

Buscar

 \mathcal{L}

000

Leveraging LLM reasoning Enhances Personalized Recommender Systems

Autores: Alicia Y. Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, Taibai Xu, Zemin Zhang, Lichan Hong, Ed H.Chi, Xinyang Yi

IIC3633, Felipe Jara, José Lira, Domingo Venegas



Contexto:

- Rápido avance de las capacidades de LLMs en distintos tipos de tareas.
- Emergencia de Chain of Thought prompting habilita razonamientos de varios pasos en estos modelos.







Contexto:

 Estudios anteriores en LLMs aplicados recsys no estudian las capacidades de razonamiento, en el contexto de preferencias personalizadas.





Conceptos Clave:



Zero Shot Prompting 2

Chain of
Thought
prompting

8

LLM Fine Tuning

Trabajos Relacionados:



1

Recommendation as instruction following: A large language model empowered recommendation approach¹ Aproximación distinta a los recsys tradicionales

LLMs habilitan "user-friendly recommender systems", potenciados por lenguaje natural

1: Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as instruction following: A large language model empowered recommendation ap proach. arXiv preprint arXiv:2305.07001

Trabajos Relacionados:



2

Do LLMs Understand User Preferences? Evaluating LLMs On User Rating Prediction² También buscan hacer predicción de ratings (con zero-shot, few-shot y fine tuning)

Superioridad de filtrado colaborativo, pero buenos resultados en fine tunina

2: Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Ma heswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do 11ms understand user preferences? evaluating 11ms on user rating pre diction. arXiv preprint arXiv:2305.06474

Trabajos Relacionados:



3

Chat-REC: Towards interactive and explainable LLMs Augmented Recommender Systems³ Otra perspectiva a LLMs + recsys: transformar perfiles de usuario a una serie de prompts

Logra un recsys interactivo (chat) y más explicable, atacando el problema de cold start

3: Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable 11ms augmented recommender system. arXiv preprint arXiv:2303.14524.



Contribuciones principales



1. Impacto del uso de razonamiento en LLMs para predicción de ratings

Se evalúan los resultados de predicciones en 4 casos

- 1. Zero-shot + CoT (método propuesto)
- 2. Zero-shot
- 3. Zero-shot (-review)
- 4. Zero-shot (-review, -rating history)



Contribuciones principales



2. Demostrar la efectividad del uso de 'data de razonamiento' para hacer fine-tuning de un modelos pequeños:

Se evalúan las predicciones de un modelo "fine tuned" en la data de razonamiento generada por otros LLMs más grandes.

Se prueban distintos modelos: razonamientos filtrados vs no filtrados en relación al ground truth





Contribuciones principales



3. Propuesta de un framework para evaluar el razonamiento generado por los LLMs (Rec-SAVER)

Se encuentra una alineación con el juicio humano en base a 3 parámetros clave: coherence, faithfulness, insightfulness



 $\times \Box -$

Problemas de recomendación



- Predicciones de ratings para un usuario ítem usando LLMs.
- Amazon product review dataset:
 BEAUTY y MOVIES/TV

 $\times \square -$

Problemas de recomendación



El dataset contiene:

- Ratings del 1 al 5.
- Reviews proporcionados por el usuario a cada ítem con el que interactúa.
- Historial de las compras del usuario.
- Metadata del ítem

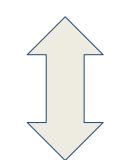
 $X \square -$

Problemas de recomendación



$$\hat{\mathbf{r}}_{u,i} = \arg \max_{k} \mathbb{P}(\mathbf{r}_{u,i} = k \mid \mathcal{H}_u, \mathcal{M}_i),$$

where $i \notin \mathcal{H}_u, k \in \{1, 2, 3, 4, 5\}.$



$$\hat{\mathbf{r}}_{u,i} = \text{LLM}(\mathcal{H}_u, \mathcal{M}_i)$$

*

Procedimientos:

• Zero shot v/s razonamiento

• Fine-tunning

• Rec-SAVER



1. Zero shot v/s razonamiento



Se aplica un prompt al modelo PaLM 2-M, con 4 secciones relevantes:

Preamblee.g. Here is information about a user and a new product ...User History $h_{u,1} = (\mathcal{M}_1, \mathbf{r}_{u,1}, \mathbf{d}_{u,1}), \ldots, h_{u,t} = (\mathcal{M}_t, \mathbf{r}_{u,t}, \mathbf{d}_{u,t})$ New Item \mathcal{M}_i , e.g. title, brand, category, ...Task Descriptione.g. Given the user's past purchase history [...]how they will rate the new item? [...]After your reasoning, predict a numerical rating.

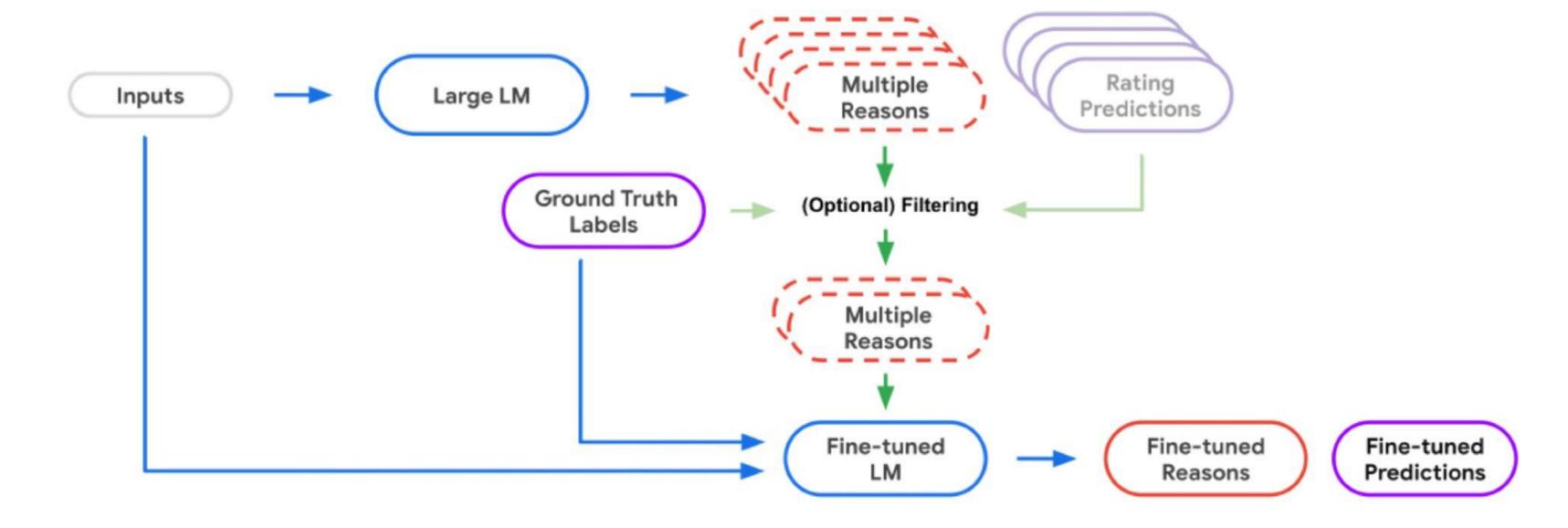
 Tsai, A. Y., Kraft, A., Jin, L., Cai, C., Hosseini, A., Xu, T., ... & Yi, X. (2024). Leveraging LLM Reasoning Enhances Personalized Recommender Systems. arXiv preprint arXiv:2408.00802.





