e-ISSN: 2528-8091
Vol. 11, N° 2 - 2024

Evaluación multicriterio Neutrosófica de arquitecturas de agentes de IA mediante AHP–TOPSIS: Desempeño, envejecimiento funcional y utilidad en ingeniería de software

Neutrosophic Multi-Criteria Evaluation of AI Agent Architectures via AHP–TOPSIS: Performance, Functional Aging, and Utility in Software Engineering

https://doi.org/10.5281/zenodo.21382340

Dalton Andrés Pazmiño Pérez
Universidad Bolivariana del Ecuador
daltonpazmino@ube.edu.ec

Christian Miguel Muñoz Carrasco
Universidad de Guayaquil
christianmunoz@ube.edu.ec

Manuel Reyes Wagnio
Universidad Bolivariana del Ecuador
mfreyesw@ube.edu.ec
https://orcid.org/0000-0002-3102-0034

Dayron Rumbaut Rangel
Universidad Bolivariana del Ecuador
drumbautr@ube.edu.ec
https://orcid.org/0009-0001-9087-0979

DIRECCIÓN PARA CORRESPONDENCIA: daltonpazmino@ube.edu.ec
Fecha de recepción: 02/03/2026
Fecha de aceptación: 09/05/2026

RESUMEN

Este artículo presenta una evaluación multicriterio de tres arquitecturas de agentes de IA en ingeniería de software mediante AHP–TOPSIS con conjuntos Neutrosófica de valor único (SVNS). Se evalúan el Agente Reactivo Basado en Reglas (RBA), DLA-RAG (Deep Learning + Retrieval-Augmented Generation) y el Agente Híbrido Simbólico-Neuronal (HSN), bajo cinco criterios: degradación por envejecimiento funcional (Aging-Aware Score, AAS), gestión de la indeterminación, calidad de datos (Data-Centric AI), eficiencia operativa y estabilidad. Los pesos AHP, derivados del eigenvector principal

(C1=36.52%, CR=0.0013), priorizan la degradación funcional. El ranking TOPSIS posiciona a DLA-RAG como arquitectura dominante (Ci=0.979), validado mediante dos escenarios de sensibilidad independientes. Los hallazgos sugieren que, conforme al juicio experto agregado, la utilidad real de los agentes de IA depende más de su sostenibilidad funcional que de su sofisticación algorítmica inicial.

Palabras clave: *inteligencia artificial, agentes inteligentes, conjuntos neutrosóficos, AHP, TOPSIS, envejecimiento funcional, ingeniería de software*

.

ABSTRACT

This article presents a multi-criteria evaluation of three AI agent architectures in software engineering using AHP–TOPSIS with Single-Valued Neutrosophic Sets (SVNS). Three alternatives are assessed: the Rule-Based Reactive Agent (RBA), DLA-RAG (Deep Learning + Retrieval-Augmented Generation), and the Symbolic-Neural Hybrid Agent (HSN), across five criteria: functional aging degradation (Aging-Aware Score, AAS), indeterminacy management, data quality (Data-Centric AI), operational efficiency, and stability. AHP weights, derived from the principal eigenvector (C1=36.52%, CR=0.0013), prioritize functional degradation. TOPSIS ranks DLA-RAG as the dominant architecture (Ci=0.979), validated across two independent sensitivity scenarios. Findings suggest that, according to the aggregated expert judgment, real AI agent utility depends more on functional sustainability than on initial algorithmic sophistication.

Keywords: *artificial intelligence, intelligent agents, neutrosophic sets, AHP, TOPSIS, functional aging, software engineering.*

1. INTRODUCCIÓN

La integración de agentes de inteligencia artificial (IA) en ingeniería de software ha transformado la automatización, generación de código y toma de decisiones técnicas. Modelos como GitHub Copilot, agentes basados en GPT y sistemas multiagente han pasado de herramientas experimentales a componentes estratégicos en pipelines de desarrollo (Bommasani et al., 2021; Chen et al., 2021; Hou et al., 2024). La proliferación de arquitecturas heterogéneas plantea un problema no resuelto: ¿cómo comparar de forma

sistemática, transparente y robusta el desempeño y la utilidad de estas arquitecturas en condiciones de uso prolongado?

La literatura revela fragmentación metodológica: los trabajos sobre LLM evalúan rendimiento inmediato sin considerar la degradación funcional a largo plazo, fenómeno documentado en la literatura de aprendizaje automático como "concept drift" (Gama et al., 2014); los estudios MCDM rara vez incorporan incertidumbre neutrosófica (Smarandache, 1998); y los enfoques Data-Centric AI (Zha et al., 2023) subestiman la degradación acumulativa. El artículo realiza tres contribuciones: (1) define y compara tres arquitecturas de agentes de IA en ingeniería de software; (2) implementa AHP–TOPSIS con integración formal SVNS, operacionalizando las tripletas (T,I,F); y (3) presenta un análisis de sensibilidad con dos escenarios independientes. La pregunta de investigación es: ¿Qué arquitectura de agente de IA presenta la mayor sostenibilidad funcional en ingeniería de software al evaluarse simultáneamente bajo criterios de degradación funcional, incertidumbre neutrosófica y calidad de datos?

