# LLMs como evaluadores de sistemas recomendadores

Tomás Vergara Browne
tomvergara@uc.cl
Pontificia Universidad Católica de Chile

Sebastián Pérez Masri
sperezmasri@uc.cl
Pontificia Universidad Católica de Chile

Tomás Fouyet Ovalle
tfouyet@uc.cl
Pontificia Universidad Católica de Chile

## ABSTRACT

Esta investigación explora el uso de Modelos de Lenguaje de Gran Escala (LLMs) como evaluadores de sistemas recomendadores, enfocándose en métricas no numéricas, especialmente en la serendipia, entendida como una sorpresa agradable. Se utilizaron cuatro LLMs, incluyendo ChatGPT, LLAMA-2-7b, Mistral-7b y Neural Chat 7b, para valorar la serendipia en recomendaciones de películas mediante prompts estandarizados. A pesar de los esfuerzos, la correlación de ChatGPT con evaluaciones humanas resultó débil. Sin embargo, en el contexto de recomendaciones de recetas, ChatGPT mostró correlaciones más fuertes. Análisis adicionales con LLAMA-2-7b, Mistral-7b y Neural Chat 7b revelaron una una correlación negativa general entre las predicciones de los modelos y las evaluaciones humanas. Estos hallazgos destacan la complejidad de usar LLMs para evaluar métricas subjetivas y resaltan los desafíos de alinear evaluaciones de IA con la percepción humana. Este estudio contribuye al entendimiento de las capacidades y limitaciones de los LLMs en sistemas de recomendación y sugiere la necesidad de métodos más sofisticados o enfoques específicos para mejorar la evaluación de la serendipia. Finalmente, hicimos público nuestro código en un repositorio para fomentar la reproducibilidad.[1]

**ACM Reference Format:**
Tomás Vergara Browne, Sebastián Pérez Masri, and Tomás Fouyet Ovalle. 2023. LLMs como evaluadores de sistemas recomendadores. In *Sistemas Recomendadores UC (IIC3633), December, 2023, Santiago, Chile.* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCCIÓN

Con el creciente uso de los Modelos de Lenguaje (LLM), hemos identificado un potencial significativo para su aplicación como evaluadores de sistemas recomendadores, especialmente en métricas que no se pueden expresar numéricamente. Consideramos que, en particular, para el caso de *serendipity*, un modelo de lenguaje podría desempeñarse de manera comparable a la evaluación realizada por un humano. Entendemos la serendipia como una sorpresa agradable. Decidimos utilizar cuatro LLMs para este estudio, con la expectativa de que esta variedad generaría resultados más completos y detallados.

[1]https://github.com/tvergara/RecSysProject

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*IIC3633, December, 2023, Pontificia Universidad Católica de Chile*
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

## 2 METODOLOGÍA

En este estudio, se adopta una metodología generalizada para evaluar la serendipia en sistemas de recomendación utilizando varios LLMs. La investigación se enfoca en cómo estos modelos avanzados pueden interpretar y evaluar la calidad de las recomendaciones basándose en el concepto de serendipia, entendida como la agradable sorpresa en una recomendación.

Para lograr esto construimos prompts con una estructura estandarizada que consiste en lo siguiente: una introducción que establece el contexto de la recomendación y la definición de serendipia. Luego, se proporciona una lista de películas previamente vistas y calificadas por el usuario, detallando elementos como el título, director, actores principales y géneros, seguido de la calificación que el usuario dio a cada película. El prompt se completa indicando las películas recomendadsa, incorporando detalles similares a los ya mencionados. Por último se le indica al LLM que debe evaluar la serendipia de la recomendación en una escala de 1 a 5. Un ejemplo de un prompt se puede encontrar en el apéndice C.

Con respecto a los modelos de lenguaje escogidos, decidimos utilizar ChatGPT (3.5), Llama-2-7b, Mistral-7b y Neural Chat 7b (un modelo fine tuneado en seguir instrucciones a partir de Mistral-7b). El primero debido a que es uno de los modelos del enguaje más reconocidos y utilizados a nivel mundial. Los otros tres, ya que nos permiten contar con una mayor diversidad, además de contar con ser open source.