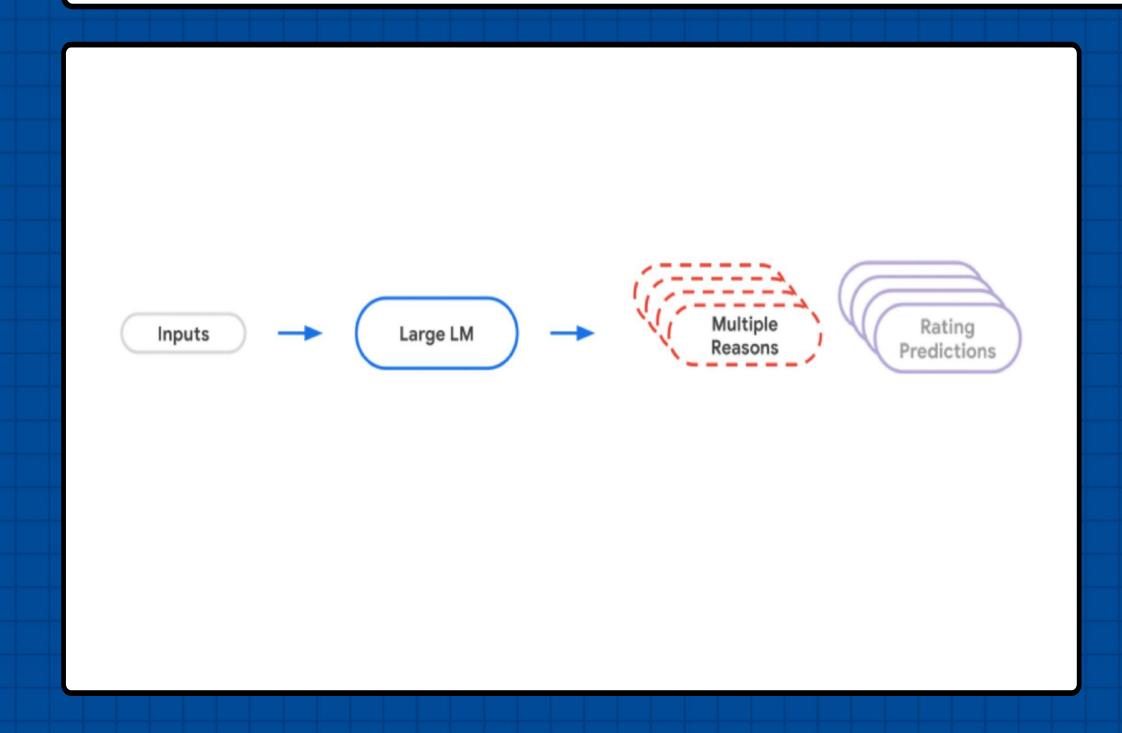


2. Fine-Tunning



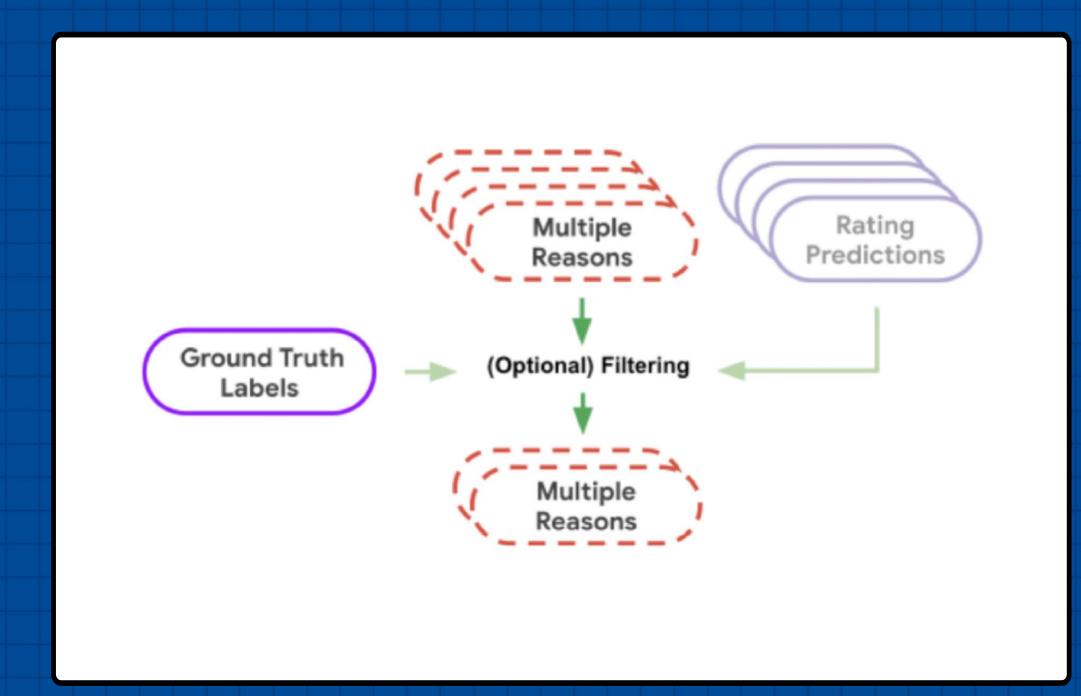
• Tsai, A. Y., Kraft, A., Jin, L., Cai, C., Hosseini, A., Xu, T., ... & Yi, X. (2024). Leveraging LLM Reasoning Enhances Personalized Recommender Systems. *arXiv preprint arXiv:2408.00802*.

<u>Paso 1:</u> fetch



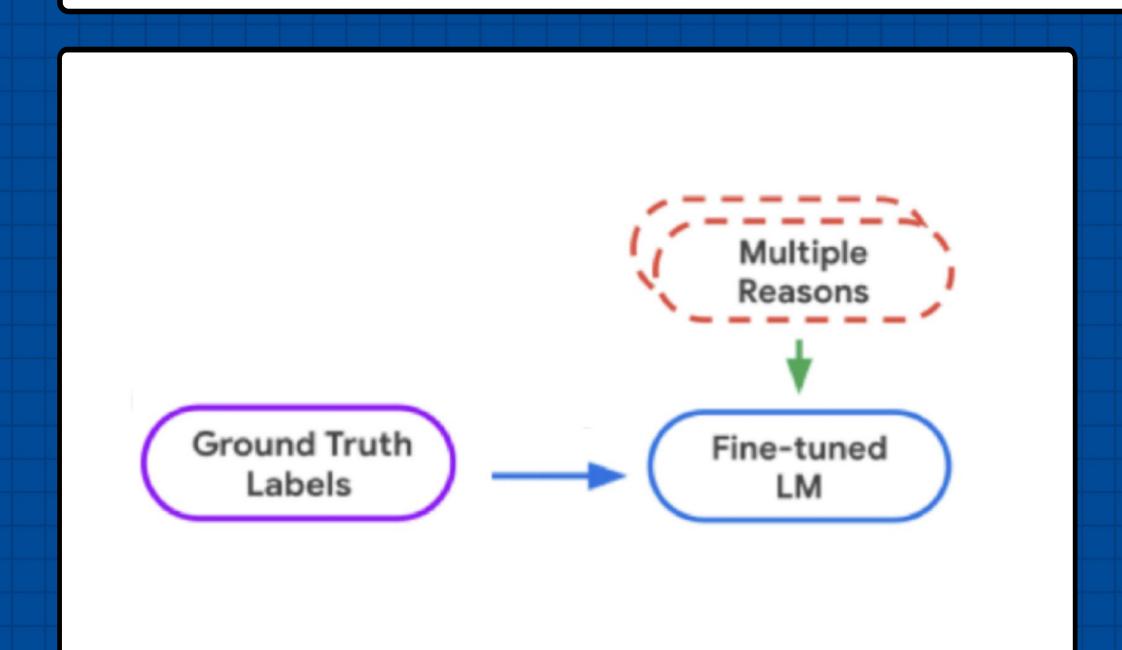
Generan
 razonamientos y
 predicciones
 mediante el mismo
 promp que en
 procedimiento
 anterior al modelo
 FlanT5-XL

<u>Paso 2:</u> filtrado



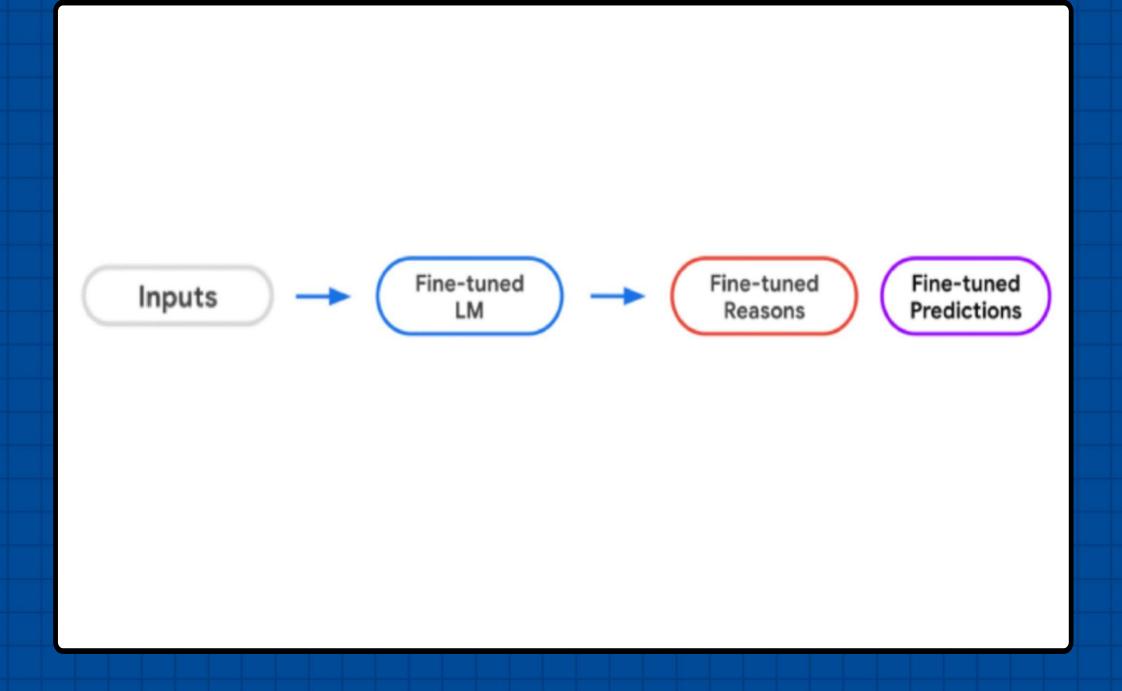
- 5-class: se descarta el razonamiento generado si el rating predicho asociado no es igual al rating real.
- Binary: se distinguen entre ítems 'relevantes' y 'no-relevantes'. Se descarta el razonamiento si la clasificación del ítem no coincide entre el predicho y el real.
- 1-Off: se descartan los razonamientos donde el rating asociado tiene una diferencia mayor a 1, con el rating real.

<u>Paso 3:</u> fine-tunning



- Se alimenta una versión igual o más pequeña del LLM
- Se utilizaron modelos small, base, large y XL.

<u>Paso 4:</u> uso



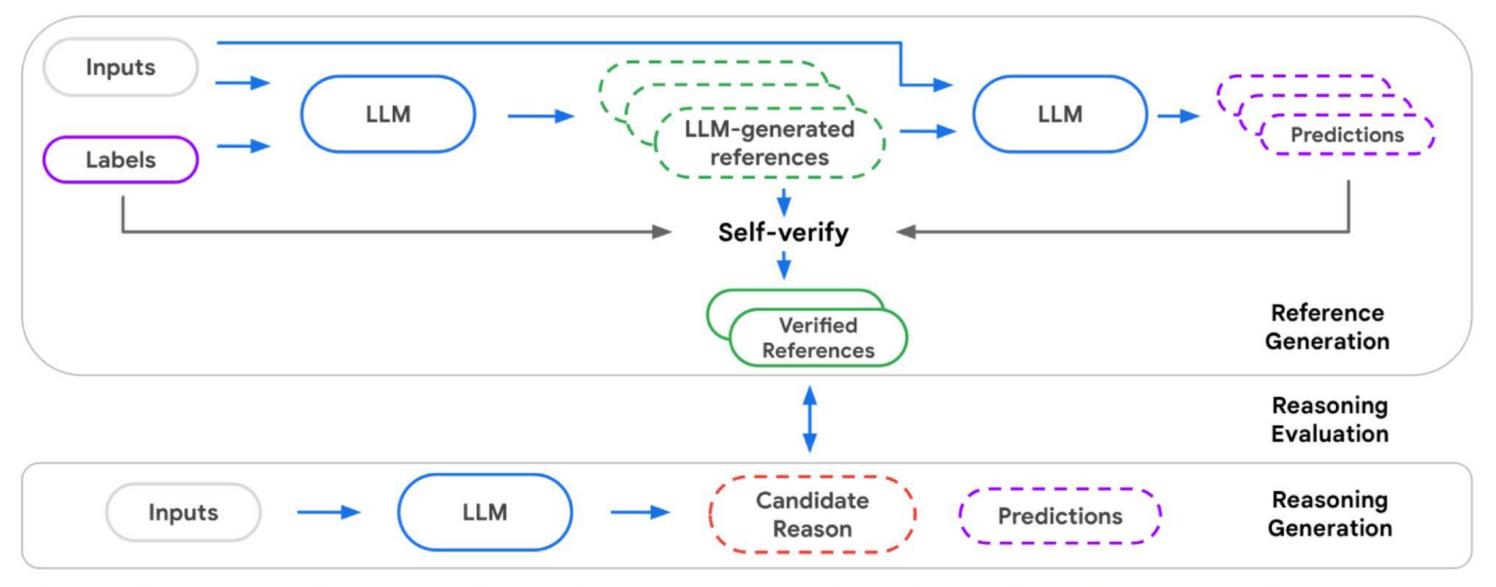
 Terminado el fine-tunning se hacen prompts idénticos al primer procedimiento.





