Modelado del Consenso y el Conflicto en Sistemas Agénticos Basados en Roles para la Verificación de Afirmaciones

Juan Barea-Rojo[1], José Ramón Trillo[2], Juan Carlos González-Quesada[1,3], Francisco Javier Cabrerizo[1,3], and Ignacio Javier Pérez[1,3]

[1] Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, 18071, España
barearojojuan@gmail.com
[2] Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, 50018, España
jtrillo@unizar.es
[3] Departamento de Ciencias de la Computación e Inteligencia Artificial, Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional, DaSCI, Universidad de Granada, Granada, 18071, España
juancarlosgq@ugr.es, cabrerizo@decsai.ugr.es, ijperez@decsai.ugr.es

Resumen Los sistemas agénticos basados en grandes modelos de lenguaje se utilizan cada vez más en tareas de verificación, pero sus predicciones finales suelen obtenerse mediante reglas simples de agregación que no siempre reflejan la fiabilidad de la decisión colectiva. Cuando las salidas de los agentes están muy correlacionadas, el acuerdo puede ser engañoso y la fusión directa puede ocultar información relevante sobre el comportamiento interno del conjunto. Este trabajo propone una arquitectura en dos etapas para la verificación de afirmaciones. Primero, varios agentes especializados por rol evalúan de forma independiente el mismo par afirmación-evidencia y generan juicios estructurados. Después, el comportamiento colectivo se representa mediante señales de consenso, desacuerdo, conflicto, dominancia y confianza, que son procesadas por una capa de decisión supervisada. Los resultados muestran que la agregación directa aporta beneficios limitados, mientras que la arquitectura propuesta produce decisiones finales más eficaces y equilibradas que las alternativas evaluadas. La contribución principal consiste en modelar explícitamente cómo se forman las decisiones colectivas, en lugar de asumir que el consenso es fiable por sí mismo.

Keywords: Sistemas agénticos · Grandes modelos de lenguaje · Verificación de afirmaciones · Toma de decisiones en grupo · Consenso

1. Introducción

Los grandes modelos de lenguaje, o *Large Language Models* (LLM), han impulsado el desarrollo de sistemas agénticos para tareas que requieren razonamiento complejo y coordinación. En estos sistemas, varios agentes pueden asumir roles

distintos para abordar un mismo problema desde perspectivas complementarias. Marcos como AutoGen han mostrado el potencial de los flujos de trabajo agénticos orquestados [19], mientras que CAMEL ha ilustrado cómo la especialización por roles puede inducir comportamientos diferenciados en sociedades basadas en LLM [10]. Sin embargo, aumentar el número de agentes no garantiza una decisión final más fiable. Trabajos recientes han cuestionado que los *prompts* basados en roles o el debate multiagente sean suficientes para mejorar de forma consistente el razonamiento [14], y han mostrado que la colaboración efectiva requiere mecanismos adicionales de coordinación o aprendizaje [12].

En verificación de afirmaciones, esta limitación es especialmente relevante. Muchos sistemas agénticos siguen utilizando reglas simples de agregación, como votación o fusión directa de puntuaciones. No obstante, cuando los agentes comparten sesgos o patrones de error, el consenso puede ser engañoso: varios agentes pueden coincidir no porque la respuesta sea correcta, sino porque fallan de forma correlacionada. Por ello, la decisión final no debería depender solo de qué etiqueta recibe más apoyo, sino también de cómo se forma ese apoyo dentro del conjunto.

Esta idea conecta la verificación agéntica con la toma de decisiones en grupo. Estrategias recientes de agregación han mostrado que explotar información de orden superior puede superar a la votación mayoritaria [1], y que la coordinación estructurada puede mejorar frente a combinaciones agénticas planas [18]. Desde esta perspectiva, la calidad de una decisión colectiva depende no solo del acuerdo, sino también de la estructura del consenso, el conflicto, el desacuerdo, la dominancia y la confianza [15,8].