## 3 RESULTADOS

Decidimos usar los datasets varios datasets abiertos con evaluaciones humanas de la serendipia de recomendaciones, con el objetivo de mostrar el potencial que ChatGPT tiene como un evaluador automático de estas métricas. En particular, usamos el dataset de Serendipity 2018 [2] (un dataset creado por el equipo de MovieLens, en donde evaluan la serendipia con ratings humanos para recomendaciones de películas) y el dataset de Serendipty TaoBao [3] (en donde tienen evaluaciones humanas de serendipia en items recomendados por la página de Alibaba). Sin embargo, nos encontramos con los siguientes problemas:

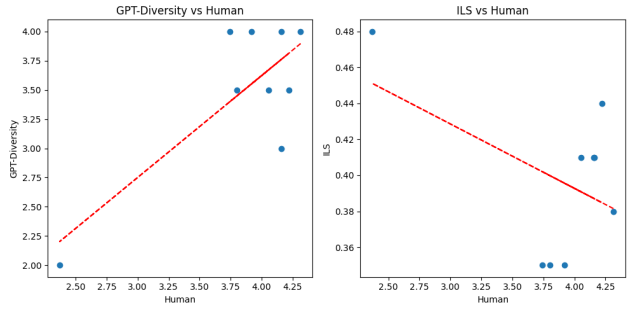
- Para el dataset de Serendipity 2018, evaluamos usando sólo un subset de 100 de las recomendaciones usando ChatGPT (por temas de costo). En el dataset evaluaban en una escala Likert del 1 al 5 varias preguntas con respecto al serendipity. A pesar de horas de probar distintos prompts, y de mostrar la información histórica del usuario de manera distinta, no logramos una correlación importante entre la evaluación de serendipity y la evaluación de ChatGPT (el mejor resultado fue una correlación de 0.08). Usar más de los datos, o usar GPT-4 nos iba a exceder los costos más de lo que es razonable

para el contexto de un ramo, por lo que desistimos de seguir intentando.

- Para el dataset de TaoBao, ellos evaluaron del 1 al 5 la recomendación de productos en una página de e-commerce de indonesia (Alibaba). Si bien ellos en el paper establecían que habían liberado el dataset para que más investigadores puedan trabajar sobre sus datos, no estaba claro el disclaimer que ellos anonimizaron los productos recomendados y el historia, por lo que no teníamos manera de obtener información que permita a un modelo de lenguaje juzgar el serendipity de la recomendación.

A partir de estos resultados, buscamos más datasets, de otras métricas que nos podrían ayudar. Encontramos el dataset creado en [1], en el que usuarios evaluan la diversidad de una lista recomendada, tanto en el ámbito de películas como en el de recetas de cocina. Para estos datos, sí logramos obtener resultados razonables.

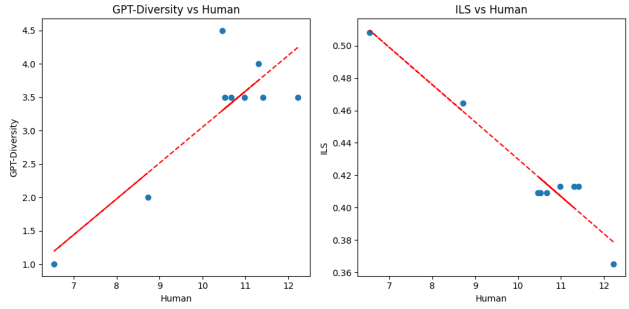
Primero, en el ámbito de recetas (en el paper es llamado "study 1b"), obtuvimos el siguiente gráfico.




Vemos que hay una clara correlación entre la percepción de ChatGPT y de los usuarios.

