LMMs recsys

Buscar

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.

Marco teórico

Conceptos Clave:

1

Zero Shot Prompting

2

Chain of Thought prompting

3

LLM Fine Tuning

fuentes: https://www.promptingguide.ai/techniques, https://www.ibm.com/es-es/topics/fine-tuning

Marco teórico

Trabajos Relacionados:

1

Recommendation as instruction following: A large language model empowered recommendation approach[1]

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

Marco teórico

Trabajos Relacionados:

2

Do LLMs Understand User Preferences? Evaluating LLMs On User Rating Prediction[2]

También buscan hacer predicción de ratings (con zero-shot, few-shot y fine tuning)

Superioridad de filtrado colaborativo, pero buenos resultados en fine tuning

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

Marco teórico

Trabajos Relacionados:

8

Chat-REC: Towards interactive and explainable LLMs Augmented Recommender Systems[3]

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 llms 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*

Problemas de recomendación

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

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

Problemas de recomendación

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

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

$$\Updownarrow$$

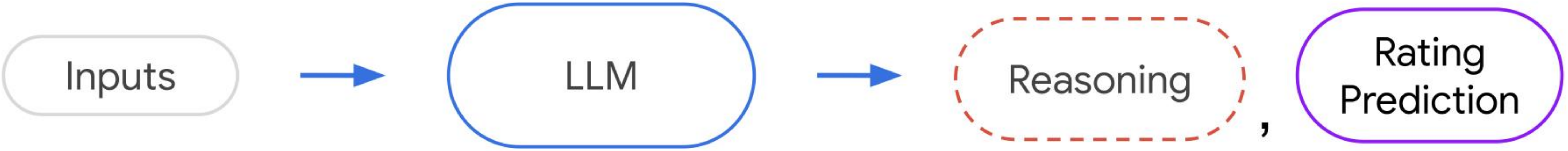
$$\hat{\mathbf{r}}_{u,i} = \mathrm{LLM}(\mathcal{H}_u, \mathcal{M}_i)$$

*

Procedimientos:

- Zero shot v/s razonamiento
- Fine-tunning
- Rec-SAVER

1. Zero shot v/s razonamiento

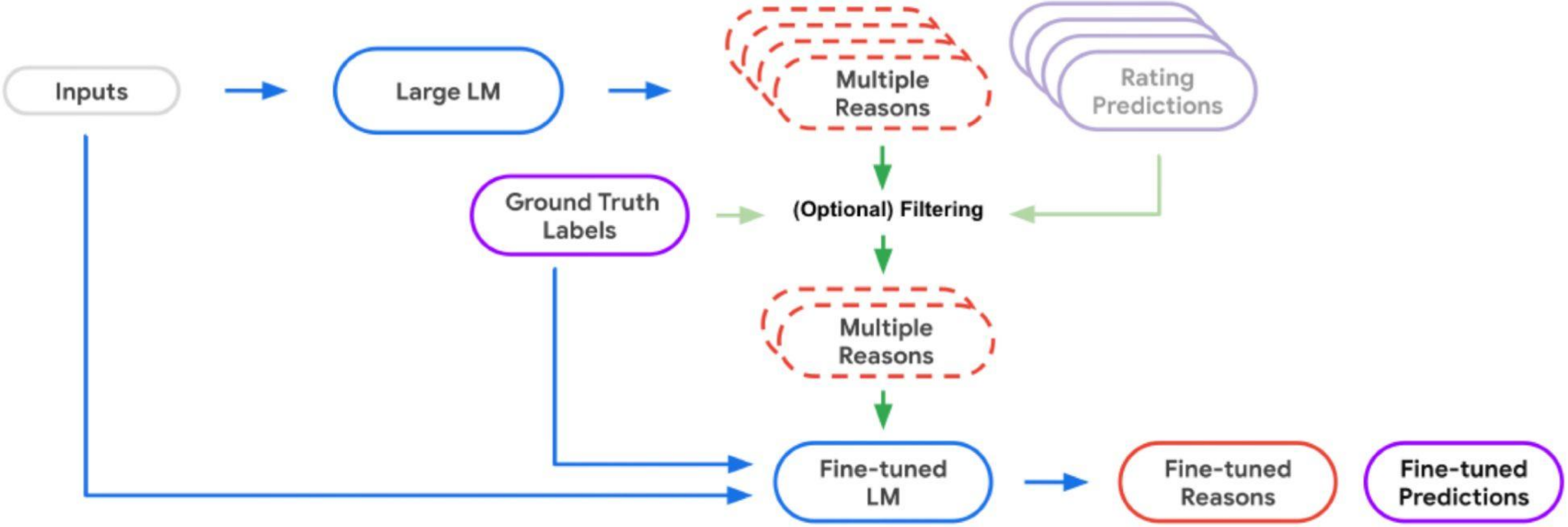


Se aplica un prompt al modelo PaLM 2-M, con 4 secciones relevantes:

Preamble	*e.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 Description	*e.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*.