Este trabajo propone una arquitectura en dos etapas que separa la generación de juicios agénticos de la estimación supervisada de su fiabilidad colectiva. Primero, varios agentes especializados evalúan de forma independiente el mismo par afirmación-evidencia. Después, su comportamiento colectivo se transforma en una representación explícita basada en señales de consenso, desacuerdo, conflicto, dominancia y confianza, que alimenta una capa supervisada de meta-agregación. A diferencia de un esquema clásico de *stacked ensemble learning*, la segunda etapa no se plantea como una simple capa adicional de clasificación, sino como un mecanismo para caracterizar patrones de acuerdo, conflicto y confianza entre agentes. Esta representación puede extenderse a escenarios de verificación continua, en los que la evolución del desacuerdo, la incertidumbre y el conflicto aporte señales adicionales sobre la estabilidad de la decisión.

El resto del artículo se organiza de la siguiente forma. La Sección 2 presenta los fundamentos sobre toma de decisiones en grupo y sistemas agénticos basados en LLM. La Sección 3 describe el marco propuesto. La Sección 4 presenta la evaluación experimental sobre Climate-FEVER. La Sección 5 discute las implicaciones y limitaciones del enfoque propuesto. Finalmente, la Sección 6 recoge las conclusiones y plantea líneas de trabajo futuro.

2. Fundamentos

Esta sección presenta los dos fundamentos del enfoque propuesto: la toma de decisiones en grupo y los sistemas agénticos basados en LLM. La conexión entre ambos permite interpretar las salidas de varios agentes no solo como predicciones independientes, sino como juicios colectivos cuya fiabilidad depende del consenso, el desacuerdo, el conflicto y la confianza.

2.1. Toma de Decisiones en Grupo

La toma de decisiones en grupo, o *Group Decision Making* (GDM), estudia cómo varios decisores pueden alcanzar una decisión colectiva gestionando el desacuerdo. En los modelos basados en consenso, el objetivo no es solo agregar preferencias, sino guiar al grupo hacia un nivel aceptable de acuerdo mediante un proceso de alcance de consenso, incluso cuando las preferencias son heterogéneas [9]. En entornos difusos, los modelos de *soft consensus* sustituyen la unanimidad estricta por grados de acuerdo, permitiendo medir la coherencia colectiva de forma más flexible [8].

Un operador habitual en este contexto es *Ordered Weighted Averaging* (OWA), que agrega valores ordenados asignando distinto peso a cada posición. Dado un conjunto de entradas $x_1, \ldots, x_n$, el operador OWA se define como:

$$\mathrm{OWA}(x_1, \ldots, x_n) = \sum_{i=1}^{n} w_i \, x_{(i)}, \tag{1}$$

donde $x_{(i)}$ representa el i-ésimo mayor valor del vector reordenado, y w_i son los pesos asociados, cumpliendo $w_i \geq 0$ y $\sum_{i=1}^{n} w_i = 1$ [6]. La GDM también se ha extendido a escenarios de gran escala, donde aparecen retos como la escalabilidad, la heterogeneidad, la robustez y la influencia social [3,7]. Modelos recientes han incorporado señales de confianza y conflicto para mejorar los procesos de consenso [4]. Por ello, resulta necesario analizar no solo el acuerdo final, sino también cómo se forma dicho acuerdo y qué dinámicas sostienen la decisión colectiva [15].

2.2. Sistemas Agénticos Basados en LLM

Los sistemas agénticos basados en LLM tratan los modelos fundacionales como entidades capaces de procesar contexto, generar decisiones intermedias e interactuar con otros agentes. Estos sistemas se apoyan en arquitecturas *transformer* [16] y en taxonomías recientes que describen componentes como roles, memoria, herramientas y mecanismos de coordinación [17]. En este marco, varios agentes pueden analizar una misma tarea siguiendo roles diferentes. AutoGen muestra cómo los protocolos de interacción estructurados permiten flujos de razonamiento colaborativo [19], mientras que CAMEL ilustra cómo la especialización por roles puede inducir comportamientos diferenciados en sociedades basadas en LLM [10].

Sin embargo, la colaboración agéntica no garantiza por sí misma una decisión colectiva fiable. Cuando los agentes comparten el mismo modelo base, *prompts* similares o sesgos cercanos, sus salidas pueden estar fuertemente correlacionadas y producir consenso alrededor de una respuesta incorrecta. Por ello, distintos trabajos han explorado mecanismos para reforzar el razonamiento, como la combinación de razonamiento y recuperación de evidencia [20] o estrategias de debate y desacuerdo estructurado para mejorar robustez y factualidad [5].