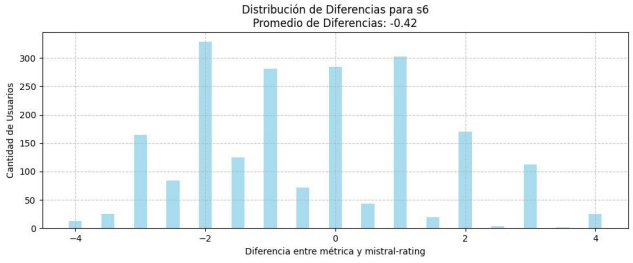
Además, tenemos evidencia para afirmar que el juicio de ChatGPT es un mejor predictor para el juicio de los usuarios que la métrica automática usada en el paper, Intra-List Similarity (ILS). Esto lo podemos afirmar porque al hacer la regresión lineal obtenemos el $p$-value asociado de si las variables están o no correlacionadas, y vemos que el valor asociado a ChatGPT es considerablemente menor al de ILS (0.039 vs 0.487, Apéndice A).

Sin embargo, al evaluar en el otro dataset del paper, en el ámbito de películas (lo que ellos llaman "study 1a"), no encontramos los mismos resultados.

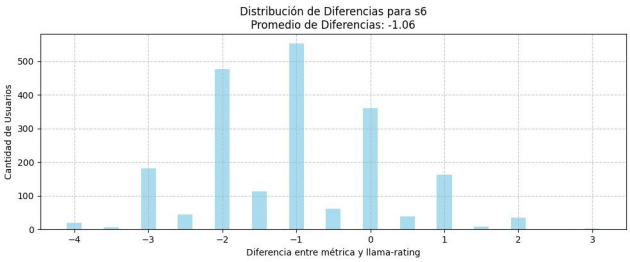



Si bien ChatGPT pareciera estar correlacionado con la percepción de los usuarios, ILS es claramente un mejor predictor. De nuevo, lo podemos ver con los $p$-values asociados a las regresiones lineales, en donde ILS es bástamente menor en esta métrica (0.005 vs 0.518, apéndice B). Esto muestra que ILS es considerablemente superior a ChatGPT como predictor de la percepción humana en este dominio.

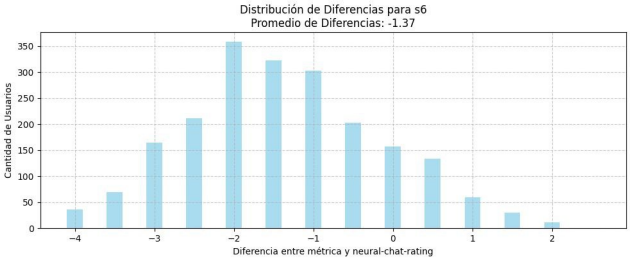
Dada la complejidad de los resultados obtenidos con ChatGPT, decidimos explorar otras opciones de evaluación. Para ello, evaluamos otro conjunto de datos utilizando tres modelos adicionales: LLAMA-2-7b, Mistral-7b y Neural Chat 7b. Se siguió un procedimiento similar al utilizado con ChatGPT, generando consultas específicas que solicitaban la predicción de la serendipity de una recomendación de película para un usuario, utilizando una escala de 1 a 5 y proporcionando una lista de películas previamente vistas por el usuario, junto con sus calificaciones correspondientes. Este enfoque nos permitió ampliar nuestro análisis y obtener una perspectiva más completa de cómo diferentes modelos abordan la evaluación de la serendipity en recomendaciones de películas.
Después de completar el entrenamiento de nuestros modelos y obtener métricas específicas en relación con la serendipia, emprendimos la tarea de establecer conexiones significativas entre los resultados generados por los modelos y las respuestas proporcionadas por los usuarios. Al examinar detenidamente el conjunto de preguntas que los usuarios respondían, nos llamó la atención una en particular: *I was (or, would have been) surprised that MovieLens picked this movie to recommend to me.*. Esta pregunta, calificada por los usuarios en una escala del 1 al 5, parecía alinearse estrechamente con nuestra intención de predecir la serendipia. Decidimos realizar un análisis preliminar de los resultados, calculando la diferencia entre el valor obtenido a través de la métrica del modelo (*modelo-rating*) y la calificación proporcionada por el usuario para la pregunta mencionada anteriormente. Además de esta pregunta, también optamos por comparar con otras dos preguntas que abordaban aspectos relacionados con la serendipia: *This is the type of movie I would not normally discover on my own; I need a recommender system like MovieLens to find movies like this one* y *This movie is different (e.g., in style, genre, topic) from the movies I usually watch.*




