TAXAL: explicabilidad cognitiva, funcional y causal para LLMs agénticos

David Herrera-Poyatos[1], Carlos Peláez-González[1], Cristina Zuheros[1], Virilo Tejedor[1], Rosana Montes[2], and Francisco Herrera[1,3]

[1] Department of Computer Science and Artificial Intelligence, Andalusian Institute of Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain
divadhp@ugr.es
[2] Department of Software Engineering, Andalusian Institute of Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain
[3] ADIA Lab, Abu Dhabi, United Arab Emirates

Abstract. Este trabajo recoge la contribución principal del artículo *TAXAL Framework: Triadic Fusion of Cognitive, Functional, and Causal Dimensions for Explainability in Agentic LLMs*, publicado en *Information Fusion* [1]. La aportación central es TAXAL (*Triadic Alignment for eXplainability in Agentic LLMs*), un meta-marco operacional para diseñar, organizar y evaluar explicaciones en modelos de lenguaje con comportamiento agéntico. TAXAL articula la explicabilidad mediante tres dimensiones complementarias: *cognitiva* (comprensión del usuario), *funcional* (utilidad en el flujo de trabajo) y *causal* (trazabilidad y contrastabilidad del razonamiento). El marco adapta las explicaciones a desarrolladores, auditores, expertos de dominio y usuarios finales. Se valida empíricamente en un escenario médico con 200 viñetas clínicas de MedQA y 800 explicaciones evaluadas mediante un protocolo de tres capas. Los resultados muestran que TAXAL permite modular explicaciones de forma controlada según audiencia y objetivo explicativo, contribuyendo a una visión de la explicabilidad como infraestructura sociotécnica para sistemas de IA agéntica supervisables y responsables.

Keywords: Explicabilidad · XAI · Modelos de lenguaje · LLMs agénticos · Interacción humano-IA · TAXAL.

1 Motivación

Los modelos de lenguaje de gran tamaño han pasado de utilizarse principalmente en tareas de generación de texto, resumen o traducción a integrarse en sistemas de asistencia experta, razonamiento y toma de decisiones. En paralelo, el desarrollo de LLMs capaces de planificar, invocar herramientas y ejecutar tareas en varios pasos ha desplazado parte del debate desde la explicación de una salida aislada hacia la explicación de comportamientos más complejos.

Este cambio es especialmente relevante en dominios de alto impacto. En salud, derecho, educación o servicios públicos, una explicación inadecuada puede

inducir sobreconfianza, dificultar la revisión humana o impedir que una decisión sea cuestionada. Además, los distintos actores implicados —desarrolladores, auditores, expertos de dominio y usuarios finales— no necesitan la misma explicación ni con el mismo nivel de detalle, trazabilidad o lenguaje. Aunque la literatura en inteligencia artificial explicable ha propuesto atribuciones, explicaciones contrafactuales, razonamientos en lenguaje natural, interfaces conversacionales y mecanismos de auditoría, estas aproximaciones suelen centrarse en una parte concreta del problema. En LLMs agénticos, las explicaciones deben coordinar comprensión humana, utilidad práctica y trazabilidad del razonamiento. TAXAL responde a esta necesidad conectando diseño de explicaciones, evaluación multidimensional y despliegue sensible al rol del usuario.

2 El marco TAXAL

TAXAL propone una visión triádica de la explicabilidad en modelos de lenguaje agénticos. Su objetivo no es sustituir técnicas concretas de XAI, sino ofrecer una estructura que permita seleccionarlas, combinarlas y evaluarlas de acuerdo con el contexto de uso, el riesgo de la decisión y el perfil del usuario.

La **dimensión cognitiva** se refiere a la capacidad de la explicación para ser entendida por quien la recibe. Incluye el lenguaje utilizado, el nivel de detalle, la carga cognitiva y la adecuación al conocimiento previo del usuario. Una explicación para un paciente no debería emplear la misma terminología que una dirigida a un especialista clínico o a un auditor técnico.

La **dimensión funcional** atiende al papel de la explicación dentro de una tarea concreta: apoyar una decisión, detectar un fallo, depurar un sistema, justificar una recomendación o documentar un proceso. Desde esta perspectiva, no se evalúa solo por su claridad, sino por su utilidad dentro del flujo de trabajo.

La **dimensión causal** hace referencia a la trazabilidad y contrastabilidad del razonamiento asociado a la salida del modelo: análisis contrafactuales, sensibilidad a cambios en la entrada, reconstrucción de pasos de razonamiento o identificación de evidencias relevantes. Esta dimensión resulta especialmente importante en contextos regulados o de alto riesgo.

Estas tres dimensiones no son independientes. Una explicación técnicamente detallada puede ser útil para auditoría, pero incomprensible para un usuario final; una explicación sencilla puede facilitar la comunicación, pero ocultar matices necesarios para la revisión profesional. TAXAL hace explícitas estas tensiones y las aborda mediante explicaciones sensibles al rol del usuario, conectando explicabilidad con confianza, contestabilidad y responsabilidad.

