Generating Data Augmentation Queries Using Large Language Models

Christopher Buss, Jasmin Mousavi, Mikhail Tokarev, Mahdis Safari, Arash Termehchy, David Maier, Stefan Lee

Oregon State University · Portland State University

1. Drug Repositioning Can Save Lives

Castleman Disease (Rare)
- Causes severe inflammation
- No known treatment

Too expensive to develop new drug

Treat with an existing drug!

… but which one?

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis
Enbrel	TNF inhibitor	plaque psoriasis

Local Data Source

Humira treats inflammation!

… but what about its risks?

Goal: augment local entity with external information

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment …
Etanercept	Etanercept binds …

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with …
$C_{2224}H_{3475}N_{621}O_{698}S_{26}$	There are …

relevant external entities

Augmented local entity:

brand_name	class	uses	adverse_effects	mechanisms
Humira	TNF inhibit.	rheumatoid …	After treatment …	Binds with …

… must extract these external entities!

Querying sources *manually*
- Find correct queries!

Generic Drugs

Humira | Search | 0 results

No relevant entities!

Bio Compounds

rheumatoid arthritis | Search | 436 results

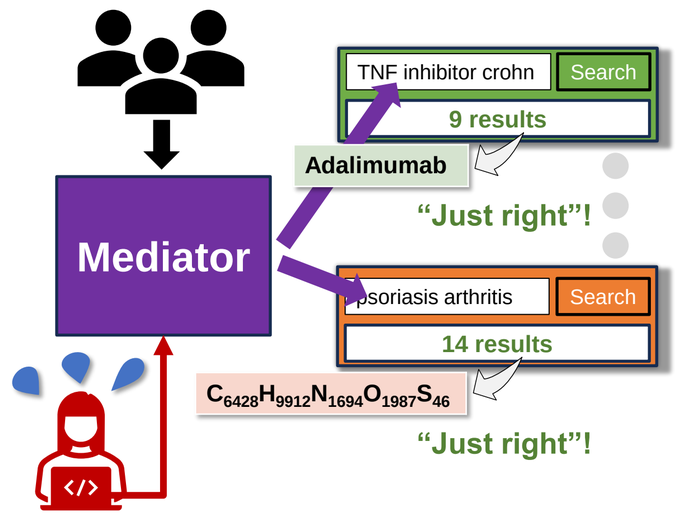
Many non-relevant entities!

Many data sources + data heterogeneity = much trial and error!