El gráfico de barras representa la diferencia entre el valor de serendipia entregado por el modelo Mistral-7B y las respuestas del usuario a las preguntas mencionadas anteriormente. Destaca que el promedio de estas diferencias no se aparta significativamente de cero, sugiriendo que inicialmente creíamos que el modelo tenía la capacidad de proporcionar resultados similares a las respuestas de los usuarios en relación con las películas. Este hallazgo nos llevó a considerar que los resultados de serendipia entregados por el modelo Mistral-7B podrían ser más prometedores que los obtenidos con ChatGPT.



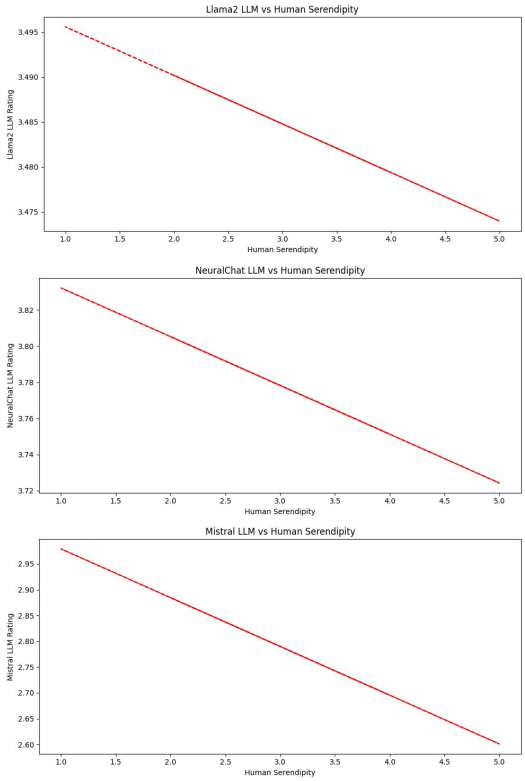

En el caso del modelo LLAMA-2-7B, los resultados presentan una mayor variación, indicando una medida de serendipia menos precisa en comparación con el modelo Mistral-7B. A pesar de ello, los resultados aún son notables, ya que el promedio de las diferencias apenas se aleja en 1 de la escala entre la métrica asociada y la pregunta formulada al usuario.




Finalmente, observamos que el modelo Neural Chat 7B proporcionó resultados más distantes en comparación con los otros dos modelos, indicando que su evaluación de serendipia fue la menos precisa entre los tres. Pareciera tener un sesgo relativamente consistente a subestimar al serendipia percibida por el usuario.

Interesantemente, los resultados de Neural Chat 7B se parecen más hacia una distribución normal, mientras que los modelos anteriores parecían tener gaps. Es posible que dado que Neural Chat 7B es un modelo finetuneado en seguir instrucciones, pueda lograr realmente evaluar en un juicio más del estilo de los humanos. Los modelos que no tienen esta etapa de finetuning, mostraban un sesgo hacia evitar los valores decimales en las respuestas (posiblemente tienen este sesgo debido a que fueron entrenados para predecir la siguiente palabra de corpuses enormes de texto, en donde los números enteros puede que sean más comunes que los números decimales).

Luego de abordar el problema de esta manera, decidimos graficar las pendientes obtenidas al hacer la correlacion de los resultados obtenidos en cada modelos con los valores reales de serendipia que entregaron los usuarios. Para los tres modelos obtuvimos un resultados bastantes similares.




Podemos ver que para los tres modelos, obtenemos un grafico de correlación ligeramente negativa. Esto indica que la tendencia del resultado entregado por el modelo se comporta de manera inversa a los resultados obtenidos por los usuarios. En este caso específico, se está sugiriendo que a medida que la percepción de serendipia por parte de los humanos aumenta (según sus respuestas a preguntas específicas), las predicciones de los modelos tienden a disminuir, y cuando la percepción humana de serendipia disminuye, las predicciones del modelo tienden a aumentar.