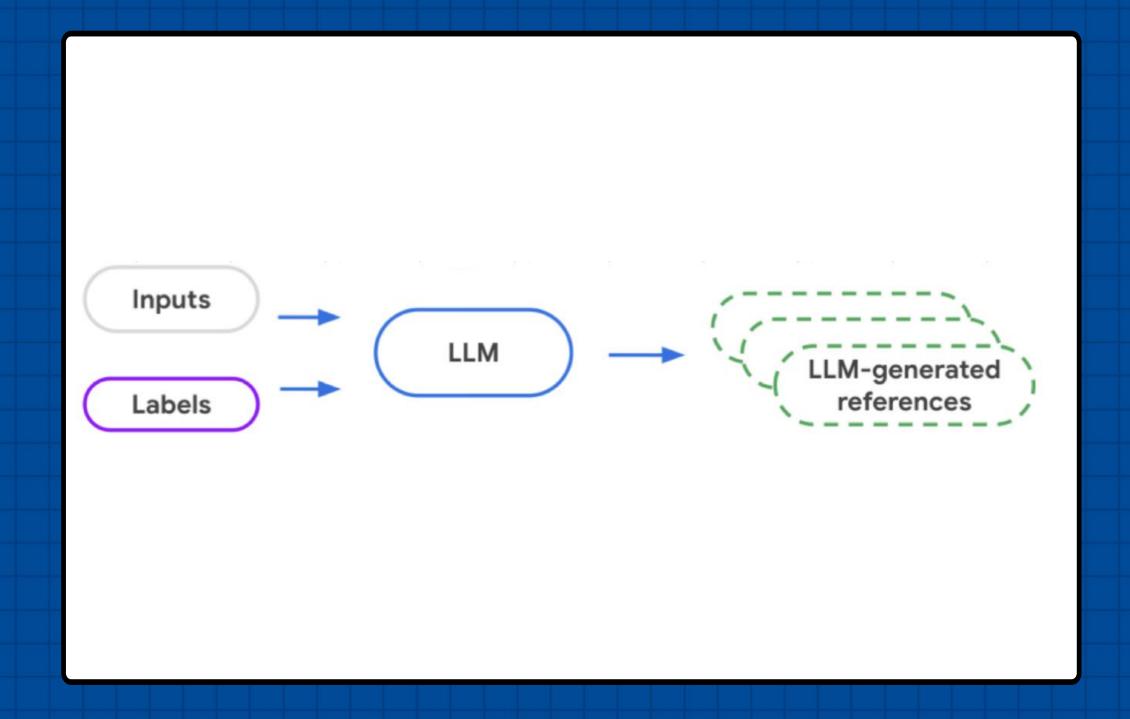


3. Rec-SAVER



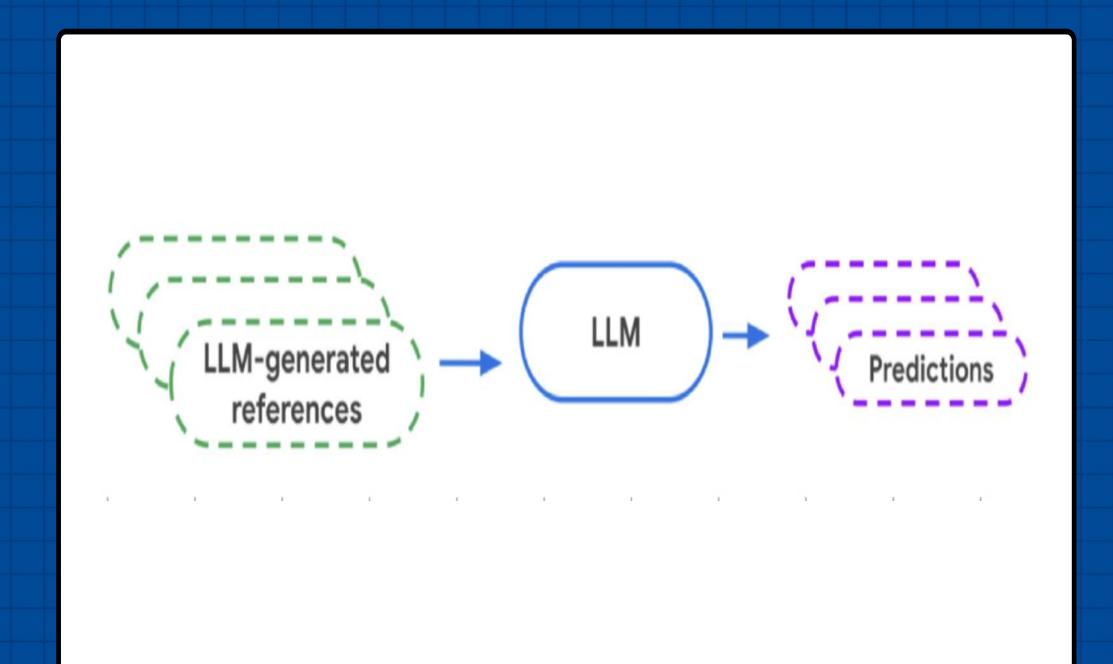
• Tsai, A. Y., Kraft, A., Jin, L., Cai, C., Hosseini, A., Xu, T., ... & Yi, X. (2024). Leveraging LLM Reasoning Enhances Personalized Recommender Systems. *arXiv preprint arXiv:2408.00802*.

<u>Paso 1:</u> fetch



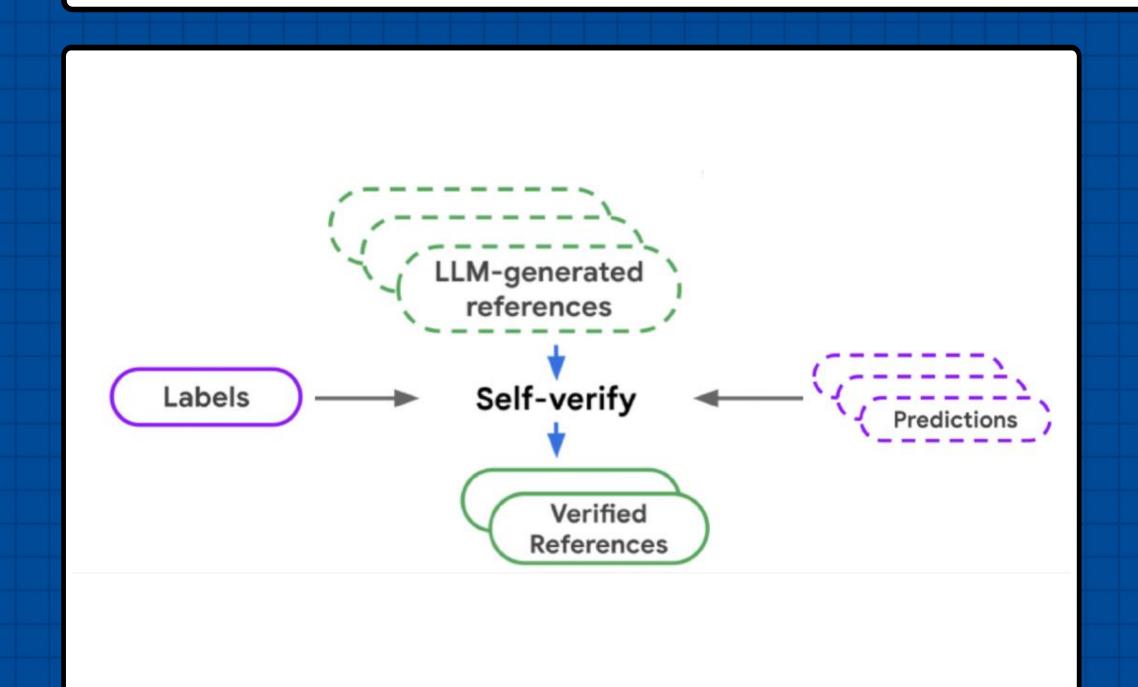
- Generar razonamientos a través de un LLM
- Inputs: metadata del ítem, historial del usuario y rating real.
- Se le pide al modelo entregar un razonamiento del rating entregado

<u>Paso 2:</u> predicción



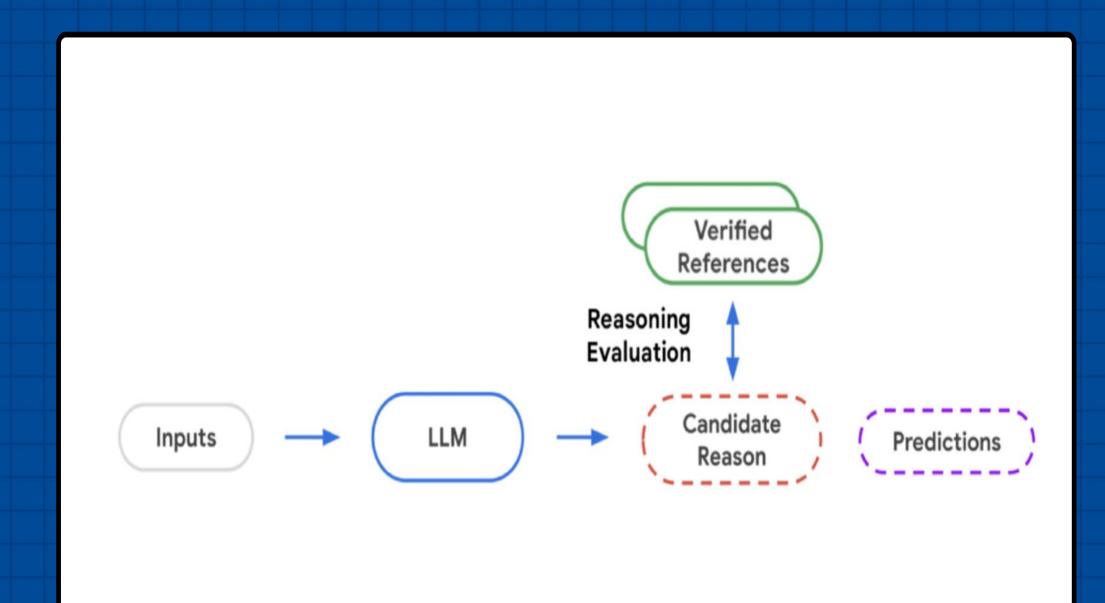
 Predecir rating en base a los razonamientos

<u>Paso 3:</u> auto-verificación



- Se contrastan predicciones con labels reales
- Criterio "5-class"
- Se verifican las referencias

<u>Paso 4:</u> evaluación



- Se evalúa la calidad de razonamiento generado por otro modelo.
- Se utilizan métricas NLG



Verificar calidad de los razonamientos



• Experimento A (Human Judgement Alignment Analysis)



Experimento B (Two-sample T-test)



• Experimento C (Effectivness of self-verification)



• Experimento D (Analysis of Reasoning Quality)

