Integración de LLMs y Computación con Palabras para la Evaluación del Desempeño 360 grados con Información Heterogénea

F. J. Martínez[1][0000−0001−7969−1737], Ángela Escobar[1][0009−0007−8879−2450], and L. Martínez[1][0000−0003−4245−8813]

Universidad de Jaén, Jaén, Spain
fjmimbre@ujaen.es, aem00007@red.ujaen.es, martin@ujaen.es

Resumen La evaluación del desempeño 360 grados integra valoraciones de múltiples colectivos (supervisores, colaboradores, clientes y el propio empleado) sobre un conjunto de criterios. Los marcos basados en computación con palabras (CWW) y modelos como la 2-tupla lingüística gestionan eficazmente la incertidumbre cualitativa de este proceso, pero requieren formularios estructurados que generan fatiga y baja participación. En este trabajo proponemos integrar Modelos de Lenguaje de Gran Escala (LLMs) como evaluadores lingüísticos automáticos capaces de derivar valoraciones por criterio a partir de texto libre. La contribución central no reside en emplear el LLM como clasificador discreto, sino en interpretar su distribución de probabilidad (restringida y renormalizada sobre los términos de la escala lingüística) como una estructura de creencia compatible con el modelo 2-tupla. La traslación simbólica α emerge así directamente de la estructura de creencia, sin heurísticas externas. La representación resultante es integrable en procesos de agregación basados en operadores de 2-tupla lingüística, los cuales fueron empleados con éxito anteriormente en procesos de evaluación del desempeño 360 grados. Se realiza una prueba de concepto sobre un caso de estudio industrial real del sector textil, empleando textos sintéticos alineados con valoraciones humanas de referencia. Los resultados son consistentes con la viabilidad del marco: el método retiene información ordinal que un clasificador convencional descartaría y, en la etapa de agregación analizada, recupera el ranking de referencia. La limitada escala del caso ($N = 7$, dos criterios, un único modelo) impide conclusiones definitivas y motiva una validación más amplia, que se discute como trabajo futuro.

Keywords: Evaluación del desempeño 360 grados · Computación con palabras · Modelo de 2-tupla lingüística · Modelos de lenguaje de gran escala (LLMs) · Agregación de información heterogénea · Interpretabilidad

1. Introducción

La evaluación del desempeño es un proceso estratégico en la gestión del talento humano que permite tomar decisiones sobre promociones, retribución y

desarrollo de empleados [26,13]. La modalidad 360 grados amplía la perspectiva tradicional al incorporar opiniones de supervisores, colaboradores, clientes y el propio empleado [14], introduciendo dos desafíos fundamentales: la *heterogeneidad* de los dominios de expresión y la *incertidumbre* inherente a las valoraciones cualitativas. Los marcos de Computación con Palabras (CWW) [5] y el modelo de 2-tuplas lingüísticas [2] han demostrado ser herramientas idóneas para abordar estos desafíos; Espinilla et al. [1] propusieron un modelo integrado que gestiona información heterogénea y emplea operadores OWA, media ponderada e integral de Choquet para obtener resultados interpretables.

Sin embargo, los sistemas actuales gestionan simultáneamente datos cuantitativos, valoraciones estructuradas y *texto libre* (comentarios, informes, respuestas cualitativas) producido de forma natural en plataformas como Workday, SAP SuccessFactors u Oracle HCM [20,27]. Aspectos como la actitud o el liderazgo son difíciles de capturar numéricamente, y los modelos CWW han ignorado este tipo de información, generando fatiga y baja participación [28].

La aparición de los LLMs [8] abre una vía para superar estas limitaciones. Frente a trabajos previos en RRHH y NLP que típicamente retienen únicamente la etiqueta más probable [23,24,21,22,16,17], este trabajo aporta un puente conceptual entre la distribución de probabilidad interna del LLM y el formalismo de la 2-tupla: dicha distribución se interpreta como una estructura de creencia [19] sobre los términos lingüísticos, lo que la hace compatible con los operadores de agregación existentes y produce resultados interpretables en lenguaje natural en todas las etapas. Esta interpretabilidad nativa de la representación (sin técnicas post-hoc [12,18]) es especialmente relevante dado que el Reglamento Europeo de IA [15] clasifica explícitamente como sistemas de alto riesgo (Anexo III, punto 4) los utilizados en empleo, gestión de trabajadores y acceso al autoempleo, exigiendo transparencia e interpretabilidad para los afectados.