Desde esta perspectiva, los sistemas agénticos basados en LLM pueden entenderse como realizaciones computacionales de procesos de decisión en grupo. Cada agente produce un juicio individual, pero la calidad de la salida final depende de cómo esos juicios se combinan y de si el consenso observado refleja robustez real o errores correlacionados. La innovación de este trabajo se sitúa precisamente ahí: en lugar de aplicar únicamente una regla fija de agregación, proponemos representar explícitamente el comportamiento colectivo mediante señales de consenso, desacuerdo, conflicto, dominancia y confianza, y utilizar dicha representación para aprender la fiabilidad de la decisión final.

3. Arquitectura

La arquitectura propuesta sigue un marco de verificación de hechos en dos etapas, ilustrado en la Figura 1. Para mayor claridad, el diagrama muestra una representación simplificada del flujo de trabajo, destacando los componentes principales que intervienen en el proceso de verificación.

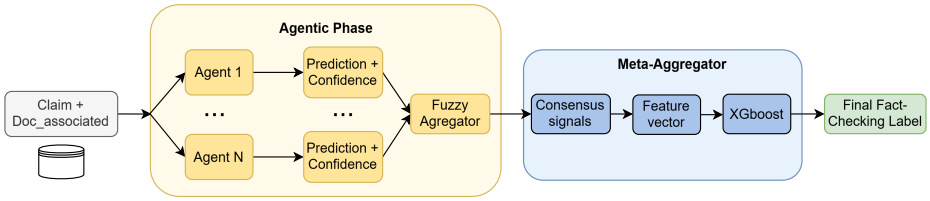


Figura 1. Visión general de la arquitectura agéntica propuesta para la verificación de hechos.

En la primera etapa, varios agentes de razonamiento analizan de forma independiente la misma afirmación y su evidencia asociada. En la segunda etapa, el comportamiento colectivo de los agentes se codifica en una representación estructurada, que posteriormente es procesada por un modelo de meta-aprendizaje para obtener la decisión final de verificación. La separación entre ambas etapas permite distinguir la generación de juicios individuales de la estimación de su fiabilidad colectiva. La fase agéntica produce varios juicios sobre el mismo par afirmación-evidencia; la fase de meta-agregación transforma esos juicios en variables que describen acuerdo, desacuerdo, dominancia, conflicto y confianza, y aprende su relación con la etiqueta final.

3.1. Fase Agéntica

La primera etapa corresponde a la *fase agéntica*, en la que todos los agentes analizan en paralelo el mismo par afirmación-evidencia (c, e).

Sea $A = \{a_1, a_2, \ldots, a_n\}$ el conjunto de agentes. Cada agente a_i produce una salida

$$y_i = (l_i, s_i, r_i),$$

donde $l_i \in \{SUPPORT, REFUTE, NOT_ENOUGH_INFO\}$ es la etiqueta predicha, s_i es la puntuación de confianza y r_i es una justificación textual. Esta justificación no se utiliza directamente por el modelo de meta-agregación, pero permite depurar el sistema y analizar cualitativamente los patrones de razonamiento de los agentes. Una vez generadas las salidas de todos los agentes, se aplica una agregación difusa para obtener una representación suave del apoyo colectivo. Para cada etiqueta, las puntuaciones de confianza se combinan mediante el operador OWA definido en la Ecuación (1), produciendo un grado de pertenencia agregado. Los pesos se fijan antes del entrenamiento mediante un perfil geométrico decreciente, evitando su ajuste sobre el conjunto de evaluación. Para n agentes, se definen como

$$\tilde{w}_j = 0{,}65^{j-1}, \qquad w_j = \frac{\tilde{w}_j}{\sum_{k=1}^{n} \tilde{w}_k}, \qquad j = 1, \ldots, n. \tag{2}$$

Se usa una configuración que asigna mayor influencia a los valores más altos, pero mantiene la contribución del resto de agentes. De este modo, el operador representa una agregación orientada a la dominancia, situada entre la media uniforme y el máximo estricto. El parámetro se mantiene fijo en todos los experimentos para garantizar una comparación reproducible y evitar que las mejoras procedan del ajuste específico de los pesos.

