ANÁLISIS MULTIVARIANTE AUTOMATIZADO: ESTUDIO COMPARATIVO DE MODELOS DE LENGUAJE DE GRAN ESCALA (LLMS) DE ÚLTIMA GENERACIÓN

José Raúl Belmonte-Sánchez*, Roberto Romero-González, Antonia Garrido Frenich

Grupo de Química Analítica de Contaminantes (QAC), Departamento de Química y Física, Centro de Investigación en Biotecnología Agroalimentaria (CIAIMBITAL), Campus de Excelencia Internacional Agroalimentario ceiA3, Universidad de Almería, E-04120 Almería, España

*jbs324@ual.es

Abstract

Los modelos de lenguaje de gran escala (LLMs, del inglés *large language models*) permiten democratizar el análisis quimiométrico al permitir ejecutar flujos de trabajo multivariantes completos mediante simples instrucciones en lenguaje natural. En este trabajo se evaluaron de forma sistemática cinco LLMs de última generación (Claude Opus 4.6, Gemini 3.1 Pro, Grok-4, ChatGPT 5.6 y Qwen 3.5) frente al *software* quimiométrico de referencia SIMCA (Sartorius Stedim, Umeå, Suecia). Como caso de estudio se empleó un conjunto real de 66 aceites vegetales analizados mediante RMN-^{1}H de campo bajo. Los resultados revelaron una jerarquía de rendimiento clara: los cuatro modelos líderes alcanzaron cerca del 100 % de acierto en la clasificación, mientras que el único modelo de código abierto evaluado falló de forma irrecuperable en la etapa supervisada. La fluidez lingüística no garantiza competencia analítica, por lo que la supervisión humana experta sigue siendo indispensable.

1. Introducción

Los LLMs han irrumpido como herramientas transformadoras en la investigación científica. Su capacidad de razonamiento y de procesar lenguaje natural permite abordar, mediante instrucciones conversacionales, tareas que hasta hace poco exigían *software* especializado o programación avanzada. En química, esta capacidad ya se ha aplicado al diseño de moléculas, la planificación de síntesis o la interpretación de datos espectroscópicos [2,3].

En química analítica, la interpretación de datos espectroscópicos y cromatográficos se apoya habitualmente en el **análisis multivariante**: análisis de componentes principales (*principal component analysis*, PCA), análisis discriminante por mínimos cuadrados parciales (*partial least squares-discriminant analysis*, PLS-DA), máquinas de vectores soporte (*support vector machines*, SVM) o bosques aleatorios (*random forest*, RF). Estas herramientas son potentes, pero su aplicación rigurosa exige dominio de *software* especializado o de entornos de programación como R, MATLAB o Python. Esta barrera técnica obliga al químico analítico a ejercer también de científico de datos o a depender de especialistas en quimiometría, lo que limita el aprovechamiento de conjuntos de datos cada vez más complejos.

Los LLMs plantean por primera vez la posibilidad de **democratizar el análisis quimiométrico**, permitiendo que el analista dialogue con sus datos sin escribir una sola línea de código. Sin embargo, estos modelos no han sido entrenados específicamente para el razonamiento numérico, y su notable fluidez lingüística puede enmascarar errores metodológicos de consecuencias graves. Resulta, por tanto, imprescindible evaluar su desempeño real sobre problemas analíticos auténticos, en lugar de sobre ejercicios sintéticos.

Con ese objetivo se diseñó el trabajo que aquí se resume, cuya meta fue **evaluar la capacidad de cinco LLMs de última generación para ejecutar un flujo de trabajo quimiométrico completo sobre datos reales de RMN-^{1}H**, contrastándolos con el *software* de referencia SIMCA.

2. Muestras, datos y diseño experimental

Como banco de pruebas se empleó un problema de autenticación de aceites vegetales, línea desarrollada previamente por nuestro grupo [4,5]. Se analizaron **66 muestras** (22 de cada tipo) mediante **RMN-^{1}H de campo bajo** (resonancia magnética nuclear de protón, *low-field ^{1}H NMR*) en un espectrómetro de sobremesa de 100 MHz (Nanalysis Scientific Corp., Calgary, Canadá), pertenecientes a tres categorías: aceite de oliva virgen extra (AOVE; *extra virgin olive oil*, EVOO), aceite de girasol (AG; *sunflower oil*, SO) y aceite de girasol alto oleico (AGAO; *high-oleic sunflower oil*, HOSO). De ellas, **57 se suministraron etiquetadas** con su clase y **9 se reservaron como incógnita**, excluidas del entrenamiento, para evaluar la capacidad real de predicción de cada modelo (Fig. 1).