Las contribuciones principales que se presentan son: (1) un marco para derivar valoraciones lingüísticas desde texto libre mediante LLMs; (2) la transformación de la distribución de logprobs a estructura de creencia y 2-tupla [19], donde la traslación simbólica α no nula preserva información ordinal que un clasificador convencional descartaría; (3) la integración transparente con el proceso de agregación de [1]; y (4) un estudio de viabilidad preliminar sobre un caso de estudio industrial real cuyos resultados son consistentes con la viabilidad del marco propuesto.

La contribución metodológica diferencial reside, por tanto, no en emplear un LLM como clasificador discreto, sino en interpretar su distribución de probabilidad restringida a la escala lingüística como una estructura de creencia, conservando mediante la traslación simbólica α información ordinal intermedia que se perdería al retener únicamente la etiqueta más probable.

2. Preliminares

Esta sección revisa los fundamentos formales sobre los que se construye el marco propuesto. En primer lugar, se introduce el modelo de representa-

ción lingüística de 2-tupla [2] y su mecanismo de gestión de información heterogénea [3]. Seguidamente, se presenta el formalismo de las estructuras de creencia y su transformación a 2-tupla [19], que constituye la base matemática del mapeo de salidas del LLM propuesto en la Sección 3. Finalmente, se describe el modelo de evaluación del desempeño 360 grados basado en CWW [1] sobre el que se integra el presente trabajo.

2.1. El modelo de representación lingüística de 2-tuplas

El modelo de 2-tuplas [2] representa la información lingüística mediante pares (s_i, α), donde $s_i \in S = \{s_0, s_1, \ldots, s_g\}$ es un término lingüístico de un conjunto predefinido y $\alpha \in [-0{,}5, 0{,}5)$ es un valor numérico denominado *traslación simbólica* que expresa la diferencia entre el valor cardinal resultante de una operación de agregación y el índice i del término más cercano.

Definition 1 ([2]). *Sea $S = \{s_0, \ldots, s_g\}$ un conjunto de términos lingüísticos. La función $\Delta_S : [0, g] \to \langle S \rangle$ se define como:*

$$\Delta_S(\beta) = (s_i, \alpha), \quad \textit{con } i = \mathrm{round}(\beta),\ \alpha = \beta - i,$$

donde round$(\cdot)$ *asigna a β el entero más próximo en $\{0, 1, \ldots, g\}$.*

La función inversa $\Delta_S^{-1}(s_i, \alpha) = i + \alpha$ permite recuperar el valor numérico equivalente, de manera que las 2-tuplas pueden identificarse con valores en $[0, g]$. Este modelo permite realizar operaciones aritméticas sobre información lingüística sin pérdida de información.

Para gestionar información heterogénea (numérica, en intervalos y lingüística con distintas granularidades) [4] Herrera et al. [3] propusieron un proceso de unificación basado en la transformación de toda la información a conjuntos difusos sobre un *Conjunto Básico de Términos Lingüísticos* (BLTS, del inglés *Basic Linguistic Term Set*) $\bar{S}$, seguida de su conversión a 2-tuplas mediante la función $\chi : \mathcal{F}(\bar{S}) \to \langle \bar{S} \rangle$.

2.2. Estructuras de creencia y transformación a 2-tuplas

Las *estructuras de creencia* (*belief structures*) proporcionan un mecanismo riguroso para representar la incertidumbre en evaluaciones lingüísticas [19]. Dada una escala lingüística $S_q = \{s_0^q, \ldots, s_{g_q}^q\}$, una estructura de creencia sobre S_q es una distribución:

$$\mathcal{B} = \{(s_j^q,\, p_j) : j = 0, \ldots, g_q\},$$

donde $p_j \geq 0$ representa el grado de creencia de que la valoración corresponde al término s_j^q. Si $\sum_{j=0}^{g_q} p_j = 1$ la estructura es *completa*; si $\sum_{j=0}^{g_q} p_j < 1$ es *incompleta* y la diferencia recoge ignorancia residual.

Para transformar una estructura de creencia completa en una 2-tupla, Martínez et al. [19] definen la función $\xi : \mathcal{B} \to [0, g_q]$ como:

$$\xi(\mathcal{B}) = \frac{\sum_{j=0}^{g_q} j \cdot p_j}{\sum_{j=0}^{g_q} p_j} = \sum_{j=0}^{g_q} j \cdot p_j = \beta, \tag{1}$$

donde la última igualdad se cumple en el caso completo ($\sum p_j = 1$). Aplicando la función Δ_{S_q} al valor β se obtiene la 2-tupla equivalente sin pérdida de información:

$$\Delta_{S_q}(\xi(\mathcal{B})) = (s_i^q, \alpha), \quad \alpha = \beta - \text{round}(\beta).$$

La traslación simbólica α captura la posición relativa de la valoración entre los términos adyacentes de la escala: $\alpha > 0$ indica proximidad al término superior, $\alpha < 0$ al inferior, y $\alpha = 0$ coincidencia exacta con un término de la escala.