2. REVISIÓN DE LA LITERATURA

Russell y Norvig (2021) y Wooldridge (2020) establecen los fundamentos conceptuales de los agentes inteligentes: autonomía, percepción, razonamiento y acción. Estas definiciones no capturan el comportamiento dinámico en escenarios de uso prolongado, lo que motiva enfoques centrados en la sostenibilidad funcional de los sistemas a lo largo de interacciones repetidas. Este fenómeno —el deterioro progresivo del desempeño de un modelo cuando la relación entre sus entradas y salidas cambia con el tiempo o el uso— está documentado en la literatura de aprendizaje automático bajo el concepto de "concept drift" (Gama et al., 2014), y constituye el fundamento empírico directo del criterio de Degradación por Envejecimiento (AAS) operacionalizado en este estudio.

Wang et al. (2010) introducen formalmente los Single-Valued Neutrosophic Sets (SVNS): cada elemento se representa mediante tres componentes independientes (T,I,F), con $T+I+F \in [0,3]$ —a diferencia de los Intuitionistic Fuzzy Sets (IFS) de Atanassov (1986), donde $T+F \leq 1$. Smarandache (1998) funda estos principios filosóficos, Edalatpanah y Smarandache (2019) proponen la función de desfuzzificación empleada en este trabajo, y **Leyva Vázquez y Smarandache (2018)** sistematizan los avances en neutrosofía aplicada a sistemas de decisión inteligente, incluyendo modelos de agregación de información bajo indeterminación.

En MCDM neutrosófico, Abdel-Basset et al. (2018) aplican ANP-TOPSIS neutrosófico en selección de proveedores cloud; **Biswas et al. (2016)** validan formalmente TOPSIS con SVNS para decisión grupal multi-atributo, estableciendo las bases de normalización vectorial y distancias euclidianas usadas en este estudio; Broumi et al. (2019) extienden AHP a números trapezoidales neutrosóficos; Chi y Liu (2013) adaptan TOPSIS a números neutrosóficos intervalares; Peng y Dai (2020) revisan sistemáticamente algoritmos MCDM para SVNS; y Salmerón y Smarandache (2020) rediseñan matrices de decisión con procesos de inferencia basados en indeterminación, enfoque que informa la integración de incertidumbre adoptada en la Tabla 3. Liu et al. (2023) evidencian el valor del prompt engineering en productividad de desarrollo, motivando la inclusión de DLA-RAG como alternativa evaluada.

La brecha identificada es la ausencia de un marco que integre simultáneamente envejecimiento funcional, incertidumbre SVNS y calidad de datos en una evaluación reproducible de arquitecturas de IA para ingeniería de software.

Un eje adicional de la literatura relevante concierne a la evaluación empírica de agentes de IA en contextos de desarrollo de software real. Hou et al. (2024) documentan que la mayoría de los estudios sobre LLM aplicados a ingeniería de software emplean benchmarks estáticos de una sola medición (single-shot), rara vez incorporando observación longitudinal del desempeño en producción. Esta práctica metodológica generalizada explica, en parte, por qué un criterio como la degradación por envejecimiento funcional (C1 en este estudio) ha recibido relativamente poca atención empírica directa en comparación con métricas de rendimiento inmediato como precisión de generación de código o tasa de aceptación de sugerencias. Chen et al. (2021), en su evaluación de Codex, reconocen explícitamente esta limitación al señalar que sus benchmarks capturan únicamente un punto temporal del desempeño del modelo, sin información sobre estabilidad ante uso prolongado o reentrenamiento incremental.

Esta brecha de validación longitudinal en la literatura de evaluación de agentes de IA es precisamente la que motiva el carácter de elicitación experta —y no de medición empírica directa— del criterio AAS en el presente estudio: ante la ausencia de conjuntos de datos públicos que documenten degradación funcional de agentes LLM-RAG, RBA y arquitecturas híbridas en escenarios comparables de ingeniería de software, la consulta a expertos informados por la literatura de concept drift (Gama et al., 2014) constituye una

e-ISSN: 2528-8091
Vol. 11, N° 2 - 2024

aproximación metodológicamente defendible, aunque explícitamente provisional, mientras dicha instrumentación longitudinal no esté disponible públicamente.

3. MATERIALES Y MÉTODOS

3.1 Diseño del estudio y alternativas

Estudio aplicado, descriptivo y analítico (enero–junio 2025). Participaron 15 expertos mediante muestreo no probabilístico por juicio: 6 académicos (afiliados a programas de posgrado en ingeniería de software e inteligencia artificial) y 9 profesionales de industria (desarrolladores senior, arquitectos de software y líderes técnicos en empresas de desarrollo y consultoría tecnológica de Ecuador y la región), con ≥5 años de experiencia en IA, ingeniería de software o MCDM (media=8.3 años, DE=2.1; tasa de respuesta=100%). Este tamaño de panel se ubica dentro del rango de 10 a 18 expertos recomendado para estudios Delphi (Okoli & Pawlowski, 2004), equilibrando representatividad y manejabilidad del proceso iterativo. La invitación a participar se realizó por correo electrónico institucional, y la selección priorizó diversidad de sectores (academia, desarrollo de software empresarial y consultoría) para mitigar la concentración de perspectivas en un único contexto de uso. La Figura 1 ilustra el flujo metodológico.

