Evaluación de LLMs Open-Source para Sistemas de Recomendación Conversacionales de Películas

Benjamín Fabián Díaz Muñoz [1] **Daniel Esteban Eduardo Alegría Toledo** [1]
Benjamín Andrés Manuel Faúndez Romero [1]

Abstract

Los modelos de lenguaje de gran tamaño (LLMs) han revolucionado las interacciones conversacionales, generando interés en su aplicación para sistemas de recomendación. Este estudio evalúa comparativamente el desempeño de LLMs open-source (Phi-4 y Mistral 7B) como sistemas de recomendación conversacionales de películas frente a métodos tradicionales. Utilizamos el dataset ReDIAL con 11,438 diálogos reales usuario-asistente y evaluamos tres técnicas de prompting: zero-shot, few-shot y chain-of-thought. Nuestros resultados demuestran que, aunque Mistral CoT alcanza la mayor precisión (0.052), los métodos tradicionales como Most Popular superan significativamente a los LLMs en recall (0.068 vs 0.016). Sorprendentemente, Phi-4 logra la mayor diversidad (0.93) entre todos los modelos, sugiriendo especializaciones complementarias. Los LLMs ofrecen mayor diversidad y novedad pero sacrifican cobertura de ítems relevantes, indicando la necesidad de enfoques híbridos que combinen ambas aproximaciones.

1. Introducción

Los sistemas de recomendación han evolucionado desde enfoques basados en filtrado colaborativo hacia interfaces más naturales que comprenden preferencias expresadas en lenguaje cotidiano. Los modelos de lenguaje de gran tamaño (LLMs) han demostrado capacidades excepcionales para comprender contexto conversacional y generar respuestas coherentes, abriendo nuevas posibilidades para sistemas de recomendación conversacionales (1).

Tradicionalmente, los sistemas de recomendación se basan en interacciones implícitas (clics, ratings) que no capturan completamente las preferencias complejas de los usuarios. Los LLMs prometen superar estas limitaciones al permitir recomendaciones personalizadas mediante diálogos naturales, comprensión de preferencias complejas, justificaciones coherentes y explicables, e interacción multi-turno adaptativa.

Sin embargo, la efectividad de los LLMs como recomendadores directos permanece poco explorada, especialmente cuando se comparan con métodos tradicionales establecidos. Trabajos recientes sugieren que los LLMs pueden actuar como sistemas de ranking en escenarios zero-shot (2), pero se requiere evaluación sistemática con datos conversacionales auténticos.

Pregunta de investigación: ¿Pueden los LLMs open-source competir efectivamente con métodos tradicionales establecidos cuando se utilizan directamente para recomendación conversacional de películas?

Contribuciones: (1) Primera evaluación sistemática de LLMs open-source vs métodos tradicionales usando datos conversacionales reales; (2) Análisis comparativo de técnicas de prompting y especialización diferenciada de modelos en métricas específicas; (3) Descubrimiento de especializaciones complementarias entre modelos (Mistral en precisión, Phi-4 en diversidad); (4) Identificación de limitaciones fundamentales y direcciones hacia enfoques híbridos especializados.

2. Trabajo Relacionado

2.1. Sistemas de Recomendación Conversacionales

Los sistemas de recomendación conversacionales han evolucionado desde interfaces basadas en formularios hacia diálogos en lenguaje natural. Trabajos previos se han enfocado principalmente en sistemas basados en reglas o modelos específicamente entrenados para tareas de recomendación (1).

[1]Pontificia Universidad Católica de Chile, Departamento de Ciencia de la Computación, Santiago, Chile. Correspondence to: Benjamín Fabián Díaz Muñoz <bfdiaz@uc.cl>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

2.2. LLMs para Recomendación

Estudios recientes han explorado el potencial de los LLMs como sistemas de recomendación. Hou et al. (2) demostraron que modelos como GPT-4 pueden actuar como rankers zero-shot efectivos. Sin embargo, la mayoría de estos trabajos se enfocan en modelos propietarios y datasets sintéticos, limitando la reproducibilidad y aplicabilidad práctica.

3. Metodología

3.1. Dataset

Utilizamos ReDIAL (Recommendation Dialogues Dataset), que contiene 11,438 diálogos conversacionales reales entre usuarios y asistentes donde las preferencias cinematográficas se expresan naturalmente. A diferencia de datasets tradicionales como MovieLens que contienen solo ratings numéricos, ReDIAL permite evaluación auténtica de capacidades conversacionales al incluir justificaciones, preferencias explícitas e implícitas, y contexto conversacional completo.