3 Explicaciones adaptadas al rol del usuario

Uno de los principios centrales de TAXAL es que no existe una explicación universalmente válida. La calidad de una explicación depende de quién la recibe, para qué la necesita y en qué contexto se utiliza. En un sistema clínico de apoyo a la decisión, por ejemplo, el profesional sanitario puede necesitar marcadores

clínicos, razonamiento diferencial y análisis contrafactuales, mientras que el paciente requiere una explicación en lenguaje cotidiano que le ayude a comprender la situación y participar en la decisión sin verse sobrecargado por detalles técnicos.

Esta sensibilidad al rol permite reconciliar dos exigencias que a menudo aparecen en tensión: la accesibilidad de la explicación para usuarios no técnicos y la auditabilidad necesaria en contextos profesionales o regulados. La clave no es producir explicaciones desconectadas para cada usuario, sino organizar distintos niveles de acceso a una misma lógica explicativa. Así, TAXAL actúa como un andamiaje para decidir quién necesita una explicación, qué dimensiones deben priorizarse y cómo debe evaluarse su calidad.

4 Validación empírica

El marco TAXAL se valida empíricamente en el dominio médico con 200 viñetas clínicas de MedQA. El diseño genera 800 explicaciones bajo cuatro condiciones: base, experta, general y triádica. La evaluación se organiza en tres niveles — métricas automáticas, evaluación mediante LLM como juez (GPT-5.2) y validación cruzada por audiencias—, que producen más de 4.800 valoraciones.

Table 1. Diferenciación medible inducida por TAXAL en las tres dimensiones (MedQA). Las comparaciones por pares son significativas ($p < 0{,}001$, Mann–Whitney U).

Dimensión	**Indicador**	**Resultado**
Cognitiva	Densidad biomédica (Experto/General)	1,3×
Causal	Cadenas de razonamiento (Experto/General)	3,5×
Funcional	Accionabilidad (Triádica vs. Baseline)	3,5×
Preferencia cruzada de audiencia (experta / general)		99,0 % / 90,6 %

Los resultados (Tabla 1) muestran diferencias medibles entre condiciones y respaldan la hipótesis de que TAXAL permite modular explicaciones según sus dimensiones cognitiva, funcional y causal. La diferenciación incluye mayor densidad biomédica y una brecha de legibilidad de aproximadamente cinco grados FKGL entre las condiciones experta y general, un conteo de cadenas de razonamiento 3,5× superior en la condición experta y un aumento equivalente en la accionabilidad de las explicaciones generales. La validación cruzada confirma el resultado más relevante: las explicaciones adaptadas a una audiencia son preferidas desde la perspectiva de esa misma audiencia en el 99,0 % de los casos para la experta y el 90,6 % para la general.

La lectura principal es que la explicabilidad en LLMs agénticos puede diseñarse de manera controlada, y no únicamente generarse a posteriori como una justificación textual genérica. No obstante, las cadenas de razonamiento y los

contrafactuales generados por LLMs son señales explicativas observables, no una prueba directa de causalidad mecanicista interna. La evaluación valida la controlabilidad de la estructura de la explicación, no sus impactos posteriores —confianza real del usuario, adopción institucional o calidad de la decisión—, que quedan como trabajo futuro.

5 Discusión y conclusiones

La principal aportación de TAXAL es ofrecer una forma estructurada de pasar de la explicabilidad como concepto general a la explicabilidad como práctica de diseño. Frente a enfoques que se limitan a clasificar técnicas, TAXAL permite razonar sobre qué tipo de explicación se necesita, para quién, con qué propósito y bajo qué restricciones.

Esta perspectiva resulta especialmente relevante en sistemas de IA agéntica. Cuando un LLM participa en procesos de decisión complejos, la explicación debe cubrir no solo el resultado final, sino también el razonamiento expresado, la utilidad de la información y la posibilidad de que distintos actores revisen o cuestionen la salida. El marco propone integrar las técnicas de explicabilidad en el ciclo de vida de los sistemas agénticos, desde el diseño hasta la evaluación y el despliegue en escenarios sociotécnicos reales.

Acknowledgments. Este trabajo ha sido financiado por el proyecto IAFER, "Ethical, Responsible, and General Purpose Artificial Intelligence: Applications in Risk Scenarios" (Exp. TSI-100927-2023-1), en el marco de los Programas Enia y del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea – NextGenerationEU a través del Ministerio para la Transformación Digital y de la Función Pública. Este trabajo ha sido también parcialmente financiado por el Ministerio de Ciencia, Innovación y Universidades de España, proyecto PID2023-150070NB-I00, y por la ayuda QUAL21-14 concedida por la Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía.

Disponibilidad de materiales. Los *prompts*, los *scripts* de evaluación y las métricas están disponibles en https://github.com/divadhp/taxal-medqa-validation.

Disclosure of Interests. Los autores declaran no tener conflictos de interés relevantes para el contenido de este trabajo.

References

1. David Herrera-Poyatos, Carlos Peláez-González, Cristina Zuheros, Virilo Tejedor, Rosana Montes, Francisco Herrera, "TAXAL framework: Triadic fusion of cognitive, functional, and causal dimensions for explainability in agentic LLMs", *Information Fusion*, Volume 134, 104389, 2026.