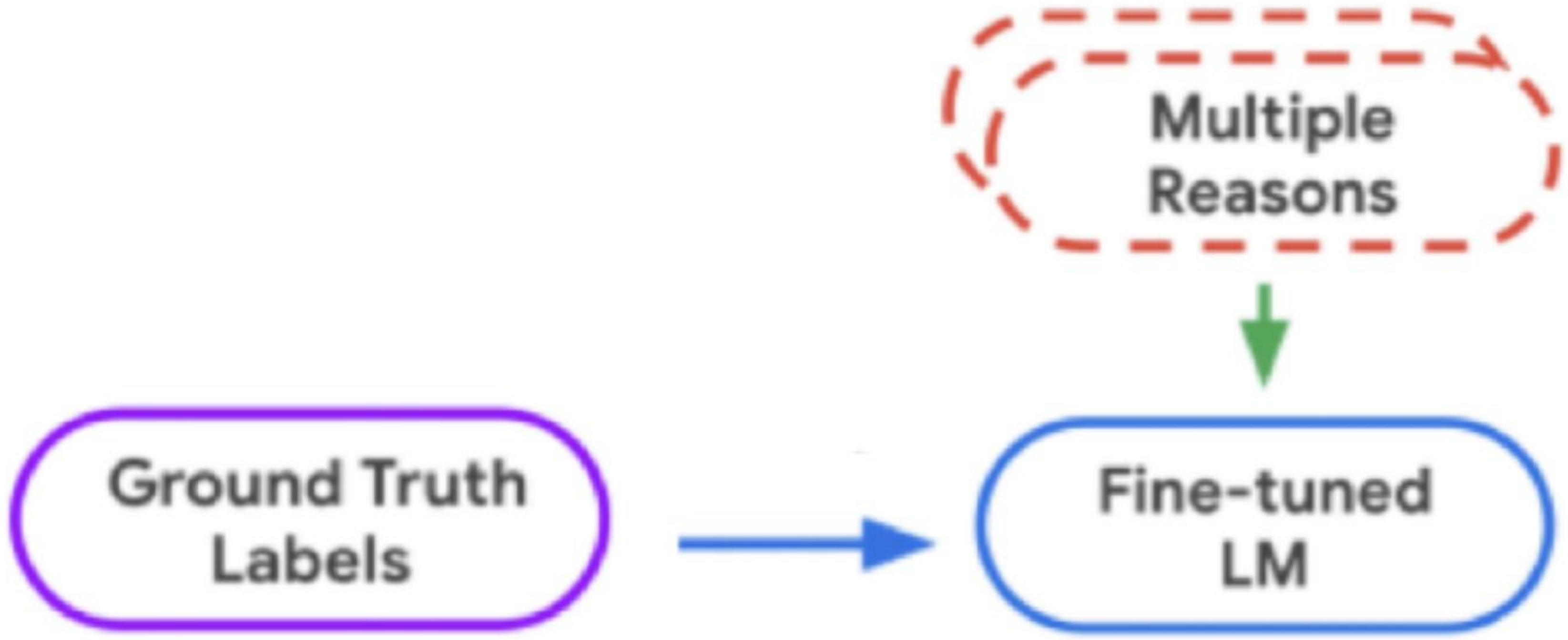
Paso 1: fetch

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

Paso 2: 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.