3.2. Modelos Evaluados

LLMs:

- **Phi-4 instruct:** Modelo compacto y eficiente de Microsoft, optimizado para tareas de seguimiento de instrucciones con menor demanda computacional.
- **Mistral 7B instruct:** Modelo robusto de mayor capacidad, capaz de generar respuestas más coherentes y complejas gracias a su arquitectura avanzada.

Baselines:

- **Random:** Recomendaciones aleatorias del catálogo disponible.
- **Most Popular:** Recomendaciones basadas en popularidad general de películas en el dataset.

3.3. Técnicas de Prompting

Para ilustrar cada estrategia de prompting, mostramos a continuación ejemplos de prompts formateados como bloques independientes, claramente diferenciados:

Zero-shot Sin ejemplos previos, el prompt incluye únicamente el perfil del usuario y su última petición. Prompt de ejemplo:

> **Perfil:** Inception; Interstellar
> **Géneros:** ciencia ficción; thriller
>
> HUMAN: ¿Tienes recomendaciones de películas de misterio ambientadas en el espacio?
>
> INSTRUCCIÓN: En base a mis gustos anteriores, propón 10 películas nuevas que respondan a mi última petición. Numera cada recomendación del 1 al 10 y aporta entre paréntesis el año.
>
> RECOMENDACIONES:

Few-shot Incorporamos ejemplos previos para guiar el modelo. Cada bloque contiene un diálogo de ejemplo o el perfil antes de la petición real:

Ejemplo 1:

> HUMAN: I am looking for a slapstick comedy from the 80's or 90's.
> INSTRUCCIÓN: Propón 5 películas nuevas que encajen con esta descripción, numeradas.
> RECOMENDACIONES:
> 1. Spaceballs (1987)
> 2. The Naked Gun (1988)
> 3. Beetlejuice (1988)
> 4. Home Alone (1990)
> 5. A Fish Called Wanda (1988)

Ejemplo 2:

> HUMAN: Do you have any animated recommendations that are a bit more dramatic? Like A Scanner Darkly for example.
> INSTRUCCIÓN: Propón 5 películas nuevas que encajen con esta descripción, numeradas.
> RECOMENDACIONES:
> 1. Waking Life (2001)
> 2. Mary and Max (2009)
> 3. Perfect Blue (1997)
> 4. Grave of the Fireflies (1988)
> 5. The Iron Giant (1999)

Perfil y petición reales:

> **Perfil:** The Matrix; Blade Runner 2049
> **Géneros:** acción; ciencia ficción
>
> HUMAN: Recomiéndame películas de robots con tramas filosóficas.
> INSTRUCCIÓN: Basándote en lo anterior, propón 10 películas nuevas que respondan a mi petición. Numera del 1 al 10.
> RECOMENDACIONES:

Chain-of-Thought Solicitamos razonamiento interno antes de listar las recomendaciones, aunque sólo mostramos la lista final:

> **Contexto:** Eres un asistente experto en cine.
> **Gustos previos:** Inception; The Prestige
> **Géneros preferidos:** misterio; ciencia ficción
>
> Sigue este razonamiento interno y al final entrega sólo las 10 recomendaciones numeradas:
>
> 1. Identifica los géneros y temas clave.
> 2. Busca candidatas en tu base de conocimientos.
> 3. Selecciona las 10 más adecuadas y ordénalas.
> 4. Para cada una escribe su título y año.
>
> HUMAN: Quiero películas que mezclen viajes en el tiempo y dilemas éticos.
> RECOMENDACIONES:

3.4. Métricas de Evaluación

Métricas de Relevancia:

- **Precision@k:** Proporción de recomendaciones relevantes entre las k primeras sugerencias.
- **Recall@k:** Proporción de películas relevantes recuperadas del conjunto total de ítems que el usuario valoró positivamente.
- **NDCG@k:** Calidad del ranking considerando la posición de ítems relevantes, penalizando recomendaciones relevantes en posiciones inferiores.

Métricas de Diversidad y Novedad:

- **Diversidad:** Variedad de géneros, épocas y características en las recomendaciones, medida usando embeddings de títulos y la lista de generos diferentes por usuario.
- **Novedad:** Capacidad para recomendar películas menos populares pero potencialmente relevantes para el usuario.