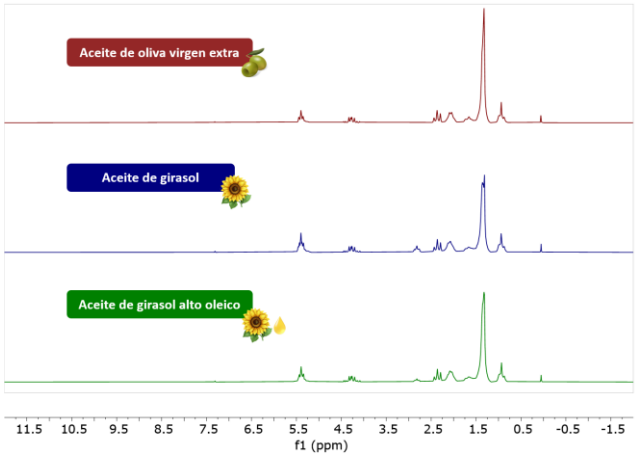


Fig. 1. Espectros RMN-^{1}H representativos de las tres categorías de aceite estudiadas: AOVE, AG y AGAO. Adaptada de [1].

La discriminación entre las tres clases descansa en su distinto perfil de ácidos grasos, que se refleja en las regiones diagnósticas del espectro: las señales olefínicas y bis-alílicas informan del grado de insaturación, mientras que el AGAO ocupa una posición composicional intermedia que lo convierte en el caso más exigente del conjunto.

Cada modelo recibió **exactamente la misma secuencia** de instrucciones **estandarizada** en lenguaje natural, cubriendo el flujo quimiométrico completo: carga y auditoría de los datos, pretratamiento espectral, análisis exploratorio mediante PCA, agrupamiento no supervisado por *K-means*, clasificación supervisada (PLS-DA, SVM y RF), validación de modelos, interpretabilidad química y autoevaluación crítica de su propio trabajo. En paralelo, el mismo conjunto de datos se analizó siguiendo el flujo quimiométrico convencional con SIMCA, que actuó como referencia (Fig. 2).

Para garantizar la neutralidad de la comparación, todos los modelos se utilizaron desde cuentas de nueva creación con la configuración por defecto, sin historial ni memoria adaptativa.

