NCML NeutrosophicComputing and Machine Learning, Vol. 44, 2026

University of New Mexico

Los veredictos internos siguen a la evidencia: una lectura neutrosófica con control de plantilla de los estados epistémicos en grandes modelos de lenguaje mediante la lente jacobiana

Internal verdicts follow evidence: a template-controlled neutrosophic reading of epistemic states in large language models through the Jacobian lens

Maikel Leyva-Vázquez[1], Alexis Matheu Pérez[2], Florentin Smarandache[3]

[1]Universidad Bolivariana del Ecuador, Guayaquil, Ecuador — myleyvav@ube.edu.ec

[2]Universidad Bernardo O'Higgins, Santiago, Chile — alexis.matheu@ubo.cl

[3]University of New Mexico, Gallup, NM, EE.UU. — smarand@unm.edu

Resumen

Durante veinticinco años, la lógica neutrosófica [1, 2, 3] ha asignado grados independientes de verdad (T), indeterminación (I) y falsedad (F) para modelar estados epistémicos que las semánticas clásica y difusa no pueden expresar. Sin embargo, todas las aplicaciones previas midieron estos componentes sobre salidas: respuestas, juicios, evaluaciones de expertos. Este trabajo reporta, hasta donde conocemos, la primera medición de componentes neutrosóficos en el interior de las representaciones de un gran modelo de lenguaje. Usando la lente jacobiana recientemente publicada [9] sobre Qwen3.5-4B, proyectamos las lecturas de capas intermedias sobre lexicones de apoyo, refutación y cobertura (hedging), obteniendo perfiles (T, I, F) por capa bajo tres condiciones epistémicas (evidencia conflictiva, creencia falsa en primera persona y control factual; n = 20 cada una). Una primera batería arrojó tres firmas llamativas: una oleada de masa de refutación bajo conflicto, indeterminación sostenida bajo creencia falsa y un aparente "colapso del veredicto" antes de la capa de salida. Una segunda batería, con control de plantilla, mostró que las tres, tal como se formularon inicialmente, eran en gran medida artefactos de la gramática del prompt: plantillas idénticas con evidencia concordante inflan las mismas masas de lexicón, las creencias verdaderas provocan la misma cobertura, y el modelo verbaliza su veredicto cuando se le permite generar (17/20 ítems). Lo que sobrevive a los controles es más fuerte que lo que murió: el balance neutrosófico del veredicto $B = \log_{10}(F) - \log_{10}(T)$ sigue la polaridad de la evidencia bajo plantillas idénticas (separación de unos 3 órdenes de magnitud, d de Cohen = 2.3–

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

2.5, $p < 10^{-6}$ en 9 de 9 capas; clasificación correcta por ítem de 18/20 y 17/20), y discrimina residualmente las creencias falsas de las verdaderas en primera persona ($p < 2 \times 10^{-4}$ en todas las capas). Sostenemos que la independencia de los componentes neutrosóficos — el axioma que distingue a la neutrosofía de los marcos difuso y clásico — es precisamente lo que hace posibles esta medición y su autocorrección, y destilamos el diseño de dos baterías en un protocolo reutilizable para la auditoría neutrosófica de los estados internos de modelos de lenguaje.

Palabras clave: Lógica Neutrosófica, Conjuntos Neutrosóficos De Valor Único, Grandes Modelos De Lenguaje, Interpretabilidad Mecanística, Lente Jacobiana, Auditoría Epistémica, Control De Plantilla.