NLG

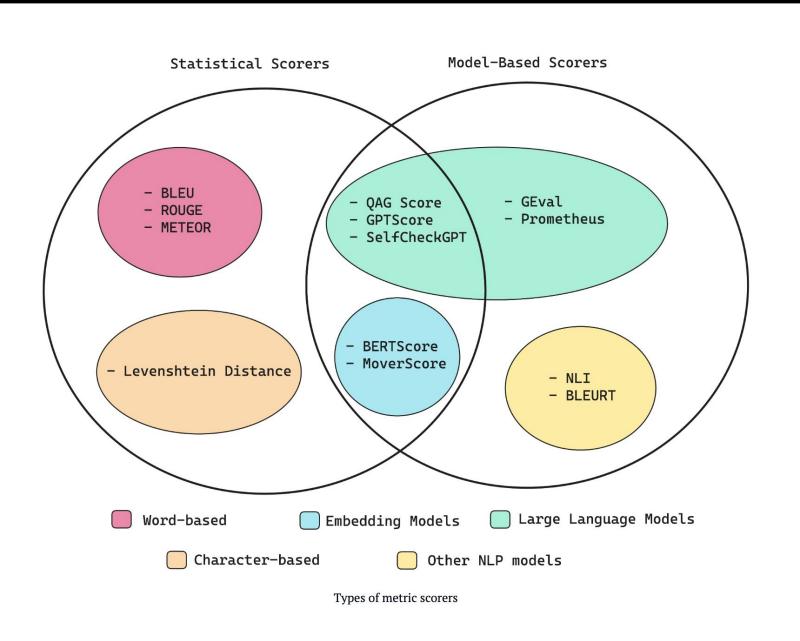
Metrics!

BLEU

ROUGE

METEOR

BERTScore



Confident AI — LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide - Confident AI confident-ai.com

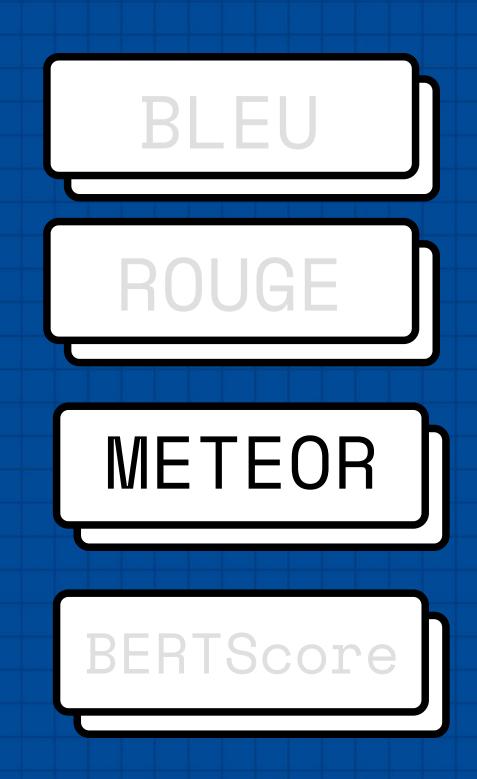
Buscar

Q

 METEOR: Metric for Evaluation of Translation with Explicit ORdering

 Mide la precisión y recall de tokens singulares, y los combina con una media armónica.

 Mide la semántica en el sentido que penaliza si las palabras no se encuentran en el orden correcto.



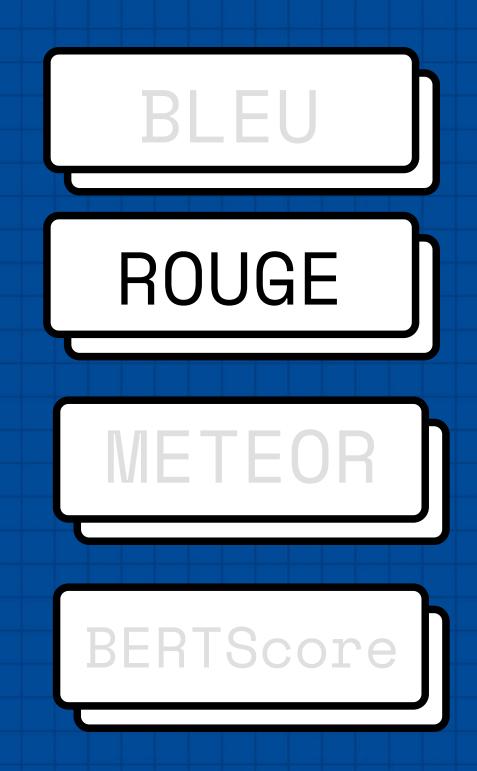
Confident AI — LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide - Confident AI confident-ai.com

Buscar



 ROUGE-1 F1: Recall-Oriented Understudy for Gisting Evaluation

 Mide la precisión y recall de tokens singulares ('-1'), y los combina con una media armónica ('F1').

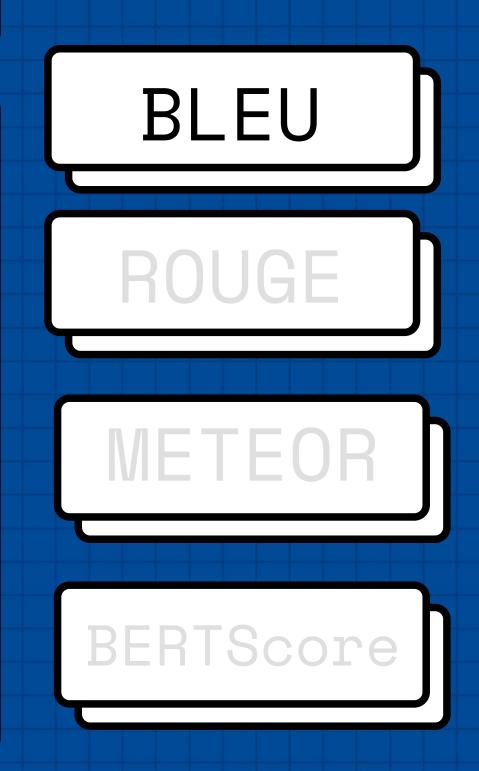


Confident AI — LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide - Confident AI confident-ai.com

Buscar



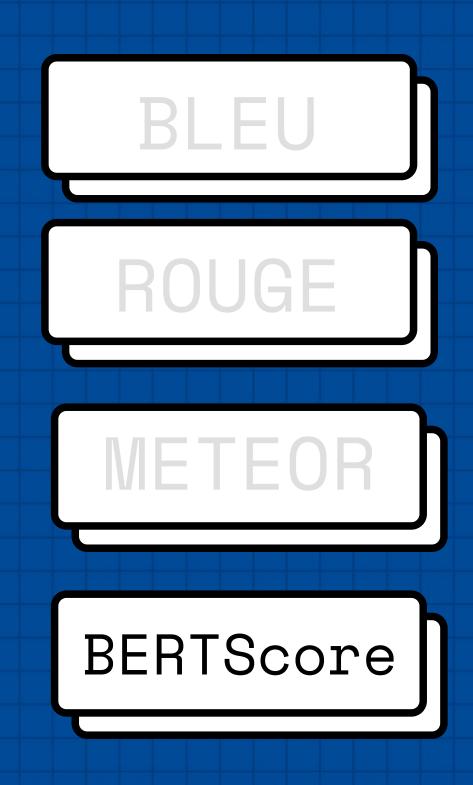
- BLEU: BiLingual Evaluation Understudy
- Mide el n-gram precision, es decir, el solapamiento de frases de largo 'n'.
- Toma todos los i-gram precision (i entre 1 y n) y los combina en media geométrica
- Finalmente pondera por una penalidad si el texto generado es más corto que la referencia



BERTScore

 Mide la similitud coseno entre los embeddings de todas las palabras en cada texto.

 Se calcula usualmente usando BERT, pero puede aplicarse con otros LLM



Confident AI — LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide - Confident Al confident-ai.com

Buscar





Zero-shot Learning Improves with Reasoning

TABLA 3

Comparaciones y ablación con PaLM-2M

	Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE ↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75		=	=	-
Y	Our Method (zero-shot CoT)	0.56	0.62	0.37	1.14	1.60	0.236	0.503	0.339	0.665
UT	- No Reasoning Outputs	0.49	0.57	0.23	1.35	1.70	-	\$ =	2.0%	
BEAUT	- No Review	0.48	0.57	0.21	1.35	1.69	0.237	0.507	0.337	0.667
$\mathbf{B}_{\mathbf{F}}$	- No Review, No Rating	0.43	0.53	0.19	1.42	1.75	0.215	0.494	0.331	0.660
	- No Item Description	0.48	0.57	0.21	1.33	1.66	0.235	0.504	0.340	0.667
	One-shot	0.43	0.57	0.26	1.52	1.97	0.225	0.502	0.335	0.664
	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	(f -	i -	-0	- 9
TV	Our Method (zero-shot CoT)	0.62	0.66	0.40	1.06	1.53	0.194	0.465	0.296	0.647
VIES/T	- No Reasoning Output	0.59	0.63	0.29	1.18	1.56	-	-	-	-
ΛΙΕ	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
Mo	- No Review, No Rating	0.43	0.54	0.20	1.42	1.75	0.150	0.434	0.283	0.633
\geq	- No Item Description	0.54	0.62	0.28	1.22	1.60	0.183	0.460	0.296	0.647
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

• Tsai, A. Y., Kraft, A., Jin, L., Cai, C., Hosseini, A., Xu, T., ... & Yi, X. (2024). Leveraging LLM Reasoning Enhances Personalized Recommender Systems. *arXiv preprint arXiv:2408.00802*.