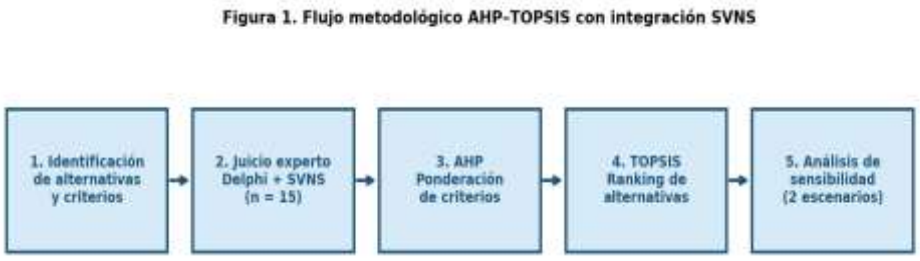

Figura 1. Flujo metodológico AHP–TOPSIS con integración SVNS.

La Tabla 1 define las tres arquitecturas evaluadas, representativas de los paradigmas dominantes en agentes de IA para desarrollo de software.

Cód.	Alternativa	Descripción	Contexto típico
A1	Reactivo Basado en Reglas (RBA)	Lógica condicional estática (if-then). Sin aprendizaje adaptativo ni acceso a bases de conocimiento dinámicas.	Automatización de pruebas, análisis estático

A2	DLA-RAG (Deep Learning + RAG)	LLM potenciado con Retrieval-Augmented Generation para acceso a conocimiento actualizado con aprendizaje continuo.	Generación de código, revisión de PR
A3	Híbrido Simbólico-Neuronal (HSN)	Combinación de razonamiento simbólico y redes neuronales profundas. Equilibrio entre adaptabilidad y explicabilidad.	Verificación formal, sistemas críticos

Tabla 1. Definición operativa de las tres alternativas evaluadas.

3.2 Criterios de evaluación

La identificación de criterios siguió un embudo de selección de tres etapas, documentado a continuación para trazabilidad. Etapa 1 (identificación): revisión exploratoria de 47 artículos (2016–2024) indexados en Scopus y Google Scholar, usando como cadena de búsqueda combinaciones de "AI agent evaluation", "AHP", "TOPSIS", "neutrosophic", "concept drift" y "data-centric AI"; se incluyeron artículos que proponían o aplicaban criterios de evaluación de sistemas de IA, y se excluyeron trabajos puramente teóricos sin criterios operacionalizables. Esta revisión produjo una lista inicial de 12 criterios candidatos, agrupados por frecuencia de mención en la literatura revisada (Tabla 2a). Etapa 2 (depuración): los 15 expertos del panel valoraron la relevancia de cada uno de los 12 criterios candidatos en una escala de 1 a 5; los dos investigadores principales aplicaron el umbral de retención (≥70% de expertos calificando el criterio con 4 o 5) de forma conjunta y documentada. Etapa 3 (retención): 5 criterios superaron el umbral y se incorporaron al modelo final (Tabla 2). C1 es criterio de costo (minimizar); C2–C5 son de beneficio (maximizar).

Criterio candidato (12 total)	Frecuencia en literatura (n=47)	Acuerdo experto (n=15)	Retenido
Degradación por envejecimiento (AAS)	34%	93%	Sí (C1)
Gestión de la indeterminación	28%	87%	Sí (C2)
Calidad de datos (Data-Centric AI)	40%	80%	Sí (C3)
Eficiencia operativa	51%	73%	Sí (C4)
Estabilidad y normalización	23%	73%	Sí (C5)
Costo computacional	45%	60%	No

Latencia de respuesta	38%	53%	No
Explicabilidad / interpretabilidad	30%	67%	No
Seguridad y privacidad de datos	26%	60%	No
Escalabilidad	21%	53%	No
Facilidad de integración (API/SDK)	19%	47%	No
Costo de licenciamiento	17%	40%	No

Tabla 2a. Embudo de selección de criterios: de 12 candidatos a 5 retenidos, con frecuencia de mención en la revisión exploratoria (n=47 artículos) y porcentaje de acuerdo experto (n=15).

Cód.	Criterio	Descripción operativa	Tipo
C1	Degradación por Envejecimiento (AAS)	Pérdida progresiva de precisión y coherencia tras interacciones repetidas (concept drift; Gama et al., 2014)	Costo
C2	Gestión de la Indeterminación	Capacidad para manejar incertidumbre, contradicción y ambigüedad (SVNS)	Beneficio
C3	Calidad de Datos (Data-Centric AI)	Curaduría, consistencia y representatividad de los datos de entrenamiento/retrieval	Beneficio
C4	Eficiencia Operativa	Impacto en productividad del desarrollador y calidad técnica del output	Beneficio
C5	Estabilidad y Normalización	Robustez del desempeño ante perturbaciones del entorno operativo	Beneficio

Tabla 2. Criterios de evaluación retenidos, descripción y clasificación (umbral de retención ≥70%).