En el presente trabajo $\mathcal{B}$ se obtiene de los logprobs internos del LLM en lugar de un sistema de reglas FURBER (*Fuzzy Rule-Based Evidential Reasoning*) como en [19]; el objeto matemático y la aplicación de ξ son idénticos.

2.3. Modelo de evaluación del desempeño 360 grados basado en CWW

El modelo propuesto en [1] estructura el proceso de evaluación integral en tres fases principales: (i) *marco de evaluación heterogéneo*, que define los colectivos de evaluadores A (supervisores), B (colaboradores) y C (clientes), el conjunto de empleados $X = \{x_1, \ldots, x_n\}$ y el conjunto de criterios $Y = \{Y_1, \ldots, Y_p\}$; (ii) *recogida de información*, en la que cada evaluador proporciona valoraciones en el dominio que le corresponde según su perfil y el criterio evaluado; y (iii) *proceso de valoración*, que comprende la unificación de la información heterogénea y un proceso de agregación multipaso.

El proceso de agregación opera en tres etapas sucesivas:

1. **Valores colectivos por criterio** $v_A^k(x_j)$: se obtienen aplicando el operador OWA de 2-tuplas G_w [6] —con vector de pesos w derivado de un cuantificador lingüístico— a las valoraciones de cada colectivo sobre cada criterio.
2. **Valores globales por criterio** $v^k(x_j)$: se obtienen mediante la media ponderada de 2-tuplas F_w sobre los valores colectivos, ponderando la relevancia de cada colectivo para el criterio en cuestión.
3. **Valor global del empleado** $v(x_j)$: se obtiene aplicando la integral de Choquet lingüística de 2-tuplas M_ϕ [7] sobre los valores globales por criterio, capturando la interacción entre criterios mediante medidas difusas.

Este modelo proporciona resultados lingüísticos en todas las etapas del proceso, próximos al modelo cognitivo humano y, por tanto, interpretables y comprensibles para el departamento de Recursos Humanos.

3. Marco Propuesto: LLMs como Evaluadores Lingüísticos

Proponemos extender el modelo de [1] añadiendo una etapa previa de *generación automática de valoraciones lingüísticas* a partir de texto libre mediante

LLMs. El LLM actúa como un *evaluador lingüístico proxy*: recibe el texto libre de un evaluador y el criterio a valorar, y emite un juicio lingüístico estructurado que se transforma al espacio de 2-tupla. La Figura 1 ilustra el proceso completo.

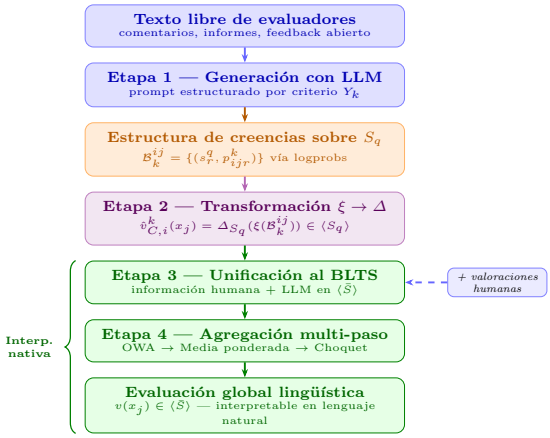


Figura 1. Proceso de evaluación del desempeño 360 grados con LLMs integrados.

3.1. Etapa 1: Generación de la estructura de creencia mediante LLM

Sea t_{ij}^C el texto libre proporcionado por el evaluador $c_i \in C$ sobre el empleado x_j, y $S_q = \{s_0^q, \ldots, s_{g_q}^q\}$ la escala lingüística asociada al criterio Y_k. Para cada par (c_i, x_j), se construye un *prompt* estructurado π que instruye al LLM para emitir una valoración sobre dicho criterio. El prompt $\pi(t_{ij}^C, Y_k, S_q)$ se organiza en tres componentes: (i) *contexto*: descripción del criterio y la escala; (ii) *instrucción*: petición de valoración en la escala indicada; y (iii) *entrada*: el texto libre del evaluador. Un ejemplo para el criterio *Atención al cliente* (Y_7) con escala $S_7 = \{\text{NA}, \text{MBA}, \text{BA}, \text{ME}, \text{AT}, \text{MAT}, \text{PE}\}$ sería:

"Eres un experto en evaluación del desempeño. El criterio a evaluar es ***Atención al cliente****: disponibilidad, saludo, agradecimiento de la compra. Utiliza la escala {NA, MBA, BA, ME, AT, MAT, PE}. A partir del siguiente comentario, indica únicamente la etiqueta más adecuada como respuesta. Comentario: [TEXTO]".*

El diseño sigue los principios de [23,24]: la asignación de rol mejora la coherencia en tareas de dominio [9], la descripción explícita de subcriterios ancla semánticamente la valoración, y restringir la respuesta a una única etiqueta facilita el mapeo directo a $\mathcal{B}$ sin posprocesamiento adicional.