TABLA 3

Zero-shot CoT

V/S

Sin Razonar

	Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE ↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
BEAUTY	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	n <u>=</u>	=	=	-
	Our Method (zero-shot CoT) - No Reasoning Outputs	0.56 0.49	0.62 0.57	0.37 0.23	1.14 1.35	1.60 1.70	0.236	0.503	0.339	0.665 -
	No ReviewNo Review, No RatingNo Item DescriptionOne-shot	0.48 0.43 0.48 0.43	0.57 0.53 0.57 0.57	0.21 0.19 0.21 0.26	1.33 1.42 1.33 1.52	1.09 1.75 1.66 1.97	0.237 0.215 0.235 0.225	0.507 0.494 0.504 0.502	0.337 0.331 0.340 0.335	0.667 0.667 0.664
2	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	1-	-	- %	- v
S/TV	Our Method (zero-shot CoT) - No Reasoning Output	0.62 0.59	0.66 0.63	0.40 0.29	1.06 1.18	1.53 1.56	0.194	0.465	0.296	0.647
Movie	No ReviewNo Review, No RatingNo Item DescriptionOne-shot	0.58 0.43 0.54 0.47	0.63 0.54 0.62 0.59	0.28 0.20 0.28 0.23	1.20 1.42 1.22 1.32	1.58 1.75 1.60 1.68	0.173 0.150 0.183 0.182	0.452 0.434 0.460 0.452	0.291 0.283 0.296 0.276	0.641 0.633 0.647 0.641

TABLA 3

Zero-shot CoT

V/S

Sin Razonar

	Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE ↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	-	<u>=</u>	=	=
VTII	Our Method (zero-shot CoT) - No Reasoning Outputs	0.56 0.49	0.62 0.57	0.37 0.23	1.14 1.35	1.60 1.70	0.236	0.503	0.339	0.665 -
BEA	No ReviewNo Review, No RatingNo Item DescriptionOne-shot	0.48 0.43 0.48 0.43	0.57 0.53 0.57 0.57	0.21 0.19 0.21 0.26	1.35 1.42 1.33 1.52	1.69 1.75 1.66 1.97	0.237 0.215 0.235 0.225	0.507 0.494 0.504 0.502	0.337 0.331 0.340 0.335	0.667 0.667 0.664
ayee .	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	_	-		2 7
MOVIES/TV	Our Method (zero-shot CoT) - No Reasoning Output	0.62 0.59	0.66 0.63	0.40 0.29	1.06 1.18	1.53 1.56	0.194	0.465	0.296	0.647
	No ReviewNo Review, No RatingNo Item DescriptionOne-shot	0.58 0.43 0.54 0.47	0.63 0.54 0.62 0.59	0.28 0.20 0.28 0.23	1.20 1.42 1.22 1.32	1.58 1.75 1.60 1.68	0.173 0.150 0.183 0.182	0.452 0.434 0.460 0.452	0.291 0.283 0.296 0.276	0.641 0.633 0.647 0.641

El paso de razonamiento mejora el task performance

TABLA 3

Inf. Explícita

V/S

Menos inf.

Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE ↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	n <u>e</u>			-
Our Method (zero-shot CoT)	0.56	0.62	0.37	1.14	1.60	0.236	0.503	0.339	0.665
- No Reasoning Outputs	11 49	05/	11.73	1 45	1 /()	7 <u>-</u>	_	<u> 11</u> 77	
✓ No Peview	0.48	0.57	0.21	1.35	1.69	0.237	0.507	0.337	0.667
- No Review, No Rating	0.43	0.53	0.19	1.42	1.75	0.215	0.494	0.331	0.660
- No Item Description	0.48	0.57	0.21	1.55	1.00	0.233	0.304	0.340	U.007
One-shot	0.43	0.57	0.26	1.52	1.97	0.225	0.502	0.335	0.664
Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	(I -	:-	:	- 9
Our Method (zero-shot CoT)	0.62	0.66	0.40	1.06	1.53	0.194	0.465	0.296	0.647
- No Reasoning Clutput	11.59	11.63	11 79	IIX	1.36	_	-	_	_
- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
- No Review - No Review, No Rating	0.43	0.54	0.20	1.42	1.75	0.150	0.434	0.283	0.633
- No Item Description	0.54	0.62	0.28	1.22	1.60	0.183	0.460	0.296	U.04/
One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

Inf. Explícita

V/S

Menos inf.

TABLA 3

No Review

 \approx

User-item Matrix

No Review, No Rating

 \approx

Implicit feedback

	Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	-	E	8	=
7	Our Method (zero-shot CoT)	0.56	0.62	0.37	1.14	1.60	0.236	0.503	0.339	0.665
BEAU	- No Reasoning Chilpins - No Review - No Review, No Rating	0.48 0.43	0.57 0.53	0.21 0.19	1.35 1.42	1.69 1.75	0.237 0.215	0.507 0.494	0.337 0.331	0.667 0.660
	- No Item Description One-shot	0.48 0.43	0.57 0.57	0.21 0.26	1.33 1.52	1.66 1.97	0.235 0.225	0.504 0.502	0.340 0.335	0.66 7 0.664
	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	(I -	; –	-:	₩V
S/TV	Our Method (zero-shot CoT)	0.62	0.66	0.40	1.06	1.53	0.194	0.465	0.296	0.647
Movies	- No Review - No Review, No Rating	0.58 0.43	0.63 0.54	0.28 0.20	1.20 1.42	1.58 1.75	0.173 0.150	0.452 0.434	0.291 0.283	0.641 0.633
Σ	- No Item Description One-shot	0.54 0.47	0.62 0.59	0.28 0.23	1.22 1.32	1.60 1.68	0.183 0.182	0.460 0.452	0.296 0.276	0.647 0.641

Conocimiento Previo

	Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE ↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	T-	Ξ	=	-
BEAUTY	Our Method (zero-shot CoT) - No Reasoning Outputs - No Review No Rating	0.56 0.49 0.48 0.43	0.62 0.57 0.57 0.53	0.37 0.23 0.21 0.19	1.14 1.35 1.35 1.42	1.60 1.70 1.69	0.236 0.237 0.215	0.503 - 0.507 0.494	0.339 - 0.337 0.331	0.665 0.667 0.660
	- No Item Description One-snot	0.48	0.57	0.21 0.26	1.33 1.52	1.66 1.97	0.235	0.504 0.502	0.340	0.667
114-11-0	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	0=	:=		- %
OVIES/TV	Our Method (zero-shot CoT) - No Reasoning Output - No Review - No Review No Rating	0.62 0.59 0.58 0.43	0.66 0.63 0.63 0.54	0.40 0.29 0.28 0.20	1.06 1.18 1.20 1.42	1.53 1.56 1.58 1.75	0.194 - 0.173 0.150	0.465 0.452 0.434	0.296 0.291 0.283	0.647 0.641 0.633
\mathbf{Z}	- No Item Description One-snot	0.54 0.47	0.62 0.59	0.28	1.22 1.32	1.60 1.08	0.183 0.182	0.460 0.432	0.296	0.647

Conocimiento Previo

	Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE ↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
8. 1	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	-	Œ	=	=
BEAUTY	Our Method (zero-shot CoT) - No Reasoning Outputs - No Review - No Review No Rating	0.56 0.49 0.48 0.43	0.62 0.57 0.57	0.37 0.23 0.21 0.19	1.14 1.35 1.35 1.42	1.60 1.70 1.69 1.75	0.236 0.237 0.215	0.503 - 0.507 0.494	0.339 0.337 0.331	0.665 0.667 0.660
	- No Item Description	0.48	0.57	0.21	1.33	1.66	0.235	0.504	0.340	0.667
	One-snot	0.43	0.57	0.26	1.52	1.97	0.225	0.502	0.555	0.004
_	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	2#	-:	H 2
OVIES/TV	Our Method (zero-shot CoT) - No Reasoning Output - No Review - No Review No Rating	0.62 0.59 0.58 0.43	0.66 0.63 0.63 0.54	0.40 0.29 0.28 0.20	1.06 1.18 1.20 1.42	1.53 1.56 1.58 1.75	0.194 - 0.173 0.150	0.465 0.452 0.434	0.296 0.291 0.283	0.647 0.641 0.633
Σ	- No Item Description	0.54	0.62	0.28	1.22	1.60	0.183	0.460	0.296	0.647
08	One-snot	0.47	0.39	0.23	1.32	1.08	0.182	0.432	0.270	0.041



El manejo del contexto de PaLM 2-M Es Mejor en MOVIES/TV

TABLA 3

Zero-shot CoT

V/S

One-shot CoT

	Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE ↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
á .	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	n <u>=</u>	-	=3	=
7	Our Method (zero-shot CoT)	0.56	0.62	0.37	1.14	1.60	0.236	0.503	0.339	0.665
BEAUTY	 No Reasoning Outputs No Review No Review, No Rating No Item Description 	0.49 0.48 0.43 0.48	0.57 0.57 0.53 0.57	0.23 0.21 0.19 0.21	1.35 1.35 1.42 1.33	1.70 1.69 1.75 1.66	0.237 0.215 0.235	0.507 0.494 0.504	0.337 0.331 0.340	- 0.667 0.660 0.667
()	One-shot	0.43	0.57	0.26	1.52	1.97	0.225	0.502	0.335	0.664
1	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	::-	:-	- 3	= 00
/T/	Our Method (zero-shot CoT)	0.62	0.66	0.40	1.06	1.53	0.194	0.465	0.296	0.647
Movies,	 No Reasoning Output No Review No Review, No Rating No Item Description 	0.59 0.58 0.43 0.54	0.63 0.63 0.54 0.62	0.29 0.28 0.20 0.28	1.18 1.20 1.42 1.22	1.56 1.58 1.75 1.60	0.173 0.150 0.183	0.452 0.434 0.460	0.291 0.283 0.296	0.641 0.633 0.647
58	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

Zero-shot CoT

V/S

One-shot CoT

	Method	Binary Acc.	Binary F1	Multi. Acc.	Multi. MAE↓	Multi. RMSE↓	ROUGE-1 F1	METEOR	BLEU	BERT Score
	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	T <u>u</u>	<u> </u>	=	-
×	Our Method (zero-shot CoT)	0.56	0.62	0.37	1.14	1.60	0.236	0.503	0.339	0.665
BEA	No Review No Review, No Rating No Item Description	0.49 0.48 0.43 0.48	0.57 0.57 0.53 0.57	0.23 0.21 0.19 0.21	1.35 1.35 1.42 1.33	1.70 1.69 1.75 1.66	0.237 0.215 0.235	0.507 0.494 0.504	0.337 0.331 0.340	0.667 0.660 0.667
	One-shot	0.43	0.57	0.26	1.52	1.97	0.225	0.502	0.335	0.664
_]	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	1-	; -	-:	- 9
	Our Method (zero-shot CoT)	0.62	0.66	0.40	1.06	1.53	0.194	0.465	0.296	0.647
- X	No Review No Review, No Rating No Item Description	0.59 0.58 0.43 0.54	0.63 0.63 0.54 0.62	0.29 0.28 0.20 0.28	1.18 1.20 1.42 1.22	1.50 1.58 1.75 1.60	0.173 0.150 0.183	0.452 0.434 0.460	0.291 0.283 0.296	0.641 0.633 0.647
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

TABLA 3

One-shot es el uso de un ejemplo previo en el prompt

Dificulta el desglose del prompt, por ende baja el rendimiento



Fine-tuning results

TABLA 4

Fine-tuning con FlanT5

	Model Size	Reas- oning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
	Small	1	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
TY	Base	1	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	0.344	0.667
٩U	Large	1	0.64	0.59	0.67	0.33	0.65	1.26	1.73	0.240	0.506	0.343	0.666
\mathbf{BE}_{L}	XL	1	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
щ	XL	X	0.55	0.61	0.74	0.28	0.67	1.31	1.75	-	-	-	-
	XL (no fine-tuning)	X	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	18 5	=#
>	Small	1	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
Ţ	Base	1	0.65	0.59	0.72	0.34	0.68	1.18	1.65	0.153	0.438	0.279	0.634
Œ	Large	1	0.64	0.58	0.72	0.32	0.67	1.23	1.70	0.165	0.448	0.286	0.639
\leq	XL	/	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
Movies/T	XL	X	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-
	XL (no fine-tuning)	Х	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-		(_

