Large Language Models
for Recommendation Systems

De la Factorización de Matrices a la IA Generativa

Universidad de La Laguna

01.12.2025

Pablo Hernández

Programando todo tipo de cosas desde 2010

Mi stack

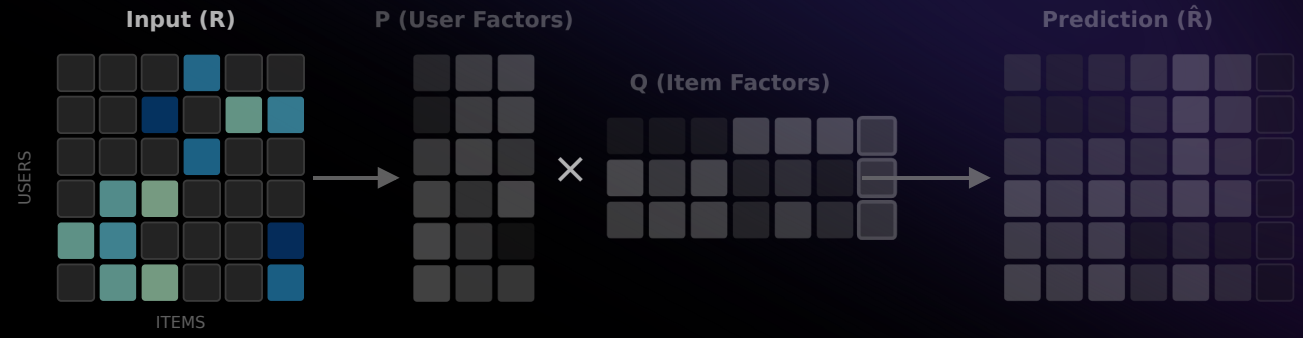
Actualmente estudiando en ULL

hadronomy.com hadronomy hadronomy

RecSys Clásico

Tiene límites claros

Cold Start & Falta de Semántica



MODEL OUTDATED: RETRAINING...

Entran los LLMs

Semántica

Entienden el contenido y el contexto, no solo coinciden IDs o etiquetas.

Zero-Shot Rec

Capacidad de recomendar sin historial previo del usuario (Cold Start).

Conversacional

El usuario puede pedir mejoras o cambios en lenguaje natural.

Explicabilidad

El modelo puede razonar y explicar por qué recomienda algo específico.

RAG & Knowledge

Uso de conocimiento externo y bases de datos vectoriales para enriquecer los datos.

Transfer Learning

Conocimiento general del mundo aplicado a tu dominio específico.

Arquitecturas

Discriminative

Embeddings

Usar el LLM (ej. BERT) para generar representaciones vectoriales densas de items y usuarios para calcular similitud.

Generative

SeqRec

Fine-tuning o Prompting (ej. GPT) para predecir el "siguiente token" como el "siguiente item" a consumir.

PROMPT TEMPLATE

```
User History: [Matrix, Inception, Interstellar]
Task: Recommend next movie.
Reasoning: User likes sci-fi & mind-bending plots.
Recommendation: _
```

THE STACK

FINE-TUNING PIPELINE

- **Unsloth** — Fast Training
- **TRL** — SFT / DPO
- **PEFT** — LoRA / QLoRA
- **Qwen 2.5** — Base Model

RETRIEVAL STACK

- **Sentence Transformers** — BERT / E5
- **Vector DB**
- **LangChain** — Orchestrator

Implementation Strate

Data: MovieLens 1M

Sliding Window Strategy: Transformamos el historial de interacciones en una conversación.

`[A, B, C] → D` `[B, C, D] → E`

Engine: Qwen 2.5 + Unsloth

Utilizamos **Unsloth** para optimizar el fine-tuning (2x más rápido, -70% VRAM) sobre una **A100** en Google Colab.

Open Notebook
Full Training & Inference Pipeline

jupyter nbviewer

JUPYTER FAQ

Open in Colab

In […

```
import os
import torch

# 1. Nuke Environment
!pip uninstall -y torch torchvision torchaudio flash-attn xformers unsloth trl pe

# 2. Install PyTorch 2.8.0 + CUDA 12.6
# We grab this from the pytorch release/test index if not yet in main,
# but assuming standard availability based on your logs.
!pip install torch==2.8.0 torchvision==0.23.0 torchaudio==2.8.0 --index-url https

# 3. Install Flash Attention 2.8.3 (Wheel for Torch 2.8)
!pip install https://github.com/Dao-AILab/flash-attention/releases/download/v2.8.

# 4. Install Xformers 0.0.32.post2 (Matches Torch 2.8)
# We force this version so Unsloth doesn't backtrack.
!pip install xformers==0.0.32.post2 --index-url https://download.pytorch.org/whl/

# 5. Install Unsloth with the PyTorch 2.8 specific tag
# [cu126onlytorch280] tells Unsloth to expect Torch 2.8 and Xformers 0.0.32.post2
!pip install "unsloth[colab-new,cu126onlytorch280] @ git+https://github.com/unslo

# 6. Install TRL & Utils
!pip install --no-deps trl peft accelerate bitsandbytes scikit-learn pandas

print("□ Environment Locked: Torch 2.8.0 | FA 2.8.3 | Unsloth [cu126onlytorch280]"
```

```
Found existing installation: torch 2.9.0+cu126
Uninstalling torch-2.9.0+cu126:
  Successfully uninstalled torch-2.9.0+cu126
Found existing installation: torchvision 0.24.0+cu126
Uninstalling torchvision-0.24.0+cu126:
  Successfully uninstalled torchvision-0.24.0+cu126
Found existing installation: torchaudio 2.9.0+cu126
Uninstalling torchaudio-2.9.0+cu126:
  Successfully uninstalled torchaudio-2.9.0+cu126
WARNING: Skipping flash-attn as it is not installed.
WARNING: Skipping xformers as it is not installed.
WARNING: Skipping unsloth as it is not installed.
WARNING: Skipping trl as it is not installed.
Found existing installation: peft 0.18.0
Uninstalling peft-0.18.0:
  Successfully uninstalled peft-0.18.0
Looking in indexes: https://download.pytorch.org/whl/cu126
Collecting torch==2.8.0
  Downloading https://download.pytorch.org/whl/cu126/torch-2.8.0%2Bcu126-cp312-cp
312-manylinux_2_28_x86_64.whl.metadata (30 kB)
```

Muchas gracias

Diapositivas y referencias en: hadronomy.com