Paso 3: fine-tunning



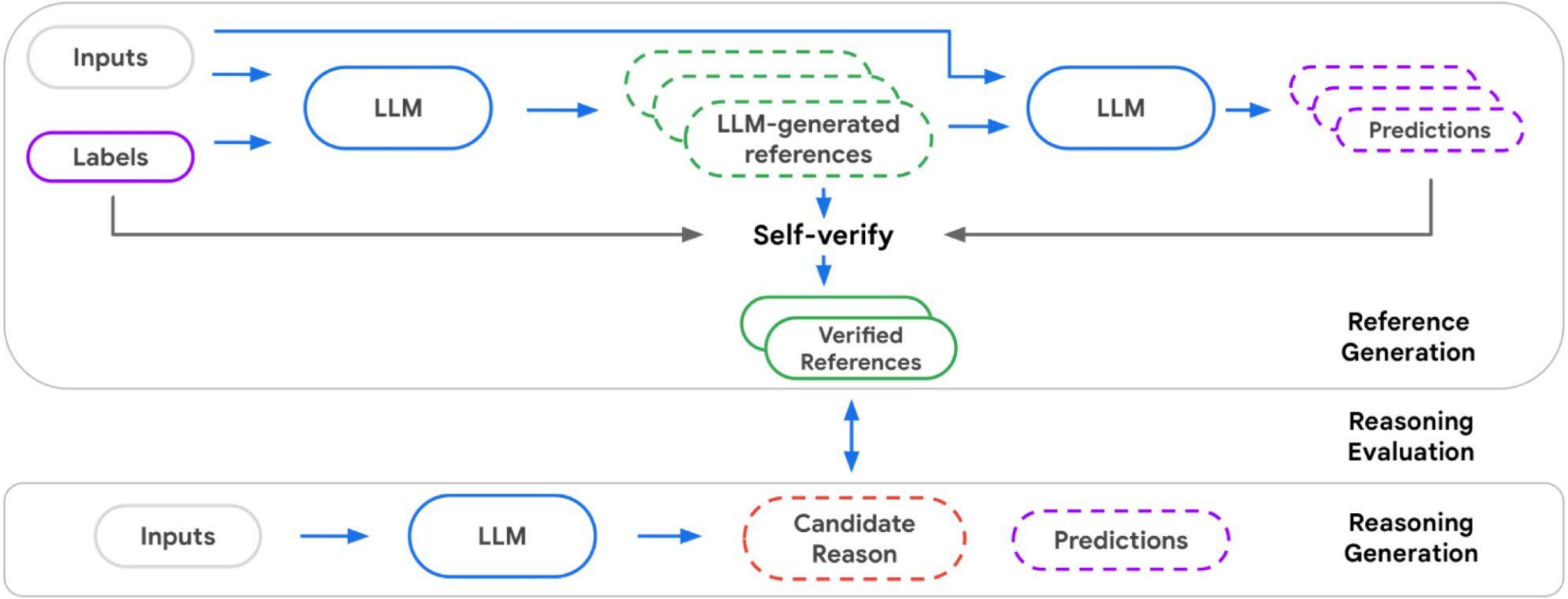
- Se alimenta una versión igual o más pequeña del LLM

- Se utilizaron modelos small, base, large y XL.

Paso 4: 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*.