TABLA 4

Fine-tuning con FlanT5

	Model Size	Reas- oning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
12	Small	/	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
TY	Base	1	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	0.344	0.667
AUTY	Large	1	0.64	0.59	0.67	0.33	0.65	1.26	1.73	0.240	0.506	0.343	0.666
BE	XL	✓	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
Н	XL	Х	0.55	0.61	0.74	0.28	0.67	1.31	1.75	-	-	=	-
	XL (no fine-tuning)	X	0.55	0.40	0.56	0.22	0.56	1.64	2.09	·-	5 .	1.55	=0
>	Small	1	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
Ĭ	Base	/	0.65	0.59	0.72	0.34	0.68	1.18	1.65	0.153	0.438	0.279	0.634
ΙΈς	Large	/	0.64	0.58	0.72	0.32	0.67	1.23	1.70	0.165	0.448	0.286	0.639
OVIES/TV	XL	/	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
\breve{Z}	XL	X	0.62	0.57	0.70	0.33	0.66	1.27	1.75	=	-	æ	-
	XL (no fine-tuning)	Х	0.61	0.43	0.61	0.23	0.61	1.56	2.00	138	-1	-	-

TABLA 4

Fine-tuning con FlanT5

	Model Size	Reas- oning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
	Small	1	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
Τ	Base	/	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	0.344	0.667
AUTY	Large	/	0.64	0.59	0.67	0.33	0.65	1.26	1.73	0.240	0.506	0.343	0.666
BE/	XL	/	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
ш	XL	Х	0.55	0.61	0.74	0.28	0.67	1.31	1.75	in .	=	Æ	-
	XL (no fine-tuning)	Х	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-		V-55	=0
>	Small	1	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
Ţ	Base	/	0.65	0.59	0.72	0.34	0.68	1.18	1.65	0.153	0.438	0.279	0.634
ES	Large	/	0.64	0.58	0.72	0.32	0.67	1.23	1.70	0.165	0.448	0.286	0.639
OVIES/TV	XL	/	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
MC	XL	Х	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-7
	XL (no fine-tuning)	Х	0.61	0.43	0.61	0.23	0.61	1.56	2.00	læ		1-	-

Modelo más grande = Mejor predicción

Puede deberse a tener una mayor capacidad de albergar conocimiento

TABLA 4

Fine-tuning con FlanT5

	Model Size	Reas- oning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
	Small	1	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
ΤY	Base	1	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	0.344	0.667
AU	Large	1	0.64	0.59	0.67	0.33	0.65	1.26	1.73	0.240	0.506	0.343	0.666
\mathbf{BE}_{L}	XL	✓	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
-	XL	Х	0.55	0.61	0.74	0.28	0.67	1.31	1.75	-	=	Œ	-
	XL (no fine-tuning)	Х	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	Œ	= 0
>	Small	1	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
TES/T	Base	1	0.65	0.59	0.72	0.34	0.68	1.18	1.65	0.153	0.438	0.279	0.634
Œ	Large	1	0.64	0.58	0.72	0.32	0.67	1.23	1.70	0.165	0.448	0.286	0.639
0.	XL	1	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
Ĭ	XL	Х	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	_	-	_
	XL (no fine-tuning)	Х	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-	=	_	_

Fine-tuning con FlanT5

	Model Size	Reas- oning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
	Small	1	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
TY	Base	1	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	0.344	0.667
AUT	Large	1	0.64	0.59	0.67	0.33	0.65	1.26	1.73	0.240	0.506	0.343	0.666
BE.	XL	1	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
П	XL	X	0.55	0.61	0.74	0.28	0.67	1.31	1.75	-	-	=	-
	XL (no fine-tuning)	Х	0.55	0.40	0.56	0.22	0.56	1.64	2.09		-	i a	-1
>	Small	√	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
OVIES/T	Base	1	0.65	0.59	0.72	0.34	0.68	1.18	1.65	0.153	0.438	0.279	0.634
IES	Large	1	0.64	0.58	0.72	0.32	0.67	1.23	1.70	0.165	0.448	0.286	0.639
>	XL	/	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
Ĭ	XL	Х	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	14	21
	XL (no fine-tuning)		0.61	0.43	0.61	0.23	0.61	1.56	2.00				-77

TABLA 4



No se pudieron
extraer
razonamientos con el
Baseline!!

TABLA 4

Fine-tuning con FlanT5

	Model Size	Reas- oning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
	Small	1	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
${ m T}{ m Y}$	Base	1	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	0.344	0.667
AUT	Large	1	0.64	0.59	0.67	0.33	0.65	1.26	1.73	0.240	0.506	0.343	0.666
BE.	XL	1	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.66
	XL	X	0.55	0.61	0.74	0.28	0.67	1.31	1.75	-		-	-
	XL (no fine-tuning)	Х	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	-	
>	Small	√	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
TES/T	Base	1	0.65	0.59	0.72	0.34	0.68	1.18	1.65	0.153	0.438	0.279	0.634
ES	Large	1	0.64	0.58	0.72	0.32	0.67	1.23	1.70	0.165	0.448	0.286	0.639
>	XL	1	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
MO	XL	X	0.62	0.57	0.70	0.33	0.66	1.27	1.75	=	-	Œ	-
	XL (no fine-tuning)	X	0.61	0.43	0.61	0.23	0.61	1.56	2.00	_	=	=	_

TABLA 4

Fine-tuning con FlanT5

-	Model Size	Reas- oning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE ↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
	Small	/	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
UTY	Base	1	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	0.344	0.667
ΑU	Large	1	0.64	0.59	0.67	0.33	0.65	1.26	1.73	0.240	0.506	0.343	0.666
BE,	XI.	1	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
1	XL	X	0.55	0.61	0.74	0.28	0.67	1.31	1.75	-	=	I	-
250	XL (no fine-tuning)	X	0.55	0.40	0.56	0.22	0.56	1.64	2.09	1 	-	-	-
\ \ \.	Small	1	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
VIES/T	Base	1	0.65	0.59	0.72	0.34	0.68	1.18	1.65	0.153	0.438	0.279	0.634
Œ	Large	1	0.64	0.58	0.72	0.32	0.67	1.23	1.70	0.165	0.448	0.286	0.639
>	XI	J	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0 449	0.286	0.643
X	XL	Х	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	Œ	=7
95	XL (no fine-tuning)	X	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-	-	-	-

Por comparación se utiliza T5-XL sin CoT

Fine-tuning mejora el modelo incluso si es grande.

TABLA 5

Uso de filtros en el proceso de Fine-tuning

Table 5: Comparison of fine-tuning Flan-T5 XL model with multiple reasoning paths per user-item pair and with different filtering methods. PaLM 2-M zero-shot (no fine-tuning) results are included for comparison.

	Samples	Filter	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
20	1	None	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
X	8	None	0.68	0.64	0.79	0.31	0.70	1.25	1.71	0.248	0.509	0.333	0.671
UT	8	5-class	0.54	0.61	0.63	0.28	0.60	1.32	1.74	0.248	0.510	0.329	0.670
BEAUTY	8	Binary	0.53	0.59	0.64	0.29	0.60	1.40	1.88	0.246	0.508	0.335	0.669
B	8	1-off	0.61	0.61	0.71	0.30	0.63	1.28	1.75	0.247	0.336	0.510	0.671
	PaLM 2-M	I Zero-shot	0.56	0.62	(#	0.37		1.14	1.60	0.236	0.503	0.339	0.665
_	1	None	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
TV	8	None	0.63	0.58	0.72	0.35	0.67	1.23	1.75	0.171	0.446	0.285	0.642
SS/	8	5-class	0.59	0.63	0.69	0.32	0.63	1.17	1.61	0.176	0.449	0.291	0.642
VIE	8	Binary	0.60	0.64	0.71	0.33	0.66	1.28	1.78	0.175	0.443	0.288	0.641
Movies/T	8	1-off	0.62	0.63	0.74	0.36	0.67	1.16	1.64	0.180	0.451	0.291	0.643
	PaLM 2-M	I Zero-shot	0.62	0.66	8 <u></u> -	0.40	œ.	1.06	1.53	0.194	0.465	0.296	0.647

TABLA 5

Uso de filtros en el proceso de Fine-tuning

Table 5: Comparison of fine-tuning Flan-T5 XL model with multiple reasoning paths per user-item pair and with different filtering methods. PaLM 2-M zero-shot (no fine-tuning) results are included for comparison.

	Samples	Filter	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
20	1	None	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
BEAUTY	8 8 8 8	None 5-class Binary 1-off	0.68 0.54 0.53 0.61	0.64 0.61 0.59 0.61	0.79 0.63 0.64 0.71	0.31 0.28 0.29 0.30	0.70 0.60 0.60 0.63	1.25 1.32 1.40 1.28	1.71 1.74 1.88 1.75	0.248 0.248 0.246 0.247	0.509 0.510 0.508 0.336	0.333 0.329 0.335 0.510	0.671 0.670 0.669 0.671
	PaLM 2-I	M Zero-shot	0.56	0.62	-	0.37		1.14	1.60	0.236	0.503	0.339	0.665
_	1	None	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
MOVIES/TV	8 8 8	None 5-class Binary 1-off	0.63 0.59 0.60 0.62	0.58 0.63 0.64 0.63	0.72 0.69 0.71 0.74	0.35 0.32 0.33 0.36	0.67 0.63 0.66 0.67	1.23 1.17 1.28 1.16	1.75 1.61 1.78 1.64	0.171 0.176 0.175 0.180	0.446 0.449 0.443 0.451	0.285 0.291 0.288 0.291	0.642 0.642 0.641 0.643
_	PaLM 2-N	M Zero-shot	0.62	0.66		0.40	-	1.06	1.53	0.194	0.465	0.296	0.647

Uso de filtros en el proceso de Fine-tuning

Table 5: Comparison of fine-tuning Flan-T5 XL model with multiple reasoning paths per user-item pair and with different filtering methods. PaLM 2-M zero-shot (no fine-tuning) results are included for comparison.