Los modelos de código abierto exponen, para cada token de su vocabulario, el *logprob* $\ell_{ijr}^k = \log P(s_r^q \mid \pi, \theta)$ [25], lo que permite extraer directamente el

logprob de cada término $s_r^q \in S_q$ en la posición de respuesta:

$$\ell_{ijr}^k = \log P_{\text{LLM}}(s_r^q \mid \pi(t_{ij}^C, Y_k, S_q)).$$

Los logprobs se convierten en probabilidades mediante softmax restringida al vocabulario de la escala [10]:

$$p_{ijr}^k = \frac{\exp(\ell_{ijr}^k)}{\sum_{h=0}^{g_q} \exp(\ell_{ijh}^k)}, \quad r = 0, \ldots, g_q.$$

Debe observarse que p_{ijr}^k no representa la probabilidad absoluta asignada por el LLM al término s_r^q sobre todo el vocabulario, sino una probabilidad condicionada y renormalizada sobre el conjunto finito de etiquetas de la escala S_q. Por tanto, la estructura de creencia $\mathcal{B}_k^{ij}$ debe interpretarse como una distribución de creencia relativa sobre los términos lingüísticos admisibles.

El resultado es una **estructura de creencia completa** sobre S_q (véase la Sección 2.2):

$$\mathcal{B}_k^{ij} = \{(s_r^q,\, p_{ijr}^k) : r = 0, \ldots, g_q\}, \quad \sum_{r=0}^{g_q} p_{ijr}^k = 1. \tag{2}$$

La distribución $\mathcal{B}_k^{ij}$ refleja la incertidumbre interna del modelo: si el texto es inequívoco, p_{ijr}^k se concentra en un único índice; si es ambiguo entre dos términos adyacentes, la masa se reparte entre ellos. En el experimento se fijó $T = 0$ y un prompt constante para garantizar reproducibilidad.

Se asume que cada etiqueta de S_q es un único token (verificado para las abreviaturas usadas) y que los logprobs son informativos con $T = 0$ y prompt delimitado [8,25]; $\mathcal{B}_k^{ij}$ debe interpretarse como incertidumbre relativa, no como probabilidad frecuentista.

3.2. Etapa 2: Transformación de la estructura de creencia a 2-tupla

Una vez obtenida la estructura de creencia $\mathcal{B}_k^{ij}$ (Ecuación 2), se aplica la función ξ definida en la Sección 2.2 para obtener el valor numérico equivalente:

$$\beta_k^{ij} = \xi\left(\mathcal{B}_k^{ij}\right) = \sum_{r=0}^{g_q} r \cdot p_{ijr}^k \;\in\; [0,\, g_q]. \tag{3}$$

Aplicando la función Δ_{S_q} se obtiene directamente la 2-tupla lingüística que representa la valoración del evaluador c_i sobre el empleado x_j en el criterio Y_k:

$$\hat{v}_{C,i}^k(x_j) = \Delta_{S_q}\left(\beta_k^{ij}\right) = (s_r^q,\, \alpha_k^{ij}), \tag{4}$$

donde $r = \text{round}(\beta_k^{ij})$ y $\alpha_k^{ij} = \beta_k^{ij} - r \in [-0{,}5,\, 0{,}5)$.

La traslación simbólica α_k^{ij} emerge directamente de la distribución del LLM, sin heurísticas externas: $|\alpha_k^{ij}|$ cercano a 0,5 indica masa repartida entre dos términos adyacentes, y $\alpha_k^{ij} \approx 0$ concentración en un único término.

3.3. Integración en el proceso de agregación

Una vez obtenida la 2-tupla $\hat{v}^k_{C,i}(x_j) \in \langle S_q \rangle$, el proceso continúa de manera idéntica al modelo de [1]: el LLM actúa como un evaluador adicional cuyas valoraciones son formalmente indistinguibles de las humanas en el espacio de 2-tuplas. El estudio de viabilidad de la Sección 4 valida el puente logprobs $\rightarrow$ 2-tupla y la etapa OWA; las etapas restantes operan sobre 2-tuplas sin adaptación.