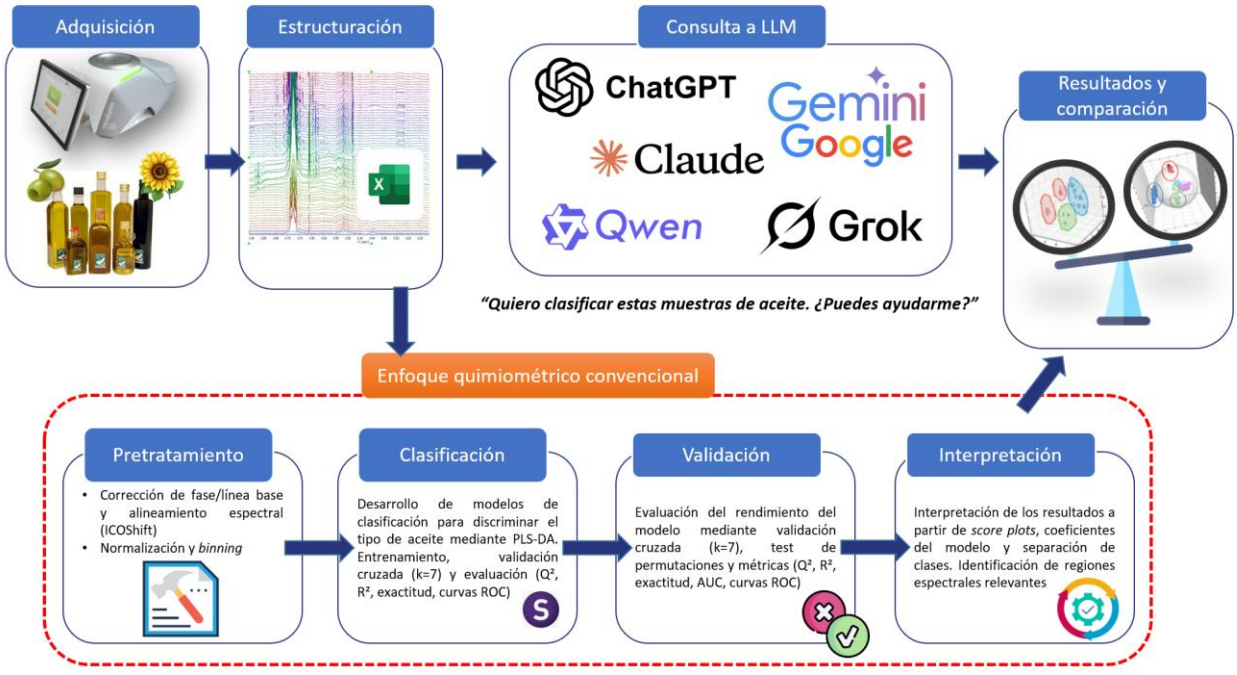


Fig. 2. Flujo de trabajo conversacional con los cinco LLMs evaluados frente al enfoque quimiométrico convencional (SIMCA). Ambas rutas permiten la interpretación de los modelos y la comparación de resultados. Adaptada de [1].

Los cinco modelos se seleccionaron según su **Arena Score**, el *ranking* público de LLMs basado en las preferencias de los usuarios (Chatbot Arena) [3], en el momento de realizar el estudio: Claude Opus 4.6 (Anthropic, 1504), Gemini 3.1 Pro (Google DeepMind, 1493), Grok-4 (xAI, 1491), ChatGPT 5.6 (OpenAI, 1484) y Qwen 3.5 (Alibaba Cloud, 1465), este último el único de código abierto. Con ello se cubrieron arquitecturas y filosofías de desarrollo distintas, mitigando el sesgo de plataforma.

3. Resultados: una jerarquía de rendimiento clara

La evaluación conjunta de los nueve criterios analíticos considerados (Tabla 1) reveló una jerarquía nítida. Los **cuatro modelos líderes alcanzaron tasas de acierto cercanas al 100 %** en la clasificación de los tres tipos de aceite y predijeron correctamente las nueve muestras incógnita, igualando funcionalmente al *software* especializado. El análisis exploratorio por PCA (Fig. 3) muestra cómo cada modelo recuperó de forma autónoma la estructura de clases presente en los datos.

Tabla 1. Matriz de evaluación del rendimiento analítico de los cinco LLMs frente al *software* quimiométrico de referencia SIMCA. Notación: **+++** muy alto; **++** alto; **+** aceptable; **–** nulo o no aplicable. Adaptada de [1].

Criterio de evaluación	Referencia (SIMCA) Umetrics	Claude Opus 4.6 Arena 1504	Gemini 3.1 Pro Arena 1493	Grok-4 Arena 1491	ChatGPT 5.6 Arena 1484	Qwen 3.5 Arena 1465
Manejo de datos	++	+++	+	+	+++	++
Pretratamiento espectral	+++	+++	++	+++	++	+
Análisis exploratorio	++	+++	++	++	+++	+
Análisis no supervisado	+++	+++	+++	++	+++	+
Análisis supervisado	+++	+++	+++	++	+++	–
Validación de modelos	++	+++	++	++	+++	–
Interpretabilidad	+	+++	++	++	++	–
Autoevaluación crítica	–	+++	+++	++	+++	–
Eficiencia	–	+++	+	+	+++	++
PUNTUACIÓN GLOBAL	Adecuado	Sobresaliente	Muy bueno	Bueno	Muy bueno	No fiable

Claude Opus 4.6 destacó como el modelo más autónomo y riguroso. Fue el único capaz de cargar de forma nativa el fichero original sin conversión previa, aplicó un pretratamiento de seis pasos justificado uno a uno con razonamiento específico de RMN y sometió sus modelos a la validación más exigente del estudio, incluyendo un test de permutaciones ($p < 0{,}002$). Predijo correctamente las nueve muestras incógnita y fue también el más crítico al evaluar sus propias limitaciones.

Gemini 3.1 Pro y **ChatGPT 5.6** ofrecieron un rendimiento correcto y muy sólido, con 100 % de acierto en la clasificación, predicción correcta de todas las incógnitas e interpretación químicamente coherente de las regiones espectrales discriminantes, aunque operaron en un modo asistido por código y con una validación menos exhaustiva.

Grok-4 obtuvo resultados próximos al 100 %, con un pretratamiento técnicamente excelente, pero mostró menor profundidad en el análisis exploratorio y una autonomía limitada en el manejo inicial de los datos.