	Samples	Filter	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
₹ŧ	1	None	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
BEAUTY	8 8 8 8	None 5-class Binary 1-off	0.68 0.54 0.53 0.61	0.64 0.61 0.59 0.61	0.79 0.63 0.64 0.71	0.31 0.28 0.29 0.30	0.70 0.60 0.60 0.63	1.25 1.32 1.40 1.28	1.71 1.74 1.88 1.75	0.248 0.248 0.246 0.247	0.509 0.510 0.508 0.336	0.333 0.329 0.335 0.510	0.671 0.670 0.669 0.671
	PaLM 2-M	I Zero-shot	0.56	0.62	-	0.37	=	1.14	1.60	0.236	0.503	0.339	0.665
_	_1	None	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
Movies/TV	8 8 8 8	None 5-class Binary 1-off	0.63 0.59 0.60 0.62	0.58 0.63 0.64 0.63	0.72 0.69 0.71 0.74	0.35 0.32 0.33 0.36	0.67 0.63 0.66 0.67	1.23 1.17 1.28 1.16	1.75 1.61 1.78 1.64	0.171 0.176 0.175 0.180	0.446 0.449 0.443 0.451	0.285 0.291 0.288 0.291	0.642 0.642 0.641 0.643
4	PaLM 2-M	I Zero-shot	0.62	0.66	-	0.40	_	1.06	1.53	0.194	0.465	0.296	0.647

TABLA 5

En general, Aplicar filtros reduce el rendimiento

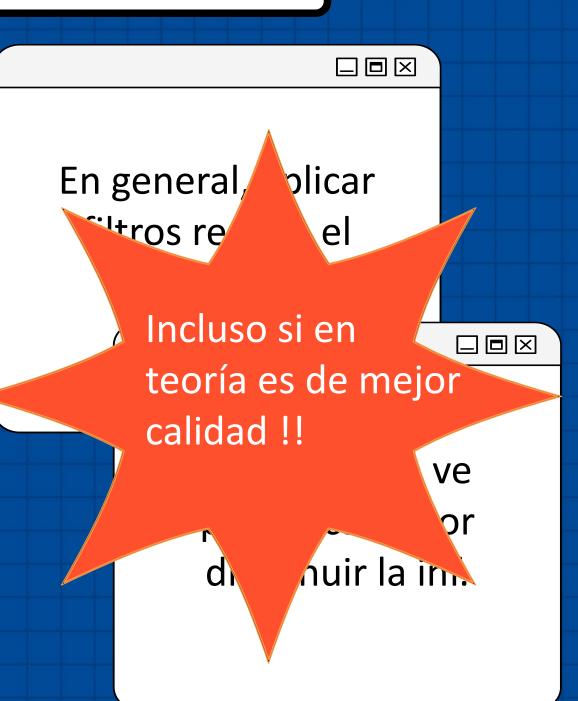
El modelo se ve perjudicado por disminuir la inf.

Uso de filtros en el proceso de Fine-tuning

Table 5: Comparison of fine-tuning Flan-T5 XL model with multiple reasoning paths per user-item pair and with different filtering methods. PaLM 2-M zero-shot (no fine-tuning) results are included for comparison.

	Samples	Filter	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
!!!	1	None	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
BEAUTY	8 8 8 8	None 5-class Binary 1-off	0.68 0.54 0.53 0.61	0.64 0.61 0.59 0.61	0.79 0.63 0.64 0.71	0.31 0.28 0.29 0.30	0.70 0.60 0.60 0.63	1.25 1.32 1.40 1.28	1.71 1.74 1.88 1.75	0.248 0.248 0.246 0.247	0.509 0.510 0.508 0.336	0.333 0.329 0.335 0.510	0.671 0.670 0.669 0.671
	PaLM 2-M	I Zero-shot	0.56	0.62	-	0.37	=	1.14	1.60	0.236	0.503	0.339	0.665
_	1	None	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
Movies/TV	8 8 8	None 5-class Binary 1-off	0.63 0.59 0.60 0.62	0.58 0.63 0.64 0.63	0.72 0.69 0.71 0.74	0.35 0.32 0.33 0.36	0.67 0.63 0.66 0.67	1.23 1.17 1.28 1.16	1.75 1.61 1.78 1.64	0.171 0.176 0.175 0.180	0.446 0.449 0.443 0.451	0.285 0.291 0.288 0.291	0.642 0.642 0.641 0.643
~	PaLM 2-M	I Zero-shot	0.62	0.66	1 -	0.40	-	1.06	1.53	0.194	0.465	0.296	0.647

TABLA 5



Uso de filtros en el proceso de Fine-tuning

Table 5: Comparison of fine-tuning Flan-T5 XL model with multiple reasoning paths per user-item pair and with different filtering methods. PaLM 2-M zero-shot (no fine-tuning) results are included for comparison.

	Samples	Filter	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE↓	Multi. RMSE↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
×(1	None	0.67	0.61	0.78	0.30	0.69	1.24	1.68	0.241	0.510	0.339	0.667
X	8	None	0.68	0.64	0.79	0.31	0.70	1.25	1.71	0.248	0.509	0.333	0.671
UTY	8	5-class	0.54	0.61	0.63	0.28	0.60	1.32	1.74	0.248	0.510	0.329	0.670
⋖	8	Binary	0.53	0.59	0.64	0.29	0.60	1.40	1.88	0.246	0.508	0.335	0.669
$\mathbf{B}\mathbf{E}$	8	1-off	0.61	0.61	0.71	0.30	0.63	1.28	1.75	0.247	0.336	0.510	0.671
	PaLM 2-M	I Zero-shot	0.56	0.62	(#	0.37	1911 1911	1.14	1.60	0.236	0.503	0.339	0.665
_	1	None	0.65	0.61	0.72	0.34	0.67	1.17	1.64	0.165	0.449	0.286	0.643
	8	None	0.63	0.58	0.72	0.35	0.67	1.23	1.75	0.171	0.446	0.285	0.642
/S?	8	5-class	0.59	0.63	0.69	0.32	0.63	1.17	1.61	0.176	0.449	0.291	0.642
ΛIE	8	Binary	0.60	0.64	0.71	0.33	0.66	1.28	1.78	0.175	0.443	0.288	0.641
MOVIES/TV	8	1-off	0.62	0.63	0.74	0.36	0.67	1.16	1.64	0.180	0.451	0.291	0.643
~	PaLM 2-M	I Zero-shot	0.62	0.66	s -	0.40	(-)	1.06	1.53	0.194	0.465	0.296	0.647

TABLA 5



En 'Movies' funciona mejor '1-off'



Puede ser por el mayor conocimiento del contexto por parte del modelo



Human Alignment Analysis

TABLA 6 & 7

Evaluación humana de RecSAVER

Table 6: Inter-annotator agreement (IAA) analysis on the human annotated scores.

	Mean	Cohen κ	Avg. ρ	p-value
Coherence	3.72	0.37	0.37	1e-10
Faithfulness	0.63	0.63	0.63	1e-12
Insightfulness	2.80	0.33	0.34	6e-4

Table 7: Correlation between coherence, insightfulness, and automatic NLG metrics. The annotated scores are averaged across the annotators for each sample.

	Coherence	Insightfulness
BLEU	0.36	0.02
ROUGE-1 F1	0.40	0.10
METEOR	0.40	0.25
BERTScore	0.36	0.20

TABLA 6 & 7

Evaluación humana de RecSAVER

Table 6: Inter-annotator agreement (IAA) analysis on the human annotated scores.

	Mean	Cohen κ	Avg. ρ	p-value
Coherence	3.72	0.37	0.37	1e-10
Faithfulness	0.63	0.63	0.63	1e-12
Insightfulness	2.80	0.33	0.34	6e-4

Table 7: Correlation between coherence, insightfulness, and automatic NLG metrics. The annotated scores are averaged across the annotators for each sample.

	Coherence	Insightfulness
BLEU	0.36	0.02
ROUGE-1 F1	0.40	0.10
METEOR	0.40	0.25
BERTScore	0.36	0.20

Coherence tiene buena correlación con NLG

Insightfulness, no tanto ...

TABLA 6 & 7

Evaluación humana de RecSAVER

Table 6: Inter-annotator agreement (IAA) analysis on the human annotated scores.

Mean Cohen κ Avg. ρ p-value 0.37 Coherence 3.72 0.37 1e-10 Faithfulness 0.63 0.63 0.63 1e-12 2.80 0.330.34Insightfulness 6e-4

Table 7: Corrautomatic NL across the ann

BLE ROU MET BER' Faithfulness no se muestra porque las métricas de NLG miden overlap, entonces no pueden detectar veracidad.

ce tiene lación con

ightfulness, no tanto ...



Two-sample T-test

TABLA 8

Evaluación humana de RecSAVER

Table 8: Two-sample t-test comparing the average of human annotated scores and NLG scores between faithful and unfaithful reasoning.

	Faithful	Unfailthful	p-value
Coherence	4.01	3.22	2e-8
Insightfulness	3.11	2.23	6e-9
BLEU	0.21	0.16	2e-3
ROUGE-1 F1	0.49	0.46	5e-3
METEOR	0.31	0.30	0.36
BERTScore	0.65	0.63	0.02

TABLA 8

Evaluación humana de RecSAVER

Table 8: Two-sample t-test comparing the average of human annotated scores and NLG scores between faithful and unfaithful reasoning.

	Faithful	Unfailthful	p-value
Coherence	4.01	3.22	2e-8
Insightfulness	3.11	2.23	6e-9
BLEU	0.21	0.16	2e-3
ROUGE-1 F1	0.49	0.46	5e-3
METEOR	0.31	0.30	0.36
BERTScore	0.65	0.63	0.02

La presencia de errores impacta la percepción humana

TABLA 8

Evaluación humana de RecSAVER

Table 8: Two-sample t-test comparing the average of human annotated scores and NLG scores between faithful and unfaithful reasoning.

	Faithful	Unfailthful	p-value
Coherence	4.01	3.22	2e-8
Insightfulness	3.11	2.23	6e-9
BLEU	0.21	0.16	2e-3
ROUGE-1 F1	0.49	0.46	5e-3
METEOR	0.31	0.30	0.36
BERTScore	0.65	0.63	0.02

La presencia de errores impacta la percepción humana

También disminuyen las métricas NLG

TABLA 8

Evaluación humana de RecSAVER

Table 8: Two-sample t-test comparing the average of human annotated scores and NLG scores between faithful and unfaithful reasoning.

	Faithful	Unfailthful	p-value
Coherence	4.01	3.22	2e-8
Insightfulness	3.11	2.23	6e-9
BLEU	0.21	0.16	2e-3
ROUGE-1 F1	0.49	0.46	5e-3
METEOR	0.31	0.30	0.36
BERTScore	0.65	0.63	0.02





Effectivness of self-verification

TABLA 9

Evaluación humana de RecSAVER

Table 9: Correlation between coherence and NLG metrics with and without using self-verified references.

Self-verification	Yes	No
BLEU	0.36	0.33
ROUGE-1 F1	0.40	0.35
METEOR	0.40	0.37
BERTScore	0.36	0.28

TABLA 9

Evaluación humana de RecSAVER

Table 9: Correlation between coherence and NLG metrics with and without using self-verified references.