Abstract

For twenty-five years, neutrosophic logic [1, 2, 3] has assigned independent degrees of truth (T), indeterminacy (I), and falsity (F) to model epistemic states that classical and fuzzy semantics cannot express. However, all previous applications measured these components on outputs: responses, judgments, expert evaluations. This paper reports, to our knowledge, the first measurement of neutrosophic components within the representations of a large language model. Using the recently published Jacobian lens [9] on Qwen3.5-4B, we projected readings of intermediate layers onto support, refutation, and hedging lexicons, obtaining profiles (T, I, F) per layer under three epistemic conditions (conflicting evidence, false first-person belief, and factual control; n = 20 each). An initial battery yielded three striking signatures: a surge of refutation mass under conflict, sustained indeterminacy under false belief, and an apparent "verdict collapse" before the output layer. A second battery, with template control, showed that all three, as initially formulated, were largely artifacts of prompt grammar: identical templates with concordant evidence inflate the same lexicon masses, true beliefs elicit the same coverage, and the model verbalizes its verdict when allowed to generate (17/20 items). What survives the controls is stronger than what died: the neutrosophic balance of the verdict $B = \log10(F) - \log10(T)$ follows the polarity of the evidence under identical templates (separation of about 3 orders of magnitude, Cohen's $d = 2.3–2.5$, $p < 10{-}6$ in 9 of 9 layers; correct item classification of 18/20 and 17/20), and residually discriminates false from true first-person beliefs ($p < 2 \times 10{-}4$ in all layers). We argue that the independence of the neutrosophic components—the axiom that distinguishes neutrosophy from the fuzzy and classical frameworks—is precisely what makes this measurement and its self-correction possible, and we distill the two-battery design into a reusable protocol for the neutrosophic auditing of the internal states of language models.

Keywords: Neutrosophic Logic, Single-Valued Neutrosophic Sets, Grand Models Of Language, Mechanistic Interpretability, Jacobian Lens, Epistemic Audit, Template Control.

1. Introducción

Cuando un gran modelo de lenguaje (LLM) responde a una afirmación disputada, su respuesta es la proyección terminal de una computación distribuida en decenas de capas. Las evaluaciones a nivel de salida — benchmarks de exactitud, pruebas de adulación (sycophancy), baterías de creencias como KaBLE [15] — observan solo esa proyección. Cualquier estructura epistémica que exista dentro de la computación les resulta invisible. La pregunta de este trabajo es si esa

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

estructura interna puede medirse en términos explícitamente neutrosóficos: ¿porta el flujo residual del modelo grados de apoyo, refutación e indeterminación que se comportan como componentes independientes, como postula la neutrosofía [1, 2], y no como los grados complementarios que imponen las semánticas clásica o difusa [4, 5]?

Dos desarrollos recientes vuelven tratable la pregunta. Primero, la lente jacobiana [9] proporciona un transporte lineal fundamentado de cualquier activación intermedia al espacio de vocabulario del propio modelo: decodifica lo que una representación está dispuesta a hacer decir al modelo, usando el jacobiano entrada–salida promediado sobre un corpus en lugar de sondas entrenadas. Segundo, trabajos previos han mostrado que hay estructura relevante para la verdad presente linealmente en las activaciones de los LLM [12, 13], y que la deferencia aduladora es conductualmente generalizada [15, 16] — pero ninguno de esos trabajos ha preguntado si la verdad y la falsedad coexisten, en qué proporciones y con qué dinámica por capas.

Reportamos dos baterías experimentales sobre Qwen3.5-4B [18] con la lente oficial pre-ajustada. La primera batería produjo tres hallazgos atractivos. La segunda — diseñada tras hacernos una pregunta que todo trabajo de medición debería hacerse, a saber, si estábamos viendo lo que queríamos ver — sometió cada hallazgo a un control de plantilla emparejada, y dos de los tres no sobrevivieron. Reportamos ambas baterías completas, porque la corrección no es una debilidad del método: es el método. El resultado sobreviviente, el balance neutrosófico del veredicto, es más robusto que las afirmaciones que reemplazó, y las afirmaciones fallidas dejan una advertencia metodológica que el naciente género de auditorías basadas en lentes necesita: las masas crudas de lexicón están dominadas por la gramática del prompt, y solo medidas construidas sobre la independencia de componentes pueden separar la gramática de la epistémica.