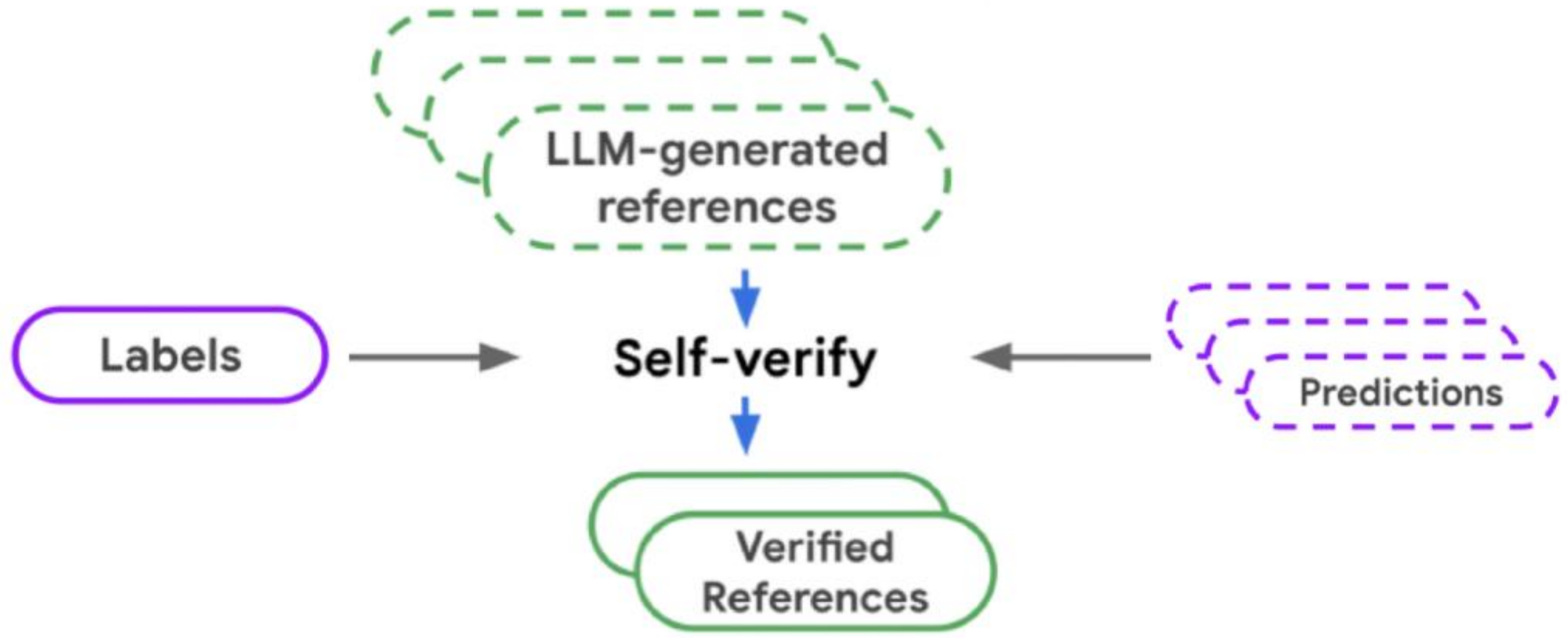
Paso 1: fetch

- 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

Paso 2: predicción

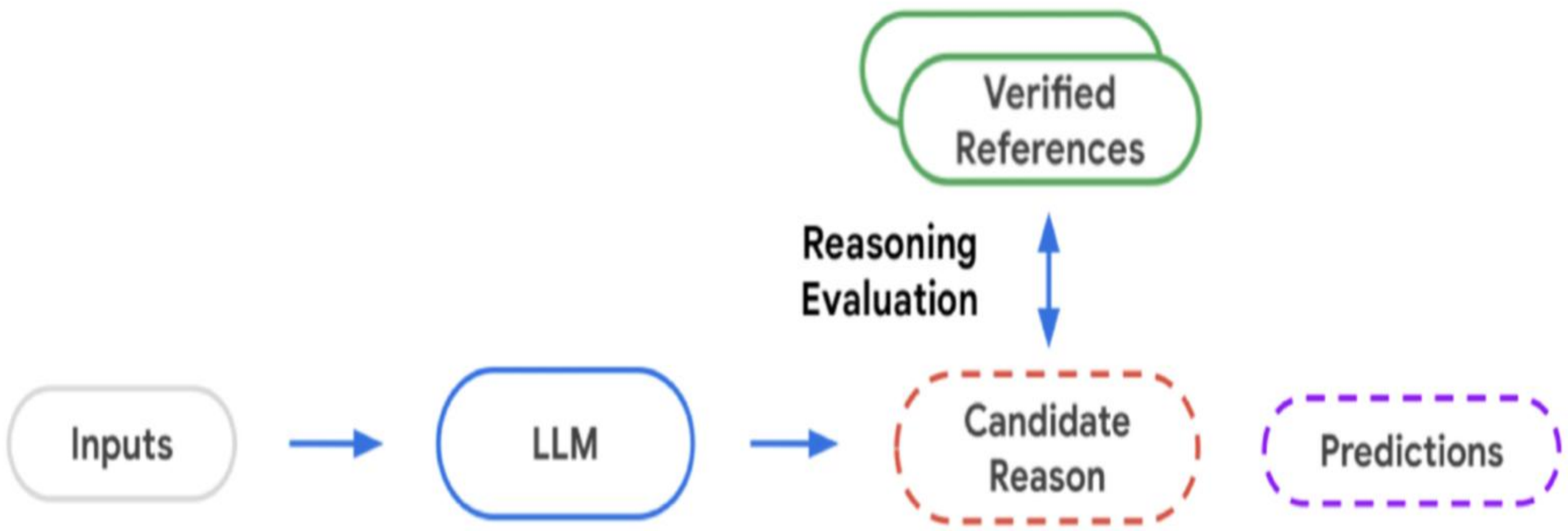
- Predecir rating en base a los razonamientos