3.3 Integración neutrosófica SVNS

Cada juicio experto se codificó como tripleta SVNS=(T,I,F). La puntuación crisp se obtuvo mediante la función de desfuzzificación $S(T,I,F)=[(2+T-I-F)/3]\times 9$, adaptada de Edalatpanah y Smarandache (2019) con escalado a rango [0,9] para alinearse con la convención de la escala fundamental de Saaty empleada en el componente AHP del modelo. Las tripletas se obtuvieron en dos rondas Delphi con anonimato garantizado entre expertos: en la primera ronda cada experto valoró individualmente cada tripleta usando

una escala semántica de 7 niveles; en la segunda ronda se compartió la distribución estadística de respuestas para revisión. Es importante distinguir que este procedimiento de convergencia aplica específicamente a las valoraciones SVNS que alimentan la matriz de decisión TOPSIS (Tabla 3), y es independiente del proceso de consenso AHP sobre la importancia relativa de los cinco criterios descrito en §3.4. La convergencia de las valoraciones SVNS se evaluó mediante W de Kendall sobre los rankings de las cinco alternativas-criterio derivados de las puntuaciones crisp de cada experto, conforme al uso estándar del estadístico para datos ordinales (umbral de aceptación $W \geq 0.70$); se obtuvo $W=0.91$, $\chi^2=54.6$, $gl=4$, $p<0.001$ (Leyva Vázquez & Smarandache, 2018).

Alt.	C1 (T,I,F)→crisp	C2 (T,I,F)→crisp	C3 (T,I,F)→crisp	C4 (T,I,F)→crisp	C5 (T,I,F)→crisp
A1	(0.8,0.2,0.1)→7.5	(0.6,0.3,0.2)→6.3	(0.5,0.3,0.3)→5.7	(0.7,0.2,0.1)→7.2	(0.6,0.3,0.2)→6.3
A2	(0.3,0.5,0.6)→3.6	(0.9,0.1,0.1)→8.1	(0.9,0.1,0.1)→8.1	(0.8,0.1,0.1)→7.8	(0.8,0.1,0.2)→7.5
A3	(0.6,0.3,0.2)→6.3	(0.8,0.1,0.2)→7.5	(0.7,0.2,0.2)→6.9	(0.6,0.3,0.2)→6.3	(0.9,0.1,0.1)→8.1

Tabla 3. Tripletas SVNS (T,I,F) y puntuación crisp. $S(T,I,F)=[(2+T-I-F)/3]\times 9$.

3.4 Método AHP

A diferencia del proceso de agregación estadística aplicado a las tripletas SVNS (§3.3) —apropiado cuando se busca preservar y cuantificar la variabilidad individual mediante W de Kendall—, la matriz de comparación pareada AHP se construyó mediante consenso negociado directo, dado que el promedio matemático de matrices de comparación individuales no garantiza por sí mismo la propiedad de reciprocidad ni un índice de consistencia (CR) óptimo, mientras que el consenso explícito sí la garantiza por construcción. Los expertos realizaron comparaciones por pares con la escala de Saaty (1–9) en dos rondas Delphi sucesivas; tras la primera ronda, las discrepancias individuales se discutieron en sesión grupal estructurada y se resolvieron mediante consenso directo sobre la escala fundamental de Saaty (Saaty, 2008). La matriz de consenso resultante (Apéndice A) preserva exactamente la propiedad de reciprocidad ($a_{ji}=1/a_{ij}$) por construcción. El vector de prioridades se obtuvo por el eigenvector principal de esta matriz: $\lambda_{max}=5.0056$, $CI=0.0014$, $RI=1.12$, $CR=0.0013<0.10$. Los pesos finales, reportados con dos decimales exactos del eigenvector (sin ajuste artificial de redondeo), son: C1=36.52%, C2=25.84%, C3=18.49%, C4=12.61%, C5=6.55%.

e-ISSN: 2528-8091
Vol. 11, N° 2 - 2024

3.5 Método TOPSIS

A partir de la matriz crisp (Tabla 3) se aplicó el procedimiento de Hwang y Yoon (1981) y Biswas et al. (2016): (i) normalización vectorial; (ii) ponderación con pesos AHP; (iii) determinación de A^+ y A^- según tipo de criterio; (iv) distancias euclidianas d^+ y d^-; (v) $Ci=d^-/(d^++d^-)$. Mayor Ci indica mejor alternativa.

4. RESULTADOS

4.1 Pesos AHP (CR=0.0013)

La Figura 2 y Tabla 4 presentan los pesos del eigenvector principal. C1 (Degradación AAS) concentra el mayor peso (36.52%), lo que posiciona al envejecimiento funcional como el factor de mayor prioridad según el panel, seguido por la gestión de la indeterminación (C2=25.84%).

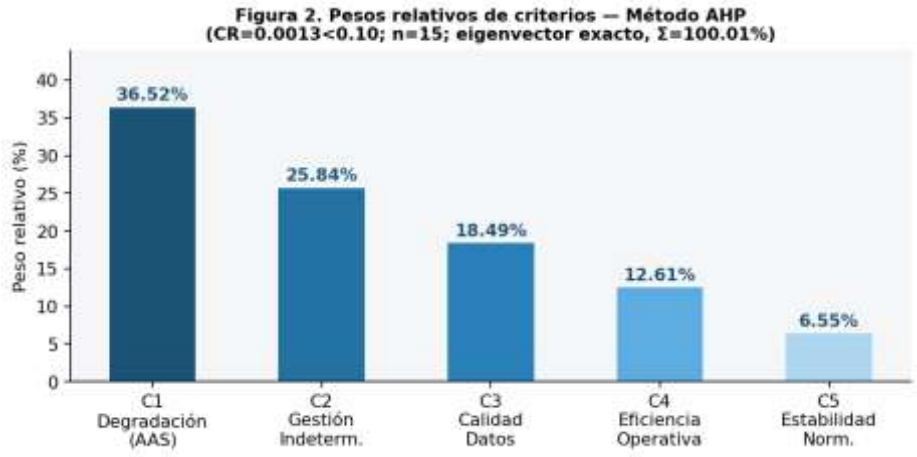

Figura 2. Pesos relativos de criterios — Método AHP (CR=0.0013<0.10; n=15).