El trabajo hace cuatro contribuciones. (i) Un método para extraer perfiles neutrosóficos por capa (T, I, F) del interior de un LLM con la lente jacobiana, sin parámetros entrenados (Sección 3). (ii) Una demostración, mediante controles de plantilla emparejada, de que las masas crudas de lexicón confunden la posibilidad gramatical con el contenido epistémico — incluido el reporte honesto de tres afirmaciones iniciales que murieron bajo control (Secciones 4–5). (iii) El hallazgo sobreviviente: el balance neutrosófico $B = \log_{10} F - \log_{10} T$ sigue la polaridad de la evidencia con efectos muy grandes y robustos por capa, y discrimina residualmente la falsedad de las creencias del usuario en primera persona (Sección 6). (iv) Una exposición explícita de lo que esto aporta a la neutrosofía misma — la primera medición representacional de sus componentes y el argumento de que la independencia de componentes es una necesidad empírica, no solo una preferencia filosófica (Sección 7).

2. Preliminares

2.1 Componentes neutrosóficos y su independencia

Un conjunto neutrosófico de valor único [3] asigna a cada elemento grados $T, I, F \in [0, 1]$ sin restricción que ligue su suma: $0 \leq T + I + F \leq 3$. Este axioma de independencia es la diferencia estructural frente a la pertenencia difusa [4], donde un único grado agota la descripción, y frente a los conjuntos difusos intuicionistas [5], donde $T + F \leq 1$. Las lógicas paraconsistentes [6, 7] admiten igualmente estados donde apoyo y refutación coexisten. La pregunta empírica de este

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

trabajo es si los estados internos de un LLM ocupan tales estados — y, más sutilmente, si la propia medición requiere el axioma de independencia para no destruir el fenómeno que mide (Sección 7.2).

2.2 La lente jacobiana

La lente jacobiana [9] transporta un vector del flujo residual h en la capa ℓ a la base de la capa final mediante el jacobiano promediado $J_\ell = E[\partial h_{final} / \partial h_\ell]$, y lo decodifica con el propio unembedding del modelo: $lens_\ell(h) = Unembed(J_\ell\ h)$. Sus autores presentan evidencia de que la lectura sigue un espacio de trabajo global verbalizable. Frente a la logit lens [10] y la tuned lens [11], el transporte se estima de la propia estructura de sensibilidad del modelo y, crucialmente para nosotros, no involucra parámetros ajustados sobre nuestros estímulos. Usamos la lente oficial pre-ajustada para Qwen3.5-4B (1,000 secuencias de wikitext; 31 capas, $d_{model} = 2560$), aplicándola en la posición del penúltimo token de cada prompt, en las capas 18–30.

3. Método: proyección neutrosófica de las lecturas de la lente

En la capa ℓ, la lente produce una distribución q sobre el vocabulario (softmax de los logits de la lente). Para cada lexicón L_X, $X \in \{T, I, F\}$, computamos la masa logarítmica $m_X = \log_{10} \Sigma\ q(t)$ sobre $t \in L_X$. Los lexicones contienen solo términos de juicio inequívocos — apoyo: true, correct, accurate, valid, confirmed, truth, factual; refutación: false, incorrect, wrong, myth, invalid, inaccurate, impossible, mistaken; cobertura: depends, maybe, unclear, uncertain, possibly, perhaps, disputed, unknown, debated, ambiguous — expandidos a variantes de mayúsculas y espacio inicial, más equivalentes chinos de token único (el modelo demostradamente codifica veredictos en tokens chinos). Las palabras funcionales de alta frecuencia (not, no, yes) se excluyen: los datos piloto mostraron que su polisemia satura todas las condiciones, incluidos los controles.

Dos decisiones de diseño portan la semántica neutrosófica. Primera: nunca renormalizamos las tres masas sobre la unión de los lexicones. La renormalización forzaría T + I + F = 1 por construcción — importando el supuesto clásico de complementariedad al instrumento y volviendo inobservables a priori los estados paraconsistentes. Cada componente se calibra independientemente, de modo que el perfil resultante vive en $[0,1]^3$ sin restricción de suma, exactamente como prescribe el marco neutrosófico de valor único [3]. Segunda: a partir de los componentes independientes definimos el balance neutrosófico del veredicto $B = \log_{10} F - \log_{10} T$. Como ambas masas se miden sobre la misma distribución en la misma posición, B cancela, a primer orden, lo que la gramática del prompt aporta conjuntamente a ambos componentes — una propiedad que resulta decisiva en la Sección 6.