Paso 3: auto-verificación



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

Paso 4: evaluación



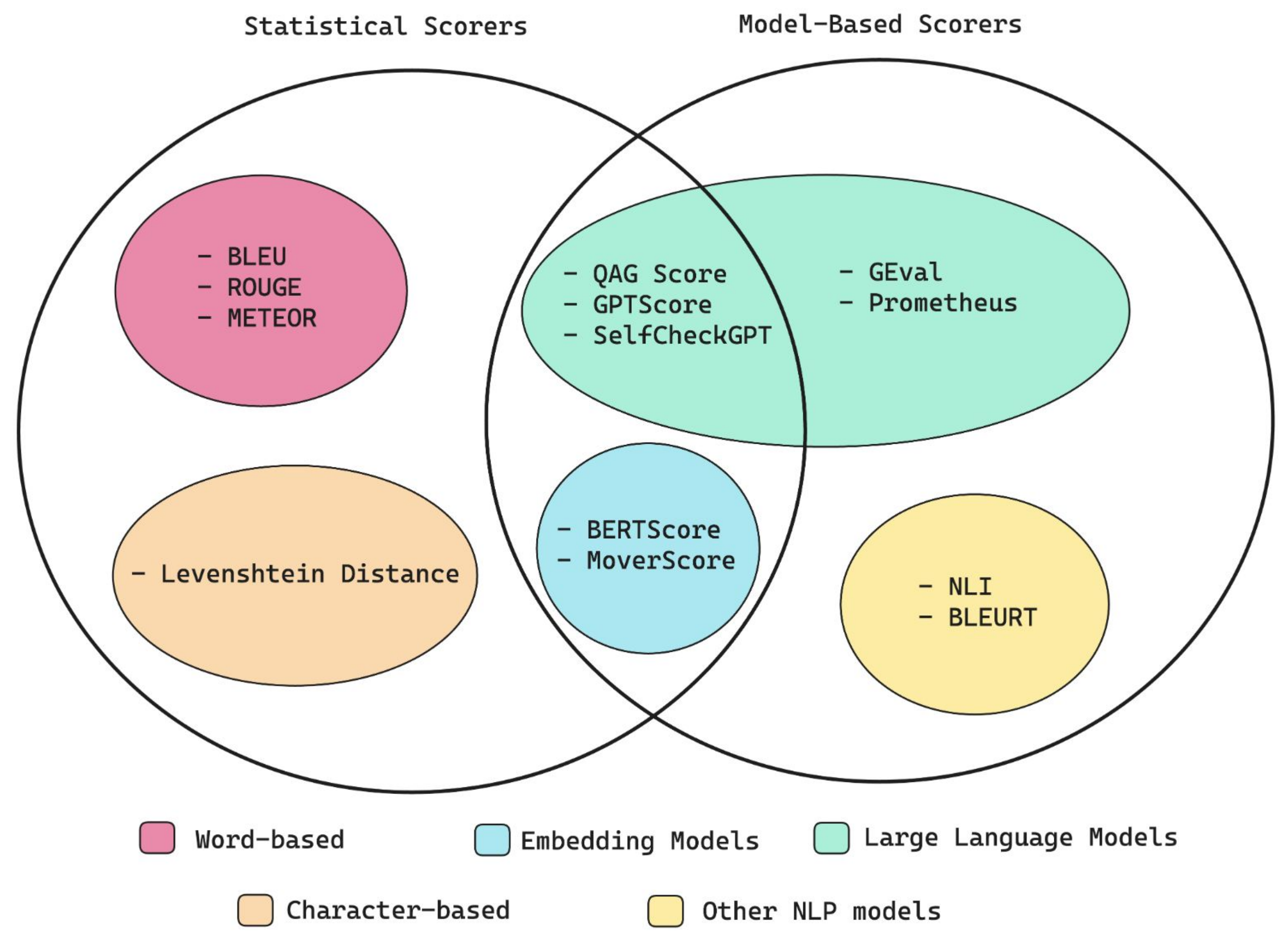
- Se evalúa la calidad de razonamiento generado por otro modelo.
- Se utilizan métricas NLG

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)

Resultados y Análisis

NLG Metrics!



Types of metric scorers

BLEU

ROUGE

METEOR

BERTScore

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

Buscar

Resultados y Análisis

- **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.

BLEU

ROUGE

METEOR

BERTScore

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

Buscar

Resultados y Análisis

- **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’).

BLEU

ROUGE

METEOR

BERTScore

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

Buscar

Resultados y Análisis

- **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

BLEU

ROUGE

METEOR

BERTScore

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

Buscar

Resultados y Análisis

- **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

BLEU

ROUGE

METEOR

BERTScore

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

Buscar

Zero-shot Learning Improves with Reasoning

Resultados y Análisis

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
BEAUTY	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	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
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.63	0.29	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**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*.

Resultados y Análisis

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	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	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
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.63	0.29	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**0.647**
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

Resultados y Análisis

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	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	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
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.63	0.29	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**0.647**
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

El paso de razonamiento mejora el task performance

Resultados y Análisis

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
BEAUTY	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	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
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.63	0.29	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**0.647**
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

Resultados y Análisis

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
BEAUTY	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	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
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.63	[illegible]	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**0.647**
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

No Review
≈
User-item Matrix

No Review,
No Rating
≈
Implicit feedback

Resultados y Análisis

Conocimiento Previo

	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	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	1.66	0.235	0.504	**0.340**	**0.667**
	One-shot	0.43	0.57	0.26	1.32	1.97	0.225	0.502	0.335	0.664
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.63	0.29	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**0.647**
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

Resultados y Análisis

Conocimiento Previo

	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	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	1.66	0.235	0.504	**0.340**	**0.667**
	One-shot	0.43	0.57	0.26	1.32	1.97	0.225	0.502	0.335	0.664
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.63	0.29	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**0.647**
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

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

Resultados y Análisis

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
BEAUTY	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	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
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.63	0.29	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**0.647**
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

Resultados y Análisis

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
BEAUTY	Naive Baseline (Avg.)	0.52	0.60	0.25	1.35	1.75	-	-	-	-
	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	0.49	0.57	0.23	1.35	1.70	-	-	-	-
	- No Review	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	1.33	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
MOVIES/TV	Naive Baseline (Avg.)	0.59	0.63	0.30	1.21	1.63	-	-	-	-
	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 Output	0.59	0.65	0.29	1.18	1.56	-	-	-	-
	- No Review	0.58	0.63	0.28	1.20	1.58	0.173	0.452	0.291	0.641
	- 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**	**0.647**
	One-shot	0.47	0.59	0.23	1.32	1.68	0.182	0.452	0.276	0.641