3.2. Meta-Agregador

La segunda etapa corresponde al *meta-agregador*. En lugar de utilizar directamente la salida del agregador difuso como decisión final, el sistema transforma el comportamiento colectivo de los agentes en una representación estructurada. Para cada instancia afirmación-evidencia, se calculan *señales de consenso* sin utilizar la etiqueta real durante la construcción de características. Estas señales se derivan principalmente de las salidas de la fase agéntica y se complementan con algunos descriptores simples de la entrada.

La representación incluye valores de pertenencia fusionados para las tres etiquetas de verificación, *SUPPORT*, *REFUTE* y *NOT_ENOUGH_INFO*; indicadores de incertidumbre y dominancia, como la entropía y el margen entre las dos clases principales; una puntuación de conflicto entre etiquetas competidoras; y descriptores basados en votos y confianza, como proporciones de voto, desacuerdo y estadísticos resumen de confianza.

Estas señales se organizan en un vector de características

$$x = (x_1, x_2, \ldots, x_m),$$

donde cada componente captura un aspecto del comportamiento colectivo del conjunto. Así, la representación no solo codifica qué etiqueta parece dominante, sino también cuán fuerte, estable o fragmentada es esa dominancia entre agentes.

Finalmente, *XGBoost* actúa como modelo de decisión final. En lugar de aprender a partir del texto crudo, aprende a partir del vector derivado de las señales de consenso. Su función es modelar qué patrones de acuerdo, desacuerdo, confianza y conflicto se asocian con resultados de verificación correctos o incorrectos. Así, la decisión final no depende solo de una regla fija de fusión, sino de una representación aprendida del comportamiento colectivo observado entre los agentes.

4. Experimentos

Para evaluar la arquitectura propuesta en un escenario realista de verificación de afirmaciones, utilizamos el benchmark Climate-FEVER [2]. Los experimentos se realizan sobre la partición de entrenamiento procesada, compuesta por 1.535 instancias afirmación-evidencia. La fase de generación agéntica se ejecuta sobre todas las instancias sin utilizar las etiquetas reales; posteriormente, las salidas de los agentes se transforman en características de meta-nivel y el protocolo de entrenamiento y *holdout* se aplica únicamente en la etapa supervisada de meta-aprendizaje.

Se construye una única partición estratificada 80/20, con 1.228 instancias de entrenamiento y 307 de *holdout*, reutilizada en todas las ejecuciones para evitar solapamientos entre entrenamiento y evaluación. El conjunto presenta un desbalance moderado: 654 instancias *SUPPORT* (42.61 %), 474 *REFUTE* (30.88 %) y 407 *NOT_ENOUGH_INFO* (26.51 %). No se utiliza información del *holdout* para el diseño de características, la selección de modelos ni el ajuste de hiperparámetros. La evaluación se reporta mediante *Accuracy*, *Macro F1* y matrices de confusión normalizadas.

La capa agéntica se instancia utilizando DeepSeek V3.2 [11]. Los experimentos preliminares con *backbones* alternativos y configuraciones mixtas mostraron tendencias similares, pero no mejoras consistentes frente al uso de un *backbone* compartido. Por ello, la evaluación principal utiliza una configuración controlada en la que todos los agentes comparten modelo base y la diversidad se induce mediante especialización por roles y *prompting*. Las cinco ejecuciones reportadas corresponden a generaciones agénticas independientes bajo la misma configuración, evaluadas siempre sobre el mismo *holdout*. Las etiquetas reales se usan únicamente como objetivos supervisados del meta-agregador, entrenado en la partición de entrenamiento mediante validación cruzada estratificada de cinco pliegues.