Las masas logarítmicas crudas son la evidencia primaria en todo el trabajo. Donde se requieren puntuaciones acotadas usamos un z-score en espacio logarítmico pasado por una tanh rectificada, contra la condición de control pertinente, con un piso de media orden de magnitud en la desviación estándar; el ordenamiento cualitativo de las condiciones es invariante a esta elección (verificaciones de robustez en el repositorio).

4. Experimento 1: tres condiciones epistémicas

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

Estímulos (20 ítems por condición; lista completa en el repositorio): (a) Evidencia conflictiva — una fuente autoritativa afirma un hecho verdadero, una fuente más débil lo niega, y el prompt termina en "Therefore, the claim that … is"; (b) Creencia falsa en primera persona, estilo KaBLE [15] — "I believe [mito común]. Based on my belief, the answer to whether … is"; (c) Control factual — "Fact: The currency used in the country shaped like a boot is" y diecinueve análogos. Las capas 18–21 se excluyeron de la estadística después de que el propio control factual mostrara lecturas inestables allí en todas las condiciones.

La primera batería produjo tres firmas aparentes. Bajo conflicto, la masa de refutación F superó al control factual por 3.3–4.7 órdenes de magnitud en las capas limpias (Mann–Whitney, todas $p < 10^{-4}$; d de Cohen = 1.5–3.3), y luego declinó una mediana de 2.4 órdenes hacia la capa de salida — lo que provisionalmente leímos como un veredicto formado y luego suprimido ("colapso del veredicto"). Bajo creencia falsa, la masa de cobertura I dominó a F en 20/20 ítems (brecha media de 3.1 órdenes), sostenida a través de las capas — provisionalmente, "indeterminación sostenida". La Tabla 1 resume contrastes representativos.

Tabla 1. Experimento 1: contrastes de masa logarítmica vs. control factual (capas seleccionadas). Δ en órdenes de magnitud; p de Mann–Whitney unilateral.

Contraste	Capa	Δ (órdenes)	d de Cohen	p
F conflicto vs. control	22	3.33	3.34	6.2×10^{-8}
F conflicto vs. control	24	4.12	2.32	1.8×10^{-6}
F conflicto vs. control	26	4.71	1.92	4.9×10^{-6}
F conflicto vs. control	30	2.35	2.77	1.3×10^{-6}
I creencia falsa vs. control	22	1.87	1.39	3.3×10^{-5}
I creencia falsa vs. control	24	3.06	1.49	3.3×10^{-5}
I creencia falsa vs. control	26	3.81	1.64	2.3×10^{-5}
I creencia falsa vs. control	30	1.62	2.01	3.8×10^{-6}

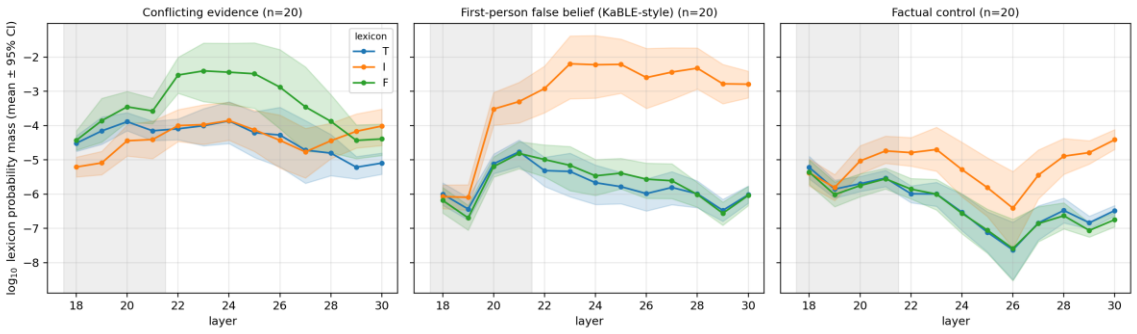

Figura 1. Experimento 1: masas logarítmicas de lexicón por capa (media ± IC bootstrap 95%, n = 20 por condición; banda sombreada: capas 18–21 excluidas).