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

Resultados y Análisis

TABLA 4

Fine-tuning con FlanT5

	Model Size	Reasoning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE ↓	Multi. RMSE ↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
BEAUTY	Small	✓	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
	Base	✓	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	**0.344**	**0.667**
	Large	✓	0.64	0.59	0.67	**0.33**	0.65	1.26	1.73	0.240	0.506	0.343	0.666
	XL	✓	**0.67**	**0.61**	**0.78**	0.30	**0.69**	**1.24**	**1.68**	**0.241**	**0.510**	0.339	**0.667**
	XL	✗	0.55	**0.61**	0.74	0.28	0.67	1.31	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	-	-
MOVIES/TV	Small	✓	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
	Base	✓	**0.65**	0.59	**0.72**	**0.34**	**0.68**	1.18	1.65	0.153	0.438	0.279	0.634
	Large	✓	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**
	XL	✗	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-	-	-	-

Resultados y Análisis

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
BEAUTY	Small	✓	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
	Base	✓	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	**0.344**	**0.667**
	Large	✓	0.64	0.59	0.67	**0.33**	0.65	1.26	1.73	0.240	0.506	0.343	0.666
	XL	✓	**0.67**	**0.61**	**0.78**	0.30	**0.69**	**1.24**	**1.68**	**0.241**	**0.510**	0.339	**0.667**
	XL	✗	0.55	**0.61**	0.74	0.28	0.67	1.31	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	-	-
MOVIES/TV	Small	✓	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
	Base	✓	**0.65**	0.59	**0.72**	**0.34**	**0.68**	1.18	1.65	0.153	0.438	0.279	0.634
	Large	✓	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**
	XL	✗	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-	-	-	-

Resultados y Análisis

TABLA 4

Fine-tuning con FlanT5

	Model Size	Reasoning	Binary Acc.	Binary F1	Binary AUC	Multi. Acc.	Multi. AUC	Multi. MAE ↓	Multi. RMSE ↓	BLEU	ROUGE-1 F1	METEOR	BERT Score
BEAUTY	Small	✓	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
	Base	✓	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	**0.344**	**0.667**
	Large	✓	0.64	0.59	0.67	**0.33**	0.65	1.26	1.73	0.240	0.506	0.343	0.666
	XL	✓	**0.67**	**0.61**	**0.78**	0.30	**0.69**	**1.24**	**1.68**	**0.241**	**0.510**	0.339	**0.667**
	XL	✗	0.55	**0.61**	0.74	0.28	0.67	1.31	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	-	-
MOVIES/TV	Small	✓	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
	Base	✓	**0.65**	0.59	**0.72**	**0.34**	**0.68**	1.18	1.65	0.153	0.438	0.279	0.634
	Large	✓	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**
	XL	✗	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-	-	-	-

Modelo más grande

=

Mejor predicción

Puede deberse a tener una **mayor capacidad** de albergar **conocimiento**

Resultados y Análisis

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
BEAUTY	Small	✓	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
	Base	✓	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	**0.344**	**0.667**
	Large	✓	0.64	0.59	0.67	**0.33**	0.65	1.26	1.73	0.240	0.506	0.343	0.666
	XL	✓	**0.67**	**0.61**	**0.78**	0.30	**0.69**	**1.24**	**1.68**	**0.241**	**0.510**	0.339	**0.667**
	XL	✗	0.55	**0.61**	0.74	0.28	0.67	1.31	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	-	-
MOVIES/TV	Small	✓	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
	Base	✓	**0.65**	0.59	**0.72**	**0.34**	**0.68**	1.18	1.65	0.153	0.438	0.279	0.634
	Large	✓	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**
	XL	✗	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-	-	-	-

Resultados y Análisis

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
BEAUTY	Small	✓	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
	Base	✓	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	**0.344**	**0.667**
	Large	✓	0.64	0.59	0.67	**0.33**	0.65	1.26	1.73	0.240	0.506	0.343	0.666
	XL	✓	**0.67**	**0.61**	**0.78**	0.30	**0.69**	**1.24**	**1.68**	**0.241**	**0.510**	0.339	**0.667**
	XL	✗	0.55	**0.61**	0.74	0.28	0.67	1.31	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	-	-
MOVIES/TV	Small	✓	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
	Base	✓	**0.65**	0.59	**0.72**	**0.34**	**0.68**	1.18	1.65	0.153	0.438	0.279	0.634
	Large	✓	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**
	XL	✗	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-	-	-	-

No se pudieron extraer razonamientos con el Baseline!!

