

Recomendación de prácticas pedagógicas mediante modelos de lenguaje

Gabriel Astudillo, Jorge Baier, Isabel Hilliger

Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Santiago, Chile.

gastudillo@uc.cl, jbaier@uc.cl e ihillige@uc.cl

Introducción

En educación superior, los docentes suelen ser reclutados como expertos en campos de conocimiento específico, sin necesariamente tener formación específica para el ejercicio de la docencia. Como resultado, muchos basan sus métodos de enseñanza en la tradición o en creencias personales (Benassi y Buskist, 2012; Groccia y Buskist, 2011). Por ello, las instituciones invierten tanto en programas de desarrollo docente, como en instrumentos para levantar información sobre el ejercicio de la docencia y entregar retroalimentación (Borda et al., 2020; Luthra et al., 2023). A pesar de que existen otros instrumentos, el más utilizado son las encuestas de evaluación aplicadas al final de cada curso.

Estas encuestas consisten en una combinación de preguntas cuantitativas sobre diferentes aspectos de la docencia, en conjunto con preguntas de texto abierto no estructurado (Parker et al., 2024), donde los estudiantes pueden describir en sus propias palabras las experiencias de aprendizaje en cada curso (McDonald et al., 2020). Trabajos previos han mostrado que, a pesar de la riqueza de información que es posible extraer de estos comentarios, derivar recomendaciones sobre posibles modificaciones en las estrategias de enseñanza-aprendizaje, que sean relevantes tanto desde una perspectiva educativa como en el contexto específico de un curso, está lejos de ser trivial (Hujala et al., 2020; Parker et al., 2024).