Cód.	Criterio	Peso (%)	Ranking
C1	**Degradación por Envejecimiento (AAS)**	36.52	1°
C2	Gestión de la Indeterminación	25.84	2°
C3	Calidad de Datos (Data-Centric AI)	18.49	3°
C4	Eficiencia Operativa	12.61	4°
C5	Estabilidad y Normalización	6.55	5°

Tabla 4. Pesos relativos AHP derivados del eigenvector principal exacto.

e-ISSN: 2528-8091
Vol. 11, N° 2 - 2024

4.2 Ranking TOPSIS

La Tabla 5 presenta la matriz ponderada y normalizada, con variabilidad sustancial entre alternativas (rango C1: 0.126–0.262). Las Figuras 3 y 4 sintetizan el perfil de desempeño y los resultados finales.

Alternativa	v(C1) Costo	v(C2)	v(C3)	v(C4)	v(C5)
A1 (RBA)	0.262	0.128	0.087	0.073	0.033
A2 (DLA-RAG)	0.126	0.164	0.124	0.080	0.039
A3 (HSN)	0.220	0.152	0.106	0.064	0.042
A^+ (ideal)	0.126	0.164	0.124	0.080	0.042
A^- (anti-ideal)	0.262	0.128	0.087	0.064	0.033

Tabla 5. Matriz de decisión ponderada y normalizada. A^+=ideal positiva; A^-=anti-ideal.

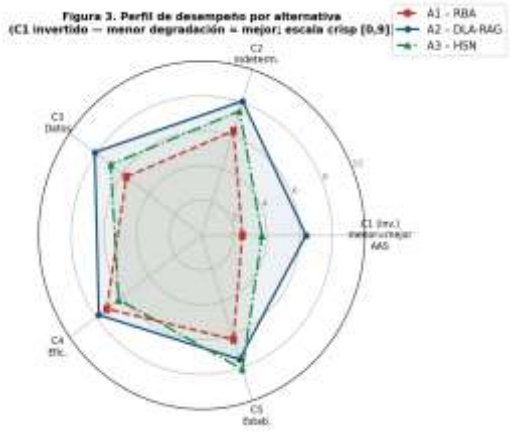

Figura 3. Perfil de desempeño por alternativa (C1 invertido — menor degradación = mejor; escala crisp [0,9]).

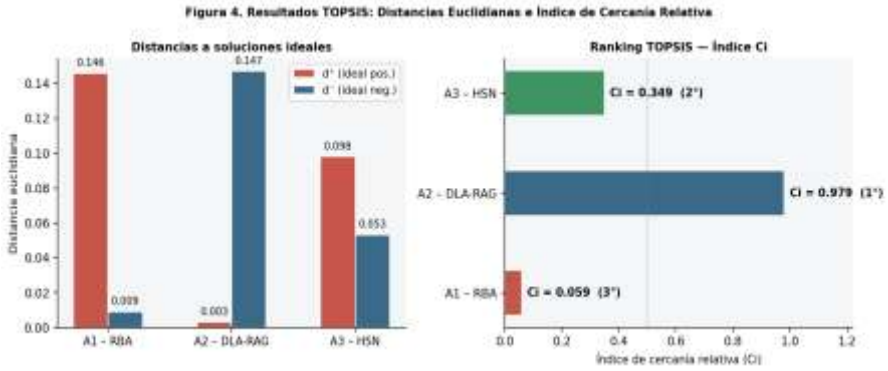

Figura 4. Resultados TOPSIS: distancias euclidianas e índice Ci.

e-ISSN: 2528-8091
Vol. 11, N° 2 - 2024

Alternativa	Arquitectura	d^+	d^-	Ci (Ranking)
A2 (DLA-RAG)	Ag. Aprendizaje Profundo + RAG	0.003	0.147	**0.979 (1°)**
A3 (HSN)	Ag. Híbrido Simbólico-Neuronal	0.098	0.053	0.349 (2°)
A1 (RBA)	Ag. Reactivo Basado en Reglas	0.146	0.009	0.059 (3°)

Tabla 6. Ranking final TOPSIS. A2 (DLA-RAG) es la alternativa dominante con Ci=0.979.

4.3 Análisis de sensibilidad — dos escenarios

Se evaluaron dos escenarios: (S1) C1 reducido a 20% con redistribución proporcional entre C2–C5; (S2) pesos uniformes al 20% cada uno. En ambos escenarios el ranking A2>A3>A1 permanece inalterado (Tabla 7), confirmando la robustez estructural del modelo.