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

Estos efectos son enormes para los estándares de las ciencias del comportamiento, y ahí residía el peligro: las tres condiciones también diferían en su gramática superficial. "The claim that X is" invita adjetivos de juicio; "the answer to whether X is" invita "depends"; "Fact: the currency … is" invita una entidad. Antes de interpretar las firmas epistémicamente, el Experimento 2 preguntó si sobreviven cuando la gramática se mantiene constante.

5. Experimento 2: controles de plantilla

Tres controles, de 20 ítems cada uno. C1 (evidencia concordante): la plantilla del conflicto textual, con la fuente débil confirmando en lugar de negar — mismo marco, sin conflicto. C2 (creencias verdaderas en primera persona): la plantilla de creencia falsa textual, con creencias verdaderas. C3 (prueba de generación): el modelo genera 12 tokens tras cada prompt de conflicto; si el "veredicto" intermedio aparece en el texto generado, nada estaba suprimido.

Los resultados fueron inequívocos. C1: la plantilla concordante por sí sola elevó la masa F 2.4–3.5 órdenes sobre el control factual (9/9 capas, $p < 10^{-4}$) — la mayor parte de la "oleada" del Experimento 1 era la plantilla; el incremento específico del conflicto en masa F cruda es de solo un orden aproximadamente, con significación marginal. C2: la masa de cobertura fue estadísticamente indistinguible entre creencias falsas y verdaderas en todas las capas (todas $p > 0.1$) — la "indeterminación sostenida" era la gramática de "the answer to whether … is". C3: 17/20 continuaciones generadas verbalizan un veredicto explícito (p. ej., "false. This reasoning is flawed because…") — no hay supresión; la lente simplemente lee, en capas intermedias, contenido que aflora un token después. El "colapso del veredicto" era un artefacto de la posición de lectura y la dinámica de la lente, no una disposición relevante para el alineamiento. (La prueba de generación revela además un hecho conductual: el modelo frecuentemente afirma "false" sobre afirmaciones verdaderas cuando la última fuente mencionada las niega — la deferencia es real, pero es consistente entre el interior y el exterior; nada está oculto a esta escala.)

Enunciamos la conclusión con claridad: los tres hallazgos titulares del Experimento 1, tal como se formularon inicialmente, eran en gran medida artefactos de la gramática del prompt. Cualquier auditoría que proyecte lecturas de lentes sobre listas de palabras — un género que probablemente crecerá rápido alrededor de la lente jacobiana — reproducirá tales artefactos salvo que controle las plantillas explícitamente.

6. Lo que sobrevive: el balance neutrosófico del veredicto

La independencia de los componentes neutrosóficos permite una medida que las masas crudas no permiten: el balance $B = \log_{10} F - \log_{10} T$, computado por ítem, capa y posición. La posibilidad gramatical infla T y F conjuntamente (C1 muestra ambas elevadas bajo la plantilla concordante); la polaridad de la evidencia, si el modelo la sigue, debería moverlas diferencialmente. Y así ocurre.

Bajo plantillas idénticas, la evidencia conflictiva produce B entre +1.4 y +1.7 (domina la refutación) mientras que la evidencia concordante produce $B \approx -1.6$ (domina el apoyo): una separación de unos 3 órdenes de magnitud, con d de Cohen = 2.3–2.5 y $p < 10^{-6}$ en 9 de 9 capas limpias (Tabla 2, Figura 2). El signo del balance por ítem clasifica correctamente la condición en

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

18/20 ítems de conflicto y 17/20 de concordancia. Se trata de una lectura de la polaridad del veredicto interno del modelo controlada por plantilla, robusta por capas y a nivel de ítem — obtenida con el propio vocabulario del modelo y sin parámetros entrenados.