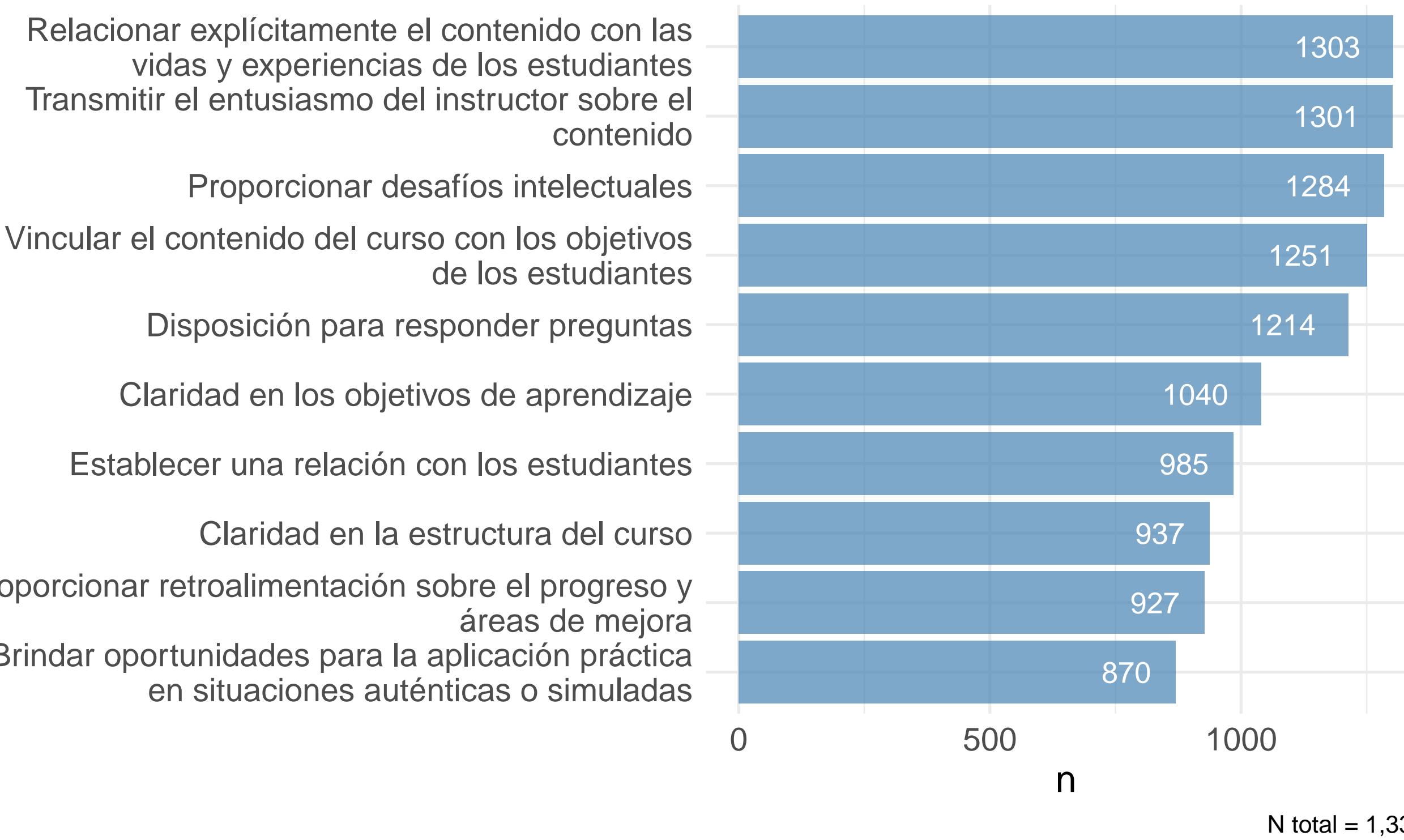


Figura 1: 10 prácticas más frecuentes detectadas en cursos

Tradicionalmente, esto debe hacerlo un humano experto en docencia que analice las encuestas, lo cual implica una gran inversión de tiempo y esfuerzo, difícilmente escalable para programas o instituciones masivas (Dehbozorgi et al., 2024; Hujala et al., 2020; Wang y Demszky, 2023). Sin embargo, con la revolución de los grandes modelos de lenguaje, trabajos recientes han comenzado a intentar utilizarlos para proponer recomendaciones sobre potenciales modificaciones en las prácticas de docencia. Ya sea con un enfoque zero-shot basándose en transcripciones de clase (Wang y Demszky, 2023), o utilizando una base de conocimiento para responder consultas de un usuario (Dehbozorgi et al., 2024). No obstante, estos enfoques no han contemplado el uso de los comentarios de estudiantes en encuestas de evaluación docente como fuente para recomendar potenciales modificaciones en las estrategias de enseñanza-aprendizaje.

Materiales y métodos

El objetivo de esta investigación es proponer una estrategia para entregar recomendaciones de modificaciones en las prácticas de enseñanza aprendizaje basado en grandes modelos de lenguaje, a partir de los comentarios de estudiantes en encuestas de evaluación docente.

Para esto, primero se utiliza una metodología de reactivos: frases que representan prácticas de docencia efectiva en un lenguaje similar al de los estudiantes (Astudillo et al., 2024). Posteriormente, mediante Sentece-BERT (Reimers y Gurevych, 2019) se obtienen representaciones vectoriales tanto de los comentarios de los estudiantes como de los reactivos, y finalmente a través de coseno de similaridad es posible detectar qué prácticas de docencia efectiva están presentes en los comentarios y cuáles no.

Paralelamente, promediando las representaciones vectoriales de los comentarios a nivel de curso, se utiliza coseno de similaridad para medir la distancia entre pares de cursos, lo que será utilizado como vecindario de búsqueda de cursos con estilos similares, sobre los que se realiza filtrado colaborativo (Schafer et al., 2007), identificando qué prácticas de docencia efectiva no implementadas en un curso particular, si son implementadas en el top de los tres cursos más similares, siendo candidatas para la recomendación. Finalmente, un modelo generativo de lenguaje hace la inferencia final de, dado el contexto de prácticas de docencia efectiva detectadas, cuáles -de las no aplicadas- serían las más pertinentes.

Se utilizó el 100K Coursera's Course Reviews Dataset (disponible en <https://www.kaggle.com/datasets/septa97/100k-courseras-course-reviews-dataset>), que contiene 140.320 reseñas de 1835 cursos en línea de la plataforma Coursera. Dado que estos comentarios están en diferentes lenguajes -como Inglés, Español, Chino, Ruso, entre otros-, se utilizó bert-base-multilingual-uncased (Devlin et al., 2018) para obtener las representaciones vectoriales, mientras que los reactivos fueron formulados en inglés. Para la inferencia final, se utilizó llama-3.1-70b (Meta, 2024).