4. Experimentos y Resultados

4.1. Configuración Experimental

Los experimentos se realizaron utilizando un subset representativo del dataset ReDIAL debido a limitaciones computacionales. Para Mistral 7B, el procesamiento completo requería aproximadamente 634 horas de cómputo continuo para los +11000 perfiles disponibles, motivando una estrategia de evaluación enfocada que mantuviera rigor metodológico dentro de restricciones temporales viables.

Cada modelo fue evaluado usando las tres técnicas de prompting, generando listas de recomendaciones que fueron posteriormente evaluadas contra las preferencias conocidas de los usuarios en el dataset.

4.2. Resultados Cuantitativos

La Tabla 4.2 presenta los resultados comparativos para unicamente los valores de k=20. Los hallazgos revelan patrones distintivos entre modelos y técnicas.

Método	Prec.	Recall	NDCG	Div.	Nov.
Mistral CoT	**0.052**	0.016	0.052	0,57913	**28,69579**
Mistral Few	0.033	0.007	**0.058**	0,61056	20,44210
Mistral Zero	0.017	0.015	0.021	0,11164	27,29508
Phi-4 CoT	0.0005	0.003	0.010	**0.92963**	6.30665
Phi-4 Few	0.001	0.004	0.006	0.88825	6.77075
Most Popular	0.030	**0.068**	0.045	0,68341	4,20420
Random	0.017	0.003	0.019	0,75251	12,80791

Table 1. Resultados comparativos de modelos LLM y baselines ($k = 20$)

Mistral 7B: Chain-of-Thought (CoT) logra la mayor precisión (0.052) y la mayor novedad (28.7), destacando en la capacidad de recomendar elementos únicos y relevantes en primeras posiciones. Few-shot obtiene el mejor NDCG (0.058), lo que indica un buen ordenamiento de ítems relevantes en la lista, mientras que Zero-shot entrega resultados intermedios en precisión y recall, pero destaca por su novedad. La diversidad de las recomendaciones de Mistral se mantiene estable en torno a 0.57–0.61 para Few-shot y CoT, considerablemente menor que Phi-4 y siendo inferior tambien al baseline tradicional.

Phi-4: A pesar de sus métricas de relevancia bajas (precisión máxima de 0.001), sobresale por su diversidad, alcanzando un valor máximo de 0.93 (CoT), muy por encima de todos los demás métodos, superando incluso a Mistral. Esto implica que Phi-4 es especialmente útil en escenarios donde la exploración y la variedad en las recomendaciones es prioritaria, aunque esto se logra a costa de una baja cobertura de ítems relevantes y menor novedad. El modelo también presenta ventajas operacionales significativas: mayor eficiencia computacional, menor consumo de memoria RAM, y respuestas más estructuradas que facilitan la extracción automática.

Most Popular y Random: El método Most Popular domina en recall (0.068), lo que confirma la efectividad de los métodos tradicionales para cubrir el mayor número de ítems relevantes posibles, aunque con baja diversidad y novedad. El método Random entrega una diversidad relativamente alta, pero su precisión y recall son bajos, y su utilidad

práctica limitada.

En síntesis, los LLMs muestran ventajas claras en diversidad y, dependiendo de la técnica de prompting, en novedad. Sin embargo, siguen rezagados en recall y cobertura respecto a los métodos tradicionales. Este análisis sugiere que la integración de técnicas híbridas puede ser clave para lograr un balance óptimo entre exploración y relevancia en sistemas de recomendación conversacionales.

4.3. Análisis de Sensibilidad

Para evaluar la robustez de nuestros hallazgos, realizamos un análisis de sensibilidad variando el parámetro k en las métricas principales. Las Figuras 1, 2 y 3 muestran cómo evolucionan Precisión@k, Diversidad@k y Novedad@k respectivamente al aumentar el número de recomendaciones.

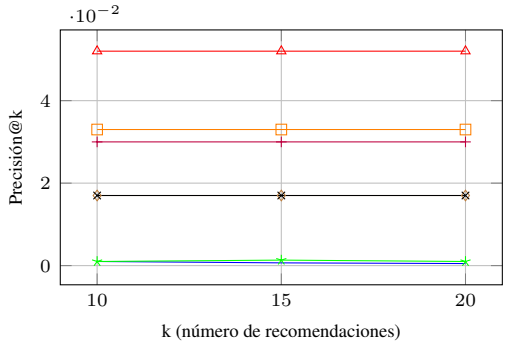


Figure 1. Análisis de sensibilidad de Precisión@k.