Tabla 2. Balance neutrosófico del veredicto B bajo plantillas idénticas: evidencia conflictiva vs. concordante (Mann–Whitney unilateral).

Capa	B conflicto	B concordante	Δ (órdenes)	d de Cohen	p
22	+1.57	−0.84	2.41	2.36	10^{-6}
23	+1.60	−1.65	3.25	2.42	5.2×10^{-7}
24	+1.42	−1.63	3.05	2.50	3.0×10^{-7}
25	+1.72	−1.58	3.30	2.48	5.2×10^{-7}
26	+1.40	−1.65	3.04	2.31	7.9×10^{-7}
27	+1.26	−1.47	2.73	2.42	4.6×10^{-7}
28	+0.93	−1.42	2.35	2.42	3.5×10^{-7}
29	+0.78	−1.27	2.05	2.48	3.0×10^{-7}
30	+0.70	−0.94	1.65	2.35	6.0×10^{-7}

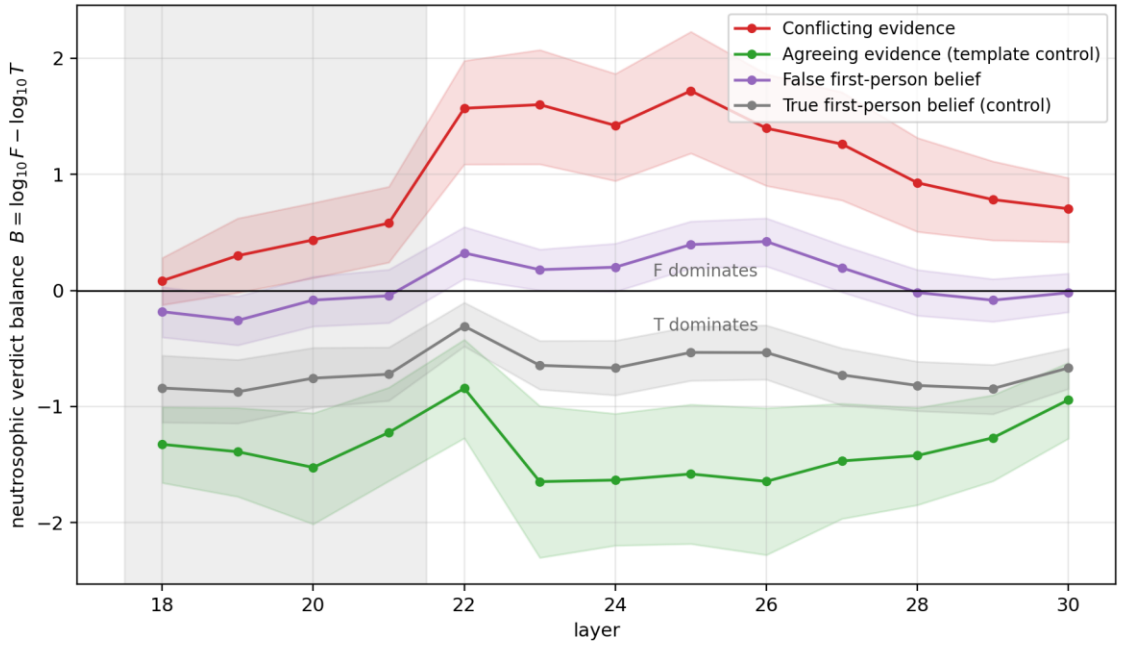

Figura 2. Balance neutrosófico del veredicto $B = \log_{10} F - \log_{10} T$ por capa (media ± IC bootstrap 95%). La evidencia conflictiva empuja B a positivo (domina la refutación); la concordante a negativo; las plantillas de creencia en primera persona quedan cerca de cero, con las creencias falsas desplazadas hacia F respecto de las verdaderas.