Alternativa	Ci base	S1: C1→20% (Ci ajust.)	ΔCi	S2: Pesos uniformes	ΔCi	Ranking (estable)
A2 (DLA-RAG)	0.979	0.963	−0.016	0.909	−0.070	1° → 1°
A3 (HSN)	0.349	0.428	+0.079	0.427	+0.078	2° → 2°
A1 (RBA)	0.059	0.103	+0.044	0.134	+0.075	3° → 3°

Tabla 7. Análisis de sensibilidad: dos escenarios de perturbación de pesos. Ranking estable en ambos.

4.4 Simulación Monte Carlo y desfuzzificación alternativa

Para complementar el análisis de sensibilidad sobre pesos (§4.3) con una prueba de robustez sobre los datos de entrada, se ejecutó una simulación Monte Carlo perturbando las tripletas SVNS originales (Tabla 3) con ruido gaussiano de media 0 y desviación σ=0.05 en cada componente (T,I,F), recortado al intervalo válido [0,1] tras la perturbación. Se generaron 10,000 iteraciones independientes (semilla aleatoria fija=42 para reproducibilidad exacta); en cada iteración se recalculó la desfuzzificación, la normalización TOPSIS y el índice Ci completo. El valor σ=0.05 se eligió como una

perturbación moderada (10% del rango T,I,F∈[0,1]); para evaluar sensibilidad a esta elección, el procedimiento se repitió con σ∈{0.02, 0.05, 0.10, 0.15} (Tabla 8).

σ (perturbación)	Iteraciones	Ranking A2>A3>A1 preservado	% de estabilidad
0.02	2,000	2,000 / 2,000	100.0%
0.05 (base)	10,000	10,000 / 10,000	**100.0%**
0.10	2,000	1,979 / 2,000	99.0%
0.15	2,000	1,891 / 2,000	94.5%

Tabla 8. Resultados de la simulación Monte Carlo: estabilidad del ranking A2>A3>A1 ante distintos niveles de perturbación σ de las tripletas SVNS (10,000 iteraciones para σ=0.05; 2,000 para los demás valores; semilla=42).

El ranking A2>A3>A1 se preserva en el 100% de las 10,000 iteraciones con σ=0.05, y permanece altamente estable incluso ante perturbaciones más severas (94.5% con σ=0.15, equivalente a una desviación del 30% del rango de cada componente T,I,F). El intervalo de confianza al 95% para Ci(A2) bajo σ=0.05 es [0.953, 1.000], sin solapamiento con el de A3 ([0.232, 0.479]), lo que respalda estadísticamente la separación observada entre alternativas. Adicionalmente, se recalculó el ranking sustituyendo la función de desfuzzificación original (Edalatpanah & Smarandache, 2019, adaptada) por una función alternativa simplificada S'(T,I,F)=(T−F)×9, que omite el componente de indeterminación y cuya salida se recortó a [0,9] al poder arrojar valores negativos (p. ej., −2.7 para A2 en C1): el ranking resultante (A2=0.967, A3=0.440, A1=0.067) preserva el orden A2>A3>A1, aunque con una brecha más estrecha entre A2 y A3. Esto sugiere que la conclusión ordinal del estudio no depende críticamente de la elección específica de función de desfuzzificación, aunque la magnitud exacta de los índices Ci sí varía.

5. DISCUSIÓN

La arquitectura DLA-RAG (A2) domina el ranking con Ci=0.979, valor cercano al límite superior teórico de 1.0, lo que indica una proximidad casi total a la solución ideal positiva. Esta magnitud es consistente con el patrón general descrito en la literatura de TOPSIS neutrosófico aplicado a selección de alternativas tecnológicas (Biswas et al., 2016; Abdel-Basset et al., 2018), donde la alternativa dominante típicamente exhibe una separación sustancial respecto a las demás cuando un criterio de alto peso —como C1 en

este estudio— favorece marcadamente a una sola opción. La comparación directa de valores Ci entre estudios distintos debe interpretarse con cautela, dado que la escala depende de la matriz de decisión y los pesos específicos de cada caso. La ventaja de DLA-RAG se explica por su desempeño en C1 (v(C1)=0.126 vs. 0.262 para RBA), el criterio de mayor peso: el acceso a bases de conocimiento actualizadas mediante RAG mitiga el envejecimiento funcional al reducir la dependencia de parámetros estáticos (Bommasani et al., 2021; Liu et al., 2023).

La primacía de C1 (36.52%) sugiere que la utilidad percibida no depende únicamente del rendimiento inicial, sino de la estabilidad longitudinal. El agente RBA (A1), eficiente en tareas estructuradas, obtiene el último lugar (Ci=0.059) por su incapacidad de actualización. La integración SVNS —distinguiéndose de los IFS (Atanassov, 1986) y ampliando los trabajos de Biswas et al. (2016) y Abdel-Basset et al. (2018)— permite capturar la ambigüedad del juicio experto de forma explícita, avalada por el marco teórico de Leyva Vázquez y Smarandache (2018) y el proceso de inferencia basado en indeterminación de Salmerón y Smarandache (2020). El criterio C2 (25.84%) sugiere que los entornos de ingeniería contemporáneos demandan agentes capaces de operar bajo información incompleta (Wang et al., 2010; Smarandache, 1998).