Resultados y Análisis

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
BEAUTY	Small	✓	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
	Base	✓	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	**0.344**	**0.667**
	Large	✓	0.64	0.59	0.67	**0.33**	0.65	1.26	1.73	0.240	0.506	0.343	0.666
	XL	✓	**0.67**	**0.61**	**0.78**	0.30	**0.69**	**1.24**	**1.68**	**0.241**	**0.510**	0.339	**0.667**
	XL	✗	0.55	**0.61**	0.74	0.28	0.67	1.31	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	-	-
MOVIES/TV	Small	✓	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
	Base	✓	**0.65**	0.59	**0.72**	**0.34**	**0.68**	1.18	1.65	0.153	0.438	0.279	0.634
	Large	✓	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**
	XL	✗	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.61	0.43	0.61	0.23	0.61	1.56	2.00	-	-	-	-

Resultados y Análisis

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
BEAUTY	Small	✓	0.62	0.53	0.65	0.30	0.63	1.35	1.84	0.225	0.499	0.342	0.663
	Base	✓	0.59	0.47	0.66	0.27	0.64	1.37	1.83	0.239	0.507	**0.344**	**0.667**
	Large	✓	0.64	0.59	0.67	**0.33**	0.65	1.26	1.73	0.240	0.506	0.343	0.666
	XL	✓	**0.67**	**0.61**	**0.78**	0.30	**0.69**	**1.24**	**1.68**	**0.241**	**0.510**	0.339	**0.667**
	XL	✗	0.55	**0.61**	0.74	0.28	0.67	1.31	1.75	-	-	-	-
	XL (no fine-tuning)	✗	0.55	0.40	0.56	0.22	0.56	1.64	2.09	-	-	-	-
MOVIES/TV	Small	✓	0.60	0.55	0.70	0.33	0.66	1.23	1.71	0.137	0.423	0.272	0.627
	Base	✓	**0.65**	0.59	**0.72**	**0.34**	**0.68**	1.18	1.65	0.153	0.438	0.279	0.634
	Large	✓	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**
	XL	✗	0.62	0.57	0.70	0.33	0.66	1.27	1.75	-	-	-	-
	XL (no fine-tuning)	✗	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.

Resultados y Análisis

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
BEAUTY	1	None	0.67	0.61	0.78	0.30	0.69	**1.24**	**1.68**	0.241	**0.510**	**0.339**	0.667
	8	None	**0.68**	**0.64**	**0.79**	**0.31**	**0.70**	1.25	1.71	**0.248**	0.509	0.333	**0.671**
	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
	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 Zero-shot		0.56	0.62	-	0.37	-	1.14	1.60	0.236	0.503	0.339	0.665
MOVIES/TV	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
	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
	8	Binary	0.60	**0.64**	0.71	0.33	0.66	1.28	1.78	0.175	0.443	0.288	0.641
	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 Zero-shot		0.62	0.66	-	0.40	-	1.06	1.53	0.194	0.465	0.296	0.647

Resultados y Análisis

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
BEAUTY	1	None	0.67	0.61	0.78	0.30	0.69	**1.24**	**1.68**	0.241	**0.510**	**0.339**	0.667
	8	None	**0.68**	**0.64**	**0.79**	**0.31**	**0.70**	1.25	1.71	**0.248**	0.509	0.333	**0.671**
	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
	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 Zero-shot		0.56	0.62	-	0.37	-	1.14	1.60	0.236	0.503	0.339	0.665
MOVIES/TV	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
	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
	8	Binary	0.60	**0.64**	0.71	0.33	0.66	1.28	1.78	0.175	0.443	0.288	0.641
	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 Zero-shot		0.62	0.66	-	0.40	-	1.06	1.53	0.194	0.465	0.296	0.647

Resultados y Análisis

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
BEAUTY	1	None	0.67	0.61	0.78	0.30	0.69	**1.24**	**1.68**	0.241	**0.510**	**0.339**	0.667
	8	None	**0.68**	**0.64**	**0.79**	**0.31**	**0.70**	1.25	1.71	**0.248**	0.509	0.333	**0.671**
	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
	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 Zero-shot		0.56	0.62	-	0.37	-	1.14	1.60	0.236	0.503	0.339	0.665
MOVIES/TV	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
	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
	8	Binary	0.60	**0.64**	0.71	0.33	0.66	1.28	1.78	0.175	0.443	0.288	0.641
	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 Zero-shot		0.62	0.66	-	0.40	-	1.06	1.53	0.194	0.465	0.296	0.647

En general, Aplicar filtros reduce el rendimiento

El modelo se ve perjudicado por disminuir la inf.