Self-verification	Yes	No
BLEU	0.36	0.33
ROUGE-1 F1	0.40	0.35
METEOR	0.40	0.37
BERTScore	0.36	0.28

TABLA 9

Evaluación humana de RecSAVER

Table 9: Correlation between coherence and NLG metrics with and without using self-verified references.

Self-verification	Yes	No
BLEU	0.36	0.33
ROUGE-1 F1	0.40	0.35
METEOR	0.40	0.37
BERTScore	0.36	0.28

Razonamientos
verificados se
correlacionan con
mejor NLG

TABLA 9

Evaluación humana de RecSAVER

Table 9: Correlation between coherence and NLG metrics with and without using self-verified references.

Self-verification	Yes	No
BLEU	0.36	0.33
ROUGE-1 F1	0.40	0.35
METEOR	0.40	0.37
BERTScore	0.36	0.28

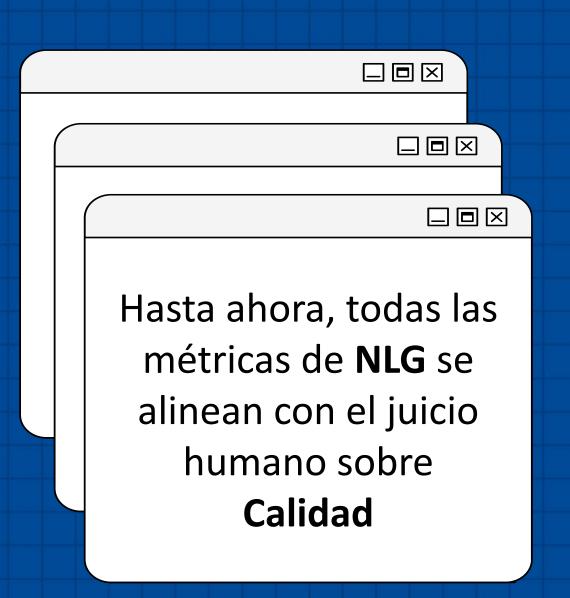
Quiere decir que filtrar aporta credibilidad

TABLA 9

Evaluación humana de RecSAVER

Table 9: Correlation between coherence and NLG metrics with and without using self-verified references.

Self-verification	Yes	No
BLEU	0.36	0.33
ROUGE-1 F1	0.40	0.35
METEOR	0.40	0.37
BERTScore	0.36	0.28





Analysis of Reasoning Quality

TABLA 11

Evaluación humana de RecSAVER

Table 11: Reasoning quality associated with correct and incorrect rating predictions for PaLM 2-M zero-shot and Flan-T5 XL fine-tuned (1 sample per example, no filtering) models.

	Model	Correct Prediction	BLEU	ROUGE-1 F1	METEOR	BERT Score
TY	PaLM 2-M PaLM 2-M	Yes No	0.260 0.221	0.522 0.491	0.342 0.336	0.666 0.665
BEAUTY	Flan-T5 XL	Yes	0.254	0.515	0.342	0.667
	Flan-T5 XL	No	0.235	0.508	0.338	0.666
MOVIES/TV	PaLM 2-M	Yes	0.204	0.480	0.306	0.648
	PaLM 2-M	No	0.187	0.455	0.290	0.647
	Flan-T5 XL Flan-T5 XL	Yes No	0.177 0.159	0.457 0.444	0.292 0.283	0.644 0.642

Evaluación humana de RecSAVER

	Model	Correct Prediction	BLEU	ROUGE-1 F1	METEOR	BERT Score
X	PaLM 2-M	Yes	0.260	0.522	0.342	0.666
BEAUTY	PaLM 2-M	No	0.221	0.491	0.336	0.665
\mathbf{B}_{E}	Flan-T5 XL	Yes	0.254	0.515	0.342	0.667
	Flan-T5 XL	No	0.235	0.508	0.338	0.666
TV	PaLM 2-M	Yes	0.204	0.480	0.306	0.648
MOVIES/TV	PaLM 2-M	No	0.187	0.455	0.290	0.647
	Flan-T5 XL	Yes	0.177	0.457	0.292	0.644
\mathbf{Z}	Flan-T5 XL	No	0.159	0.444	0.283	0.642

TABLA 11

TABLA 11

Evaluación humana de RecSAVER

	Model	Correct Prediction	BLEU	ROUGE-1 F1	METEOR	BERT Score
>	PaLM 2-M	Yes	0.260	0.522	0.342	0.666
BEAUTY	PaLM 2-M	No	0.221	0.491	0.336	0.665
m BE	Flan-T5 XL	Yes	0.254	0.515	0.342	0.667
	Flan-T5 XL	No	0.235	0.508	0.338	0.666
ΓV	PaLM 2-M	Yes	0.204	0.480	0.306	0.648
MOVIES/TV	PaLM 2-M	No	0.187	0.455	0.290	0.647
	Flan-T5 XL	Yes	0.177	0.457	0.292	0.644
M	Flan-T5 XL	No	0.159	0.444	0.283	0.642



NLG mejoran cuando se asocian a una buena predicción

TABLA 11

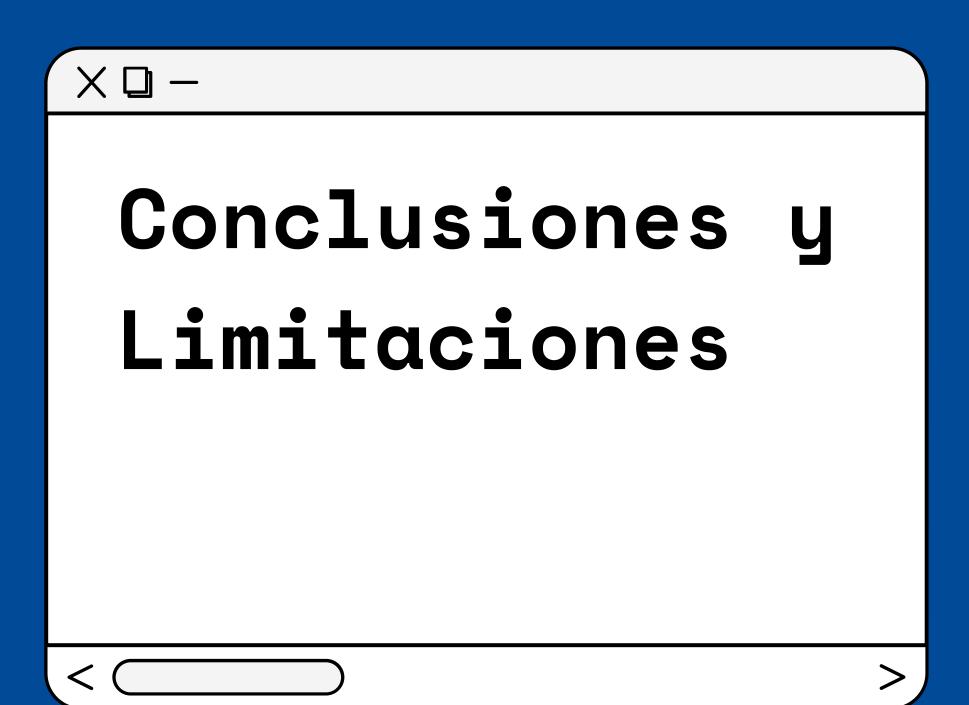
Evaluación humana de RecSAVER

	Model	Correct Prediction	BLEU	ROUGE-1 F1	METEOR	BERT Score
	PaLM 2-M	Yes	0.260	0.522	0.342	0.666
BEAUTY	PaLM 2-M	No	0.221	0.491	0.336	0.665
\mathbf{B}_{E}	Flan-T5 XL	Yes	0.254	0.515	0.342	0.667
	Flan-T5 XL	No	0.235	0.508	0.338	0.666
$^{\prime}\mathrm{TV}$	PaLM 2-M	Yes	0.204	0.480	0.306	0.648
MOVIES/TV	PaLM 2-M	No	0.187	0.455	0.290	0.647
	Flan-T5 XL	Yes	0.177	0.457	0.292	0.644
\geq	Flan-T5 XL	No	0.159	0.444	0.283	0.642

NLG mejoran cuando se asocian a una buena predicción

Pero las diferencias no son muy grandes

. . .

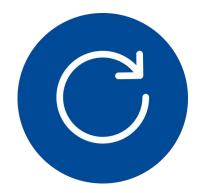


conclusiones presentadas por el paper



Conclusiones

presentadas por el paper:



1

<u>Razonamiento de LLMs</u>

- LLMs hacen mejores predicciones si se les pide razonar.
- Dependencia de data explícita (user reviews) para generar buenas predicciones.
- Dan un insight a las **razones** de por qué un usuario interactúa con un ítem.
- Necesidad de prompts sencillos (Zero-shot > One-shot)

Impacto de training data

- La calidad de pre-entrenamiento influye mucho en el desempeño de predicciones (PaLM vs Flan)
- Presencia de datos del dominio trabajado en el training data también tiene impacto.
- En general, hubieron mejores resultados para MOVIES/TV que BEAUTY

2

Conclusiones

presentadas por el paper:



8

<u>Impacto de Fine Tunina</u>

- Mejores rendimientos en el task de predicción. (Flan normal vs Flan fine tuned)
- Importancia de entrenar con razonamientos de calidad, pero necesidad de varios ejemplos.
- Abre la posibilidad de usar modelos pequeños, reduciendo costos de inferencia por predicción.

Rec-SAVER

- Validación de la generación automática de referencias auto-validadas, según evaluación humana (excepto insightfulness)
- Buena predicción → Calidad razonamiento producido.
- Es viable usar referencias generadas por Rec-SAVER para evaluar el razonamiento de otros LLMs en el mismo task.

4

 $\times \Box -$ Conclusiones y Limitaciones

limitaciones y discusión



Limitaciones

y discusión



П

Hubiese sido bueno comparar LLMs vs Filtrado Colaborativo (u otros recsys tradicionales)

4

Poca claridad de las métricas NLG utilizadas para evaluar razonamientos ¿que significa específicamente?

٧_

Manejo manual de *leaks*de ground truth en
Rec-SAVER (dificulta el
filtrado completo de
razonamientos)

5

Diferencias muy pequeñas
al comparar task
performance con la calidad
del razonamiento ¿Existe
realmente una correlación?

No se logra aplicar zero-shot CoT a FlanT5 XL sin fine-tuning (falta de un baseline importante)

Diferencia en calidad entre modelos usados para cada método no permite una buena comparación zero-shot CoT vs fine tuning



Referencias

Tablas 3,4,5,6,7,8,9 y 11:

Tsai, A. Y., Kraft, A., Jin, L., Cai, C., Hosseini, A., Xu, T., ... & Yi, X. (2024). Leveraging LLM Reasoning Enhances Personalized Recommender Systems. arXiv preprint arXiv:2408.00802.