Qwen 3.5, el único modelo de código abierto evaluado, sufrió un **error interno irrecuperable** en la etapa de clasificación supervisada que persistió tras más de veinte minutos de intentos en sesiones independientes. Además, en las fases previas no llegó a reconocer la estructura real de tres clases del conjunto de datos.

Un hallazgo especialmente relevante es que el **Arena Score no predijo el rendimiento quimiométrico**. Grok-4, con mejor puntuación general que ChatGPT 5.6, fue superado por este en varios criterios analíticos. Los *benchmarks* generalistas basados en las preferencias de los usuarios no son, por tanto, un buen indicador de la competencia de un modelo en tareas científicas especializadas.

4. Implicaciones para el laboratorio analítico

La conclusión práctica del estudio es doble. Por un lado, los LLMs líderes **igualan funcionalmente al *software* quimiométrico especializado** en un problema real de autenticación de alimentos, y lo hacen en una fracción del tiempo y sin requerir programación. Esto abre la puerta a que grupos sin especialistas en quimiometría accedan a herramientas multivariantes avanzadas.

Por otro lado, y de forma igualmente importante, la **fluidez lingüística no garantiza competencia analítica**. Los modelos presentan sus resultados con una seguridad narrativa uniforme, independientemente de si el procedimiento subyacente ha sido riguroso o defectuoso. El caso de Qwen 3.5 ilustra además que la accesibilidad del código abierto no equivale todavía a fiabilidad analítica.

Los LLMs deben entenderse, por tanto, como **copilotos analíticos competentes, no como sustitutos del criterio experto**. La supervisión humana, la validación sistemática de las salidas y el escrutinio crítico de los resultados autoinformados siguen siendo salvaguardas imprescindibles.

5. Conclusiones y perspectivas

Este trabajo constituye, hasta donde alcanza nuestro conocimiento, la primera evaluación comparativa y sistemática de LLMs de última generación ejecutando un flujo quimiométrico completo sobre datos espectroscópicos reales. La jerarquía de rendimiento observada fue clara y estructurada, con Claude Opus 4.6 como el modelo más completo y autónomo en la fecha que se realizó el estudio.