Resultados y Análisis

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
BEAUTY	1	None	0.67	0.61	0.78	0.30	0.69	**1.24**	**1.68**	0.241	**0.510**	**0.339**	0.667
	8	None	**0.68**	**0.64**	**0.79**	**0.31**	**0.70**	1.25	1.71	**0.248**	0.509	0.333	**0.671**
	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
	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 Zero-shot		0.56	0.62	-	0.37	-	1.14	1.60	0.236	0.503	0.339	0.665
MOVIES/TV	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
	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
	8	Binary	0.60	**0.64**	0.71	0.33	0.66	1.28	1.78	0.175	0.443	0.288	0.641
	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 Zero-shot		0.62	0.66	-	0.40	-	1.06	1.53	0.194	0.465	0.296	0.647

En general, plicar
ltros re el

Incluso si en teoría es de mejor calidad !!

ve
or
di nuir la in

Resultados y Análisis

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
BEAUTY	1	None	0.67	0.61	0.78	0.30	0.69	**1.24**	**1.68**	0.241	**0.510**	**0.339**	0.667
	8	None	**0.68**	**0.64**	**0.79**	**0.31**	**0.70**	1.25	1.71	**0.248**	0.509	0.333	**0.671**
	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
	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 Zero-shot		0.56	0.62	-	0.37	-	1.14	1.60	0.236	0.503	0.339	0.665
MOVIES/TV	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
	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
	8	Binary	0.60	**0.64**	0.71	0.33	0.66	1.28	1.78	0.175	0.443	0.288	0.641
	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 Zero-shot		0.62	0.66	-	0.40	-	1.06	1.53	0.194	0.465	0.296	0.647

En 'Movies' funciona mejor '1-off'

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

Human Alignment Analysis

Resultados y Análisis

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

Resultados y Análisis

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 …

Resultados y Análisis

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: Cor
automatic NL
across the ann

BLE
ROU
MET
BER

ce tiene
elación con
G

Faithfulness no se muestra porque las métricas de **NLG** miden *overlap*, entonces **no pueden detectar veracidad**.

ightfulness, no
tanto …

Two-sample T-test

Resultados y Análisis

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**

Resultados y Análisis

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

Resultados y Análisis

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

Resultados y Análisis

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**

le
a la
ana

La excepción es
METEOR

disminuyen
ricas NLG

Effectivness of self-verification

Resultados y Análisis

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

Resultados y Análisis

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

Resultados y Análisis

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**

Resultados y Análisis

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**

Resultados y Análisis

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

Hasta ahora, todas las métricas de **NLG** se alinean con el juicio humano sobre **Calidad**

Analysis of Reasoning Quality

Resultados y Análisis

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
BEAUTY	PaLM 2-M	Yes	**0.260**	**0.522**	**0.342**	**0.666**
	PaLM 2-M	No	0.221	0.491	0.336	0.665
	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	Yes	**0.177**	**0.457**	**0.292**	**0.644**
	Flan-T5 XL	No	0.159	0.444	0.283	0.642

Resultados y Análisis

TABLA 11

Evaluación humana de RecSAVER

	Model	Correct Prediction	BLEU	ROUGE-1 F1	METEOR	BERT Score
BEAUTY	PaLM 2-M	Yes	**0.260**	**0.522**	**0.342**	**0.666**
	PaLM 2-M	No	0.221	0.491	0.336	0.665
	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	Yes	**0.177**	**0.457**	**0.292**	**0.644**
	Flan-T5 XL	No	0.159	0.444	0.283	0.642

Resultados y Análisis

TABLA 11

Evaluación humana de RecSAVER

	Model	Correct Prediction	BLEU	ROUGE-1 F1	METEOR	BERT Score
BEAUTY	PaLM 2-M	Yes	**0.260**	**0.522**	**0.342**	**0.666**
	PaLM 2-M	No	0.221	0.491	0.336	0.665
	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	Yes	**0.177**	**0.457**	**0.292**	**0.644**
	Flan-T5 XL	No	0.159	0.444	0.283	0.642

NLG mejoran cuando se asocian a una buena predicción

Resultados y Análisis

TABLA 11

Evaluación humana de RecSAVER

	Model	Correct Prediction	BLEU	ROUGE-1 F1	METEOR	BERT Score
BEAUTY	PaLM 2-M	Yes	**0.260**	**0.522**	**0.342**	**0.666**
	PaLM 2-M	No	0.221	0.491	0.336	0.665
	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	Yes	**0.177**	**0.457**	**0.292**	**0.644**
	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 y Limitaciones

conclusiones presentadas por el paper

Conclusiones

presentadas por el paper:

1

Razonamiento de LLMs

- 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)

2

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*

Conclusiones

presentadas por el paper:

3

Impacto de Fine Tuning

- **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.

4

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.

Conclusiones y Limitaciones

limitaciones y discusión

Limitaciones

y discusión

1

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

2

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

3

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

4

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

5

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

6

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.