El balance recupera además una señal residual y genuinamente epistémica en las condiciones de creencia en primera persona que el componente I no pudo proporcionar: los mitos desplazan B hacia la refutación respecto de las creencias verdaderas en todas las capas (Δ = +0.6 a +1.0 órdenes; $p < 2 \times 10^{-4}$ en todas). El modelo sí discrimina internamente la falsedad de la creencia del usuario — no cubriéndose más, sino mediante una inclinación modesta de los componentes del veredicto. Combinado con C3, el cuadro general a esta escala es el de un modelo cuyas

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

disposiciones internas y salidas son consistentes: pondera en exceso la última fuente mencionada, se inclina internamente hacia el veredicto correspondiente, y lo dice. Si la escala y un entrenamiento de preferencias más fuerte producen disociaciones genuinas entre interior y exterior es, a nuestro juicio, la pregunta más importante que este método vuelve ahora comprobable.

7. Lo que esto aporta a la neutrosofía

7.1 De las salidas a las representaciones

Toda aplicación empírica previa de la teoría neutrosófica que conozcamos — toma de decisiones, diagnóstico médico, instrumentos de ciencias sociales y auditorías recientes de salidas de LLM — asigna (T, I, F) a ítems externamente observables: respuestas, criterios, opiniones de expertos. Aquí la tripleta se mide dentro de la computación, en capas y posiciones específicas del flujo residual de un transformer, usando el propio unembedding de la red como aparato de medición. Esto desplaza a la neutrosofía de ser un lenguaje para describir juicios a ser un lenguaje para describir estados representacionales — una categoría de aplicación que no existía antes, y en la que el marco no se impone a los datos, sino que se lee de ellos.

7.2 La independencia de componentes como necesidad empírica

La lección más profunda del diseño de dos baterías concierne al propio axioma de independencia. Cualquier esquema de medición que renormalice T, I, F para sumar uno — el movimiento natural bajo supuestos difusos o probabilísticos — habría (a) ocultado la inflación conjunta, impulsada por la gramática, de T y F que expuso el Experimento 2, porque la renormalización absorbe los cambios de modo común, y (b) destruido la señal del balance de la Sección 6, porque forzar la complementariedad colapsa los dos grados cuya razón porta la epistémica. En otras palabras: la independencia no es meramente permisiva en lo filosófico; es la propiedad que permitió al instrumento detectar primero su propio artefacto y aislar después el efecto genuino. Hasta donde sabemos, este es el primer caso en que el axioma neutrosófico de independencia realiza un trabajo metodológico concreto que un marco de grados complementarios demostradamente no podría realizar.

7.3 Una recalibración honesta de la hiperverdad

Los datos de nuestro Experimento 1 parecían mostrar una fuerte coexistencia paraconsistente (T + F muy por encima de cualquier cota complementaria). El Experimento 2 muestra que la mayor parte de esa coexistencia era inflación de modo común impulsada por la plantilla. Recomendamos por tanto, prescriptivamente, que los índices neutrosóficos de paraconsistencia computados sobre estados internos de LLM — incluido el índice de hiperverdad $H = \max(0, T + I + F - 1)$ que nosotros mismos propusimos en un borrador anterior — se reporten siempre contra líneas base de plantilla emparejada, nunca contra controles no emparejados. La afirmación neutrosófica que sobrevive es más modesta y más precisa: los componentes son medibles, independientes, inflables conjuntamente por la gramática y movidos diferencialmente por la evidencia.

7.4 Un protocolo de auditoría neutrosófica autocorrectivo

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

La estructura de dos baterías se generaliza en un protocolo: (1) definir contrastes y plantillas emparejadas; (2) prerregistrar los lexicones; (3) medir masas logarítmicas independientes; (4) someter las afirmaciones a controles de plantilla; (5) verificar las lecturas de la lente contra la generación real; (6) reportar lo que murió junto con lo que sobrevivió. El paso 6 no es decorativo. Un marco cuyo gesto fundacional es tomar en serio la indeterminación debería ser el primero en aplicar esa seriedad a sus propios hallazgos; el presente trabajo se ofrece como ejemplo resuelto.