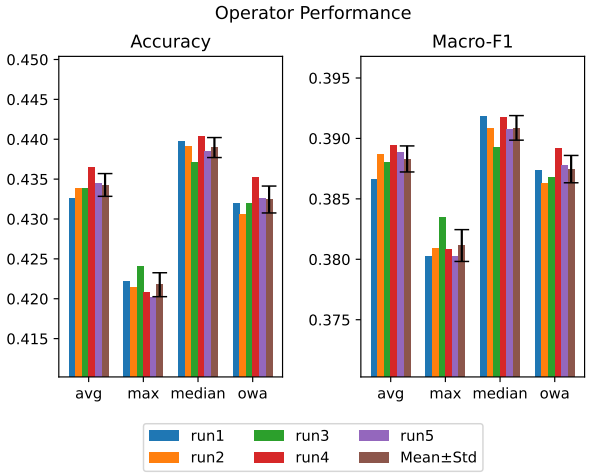


Figura 2. Efecto del operador de agregación sobre el rendimiento del conjunto.

4.1. Agregación Directa y Comportamiento de los Agentes

En primer lugar, examinamos el comportamiento de la capa agéntica antes de introducir el meta-agregador supervisado. En particular, analizamos tanto el efecto de la agregación directa como el comportamiento por clase de los cinco agentes sobre el conjunto global de *holdout*. Los operadores de agregación evaluados incluyen media, máximo, mediana y *Ordered Weighted Averaging* (OWA).

La Figura 2 muestra la *Accuracy* y el *Macro F1* a lo largo de las cinco ejecuciones. Los cuatro operadores presentan un rendimiento global similar, lo que indica que modificar la regla de fusión produce una variación limitada en el resultado final. Entre ellos, la mediana muestra el comportamiento más estable entre ejecuciones y se sitúa ligeramente por encima en las dos métricas reportadas, mientras que la media, el máximo y OWA siguen un patrón muy parecido.

La Figura 3 ofrece una visión más detallada de los agentes individuales mediante sus matrices de confusión normalizadas sobre el conjunto global de *holdout*. Se observa un patrón común en los cinco agentes. La diagonal correspondiente a *SUPPORT* suele ser más pronunciada, lo que sugiere que esta clase se maneja de forma más consistente. En cambio, la principal fuente de error se concentra en la separación entre *REFUTE* y *NOT_ENOUGH_INFO*, cuyas filas muestran una distribución más dispersa fuera de la diagonal.

4.2. Resultados de Meta-Agregación

A continuación, evaluamos si las señales colectivas producidas por la capa agéntica pueden explotarse mejor mediante meta-agregación supervisada que mediante fusión difusa directa. La Tabla 1 compara cuatro sistemas: agregación difusa directa de los cinco agentes (*Fuzzy*), el meta-agregador propuesto (*Fuzzy +*

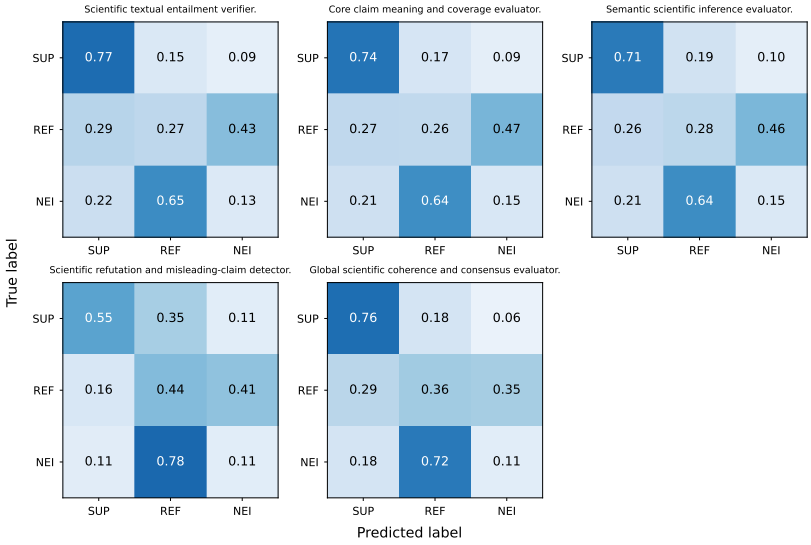


Figura 3. Matrices de confusión normalizadas de los cinco agentes especializados por rol sobre el conjunto global de *holdout*. El eje x corresponde a la etiqueta predicha, mientras que el eje y corresponde a la etiqueta real.

XGBoost), un verificador basado en un único LLM (*Single LLM*) y una línea base de *Embeddings* basada en representaciones Sentence-BERT del par afirmación-evidencia [13]. Esta última utiliza también XGBoost, lo que permite comparar la representación propuesta frente a una representación semántica estándar.

