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

Benjamín Fabián Díaz Muñoz, Daniel Esteban Eduardo Alegría Toledo, Benjamín Andrés Manuel Faúndez Romero

Pontificia Universidad Católica de Chile - Departamento de Ciencia de la Computación
IIC3633 - Sistemas Recomendadores - Profesor: Denis Parra

Motivación

Los modelos de lenguaje de gran tamaño (LLMs) han revolucionado el procesamiento de lenguaje natural, habilitando interacciones conversacionales más naturales que abren nuevas posibilidades en sistemas de recomendación. Estas tecnologías prometen recomendaciones personalizadas mediante diálogos naturales, comprensión de preferencias complejas expresadas en lenguaje cotidiano, justificaciones coherentes y explicables para cada sugerencia, e interacción multi-turno adaptativa que evoluciona con el usuario.

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

Objetivos de Investigación

Objetivo General: Evaluar comparativamente el desempeño de LLMs open-source recientes como sistemas de recomendación conversacionales de películas, utilizando datos reales de interacciones usuario-asistente.

Objetivos Específicos: Comparar la efectividad de los modelos Phi-4 y Mistral 7B frente a métodos baseline tradicionales, analizar el impacto de diferentes técnicas de prompting (zero-shot, few-shot, chain-of-thought) en la calidad de las recomendaciones generadas, evaluar la calidad conversacional usando el dataset ReDIAL con diálogos auténticos, y medir múltiples dimensiones de rendimiento incluyendo relevancia, diversidad y novedad de las recomendaciones.

Metodología

Dataset: ReDIAL contiene 11,438 diálogos conversacionales reales entre usuarios y asistentes, donde las preferencias cinematográficas se expresan naturalmente, permitiendo evaluación auténtica de capacidades conversacionales.

Modelos Evaluados: Se compararon Phi-4 instruct (modelo compacto y eficiente) y Mistral 7B instruct (modelo robusto de mayor capacidad) contra baselines tradicionales (Random, Most Popular).

Técnicas de Prompting: Se implementaron Zero-shot (sin ejemplos), Few-shot (con ejemplos del dataset) y Chain-of-Thought (razonamiento explícito paso a paso) para analizar el impacto en la calidad de recomendaciones.

Métricas de Evaluación

Métricas Cuantitativas: Se evaluaron Precision@k y Recall@k para medir relevancia inmediata, NDCG@k para calidad del ranking considerando posición, Diversidad basada en embeddings de títulos, y Novedad para capacidad de recomendar contenido menos popular.

Análisis Cualitativo: Se analizó la calidad de explicaciones generadas, naturalidad del diálogo mantenido, adecuación a preferencias del usuario y coherencia conversacional multi-turno para evaluar la experiencia integral del usuario.

Resultados Finales

Experimento completo: Evaluación exhaustiva con dataset ReDIAL

Modelo/Método	Precision@20	Recall@20	NDCG@20	Diversidad
Mistral CoT	**0.052**	0.016	0.052	Alta
Mistral Few-shot	0.033	0.007	**0.058**	Alta
Mistral Zero-shot	0.017	0.015	0.021	Alta
Phi-4 CoT	0.0005	0.0005	0.01	Media
Phi-4 Few-shot	0.001	0.001	0.006	Media
Most Popular	0.030	**0.068**	0.045	Baja
Random	0.017	0.003	0.019	Media

Hallazgos clave: Mistral CoT logra la mayor precisión, mientras que Few-shot destaca en ordenamiento (NDCG). Sin embargo, Most Popular supera a todos los LLMs en recall, manteniendo su efectividad como baseline. Los LLMs ofrecen mayor diversidad y novedad pero con menor cobertura de ítems relevantes.

Análisis Mistral 7B

Chain-of-Thought: Mayor precisión (0.052) pero valores constantes al aumentar k.

Few-shot: Mejor NDCG (0.058), superior para ordenar ítems relevantes.

Zero-shot: Precisión intermedia (0.017) pero mejor recall (0.015).

Análisis Phi-4 y Baselines

Phi-4: Métricas bajas (precisión 0.001) pero mayor eficiencia computacional.

Most Popular: Supera a LLMs en recall (0.068), validando métodos tradicionales.

Random: Baseline con métricas bajas (precision 0.017, recall 0.003).

Conclusiones e Implicaciones

Limitaciones de efectividad: Nuestros resultados demuestran que el uso directo de LLMs no constituye una mejora significativa sobre métodos tradicionales como Most Popular. Aunque Mistral CoT alcanza la mayor precisión (0.052), el baseline supera ampliamente en recall (0.068 vs 0.016), indicando que los LLMs tienen dificultades para recuperar el conjunto completo de ítems relevantes que un usuario podría apreciar.

Especialización técnica revelada: El análisis de técnicas de prompting revela patrones distintivos: Few-shot se destaca en ordenamiento de recomendaciones con el mejor NDCG, mientras que Chain-of-Thought optimiza la precisión pero muestra limitaciones en diversificación posicional. Este hallazgo sugiere que diferentes aplicaciones requieren estrategias específicas de prompting.

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, planteando cuestionamientos sobre su viabilidad como reemplazo directo de sistemas tradicionales.

Direcciones futuras: Los resultados apuntan hacia la necesidad de enfoques híbridos que combinen la eficiencia de recuperación de métodos tradicionales con las capacidades explicativas y de diversificación de los LLMs, en lugar de intentar reemplazar completamente los sistemas existentes.

Referencias

Hou, Y., Zhang, J., Lin, Z., et al. (2024). *Large Language Models are Zero-Shot Rankers for Recommender Systems.* ECIR 2024.

Yang, D., Chen, F., & Fang, H. (2024). *Behavior Alignment: A New Perspective of Evaluating LLM-based Conversational Recommendation Systems.* SIGIR 2024.