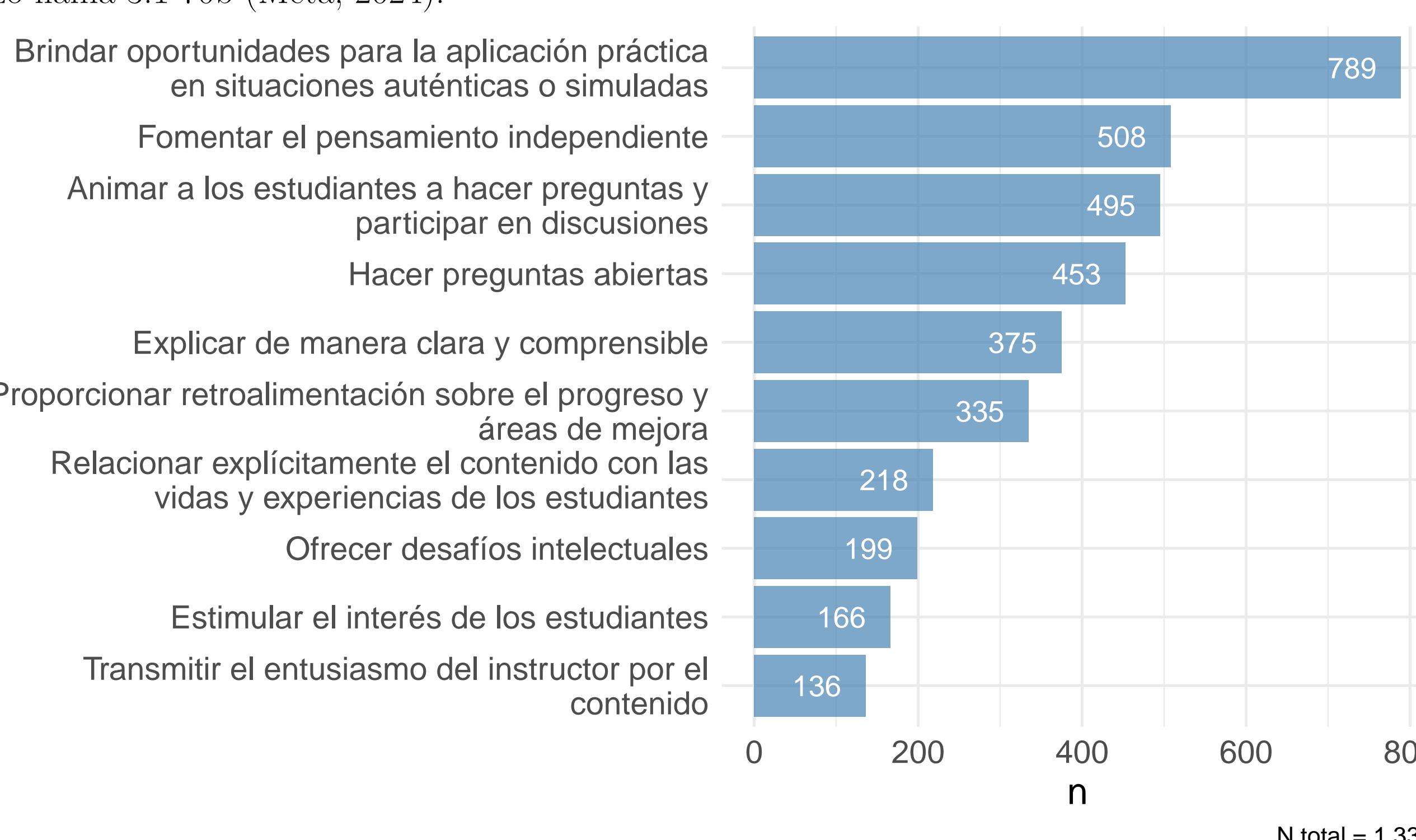


Figura 2: 10 recomendaciones más frecuentes

Finalmente, para evaluar las recomendaciones generadas por el modelo, se utilizó una metodología de LLM as a judge (Zheng et al., 2023), que consisten en el uso un segundo modelo de lenguaje -con mayor número de parámetros- para evaluar el desempeño del primero. En este caso, se el modelo evaluador fue gpt-4o-mini (OpenAI, 2024) al que se solicitó responder la rúbrica de Wang y Demszky (2023) y calificar las recomendaciones del primer modelo de lenguaje en cuatro aspectos: relevancia, fidelidad, acciónabilidad y novedad, en escala de tres categorías: bajo, medio y alto.

Resultados

De los 1835 cursos originalmente incluidos en el dataframe, 1338 fueron finalmente considerados en el análisis. En estos cursos, las prácticas de docencia efectiva más comúnmente detectadas fueron: (1) conectar explícitamente los contenidos con la vida y experiencias de los estudiantes, (2) transmitir el entusiasmo del docente sobre el contenido y (3) proveer desafíos intelectuales.

En ese contexto, las recomendaciones más comunes del modelo fueron: (1) proporcionar oportunidades para la aplicación práctica en situaciones auténticas o simuladas, (2) fomentar el pensamiento independiente y (3) animar a los estudiantes a hacer preguntas y participar en discusiones.