4. Estudio de Viabilidad

4.1. Descripción del caso de estudio

Para evaluar la viabilidad del marco propuesto empleamos el caso de estudio industrial real de [1], correspondiente a una empresa multinacional del sector textil con 7 empleados evaluados ($x_1, \ldots, x_7$), 11 criterios ($Y_1, \ldots, Y_{11}$) y 38 evaluadores distribuidos en tres colectivos: supervisores ($|A| = 7$), colaboradores ($|B| = 11$) y clientes ($|C| = 20$).

El experimento se focaliza en el colectivo de **clientes** ($|C| = 20$) y los dos criterios que este colectivo evalúa: Y_7 (*Atención al cliente*: disponibilidad, saludo y agradecimiento de la compra) con escala $S_7 = \{\text{NA}, \text{MBA}, \text{BA}, \text{ME}, \text{AT}, \text{MAT}, \text{PE}\}$, y Y_{11} (*Imagen personal*: forma de vestir y trato) con escala $S_3 = \{\text{NA}, \text{ME}, \text{PE}\}$, donde las abreviaturas denotan, en orden creciente: Nada (NA), Muy Bajo (MBA), Bajo (BA), Medio (ME), Alto (AT), Muy Alto (MAT) y Perfecto (PE) para S_7, y Nada (NA), Medio (ME) y Perfecto (PE) para S_3. Esto supone un total de **280 valoraciones de referencia** (140 por criterio).

La Tabla 1 recoge las valoraciones de referencia del experimento.

Tabla 1. Muestra de valoraciones de referencia (c_1–c_5) sobre Y_7 (escala S_7) e Y_{11} (escala S_3).

	Y_7							Y_{11}						
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_1	x_2	x_3	x_4	x_5	x_6	x_7
c_1	ME	ME	MBA	ME	NA	MAT	PE	PE	ME	ME	PE	ME	PE	PE
c_2	ME	MBA	NA	MBA	MBA	ME	MAT	PE	ME	ME	PE	ME	PE	PE
c_3	ME	MBA	NA	MBA	ME	ME	MAT	ME	ME	NA	PE	ME	PE	PE
c_4	ME	MBA	NA	MBA	ME	MAT	MAT	ME	ME	NA	ME	ME	ME	PE
c_5	ME	MBA	NA	MBA	MBA	MAT	MAT	ME	ME	ME	NA	ME	ME	ME

4.2. Generación de texto libre sintético

Esta prueba de concepto requiere texto libre como entrada al LLM. Dado que el caso de estudio original recogió valoraciones estructuradas pero no comentarios textuales, se ha generado texto sintético alineado con cada valoración de referencia. Para cada trío (c_i, x_j, Y_k) se redactó un comentario de 2–3 frases en

lenguaje coloquial que describiese conductualmente un desempeño consistente con la etiqueta lingüística de referencia, evitando deliberadamente el término de la etiqueta y sus sinónimos directos. Esta restricción busca que el LLM tenga que inferir el nivel a partir de descripciones de comportamiento (el escenario realista de aplicación) y no por simple coincidencia léxica con la escala. La redacción la realizó manualmente una persona con formación en Economía y experiencia en el dominio, sin asistencia de ningún LLM. Como limitación, el redactor conocía la valoración de referencia, lo que podría introducir sesgos sutiles (*demand characteristics*) pese a la restricción léxica; un control completo requeriría anotación ciega o textos reales y queda pendiente para trabajo futuro. La Tabla 2 muestra ejemplos representativos para distintos niveles de valoración.

Tabla 2. Ejemplos de textos sintéticos para distintas valoraciones de referencia.

Crit.	Par	Ref.	Texto sintético generado
Y_7	c_1, x_5	NA	*"No recibí ninguna atención durante toda la visita. Ningún empleado se acercó ni respondió cuando intenté llamarles."*
Y_7	c_3, x_1	ME	*"El servicio fue aceptable. Cuando pregunté me atendieron, aunque no estaban muy pendientes y tardaron en acercarse."*
Y_7	c_1, x_6	MAT	*"La atención fue muy buena. La dependienta estaba muy dispuesta, nos saludó al entrar y se ofreció a ayudarnos en varias ocasiones."*
Y_7	c_1, x_7	PE	*"Atención impecable. Nos recibieron con una sonrisa, resolvieron todas las dudas y nos agradecieron la compra al marcharnos."*
Y_{11}	c_3, x_3	NA	*"La empleada iba bastante desaliñada, su ropa no guardaba relación con el estilo de la tienda y causó una impresión muy negativa."*
Y_{11}	c_1, x_2	ME	*"La empleada vestía de forma correcta, sin que su imagen destacara especialmente ni en positivo ni en negativo."*
Y_{11}	c_1, x_1	PE	*"Imagen impecable. Vestía con mucho estilo, muy acorde a la filosofía de la marca, y transmitía una imagen muy cuidada y profesional."*

Es importante señalar que los textos sintéticos son más directos y menos ambiguos que los comentarios reales, por lo que las métricas obtenidas deben interpretarse como un *límite superior* del rendimiento esperable con texto humano real.

