EVALUACIÓN Y ANÁLISIS DE LLMS Y DATASETS PARA RECOMENDACIÓN CONVERSACIONAL

Octavio Águila, Benjamín Pedraza, Sebastián Rasmussen

Problema y motivación

La masificación de los *Large Language Models* (LLMs) en los últimos años ha generado un aumento en el interés y uso de estos modelos como sistemas recomendadores conversacionales (CRS). Sin embargo, tal como plantea [3], los datasets públicos para entrenar estos modelos poseen dos grandes problemas que limitan el avance de estos sistemas: preferencias de usuario demasiado genéricas y falta de explicaciones fundamentadas en conocimiento del dominio.

Es por ello que en el presente trabajo se busca evaluar el rendimiento de diferentes modelos entrenados con 2 datasets conversacionales: PEARL y ReDial, con el objetivo de contrastar la efectividad de incorporar contexto de preferencias del usuario y conocimiento extraído desde reseñas contra el uso de diálogos y conversaciones breves.

Metodología

Con la idea de extender la investigación de [3], utilizamos los actuales modelos *open source* que forman parte del estado del arte, realizando la experimentación con LLaMa 3.2 (3B), Qwen (1.7B) y DeepSeek LLaMa (8B).

Se le realiza *fine-tuning* a cada uno de ellos en ambos datsets por separado, utilizando el framework Unsloth junto con LoRa (Low-Rank Adaption) como técnica de *fine-tuning* eficiente [2]. Todos los modelos comparten la misma configuración de entrenamiento, con una longitud máxima de entrada de 2048 tokens y la misma random seed para asegurar la reproducibilidad.

Finalmente se realiza la evaluación de cada uno de estos tres modelos (en su versión entrenada con ReDial y PEARL), con las métricas presentadas en la sección de resultados.

Análisis de los datasets

Los datasets utilizados fueron PEARL y ReDial, que ya vienen separados parcialmente en conjuntos de entrenamiento y prueba. La siguiente tabla resume la información disponible:

Conjunto	PEARL	ReDial
Train	50,000	10,006
Test	5,000	1,342
Validation	2,277	*

Se unificó el formato de sus diálogos para asegurar comparabilidad entre los datasets y compatibilidad con los modelos pre-entrenados para chats. Se usa el template *llama-3.2*.

```
{ "content": ..., "role": ... }
```

Existe un a diferencia clave entre los dos datatasets procesados. PEARL tiene un primer mensaje con un rol que sólo participa ahí. Dicho rol es *system* que contiene las preferencias del usuario (Likes, Dislikes y Seen). Esto significa que la recomendaciones se basan más en información de los gustos del usuario previa que las que se mencionen directamente en el chat.

Resultados

Los modelos entrenados con PEARL generan respuestas más coherentes y contextuales, superando a ReDial en métricas como BLEU, ROUGE y BERTScore. También tienen mejor precisión y recall general. Aunque ReDial logra un mejor desempeño en Recall@1, lo que indica una mayor efectividad en identificar la mejor recomendación inmediata, PEARL muestra menor diversidad en las respuestas, posiblemente debido a la consistencia que aporta la información de personas y conocimiento del dominio.

En conjunto, PEARL demuestra ser un dataset más robusto y efectivo para entrenar modelos de recomendación conversacional, ya que ofrece preferencias más específicas, incorpora conocimiento del dominio y permite generar recomendaciones más personalizadas y relevantes.

Métricas con ReDial

Métrica	LLaMA	Qwen	DeepSeek
BLEU Score	0.0151	0.0154	0.0153
ROUGE-1	0.0350	0.0349	0.0348
ROUGE-2	0.0084	0.0093	0.0089
ROUGE-L	0.0302	0.0300	0.0300
Distinct-1 intra/inter	0.9764 / 0.1551	0.9764 / 0.1551	0.9764 / 0.1551
Distinct-2 intra/inter	0.8837 / 0.4488	0.8837 / 0.4488	0.8837 / 0.4488
BERTScore F1	0.6471	0.6542	0.6480
BERTScore Precisión	0.5725	0.5810	0.5727
Recall	0.7474	0.7517	0.7494
Recall@1	0.022	0.019	0.015
Novelty	0.9670	0.9630	0.9673
Self-BLEU	0.6451	0.7002	0.6458

Métricas con PEARL

Métrica	LLaMA	Qwen	DeepSeek
BLEU Score	0.0889	0.0912	0.0889
ROUGE-1	0.1486	0.1519	0.1486
ROUGE-2	0.0792	0.0806	0.0792
ROUGE-L	0.1032	0.1058	0.1039
Distinct-1 intra/inter	0.8172 / 0.1154	0.8172 / 0.1154	0.8172 / 0.1154
Distinct-2 intra/inter	0.9857 / 0.3758	0.9857 / 0.3758	0.9857 / 0.3758
BERTScore F1	0.7616	0.7653	0.7616
BERTScore Precisión	0.7039	0.7088	0.7040
Recall	0.8300	0.8319	0.8299
Recall@1	0.018	0.002	0.007
Novelty	0.8463	0.8374	0.8461
Self-BLEU	0.6810	0.6846	0.6813