Como se observa en la Figura 3, la evaluación realizada por GPT-4o-mini muestra que las recomendaciones proporcionadas por Llama 3.1-70b, basadas en un sistema de filtrado colaborativo, son en general relevantes (82.79% calificadas como "Alto") y fieles en el contexto de los comentarios de los estudiantes del curso (76.76% calificadas como "Alto"). En menor medida, también son evaluadas como accionables (66.3% calificadas como "Alto"). Sin embargo, el aspecto con menor rendimiento es la novedad de las recomendaciones: en el 81.89% de los casos, esto se evaluó como "Medio", y solo en el 1.86% como "Alto".

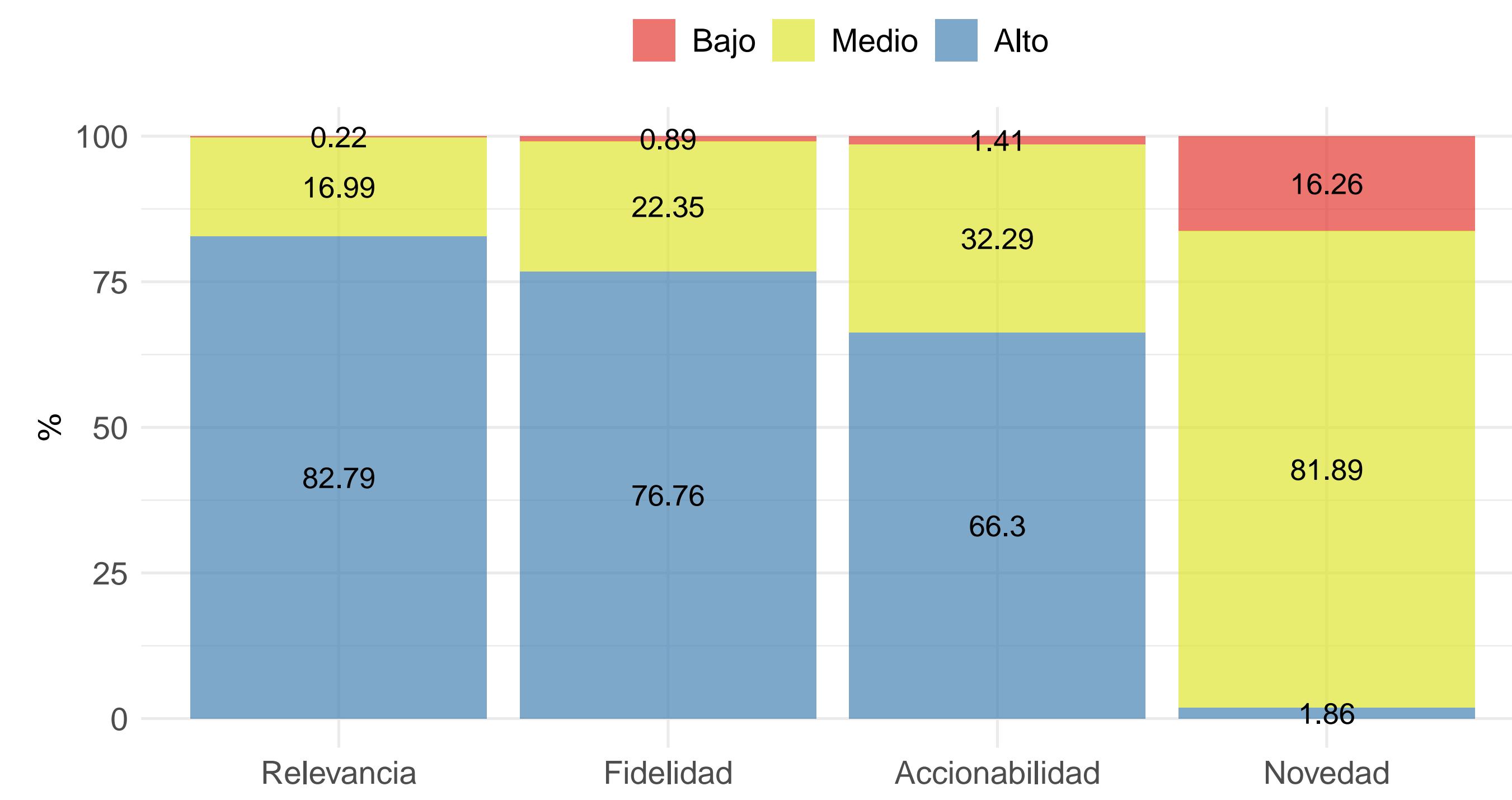


Figura 3: Evaluación por GPT-4o-mini de las recomendaciones entregadas por Llama 3.1-70b