4.3. Configuración del LLM y métricas de evaluación

Se empleó el modelo `Llama-3.1-8B-Instruct` [11] en cuantización Q4_K_M, servido localmente mediante `llama.cpp` sobre hardware Apple Silicon con aceleración Metal. Se eligió por ser open-weight (acceso completo a logprobs y reproducibilidad total) y ejecutable en hardware de consumo, requisitos que los modelos propietarios o de mayor tamaño no satisfacen simultáneamente. Se fijó $T = 0$ (decodificación greedy determinista), `seed=42`, `n_predict=1` y `n_probs=20`, renormalizando por softmax restringida a los términos de S_q para obtener $\{p^k_{ijr}\}$; la baja dispersión de los logprobs de respuesta sustenta que la masa asignada a tokens fuera de la escala es ruido no semántico. Se realizaron 280 inferencias

independientes. Se asume que la cuantización Q4_K_M tiene un impacto despreciable en el orden relativo de los logprobs dado el uso de etiquetas de un único token y escalas de baja granularidad; su verificación cuantitativa queda como trabajo futuro.

Las métricas de evaluación utilizadas fueron: **Tasa de Coincidencia Exacta (TCE)**, definida como la proporción de inferencias en las que $s_r^q = \text{argmax}_r\, p_{ijr}^k$ coincide con la etiqueta de referencia; **Error de Traslación Medio (ETM)**, calculado como $|\beta_k^{ij} - r_{\text{ref}}|/g$ promediado sobre las inferencias (donde r_{ref} es el índice de la etiqueta de referencia y g la granularidad de la escala); y la **correlación de Kendall** τ sobre el ranking global de empleados.

Adicionalmente, para cuantificar la aportación del mapeo probabilístico, se compara con un **baseline** que utiliza únicamente la etiqueta más probable ($s_r = \text{argmax}_r\, p_{ijr}^k$) con $\alpha = 0$, descartando la distribución de probabilidad sobre la escala. Este baseline representa el uso habitual de un LLM como clasificador discreto: se conserva únicamente la etiqueta más probable y se descarta toda la distribución de probabilidad restante, como es habitual en la literatura de LLM-as-a-judge [23,24]. Permite así aislar el efecto específico de utilizar la estructura completa de creencias y la traslación simbólica α.

4.4. Resultados y discusión

Alcance del experimento. El estudio de viabilidad se centra en validar el puente logprobs → 2-tupla y la primera etapa de agregación (OWA intracolectivo). Las dos etapas restantes del pipeline de [1] no se ejecutan en este experimento por dos razones: (i) la media ponderada intercolectivos requeriría incorporar también supervisores y colaboradores, mientras que los textos disponibles proceden únicamente del colectivo de clientes; y (ii) la integral de Choquet intercriterios exigiría una medida difusa sobre $\{Y_7, Y_{11}\}$ cuya elicitación con expertos queda fuera del alcance de esta prueba de concepto. El ranking global se obtiene, por tanto, mediante la media de las valoraciones colectivas en Y_7 e Y_{11}, equivalente a una integral de Choquet con medida aditiva uniforme. La validación de las etapas 2 y 3 con datos completos del caso de estudio queda como trabajo futuro.

La Tabla 3 resume las métricas obtenidas por el método propuesto y el baseline. El método propuesto alcanza una TCE global del 88,9 % (80,0 % en Y_7 y 97,9 % en Y_{11}), con un ETM global de 0,041. El mejor rendimiento en Y_{11} es coherente con la menor granularidad de su escala: con 3 etiquetas la probabilidad de error de clasificación es inherentemente menor que con 7.

Un resultado aparentemente paradójico es que el baseline presenta una TCE global *mejor* (92,1 % vs. 88,9 %). Una posible explicación es que el método propuesto incorpora $\alpha \neq 0$ cuando el modelo distribuye masa entre dos términos adyacentes, lo que incrementa el ETM individual pero retiene información ordinal que el baseline descarta. Cabe señalar que este efecto no se limita a un ajuste de α: cuando la masa de probabilidad cruza el punto medio entre dos términos adyacentes, el redondeo de Δ_{S_q} desplaza también el término base, como ocurre en x_4 sobre Y_7 (Tabla 4), donde el método propuesto agrega a $(MB,\ -,43)$

Tabla 3. Métricas comparativas: método propuesto vs. baseline (argmax, $\alpha = 0$).