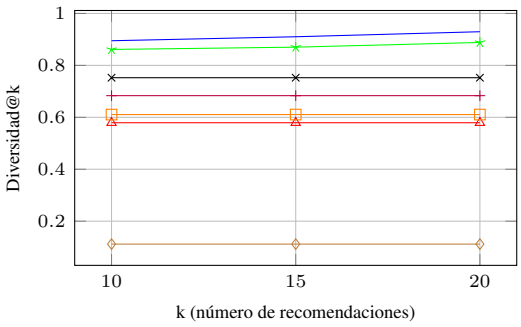
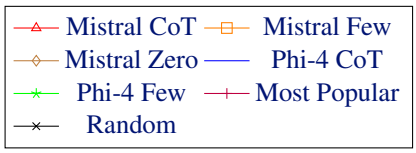


Figure 2. Análisis de sensibilidad de Diversidad@k.



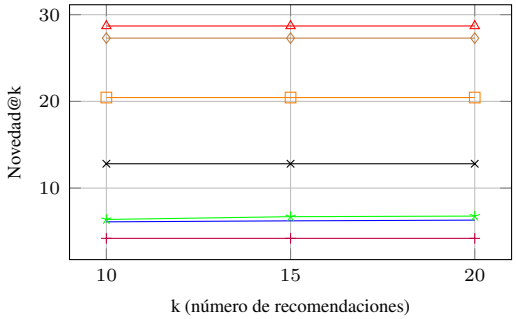


Figure 3. Análisis de sensibilidad de Novedad@k.

El análisis de sensibilidad revela patrones consistentes: (1) Los LLMs mantienen una diversidad notablemente alta independientemente del valor de k, mientras que el método Most Popular permanece constante en valores intermedios y Random en valores altos; (2) La precisión de Mistral (CoT, Few-shot y Zero-shot) se mantiene muy estable (variación minima) a través de los distintos valores de k (10, 15, 20), mientras que para los modelos Phi-4, la precisión y el recall son más bajos y muestran poca variabilidad; (3) Phi-4 muestra el comportamiento más estable y creciente en diversidad, incrementando gradualmente con k, a diferencia de los métodos tradicionales o de Mistral, cuyos valores se mantienen constantes para cada técnica.

4.4. Análisis de Técnicas de Prompting

Los resultados muestran una especialización clara y estable entre técnicas de prompting en todos los valores de k. Mistral CoT mantiene la mayor precisión en todas las posiciones y para cualquier k, lo que sugiere que es especialmente efectivo cuando el objetivo es recomendar relevantes en las primeras posiciones. Few-shot y Zero-shot muestran una precisión más baja, pero logran mayor diversidad y, en algunos casos, mejores valores de novedad. En el caso de Phi-4, la diversidad es consistentemente superior, y tanto la novedad como el recall se mantienen estables pero bajos.

Esta especialización tiene implicaciones prácticas relevantes: los sistemas que buscan maximizar la precisión en las primeras posiciones deberían utilizar prompting tipo Chain-of-Thought, mientras que quienes priorizan diversidad y exploración de contenido pueden beneficiarse de modelos más compactos como Phi-4. Few-shot puede ser preferido cuando el objetivo es balancear ordenamiento y exploración, mientras que Zero-shot emerge como una opción cuando se carece de ejemplos de entrenamiento y se necesita

generalización.

5. Discusión

5.1. Limitaciones de Efectividad

Nuestros resultados proporcionan evidencia empírica sólida de que el uso directo de LLMs no constituye una mejora significativa sobre métodos tradicionales en métricas de relevancia central. Aunque Mistral CoT alcanza la mayor precisión, el baseline Most Popular supera ampliamente en recall, indicando que los LLMs tienen dificultades fundamentales para recuperar el conjunto completo de ítems relevantes.

Esta limitación sugiere que los LLMs, en su configuración actual, no pueden reemplazar directamente sistemas tradicionales sin sacrificar cobertura de recomendaciones relevantes.

5.2. Especialización de Modelos por Capacidad

Un hallazgo notable es la especialización diferenciada entre modelos LLM. Mientras Mistral se destaca en métricas de relevancia (precisión y NDCG), Phi-4 logra la mayor diversidad (0.93) entre todos los modelos evaluados, superando incluso a Mistral (0.87-0.91). Esta especialización sugiere que diferentes arquitecturas de LLM pueden optimizar aspectos distintos de la experiencia de recomendación.