Discusión

Con el objetivo de proporcionar retroalimentación sobre cómo modificar las prácticas de enseñanza-aprendizaje en la educación superior, este trabajo propone una aproximación en dos etapas. La primera etapa consiste en un diagnóstico de la docencia basado en la detección de prácticas de docencia efectiva, utilizando la similaridad semántica entre los vectores de prácticas y reactivos. A partir de este diagnóstico, se realiza un filtrado colaborativo para obtener una lista de prácticas candidatas para recomendar a un curso. En la segunda etapa, se utiliza Llama 3.1-70b para realizar la inferencia y seleccionar las tres prácticas más pertinentes para el curso. Posteriormente, un segundo modelo de lenguaje, GPT-4o-mini, evalúa las recomendaciones del primer modelo, calificando su relevancia, precisión, acciónabilidad y novedad.

Los resultados obtenidos en la evaluación son muy similares a los obtenidos por Wang y Demszky (2023) con evaluadores humanos. En su estudio, los criterios mejor logrados por las recomendaciones del modelo también fueron relevancia, precisión y acciónabilidad. Sin embargo, a diferencia de este estudio, donde el 66% de las recomendaciones fueron evaluadas como altamente accionables, el estudio de Wang y Demszky encontró que el 76% de las recomendaciones del modelo fueron calificadas como altamente accionables.

Por otra parte, se observa un avance significativo en el criterio de novedad de las recomendaciones. Wang y Demszky (2023) reportaron que el 82% de las recomendaciones tenían una novedad baja y el 11% una novedad media. En contraste, el enfoque propuesto en este estudio mostró que la mayoría de las evaluaciones de novedad fueron calificadas como "media". A pesar de estos resultados prometedores, las diferencias en las metodologías de evaluación entre ambos estudios impiden atribuir directamente esta mejora al enfoque propuesto. Se requiere más investigación para determinar si este avance es efectivamente atribuible al enfoque utilizado en este estudio.

Referencias

- Astudillo, G., Hilliger, I. y Baier, J. (2024). How Could Be Used Student Comments for Delivering Feedback to Instructors in Higher Education? (pp. 401–408). https://doi.org/10.1007/978-3-031-64312-5_50
- Benassi, V. A. y Buskist, W. (2012). Preparing the new professoriate to teach. In Effective college and university teaching: Strategies and tactics for the new professoriate (pp. 1–8). SAGE.
- Borda, E., Schumacher, E., Hanley, D., Geary, E., Warren, S., Ipsen, C. y Stredicke, L. (2020). Initial implementation of active learning strategies in large, lecture STEM courses: lessons learned from a multi-institutional, interdisciplinary STEM faculty development program. International Journal of STEM Education, 7(1), 4. <https://doi.org/10.1186/s40594-020-0203-2>
- Dehbozorgi, N., Kunuku, M. T. y Pouriyeh, S. (2024). Personalized Pedagogy Through a LLM-Based Recommender System (pp. 63–70). https://doi.org/10.1007/978-3-031-64312-5_8
- Devlin, J., Chang, M.-W., Lee, K. y Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Groccia, J. E. y Buskist, W. (2011). Need for evidence-based teaching. New Directions for Teaching and Learning, 128.
- Hujala, M., Kuntas, A., Hynninen, T. y Arminen, H. (2020). Improving the quality of teaching by utilising written student feedback: A streamlined process. Computers & Education, 157, 103965. <https://doi.org/10.1016/j.compedu.2020.103965>
- Luthra, A., Dixit, S. y Arya, V. (2023). Evaluating the impact of faculty development on employee engagement practices in higher education: analysing the mediating role of professional development. The Learning Organization. <https://doi.org/10.1108/TLO-01-2023-0014>
- McDonald, J., Moskal, A. C. M., Goodchild, A., Stein, S. y Terry, S. (2020). Advancing text-analysis to tap into the student voice: a proof-of-concept study. Assessment & Evaluation in Higher Education, 45(1), 154–164. <https://doi.org/10.1080/02602938.2019.1614524>
- Meta, A. (2024). The LLaMA 3 Herd of Models. arXiv preprint.
- OpenAI. (2024). GPT-4o-mini [Large Language Model].
- Parker, M. J., Anderson, C., Stone, C. y Oh, Y. (2024). A Large Language Model Approach to Educational Survey Feedback Analysis. International Journal of Artificial Intelligence in Education. <https://doi.org/10.1007/s40593-024-00414-0>
- Reimers, N. y Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
- Schafer, J. Ben, Frankowski, D., Herlocker, J. y Sen, S. (2007). Collaborative Filtering Recommender Systems. In The Adaptive Web (pp. 291–324). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_9
- Wang, R. E. y Demszky, D. (2023). Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E. y Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.