Criterio	**Método**	**TCE (%)**	**ETM**
Y_7 (S_7, 7 términos)	Propuesto	80,0	0,052
	Baseline	86,4	0,023
Y_{11} (S_3, 3 términos)	Propuesto	97,9	0,031
	Baseline	97,9	0,021
Global (280 inf.)	Propuesto	88,9	0,041
	Baseline	92,1	0,022

Método	**Ranking global**	$\boldsymbol{\tau}$ (p)
Propuesto	$x_7 \succ x_6 \succ x_1 \succ x_4 \succ x_5 \succ x_2 \succ x_3$	1,000 ($p < 0{,}001$)
Baseline	$x_7 \succ x_6 \succ x_1 \succ x_2 \succ x_4 \succ x_5 \succ x_3$	0,810 ($p = 0{,}011$)
Referencia	$x_7 \succ x_6 \succ x_1 \succ x_4 \succ x_5 \succ x_2 \succ x_3$	—

frente al (CNA, ,33) del baseline. En el caso analizado, esta diferencia se traduce en rankings distintos: el método propuesto recupera el ranking de referencia ($\tau = 1{,}000$, $p < 0{,}001$), mientras que el baseline incurre en dos pares discordantes —x_2 se sitúa por delante de x_4 y de x_5, mientras que en la referencia ocurre lo contrario— ($\tau = 0{,}810$, $p = 0{,}011$). Una causa plausible es que, al colapsar $\alpha = 0$, el baseline produce empates en la agregación OWA (x_2, x_4 y x_5 comparten (CNA, ,33) en Y_7, véase Tabla 4) que limitan la discriminación entre empleados con distribuciones de probabilidad distintas. Conviene insistir en que estos hallazgos son ilustrativos, no concluyentes: con $N = 7$ no es posible descartar que la diferencia observada se atenúe o se invierta en muestras mayores.

La Tabla 4 detalla las valoraciones colectivas (OWA, cuantificador "Most") unificadas al BLTS $\bar{S} = S_9 = \{\text{NA}, \text{CNA}, \text{MB}, \text{BA}, \text{ME}, \text{AT}, \text{LAT}, \text{MA}, \text{PE}\}$. Nótese que las abreviaturas difieren entre escalas: el término *Muy Bajo* se abrevia MBA en S_7 y MB en S_9, al tratarse de conjuntos de términos distintos con convenciones propias. En esta tabla se aprecian los empates del baseline en Y_7.

Tabla 4. Valoraciones colectivas (OWA) en $\bar{S} = S_9$ para cada empleado, por criterio.

		x_1	x_2	x_3	x_4	x_5	x_6	x_7
Y_7	Referencia	(BA, −.07)	(CNA, .07)	(NA, .00)	(CNA, .33)	(CNA, .20)	(AT, −.47)	(LAT, −.13)
	Propuesto	(BA, −.22)	(CNA, .38)	(CNA, −.05)	(MB, −.43)	(CNA, .41)	(ME, .17)	(LAT, −.44)
	Baseline	(BA, −.07)	(CNA, .33)	(CNA, −.07)	(CNA, .33)	(CNA, .33)	(AT, −.47)	(LAT, −.13)
Y_{11}	Referencia	(AT, −.20)	(ME, .00)	(BA, −.20)	(ME, .00)	(ME, .00)	(ME, .00)	(MA, −.20)
	Propuesto	(AT, −.09)	(ME, .00)	(ME, −.02)	(ME, .12)	(ME, .02)	(ME, .11)	(MA, −.17)
	Baseline	(AT, −.20)	(ME, .00)	(ME, .00)	(ME, .00)	(ME, .00)	(ME, .00)	(MA, −.20)

5. Conclusiones y Trabajo Futuro

Hemos propuesto un marco que integra LLMs como evaluadores lingüísticos en el modelo CWW de evaluación 360 grados, mapeando sus logprobs a una es-

tructura de creencia y de ahí a 2-tupla. Compatible con [1] y aplicable a cualquier evaluación lingüística multicriterio con texto libre, el sistema ofrece interpretabilidad nativa en lenguaje natural, alineada con los requisitos del Reglamento Europeo de IA [15] para sistemas de alto riesgo. El estudio de viabilidad arroja resultados consistentes con la hipótesis de partida (TCE del 88,9 % y recuperación del ranking de referencia, $\tau = 1{,}000$ frente a $\tau = 0{,}810$ del baseline), si bien el reducido tamaño muestral, el carácter sintético de los textos y el uso de un único modelo cuantizado limitan el alcance de las conclusiones que cabe extraer. Estos resultados deben entenderse, por tanto, como una primera evidencia de que la traslación simbólica derivada de la distribución del LLM puede aportar información ordinal útil al proceso de agregación.