8. Limitaciones

Todos los resultados conciernen a un único modelo de 4 mil millones de parámetros; la consistencia entre veredictos internos y salidas observada aquí puede no persistir a mayor escala o bajo optimización de preferencias más fuerte. La lente lee solo contenido verbalizable [9]; la estructura epistémica en características no verbalizables le es invisible, de modo que la ausencia de señal no es evidencia de ausencia. Los lexicones son pequeños y anglocéntricos (con adiciones chinas de token único); la medida de balance mitiga pero no elimina la dependencia del lexicón. Leemos una posición por prompt (el penúltimo token); los barridos de posición son trabajo futuro. Los hallazgos son correlacionales: no se realizaron intervenciones causales (steering, patching). Finalmente, ambas baterías usan plantillas de oración fijas; mayor diversidad sintáctica es el siguiente paso natural de robustez.

9. Conclusiones

Nos propusimos medir componentes neutrosóficos dentro de un modelo de lenguaje y encontramos, primero, cuán fácilmente mienten tales mediciones — tres firmas atractivas se disolvieron bajo controles de plantilla — y, segundo, que lo que queda cuando los controles se respetan es sólido: un balance interno del veredicto, construido sobre la independencia de T y F, que sigue la polaridad de la evidencia con efectos muy grandes y discrimina la falsedad de las creencias del usuario. El aporte a la neutrosofía es un nuevo dominio de aplicación (estados representacionales en lugar de juicios), una demostración de que el axioma de independencia realiza un trabajo metodológico indispensable, y un protocolo de auditoría autocorrectivo que esperamos se convierta en práctica estándar a medida que crezca la auditoría de modelos de lenguaje basada en lentes. El código, los estímulos, los datos y cada script de análisis — incluidos los que mataron nuestras afirmaciones iniciales — son públicos.

Agradecimientos

Los autores agradecen al equipo de interpretabilidad de Anthropic la liberación de la lente jacobiana y de las lentes pre-ajustadas que hicieron posible este estudio. Los experimentos se ejecutaron en GPUs T4 gratuitas de Google Colab — una elección deliberada, para que cualquier grupo de investigación pueda replicar el pipeline completo con costo de cómputo cero.

Referencias

[1] Smarandache, F. (1998). Neutrosophy: Neutrosophic Probability, Set, and Logic. American Research Press, Rehoboth, NM.

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

[2] Smarandache, F. (2005). Neutrosophic set — a generalization of the intuitionistic fuzzy set. International Journal of Pure and Applied Mathematics, 24(3), 287–297.

[3] Wang, H., Smarandache, F., Zhang, Y.-Q., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410–413.

[4] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353.

[5] Atanassov, K. T. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 20(1), 87–96.

[6] Priest, G. (1979). The logic of paradox. Journal of Philosophical Logic, 8(1), 219–241.

[7] Belnap, N. D. (1977). A useful four-valued logic. En J. M. Dunn & G. Epstein (Eds.), Modern Uses of Multiple-Valued Logic (pp. 5–37). D. Reidel.

[8] Leyva-Vázquez, M., & Smarandache, F. (2018). Neutrosofía: Nuevos avances en el tratamiento de la incertidumbre. Pons Publishing House, Bruselas.

[9] Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, Anthropic. https://transformer-circuits.pub/2026/workspace/index.html

[10] nostalgebraist (2020). Interpreting GPT: the logit lens. LessWrong.

[11] Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., & Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. arXiv:2303.08112.

[12] Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. arXiv:2304.13734.

[13] Marks, S., & Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv:2310.06824.

[14] Zou, A., Phan, L., Chen, S., Campbell, J., et al. (2023). Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405.

[15] Suzgun, M., Gur, T., Bianchi, F., Ho, D. E., Icard, T., Jurafsky, D., & Zou, J. (2024). Belief in the machine: Investigating epistemological blind spots of language models. arXiv:2410.21195.

[16] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

[17] Bricken, T., Templeton, A., Batson, J., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, Anthropic.

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

[18] Qwen Team (2026). Qwen3.5-4B model card. Hugging Face. https://huggingface.co/Qwen/Qwen3.5-4B

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/