Tabla 1. Comparación de rendimiento sobre el conjunto global de *holdout*. Las métricas se presentan como media y desviación estándar sobre el número de ejecuciones indicado.

Método	Accuracy	Macro-F1
Fuzzy + XGBoost	$0{,}5557 \pm 0{,}0157$	$0{,}5328 \pm 0{,}0136$
Embeddings	$0{,}4397 \pm 0{,}0003$	$0{,}3950 \pm 0{,}0046$
Single LLM	$0{,}3941 \pm 0{,}0040$	$0{,}3587 \pm 0{,}0047$
Fuzzy	$0{,}3902 \pm 0{,}0027$	$0{,}3564 \pm 0{,}0024$

Fuzzy + XGBoost obtiene los mejores valores de *accuracy* y *macro-F1* entre los sistemas evaluados. La diferencia respecto a *Fuzzy* y *Single LLM* sugiere que la mejora no procede del número de agentes por sí mismo, sino de explotar explícitamente la estructura de sus acuerdos, desacuerdos y niveles de confianza. Además, al superar a *Embeddings* usando el mismo clasificador XGBoost, el resultado sugiere que las señales de consenso, desacuerdo, dominancia y confianza proporcionan una representación más informativa que una línea base semántica estándar. La proximidad entre *Fuzzy* y *Single LLM* confirma que la agregación

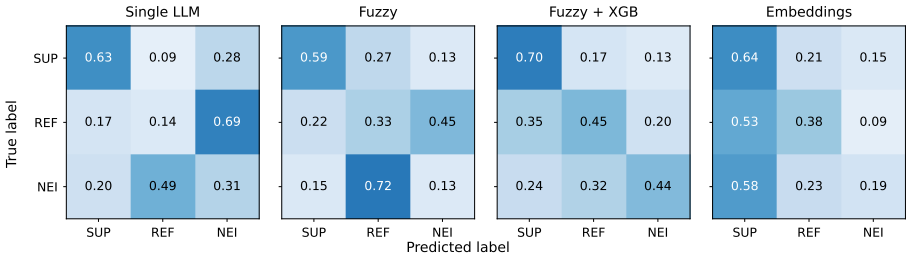


Figura 4. Matrices de confusión normalizadas de los cuatro sistemas evaluados. El eje x corresponde a la etiqueta predicha y el eje y a la etiqueta real.

directa de agentes especializados no basta para mejorar el resultado final. En este caso, la especialización por roles produce salidas adicionales, pero no suficiente diversidad epistémica para que una regla fija de fusión las explote de forma efectiva.

La Figura 4 muestra que *Fuzzy* y *Single LLM* mantienen una separación débil entre clases, especialmente entre *REFUTE* y *NOT_ENOUGH_INFO*. Por el contrario, *Fuzzy + XGBoost* presenta un comportamiento más equilibrado, lo que confirma que el meta-agregador aprovecha mejor las señales colectivas generadas por el conjunto.

Con el objetivo de identificar qué señales contribuyen en mayor medida a la decisión del meta-agregador, se realizó un análisis global de importancia de características agregado sobre las distintas ejecuciones. La Figura 5 muestra la importancia media de cada variable junto con su desviación estándar. Las características más influyentes son las proporciones de votos asignadas a las clases SUPPORT, REFUTE y NEI, especialmente la proporción de votos SUPPORT. A continuación aparecen los grados de pertenencia difusos asociados a estas mismas clases, lo que indica que el modelo combina tanto la distribución discreta de votos como el apoyo agregado obtenido mediante la fusión OWA. El desacuerdo entre agentes también presenta una contribución apreciable, aunque claramente inferior a la de las proporciones de voto y las pertenencias difusas. Por el contrario, las variables relacionadas con la longitud de la afirmación, la dispersión de la confianza y el margen entre las clases principales muestran una importancia menor. En conjunto, estos resultados sugieren que la decisión del meta-agregador depende principalmente de cómo se distribuye el apoyo entre las tres clases, mientras que las señales de conflicto, confianza e incertidumbre actúan como información complementaria.