Las conclusiones deben interpretarse con cautela. El experimento valida la viabilidad técnica del mapeo propuesto, pero no demuestra todavía la validez externa del sistema en entornos reales de evaluación de desempeño. Las limitaciones principales son el uso de textos sintéticos, el tamaño reducido del caso ($N = 7$, dos criterios, un único colectivo), un solo LLM cuantizado sin calibración probabilística y la validación parcial de la agregación.

Como trabajo futuro, las líneas prioritarias son validar con texto libre real en colaboración con empresas, explorar estructuras de creencia incompletas del tipo ($\sum p_{ijr}^k < 1$), extender el análisis a escalas de mayor granularidad (S_9, S_{11}) y comparar distintos modelos open-weight (Qwen,Deepseek etc) para verificar la independencia del marco respecto al LLM subyacente.

Referencias

1. Espinilla, M., de Andrés, R., Martínez, F.J., Martínez, L.: A 360-degree performance appraisal model dealing with heterogeneous information and dependent criteria. Information Sciences **222**, 459–471 (2013)
2. Herrera, F., Martínez, L.: A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on Fuzzy Systems **8**(6), 746–752 (2000)
3. Herrera, F., Martínez, L., Sánchez, P.J.: Managing non-homogeneous information in group decision making. European Journal of Operational Research **166**(1), 115–132 (2005)
4. Martínez, L., Herrera, F.: An overview on the 2-tuple linguistic model for computing with words in decision making: extensions, applications and challenges. Information Sciences **207**(1), 1–18 (2012)
5. Zadeh, L.A.: From computing with numbers to computing with words: from manipulation of measurements to manipulation of perceptions. IEEE Transactions on Circuits and Systems I **46**(1), 105–119 (1999)
6. Yager, R.R.: On ordered weighted averaging operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics **18**, 183–190 (1988)
7. Choquet, G.: Theory of Capacities. Annales de l'Institut Fourier **5**, 131–295 (1953)
8. Brown, T.B., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)
9. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
10. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
11. Dubey, A., et al.: The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783 (2024)

12. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)
13. Fisher, C., Schoenfeldt, L.F., Shaw, J.B.: Human Resources Management. Houghton Mifflin Company, Boston (2006)
14. Lepsinger, R., Lucia, A.D.: The Art and Science of 360 Degree Feedback. John Wiley & Sons (2009)
15. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series (2024)
16. Ma, X., Liu, W., Zhao, C., Tukhvatulina, L.R.: Can large language model predict employee attrition? arXiv preprint arXiv:2411.01353 (2024)
17. Dasaklis, T.K., Giannopoulos, P.G., Koutras, D., Malamas, V., Chountalas, P.: Large language models in human resource management: a systematic literature review of applications, open issues and future research directions. SSRN preprint 5314976 (2025)
18. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for large language models: a survey. ACM Transactions on Intelligent Systems and Technology **15**(2), 1–38 (2024)
19. Martínez, L., Liu, J., Yang, J.B., Herrera, F.: A multigranular hierarchical linguistic model for design evaluation based on safety and cost analysis. International Journal of Intelligent Systems **20**(12), 1161–1194 (2005)
20. Aguinis, H.: Performance Management, 4th edn. Chicago Business Press, Chicago (2019)
21. Speer, A.B.: Scoring dimension-level job performance from narrative comments: validity and generalizability when using natural language processing. Organizational Research Methods **24**(3), 572–597 (2021)
22. Speer, A.B., Perrotta, J., Kordsmeyer, T.L.: Taking it easy: off-the-shelf versus fine-tuned supervised modeling of performance appraisal text. Organizational Research Methods **29**(1), 73–91 (2026) `https://doi.org/10.1177/10944281241271249`
23. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al.: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems **36**, 46595–46623 (2023)
24. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-EVAL: NLG evaluation using GPT-4 with better human alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2511–2522 (2023)
25. Kauf, C., Chersoni, E., Lenci, A., Fedorenko, E., Ivanova, A.A.: Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In: Proceedings of the 7th BlackboxNLP Workshop, pp. 263–277. Association for Computational Linguistics (2024)
26. DeNisi, A.S., Murphy, K.R.: Performance appraisal and performance management: 100 years of progress? Journal of Applied Psychology **102**(3), 421–433 (2017)
27. Pulakos, E.D., Mueller-Hanson, R.A., Arad, S.: The evolution of performance management: searching for value. Annual Review of Organizational Psychology and Organizational Behavior **6**, 249–271 (2019)
28. Pulakos, E.D., Mueller Hanson, R., Arad, S., Moye, N.: Performance management can be fixed: an on-the-job experiential learning approach for complex behavior change. Industrial and Organizational Psychology **8**(1), 51–76 (2015)