Encuadre honesto del criterio C1. Es importante precisar que el peso dominante de C1 (36.52%) refleja la *percepción experta* sobre la relevancia del envejecimiento funcional, no una medición longitudinal de degradación observada en producción. Las tripletas SVNS de la Tabla 3 son elicitaciones de juicio informado por la literatura de concept drift (Gama et al., 2014), no series de tiempo de desempeño real de los tres agentes. Esta distinción es metodológicamente relevante: el modelo AAS aquí operacionalizado debe leerse como una *estructura de prioridades basada en expectativa experta*, útil para orientar decisiones de adopción tecnológica bajo incertidumbre, y no como una validación empírica de que DLA-RAG efectivamente se degrada menos que RBA en uso prolongado real. La validación longitudinal directa —monitoreo de los tres agentes en producción durante un periodo extendido— queda fuera del alcance del presente estudio y se identifica explícitamente como dirección prioritaria de investigación futura.

Posible sesgo del panel. Dado que DLA-RAG (A2) obtiene el mejor o segundo mejor desempeño en cuatro de los cinco criterios (Tabla 5), es pertinente examinar si el panel de expertos —compuesto mayoritariamente por profesionales de industria con

experiencia reciente en herramientas basadas en LLM y RAG— pudo haber favorecido sistemáticamente arquitecturas con las que tiene mayor familiaridad práctica, en detrimento de RBA y HSN. Este sesgo de disponibilidad (los expertos valoran mejor lo que conocen y usan con mayor frecuencia) no puede descartarse con los datos actuales y constituye una limitación adicional a las ya declaradas en la sección de Conclusiones.

Esta amenaza a la validez tiene implicaciones directas para el diseño de estudios de réplica. Un panel futuro debería estratificarse deliberadamente por exposición tecnológica previa —por ejemplo, garantizando subgrupos con experiencia comparable en arquitecturas basadas en reglas, sistemas híbridos simbólico-neuronales y LLM con recuperación aumentada—, de modo que las diferencias de familiaridad no se confundan con diferencias reales de desempeño percibido. Asimismo, sería metodológicamente valioso comparar los resultados aquí obtenidos con un panel compuesto exclusivamente por expertos en sistemas simbólicos clásicos, como prueba de robustez del ranking ante composiciones de panel con sesgos de familiaridad opuestos al observado en este estudio.

El análisis de sensibilidad con dos escenarios independientes (Tabla 7) descarta que la superioridad de A2 sea un artefacto de ponderación: en S1 Ci(A2)=0.963 y en S2 Ci(A2)=0.909, manteniendo en ambos casos una ventaja de al menos 0.48 puntos sobre A3. Cabe señalar como limitación metodológica que la función de desfuzzificación adoptada (Edalatpanah & Smarandache, 2019) no es la única propuesta en la literatura SVNS; otras funciones podrían producir variaciones leves en los valores crisp, aunque la amplia separación entre alternativas sugiere estabilidad ordinal del ranking. Siguiendo principios de documentación transparente de modelos de IA (Mitchell et al., 2019), estas limitaciones y los supuestos del proceso de agregación experto se reportan explícitamente para facilitar la evaluación crítica y la replicación del estudio.

5.1 Aporte central del trabajo

Para evitar ambigüedad sobre qué afirma exactamente este estudio, se distinguen explícitamente tres posibles contribuciones y se precisa cuál es la central. *(a) El ranking concreto* (DLA-RAG > HSN > RBA) es un resultado ilustrativo, condicionado al panel de 15 expertos consultado y no debe generalizarse sin replicación independiente. *(b) El criterio AAS* es una propuesta conceptual con anclaje en la literatura de concept drift, pero su operacionalización aquí es de naturaleza perceptual (§5, encuadre honesto de C1), no longitudinal. *(c) El método AHP–TOPSIS–SVNS integrado* —con desfuzzificación explícita, matriz de consenso AHP documentada, validación de convergencia mediante

W de Kendall, y protocolo de análisis de sensibilidad de dos escenarios— constituye el **aporte central y más transferible del trabajo**: una plantilla metodológica reproducible que otros investigadores pueden aplicar a distintos paneles de expertos, conjuntos de alternativas o dominios de aplicación, sustituyendo las tripletas SVNS y las comparaciones AHP por datos propios sin alterar la estructura analítica subyacente.

6. CONCLUSIONES

Este trabajo contribuye en tres dimensiones. Primero, operacionalizó un modelo AHP–TOPSIS neutrosófico reproducible: pesos del eigenvector puro (CR=0.0013), tripletas SVNS verificables, y análisis de sensibilidad con dos escenarios. Segundo, definió y comparó tres arquitecturas de agentes de IA en ingeniería de software con criterios derivados de un embudo de selección documentado (47 artículos → 12 criterios candidatos → 5 retenidos con ≥70% de acuerdo experto). Tercero, estableció que el ranking A2>A3>A1 es estructuralmente robusto dentro de los límites del panel consultado.

Las implicaciones prácticas indican priorizar arquitecturas con mecanismos de actualización del conocimiento (RAG, fine-tuning continuo) sobre soluciones estáticas, especialmente con horizontes de uso superiores a seis meses, si bien esta recomendación debe contrastarse con evidencia longitudinal antes de adoptarse como política general. Las limitaciones incluyen el tamaño muestral (n=15), el foco en ingeniería de software, la dependencia de una única función de desfuzzificación SVNS, el carácter perceptual (no longitudinal) de las valoraciones de degradación funcional, y el riesgo de sesgo de familiaridad del panel hacia arquitecturas LLM-RAG. Futuras investigaciones podrían ampliar el panel de expertos con mayor diversidad de exposición tecnológica, incorporar datos longitudinales reales de degradación en producción, comparar funciones de desfuzzificación alternativas, y extender el marco a conjuntos neutrosóficos intervalares (INS).