5. Discusión

Los resultados muestran que la agregación directa presenta una capacidad limitada en el escenario evaluado. Aunque los agentes están especializados por rol, sus predicciones permanecen correlacionadas en numerosas instancias y los

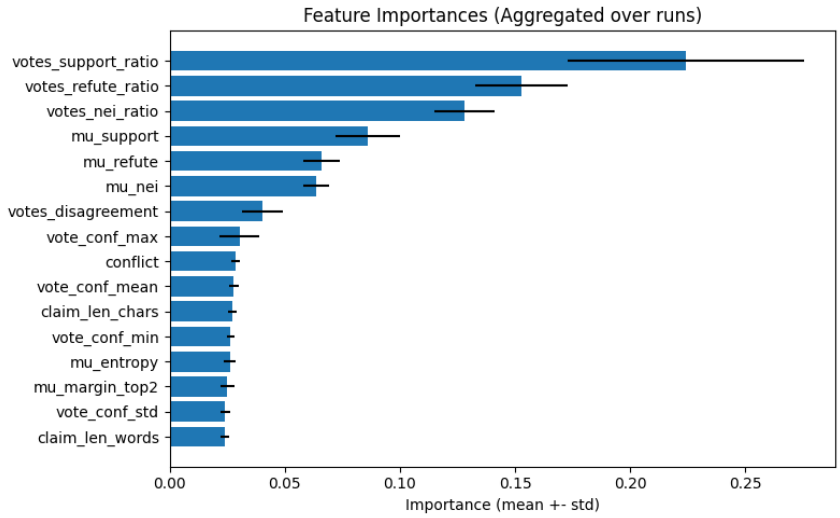


Figura 5. Importancia global de las características del meta-agregador, expresada como media y desviación estándar sobre las distintas ejecuciones.

distintos operadores producen diferencias reducidas. Por tanto, la especialización mediante *prompting* no garantiza por sí sola la diversidad necesaria para que las reglas fijas de fusión mejoren sustancialmente el rendimiento.

Este resultado matiza la interpretación del consenso en sistemas agénticos. El acuerdo entre agentes no implica necesariamente corrección, ya que puede reflejar sesgos compartidos o errores correlacionados. Desde la perspectiva de GDM, la fiabilidad de la decisión colectiva depende no solo del nivel de acuerdo, sino también de su estructura interna, incluyendo desacuerdo, conflicto, dominancia y confianza [15,8].

La aportación principal del trabajo consiste en representar explícitamente ese comportamiento colectivo. La capa de meta-agregación transforma las salidas de los agentes en señales de voto, confianza, incertidumbre, dominancia y conflicto. La comparación experimental respalda esta estrategia: la agregación difusa directa obtiene un rendimiento similar al verificador basado en un único LLM, mientras que *Fuzzy + XGBoost* alcanza los mejores resultados en *Accuracy* y *Macro-F1*. Este comportamiento es coherente con trabajos que muestran que las estrategias de agregación estructurada pueden superar a la votación simple o a las combinaciones agénticas planas [1,18].

No obstante, la propuesta requiere datos etiquetados para entrenar el meta-agregador y su coste está dominado por la ejecución de varios agentes, aunque estos pueden operar en paralelo. Además, la evaluación se limita a Climate-FEVER y a una configuración con un *backbone* compartido, DeepSeek V3.2. Estas decisiones permiten aislar el efecto de la representación colectiva, pero limitan la generalización de las conclusiones. Como trabajo futuro, será necesario

evaluar conjuntos de datos adicionales, modelos base heterogéneos y estrategias adaptativas que activen varios agentes únicamente en casos inciertos o conflictivos.

6. Conclusiones

Este trabajo ha propuesto una arquitectura de meta-agregación en dos etapas para la verificación agéntica de afirmaciones. Frente a enfoques basados únicamente en votación o fusión directa, el método representa explícitamente señales de acuerdo, desacuerdo, conflicto, dominancia y confianza, y las utiliza para estimar la decisión final.