## 4 DISCUSIÓN

La obtención y análisis de los resultados proporciona una visión integral del desempeño de diferentes modelos, y las observaciones abren puertas a diversas discusiones y reflexiones.

Primero, con respecto a ChatGPT, queda claro que la correlación con la percepción humana de serendipia no es lo suficientemente fuerte, especialmente en el ámbito de películas. Aunque existe cierta correlación, la métrica automática ILS se destaca como un predictor más robusto de la percepción humana en este dominio específico. Este hallazgo destaca la complejidad de utilizar modelos de lenguaje para evaluar la serendipia y sugiere que, aunque puedan capturar ciertos aspectos, no son perfectos en todos los contextos.

La transición hacia otros modelos, como LLAMA-2-7B, Mistral-7B y Neural Chat 7B, muestra una diversidad en los resultados. El análisis de las respuestas proporcionadas por los usuarios destaca la variabilidad en la capacidad de cada modelo para predecir la serendipia, con diferencias notables entre ellos. Además, el enfoque de comparar las respuestas de los modelos con preguntas específicas formuladas a los usuarios proporciona una perspectiva valiosa para

entender cómo cada modelo aborda la sorpresa en las recomendaciones.

La visualización de la correlación negativa en los gráficos para todos los modelos sugiere un patrón interesante. La tendencia inversa entre la percepción humana y las predicciones de los modelos podría indicar que, en algunos casos, los modelos podrían estar perdiendo. Este patrón no es exclusivo de un modelo en particular, sino que se observa en todos los modelos analizados. Esto es poco prometedor en el sentido de utilizar modelos de lenguaje como evaluadores de métricas de serendipia, pero además es un resultado interesante. ¿Por qué existe una correlación negativa consistente entre distintos modelos de lenguaje al comparar su percepción de serendipia con la percepción humana? Encontrar una explicación no es sencilla, pero lo dejamos propuesto como posible trabajo futuro en el área.

Es crucial destacar que la serendipia es una métrica intrínsecamente subjetiva, incluso para el propio humano que realiza la evaluación. La interpretación de lo que constituye una experiencia sorprendente y novedosa puede variar significativamente entre individuos. Esta subjetividad inherente complica aún más la tarea de evaluar la serendipia, ya que la misma recomendación puede generar respuestas diversas según las expectativas y preferencias de cada usuario. En este contexto, la búsqueda de un modelo que pueda capturar y cuantificar de manera precisa esta cualidad subjetiva representa un desafío sustancial en la investigación de sistemas de recomendación.

Estos resultados invitan a una discusión más profunda sobre la naturaleza de la serendipia y cómo los modelos de lenguaje pueden aproximarse a su evaluación. La discrepancia entre las predicciones de los modelos y las percepciones humanas destaca la complejidad de este problema y subraya la necesidad de métodos más sofisticados o enfoques específicos para mejorar la capacidad de los modelos para capturar la sorpresa y la novedad en las recomendaciones. Movernos hacia métricas confiables y automatizadas de sistemas recomendadores es un camino a futuro que tiene el potencial de acelerar fuertemente la investigación en sistemas recomendadores, y mostramos un resultado que muestra que no es un trabajo sencillo.

Además, la elección de modelos más avanzados, como GPT-4 o Gemini, podría ser una dirección futura para explorar si estos modelos más potentes podrían superar las limitaciones encontradas en este estudio, aunque se reconoce que esto podría conllevar mayores costos computacionales.

## REFERENCES


[1] Mathias Jesse, Christine Bauer, and Dietmar Jannach. 2023. Intra-list similarity and human diversity perceptions of recommendations: the details matter. *User Modeling and User-Adapted Interaction* 33, 4 (2023), 769–802.

[2] Denis Kotkov, Joseph Konstan, Qian Zhao, and Jari Veijalainen. 2018. Investigating serendipity in recommender systems based on real user feedback. 1341–1350. https://doi.org/10.1145/3167132.3167276

[3] Salsabila Martono, Dana Sulistyo Kusumo, Arfive Gandhi, Su Cheng Haw, and Kok Why Ng. 2023. User Evaluation of Diversity and Novelty in the Redesigned Recommender List for an Indonesian E-Commerce Platform. *Journal of System and Management Sciences* 13, 4 (2023), 615–623.


