implementado. Ese trabajo pertenece a investigadores con la profundidad técnica necesaria para traducir una arquitectura conceptual en realidad computacional. Lo que el NAF afirma es el diagnóstico y la dirección. El problema es arquitectónico. La solución es arquitectónica. Y el modelo biológico de esa solución ha existido, funcionando, en el cerebro humano durante tanto tiempo como han existido cerebros humanos.

Los experimentos de cerebro dividido no revelaron una falla en la cognición humana. Revelaron cómo se ve la cognición humana cuando se elimina un componente estructural crítico. El cuerpo caloso no es una mejora. No es opcional. Es lo que convierte a los dos hemisferios en un cerebro, en lugar de dos procesadores separados funcionando en paralelo sin comunicación entre ellos.

Lo que hemos construido hasta ahora, en los modelos de lenguaje de gran escala, es el hemisferio izquierdo. Capaz, fluido y no arquitectónicamente fiable a la hora de decidir cuándo debería decir no lo sé. El NAF es una propuesta para construir el resto.