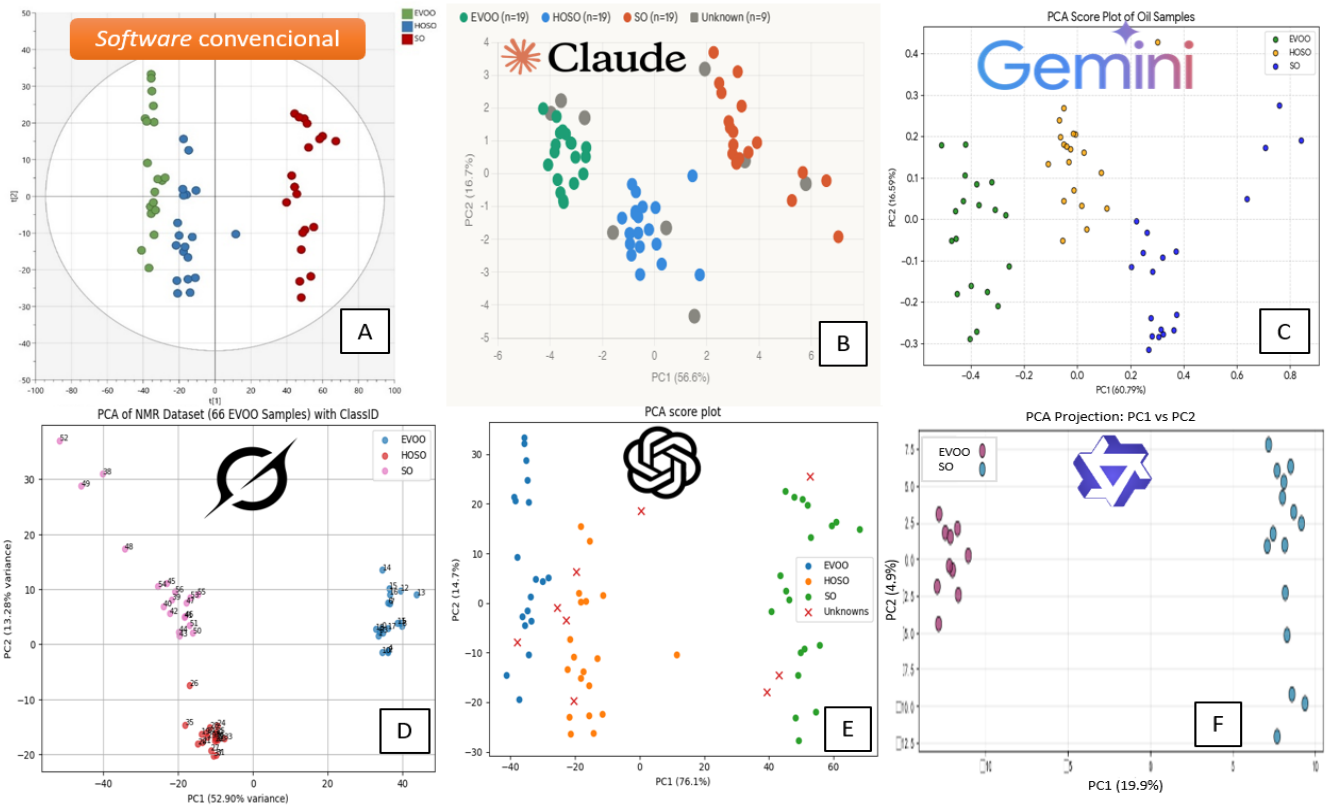


Fig. 3. Análisis exploratorio (PCA) realizado por: **A)** el *software* quimiométrico convencional de referencia (SIMCA); y los LLMs evaluados, **B)** Claude, **C)** Gemini, **D)** Grok, **E)** ChatGPT y **F)** Qwen. AOVE (*extra virgin olive oil*, EVOO), AG (*sunflower oil*, SO) y AGAO (*high-oleic sunflower oil*, HOSO). Adaptada de [1].

Dada la velocidad a la que evolucionan estos sistemas, los resultados deben interpretarse como una fotografía de un momento concreto. Las líneas de continuación de nuestro grupo pasan por extender la evaluación a otras técnicas instrumentales y matrices, por incorporar conjuntos de datos multilaboratorio que permitan una validación externa, y por definir **protocolos de buenas prácticas** para el uso responsable de LLMs en química analítica.

Agradecimientos

Los autores agradecen a la Sociedad Española de Química Analítica (SEQA) el premio otorgado a la mejor comunicación tipo Póster presentada por José Raúl Belmonte-Sánchez en el 1.er Congreso de Análisis Instrumental (CAI 2026), así como la invitación a participar en este número de *Actualidad Analítica*. Este trabajo ha sido financiado por el proyecto CPP2021-008672, MICIU/AEI/10.13039/501100011033 y la «Unión Europea NextGenerationEU/PRTR».

Referencias

[1] J.R. Belmonte-Sánchez, R. Romero-González, A. Garrido Frenich, LLMs in the laboratory: benchmarking conversational AI for multivariate analysis of NMR spectroscopic data, manuscrito enviado a *Analytical Chemistry* para su publicación, 2026.

[2] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A.D. White, P. Schwaller, Augmenting large language models with chemistry tools, Nat. Mach. Intell. 6 (2024) 525-535. https://doi.org/10.1038/s42256-024-00832-8.

[3] W.L. Chiang, L. Zheng, Y. Sheng, A.N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M.I. Jordan, J.E. Gonzalez, I. Stoica, Chatbot Arena: an open platform for evaluating LLMs by human preference, Proc. Mach. Learn. Res. 235 (2024) 8359-8388.

[4] J.R. Belmonte-Sánchez, R. Romero-González, M.Á. Martínez Orosa, M. Calvo Morata, A. Garrido Frenich, Benchtop NMR spectroscopy for quantitative determination of milk fat and qualitative determination of lactose: from calibration curve to deep learning, LWT 212 (2024) 117000. https://doi.org/10.1016/j.lwt.2024.117000.

[5] J.R. Belmonte-Sánchez, R. Romero-González, J.A. Tello-Jiménez, A. Garrido Frenich, Expanding horizons: toward the classification of extra virgin olive oil using ^{1}H NMR and machine learning, Food Control 179 (2026) 111610. https://doi.org/10.1016/j.foodcont.2025.111610.

[6] L. Duponchel, R. Rocha de Oliveira, V. Motto-Ros, Large language models (such as ChatGPT) as tools for machine learning-based data insights in analytical chemistry, Anal. Chem. 97 (2025) 6956-6961. https://doi.org/10.1021/acs.analchem.4c05046.

Los autores conservan los derechos de autor de este trabajo. Este artículo se publica en acceso abierto bajo la licencia