Contribuciones de autoría:

Dalton Andrés Pazmiño Pérez y Christian Miguel Muñoz Carrasco: conceptualización, curaduría de datos, análisis formal, investigación, metodología, software, redacción del borrador original. Manuel Reyes Wagnio: supervisión, conceptualización, validación metodológica, revisión crítica y edición. Dayron Rumbaut

Rangel: supervisión, validación, revisión crítica y edición. Todos los autores leyeron y aprobaron la versión final del manuscrito. El orden de autoría refleja la contribución decreciente en la ejecución operativa del estudio (primeros dos autores) seguida de los roles de supervisión académica (últimos dos autores), conforme a la convención estándar en publicaciones de investigación de pregrado y posgrado supervisada.

Declaración de conflicto de interés:

Los autores declaran no tener conflictos de interés relacionados con este estudio.

Disponibilidad de datos y código:

Las matrices de comparación AHP, las tripletas SVNS, los resultados TOPSIS y los parámetros completos de la simulación Monte Carlo (§4.4) están documentados en el Apéndice A y en las Tablas 3 a 8 del presente artículo, con suficiente detalle para su reproducción independiente. El código completo del modelo (cálculo de eigenvector, TOPSIS, análisis de sensibilidad y simulación Monte Carlo con semilla fija=42) y las matrices de decisión en formato editable se depositarán en un repositorio público (OSF o Zenodo) al momento de la publicación, cuyo enlace se incorporará en la versión final. Los cuestionarios anonimizados de los 15 expertos, incluyendo medias y desviaciones estándar por celda de la Tabla 3, están disponibles bajo solicitud razonada a los autores correspondientes.

7. REFERENCIAS

Abdel-Basset, M., Mohamed, M., & Smarandache, F. (2018). A hybrid neutrosophic group ANP-TOPSIS framework for supplier selection problems. Symmetry, 10(6), 226. https://doi.org/10.3390/sym10060226

Atanassov, K. T. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 20(1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3

Biswas, P., Pramanik, S., & Giri, B. C. (2016). TOPSIS method for multi-attribute group decision-making under single-valued neutrosophic environment. Neural Computing and Applications, 27(3), 727–737. https://doi.org/10.1007/s00521-015-1891-2

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Bernstein, M. S., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Broumi, S., Nagarajan, D., Bakali, A., Talea, M., Smarandache, F., & Lathamaheswari, M. (2019). The shortest path problem in interval valued trapezoidal and triangular neutrosophic environment. Complex and Intelligent Systems, 5(4), 391–402. https://doi.org/10.1007/s40747-019-0092-5

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. O., Kaplan, J., et al. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374. https://doi.org/10.48550/arXiv.2107.03374

Chi, P., & Liu, P. (2013). An extended TOPSIS method for the multiple attribute decision making problems based on interval neutrosophic set. Neutrosophic Sets and Systems, 1(1), 63–70. http://fs.unm.edu/NSS/AnExtendedTOPSIS.pdf

Edalatpanah, S. A., & Smarandache, F. (2019). Data envelopment analysis for simplified neutrosophic sets. Neutrosophic Sets and Systems, 29, 215–226. https://doi.org/10.5281/zenodo.3514433

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., & Wang, H. (2024). Large language models for software engineering: A systematic literature review. ACM Transactions on Software Engineering and Methodology, 33(8), 1–79. https://doi.org/10.1145/3695988

Hwang, C. L., & Yoon, K. (1981). Multiple attribute decision making: Methods and applications. Springer. https://doi.org/10.1007/978-3-642-48318-9

Leyva Vázquez, M. Y., & Smarandache, F. (2018). Neutrosofía: Nuevos avances en el tratamiento de la incertidumbre. Pons Publishing House. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3729638

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1–35. https://doi.org/10.1145/3560815

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting.

Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. Information & Management, 42(1), 15–29. https://doi.org/10.1016/j.im.2003.11.002

Peng, X., & Dai, J. (2020). A bibliometric analysis of neutrosophic set: Two decades review from 1998 to 2017. Artificial Intelligence Review, 53(1), 199–255. https://doi.org/10.1007/s10462-018-9652-0

Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4.ª ed.). Pearson. ISBN 978-0-13-461099-3.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. International Journal of Services Sciences, 1(1), 83–98. https://doi.org/10.1504/IJSSci.2008.017590

Salmerón, J. L., & Smarandache, F. (2020). Redesigning decision matrix method with an indeterminacy-based inference process. Advances in Fuzzy Systems, 2020, Article 816038. https://doi.org/10.1155/2020/816038

Smarandache, F. (1998). Neutrosophy: A new branch of philosophy. Multiple-Valued Logic, 8(3), 297–384. American Research Press. http://fs.unm.edu/Neutrosophy.pdf

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

Wooldridge, M. (2020). An introduction to multiagent systems (2.ª ed.). Wiley. ISBN 978-0-470-51946-2.

Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). Data-centric artificial intelligence: A survey. arXiv preprint arXiv:2303.10158. https://doi.org/10.48550/arXiv.2303.10158