El problema de Uncertainty Quantifycation

Como afirma [1], con el masificado uso de los LLMs, se vuelve crucial garantizar que las predicciones de los LLMs sean confiables. Una dimensión de la confiabilidad es la capacidad de indicar cuándo el texto generado es fiable y correcto, lo cual puede formalizarse como el problema de *uncertainty quantification*.
Uncertainty en un LLM refleja cuán seguro está el modelo de su respuesta. Según [4] existen dos tipos de incertidumbre: la incertidumbre aleatoria (*aleatoric*), inherente a la propia tarea cuando el enunciado es ambiguo o admite varias respuestas válidas, y la incertidumbre epistémica (*epistemic*), ligada a lo que el modelo desconoce, la cual disminuye si éste aprende más sobre el dominio.

Metodología para Uncertainty

Basándonos en [1], proponemos una metodología simple para el cálculo de *uncertainty* en sistemas recomendadores, testeado en PEARL para recomendar películas. Consideremos el conjunto:

$$\mathcal{C}(X) = \{C^{(1)}, \ldots, C^{(K)}\}$$

correspondiente a K aclaraciones generadas para el propmt o turno del usuario X, donde una aclaración se entiende como una reformulación del prompt original manteniendo su sentido.

Finalmente denotamos como $T^{(k)} \in \mathcal{T}$ para cada aclaración k el título de la película recomendada por el modelo y su distribución dentro de las recomendaciones como:

$$p(t) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\big[T^{(k)} = t\big]$$

midiendo la *uncertainty* aleatoria con la entropía de Shannon:

$$\hat{U}_{\text{alea}}(X) = H(T) = -\sum_{t\in\mathcal{T}} p(t)\,\log p(t)$$

Para el *uncertainty* epistémico,

$$\hat{U}_{\text{epi}}(X) = -\frac{1}{K}\sum_{k=1}^{K} \bar{\ell}\big(Y^{(k)}\big)$$

donde el modelo $Y^{(k)} = (y_1^{(k)}, \ldots, y_{N_k}^{(k)})$ es cada una de las respuestas del modelo y $\bar{\ell}$ corresponde a las log-probabilidad media por token.

Experimentos y Resultados

Se extrajeron 200 prompts ambiguos de PEARL, a los cuales se les generaron $K = 4$. aclaraciones utilizando la versión *Instruct* de Llama3. Luego, se evaluaron los modelos base Llama-3, DeepSeek y Qwen-2 con previo *fine-tunning* en PEARL. Cada aclaración se pasó una única vez al recomendador y se capturó: i) el título de la primera película nombrada y ii) la log-probabilidad media por token de la respuesta.

Para cada prompt estimamos la entropía promedio de títulos de película y el promedio negativo de la log-probabilidad, obteniendo:

Modelo	$\mathbb{E}\,[\hat{U}_{\text{alea}}]$	$\mathbb{E}\,[\hat{U}_{\text{epi}}]$
Llama-3	2.25	5.38
DeepSeek	3.89	1.01
Qwen-2	3.98	0.77

DeepSeek y Qwen arrojan mayor diversidad de títulos, lo que indica que sus respuestas varían más con pequeñas aclaraciones. Llama-3 muestra menor variabilidad, lo que lo hace más consistente. Por otrolado, Qwen exhibe la mayor confianza ($\hat{U}_{\text{epi}} \approx 0.8$), seguido de DeepSeek. Llama-3 genera muchas secuencias con probabilidad numéricamente nula (`-inf`) y, por ende, una incertidumbre epistémica elevada (5.4).

Discusión

Una de las grandes limitaciones de nuestro trabajo es el uso de modelos con una limitada cantidad de parámetros debido a costo computacional disponible. Esto hace que la tanto la generación de recomendaciones como la reformulación de preguntas sea propensa a alucinaciones o imprecisiones. Por útlimo, en el caso del *uncertainty*, esto también afectó en la cantidad de prompts generados, lo que hace que el estudio sea limitado.

Referencias

[1] Bairu Hou et al. *Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling*. 2024. arXiv: 2311.08718 [cs.CL]. URL: https://arxiv.org/abs/2311.08718.

[2] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[3] Minjin Kim et al. "Pearl: A Review-driven Persona-Knowledge Grounded Conversational Recommendation Dataset". In: Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1105–1120. DOI: 10.18653/v1/2024.findings-acl.65. URL: https://aclanthology.org/2024.findings-acl.65/.

[4] Armen Der Kiureghian and Ove Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009). Risk Acceptance and Risk Communication, pp. 105–112. ISSN: 0167-4730. DOI: https://doi.org/10.1016/j.strusafe.2008.06.020. URL: https://www.sciencedirect.com/science/article/pii/S0167473008000556.