## A APPENDIX A

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | human | R-squared: | 0.639 |
| Model: | OLS | Adj. R-squared: | 0.518 |
| Method: | Least Squares | F-statistic: | 5.303 |
| Date: | Thu, 02 Nov 2023 | Prob (F-statistic): | 0.0472 |
| Time: | 13:03:10 | Log-Likelihood: | −2.9184 |
| No. Observations: | 9 | AIC: | 11.84 |
| Df Residuals: | 6 | BIC: | 12.43 |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | −0.6812 | 2.956 | −0.230 | 0.825 | −7.913 | 6.551 |
| gpt-diversity | 0.8825 | 0.335 | 2.632 | 0.039 | 0.062 | 1.703 |
| ils | 3.6501 | 4.935 | 0.740 | 0.487 | −8.425 | 15.725 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.154 | Durbin-Watson: | 2.261 |
| Prob(Omnibus): | 0.562 | Jarque-Bera (JB): | 0.436 |
| Skew: | 0.518 | Prob(JB): | 0.804 |
| Kurtosis: | 2.706 | Cond. No. | 155. |

## B APPENDIX B

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | human | R-squared: | 0.937 |
| Model: | OLS | Adj. R-squared: | 0.915 |
| Method: | Least Squares | F-statistic: | 44.33 |
| Date: | Sun, 05 Nov 2023 | Prob (F-statistic): | 0.000255 |
| Time: | 15:36:02 | Log-Likelihood: | −4.6177 |
| No. Observations: | 9 | AIC: | 15.24 |
| Df Residuals: | 6 | BIC: | 15.83 |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 24.7111 | 4.384 | 5.637 | 0.001 | 13.985 | 35.437 |
| gpt-diversity | 0.2159 | 0.314 | 0.687 | 0.518 | −0.553 | 0.985 |
| ils | −35.6996 | 8.245 | −4.330 | 0.005 | −55.876 | −15.524 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.543 | Durbin-Watson: | 3.144 |
| Prob(Omnibus): | 0.762 | Jarque-Bera (JB): | 0.533 |
| Skew: | 0.279 | Prob(JB): | 0.766 |
| Kurtosis: | 1.947 | Cond. No. | 200. |

## C APPENDIX C

Este es un ejemplo del prompt utilizado en el dataset de Serendipity 2018.

```
We are interested in predicting the
    ↪ serendipity of a movie recommendation
    ↪  for a user. We understand
    ↪ serendipity as the pleasent surprise
    ↪ in the recommendation. We will
    ↪ evaluate in a scale of 1 to 5. The
    ↪ following is a list of movies the
    ↪ user watched, and its corresponding
    ↪ rating.

Title: Baywatch (2017)
Directed by: Seth Gordon
Starring: Dwayne Johnson,Zac Efron,Alexandra
    ↪  Daddario,Kelly Rohrbach,Priyanka
    ↪ Chopra
Genres: Action,Comedy
The user rated it as: 1.0/5

Title: Kingsman: The Golden Circle (2017)
Directed by: Matthew Vaughn
```

```
Starring: Taron Egerton,Julianne Moore,Mark
    ↪ Strong,Sophie Cookson,Colin Firth
Genres: Action,Adventure,Comedy
The user rated it as: 3.0/5

Title: Atomic Blonde (2017)
Directed by: David Leitch
Starring: Charlize Theron,James McAvoy,Sofia
    ↪  Boutella,John Goodman,Toby Jones
Genres: Thriller
The user rated it as: 5.0/5

The following movie was recommended to the
    ↪ user:

Title: High Sierra (1941)
Directed by: Raoul Walsh
Starring: Ida Lupino, Humphrey Bogart, Alan
    ↪ Curtis, Arthur Kennedy
Genres: Crime,Drama,Film-Noir,Thriller
The user rated it as: 4.0/5

What would you rate the serendipity of this
    ↪ recommendation in a scale of 1 to 5?
    ↪ Please, first give your thoughts, and
    ↪  afterwards answer in the format "
    ↪ Serendipity: {value}"
```