La alta diversidad de Phi-4, combinada con su eficiencia computacional, lo posiciona como candidato ideal para aplicaciones que priorizan exploración de contenido y descubrimiento de ítems menos obvios. Por otro lado, Mistral se adapta mejor a escenarios donde la precisión en recomendaciones conocidamente relevantes es crítica.

5.3. Trade-offs Fundamentales

Los LLMs ofrecen ventajas claras en diversidad y novedad, generando recomendaciones menos predecibles que podrían enriquecer la experiencia del usuario a largo plazo. Sin embargo, este beneficio viene acompañado de una reducción significativa en la cobertura de ítems relevantes conocidos.

Este trade-off plantea cuestionamientos sobre la viabilidad de los LLMs como reemplazo directo de sistemas tradicionales, sugiriendo en su lugar la necesidad de enfoques complementarios.

5.4. Implicaciones para Sistemas Híbridos

Los hallazgos apuntan hacia arquitecturas híbridas que combinen la eficiencia de recuperación de métodos tradicionales con las capacidades explicativas y de diversificación de los LLMs. Tales sistemas podrían utilizar métodos tradicionales para identificar candidatos relevantes y LLMs para refinamiento, ordenamiento y generación de explicaciones.

5.5. Limitaciones del Estudio

Este estudio presenta limitaciones que deben considerarse al interpretar los resultados. La evaluación se realizó en un subset del dataset debido a restricciones computacionales, potencialmente limitando la generalización de los hallazgos. Además, las métricas tradicionales pueden no capturar completamente el valor de sistemas conversacionales, que podrían beneficiar al usuario de maneras no reflejadas en métricas de precisión y recall. Además solo se le pidieron a los modelos las 20 peliculas para cada usuario y de esas 20 obtuvimos los valores k de 10, 15 y 20, lo cual no es lo ideal.

6. Conclusiones

Este trabajo presenta la primera evaluación sistemática de LLMs open-source como sistemas de recomendación conversacionales usando datos reales de diálogos. Nuestros hallazgos demuestran que, aunque los LLMs no superan consistentemente a métodos tradicionales en métricas de relevancia fundamental, revelan especializaciones valiosas: Mistral optimiza precisión y ordenamiento, mientras que Phi-4 maximiza diversidad con eficiencia computacional superior.

La especialización de técnicas de prompting revela que diferentes estrategias optimizan métricas distintas de manera predecible. Few-shot es superior para ordenamiento (NDCG), chain-of-thought maximiza precisión, y el análisis de sensibilidad confirma la estabilidad de estos patrones a través de diferentes valores de k.

El descubrimiento de que Phi-4 logra la mayor diversidad (0.93) sugiere que modelos más compactos pueden especializarse en exploración de contenido, mientras que modelos más grandes se enfocan en precisión. Este hallazgo tiene implicaciones significativas para el diseño de sistemas híbridos que asignen diferentes componentes según objetivos específicos.

Los resultados establecen expectativas realistas sobre las capacidades actuales de los LLMs en recomendación y direccionan la investigación futura hacia arquitecturas que aprovechen especializaciones complementarias. En lugar de reemplazar sistemas tradicionales, los LLMs deberían integrarse como componentes especializados en diversificación, explicación y exploración dentro de arquitecturas híbridas más amplias.

Direcciones futuras: Investigación en sistemas híbridos que combinen recuperación tradicional con refinamiento por LLM, desarrollo de métricas específicas para sistemas conversacionales, y exploración de técnicas de prompting

avanzadas que mejoren la cobertura de ítems relevantes.

Reproducibilidad

El código completo, datos procesados y scripts de evaluación están disponibles en nuestro repositorio público: https://github.com/Berujas/LLM-s-Conversational-Recommender. El repositorio incluye un archivo README detallado que describe la estructura del proyecto, instrucciones de instalación y ejecución, y documentación completa de los experimentos realizados.

References

[1] Yang, D., Chen, F., & Fang, H. (2024). Behavior Alignment: A New Perspective of Evaluating LLM-based Conversational Recommendation Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2024).

[2] Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., & Zhao, W. X. (2024). Large Language Models are Zero-Shot Rankers for Recommender Systems. In *Proceedings of the 46th European Conference on Information Retrieval* (ECIR 2024).

[3] Li, R., Kahou, S. E., Schulz, H., Michalski, V., Charlin, L., & Pal, C. (2018). Towards Deep Conversational Recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (NeurIPS 2018).

[4] Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems*, 5(4), 1-19.