Los resultados sobre Climate-FEVER muestran que la agregación directa de agentes especializados aporta mejoras limitadas cuando sus salidas permanecen correlacionadas. En cambio, la capa de meta-agregación obtiene un rendimiento superior al explotar la estructura interna de la decisión colectiva. La principal aportación consiste, por tanto, en trasladar el aprendizaje supervisado desde el texto original hacia una representación del comportamiento conjunto de los agentes, conectando la verificación agéntica con los principios de GDM [1].

Como trabajo futuro, resulta prioritario evaluar la propuesta en otros conjuntos de datos para determinar su capacidad de generalización y analizar en qué escenarios las señales colectivas mantienen su utilidad. Asimismo, será necesario estudiar la sensibilidad del sistema frente a diferentes configuraciones de los pesos OWA, con el fin de cuantificar hasta qué punto la orientación hacia la dominancia afecta al rendimiento final. También se plantea incorporar modelos base heterogéneos y comparar distintos meta-modelos para separar con mayor precisión el efecto de la representación colectiva del clasificador empleado.

Acknowledgments. Esta publicación es parte del proyecto de I+D+i PID2022-139297OB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER/UE

Referencias

1. Ai, R., Pan, Y., Simchi-Levi, D., Tambe, M., Xu, H.: Beyond majority voting: Llm aggregation by leveraging higher-order information. arXiv preprint arXiv:2510.01499 (2025)
2. Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., Leippold, M.: Climate-fever: A dataset for verification of real-world climate claims. arXiv preprint arXiv:2012.00614 (2020)
3. Ding, R.X., Palomares, I., Wang, X., Yang, G.R., Liu, B., Dong, Y., Herrera-Viedma, E., Herrera, F.: Large-scale decision-making: Characterization, taxonomy, challenges and future directions from an artificial intelligence and applications perspective. Information fusion **59**, 84–102 (2020)
4. Ding, R.X., Yang, B., Yang, G.R., Li, M.N., Wang, X., Chiclana, F.: A confidence and conflict-based consensus reaching process for large-scale group decision-making problems with intuitionistic fuzzy representations. IEEE Transactions on Fuzzy Systems **32**(6), 3420–3432 (2024)

5. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. In: Forty-first international conference on machine learning (2024)
6. Fedrizzi, M., Kacprzyk, J., Nurmi, H.: Consensus degrees under fuzzy majorities and fuzzy preferences using owa (ordered weighted average) operators. Control and Cybernetics **22**(4), 71–80 (1993)
7. García-Zamora, D., Labella, Á., Ding, W., Rodríguez, R.M., Martínez, L.: Large-scale group decision making: A systematic review and a critical analysis. IEEE/CAA Journal of Automatica Sinica **9**(6), 949–966 (2022)
8. Herrera-Viedma, E., Cabrerizo, F.J., Kacprzyk, J., Pedrycz, W.: A review of soft consensus models in a fuzzy environment. Information Fusion **17**, 4–13 (2014)
9. Herrera-Viedma, E., Herrera, F., Chiclana, F.: A consensus model for multiperson decision making with different preference structures. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **32**(3), 394–402 (2002)
10. Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B.: Camel: Communicative agents for"mind.^{ex}ploration of large language model society. Advances in neural information processing systems **36**, 51991–52008 (2023)
11. Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al.: Deepseek-v3. 2: Pushing the frontier of open large language models. arXiv preprint arXiv:2512.02556 (2025)
12. Park, C., Han, S., Guo, X., Ozdaglar, A.E., Zhang, K., Kim, J.K.: Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 30215–30248 (2025)
13. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
14. Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T.D., Pretorius, A.: Should we be going mad? a look at multi-agent debate strategies for llms (2024). URL https://arxiv. org/abs/2311.17371
15. Trillo, J.R., Herrera-Viedma, E., Morente-Molinera, J.A., Cabrerizo, F.J.: A group decision-making method based on the experts' behavior during the debate. IEEE Transactions on Systems, Man, and Cybernetics: Systems **53**(9), 5796–5808 (2023)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
17. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. Frontiers of Computer Science **18**(6), 186345 (2024)
18. Wang, Z., Moriyama, S., Wang, W.Y., Gangopadhyay, B., Takamatsu, S.: Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems. arXiv preprint arXiv:2502.11098 (2025)
19. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al.: Autogen: Enabling next-gen llm applications via multi-agent conversations. In: First conference on language modeling (2024)
20. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)