Use a Mediator
- Submits **"just right"** queries for the user

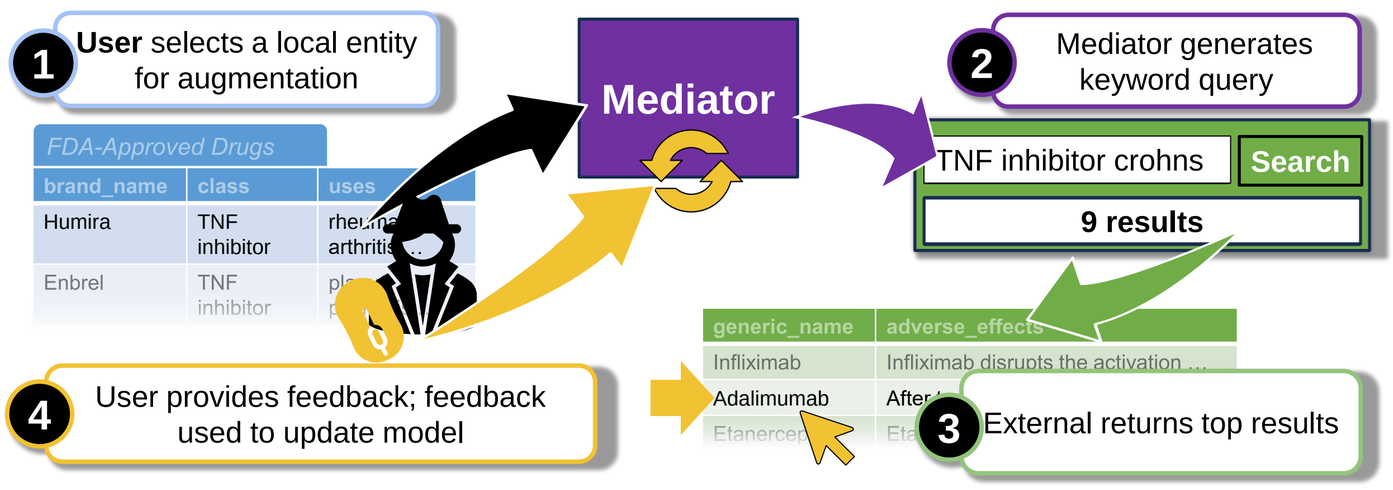
2. Existing Work

Programmers write **Mediators** by hand



NIH-funded consortium of systems (~14) → **$923M / year development and maintenance**

3. Online Autonomous Querying



Over time, **model** refines its estimator of **query effectiveness**

4. What Makes a Good Model?

Short-Run: find good queries quickly, so users remain engaged

Long-Run: continue to improve as feedback is provided

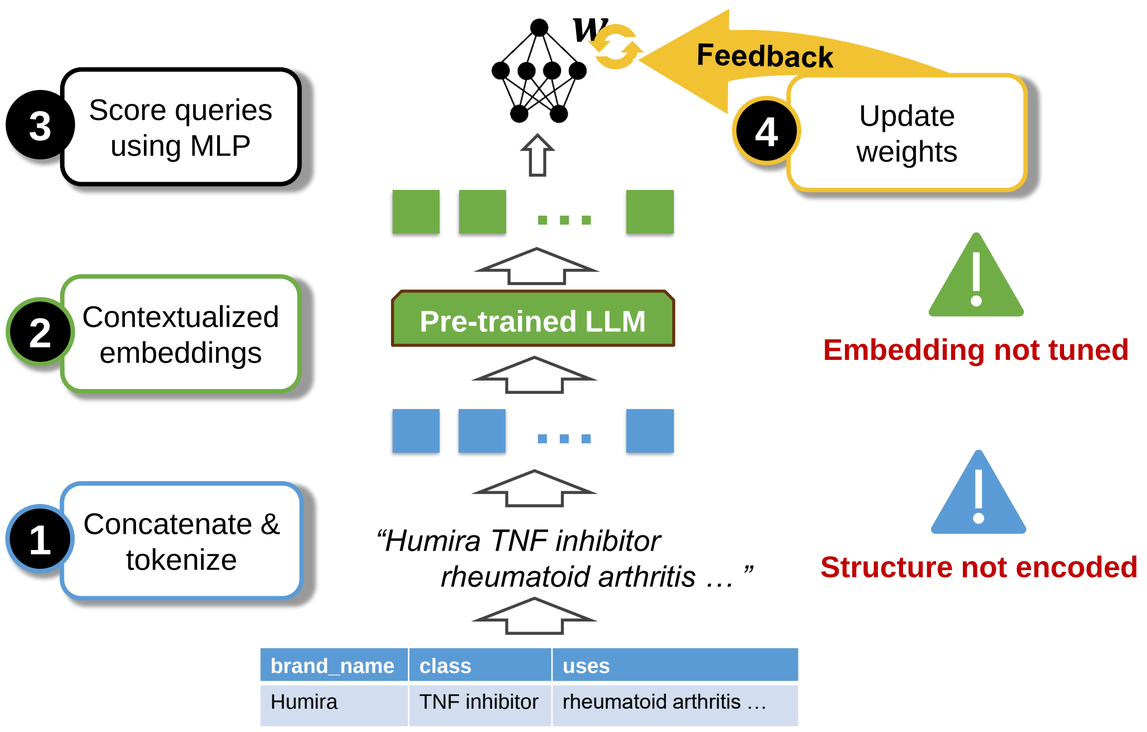
A strong prior for a variety of data… *that can adapt to domain-specific sources.*

Leverage a pre-trained LLM prior
- Learn over its rich encodings

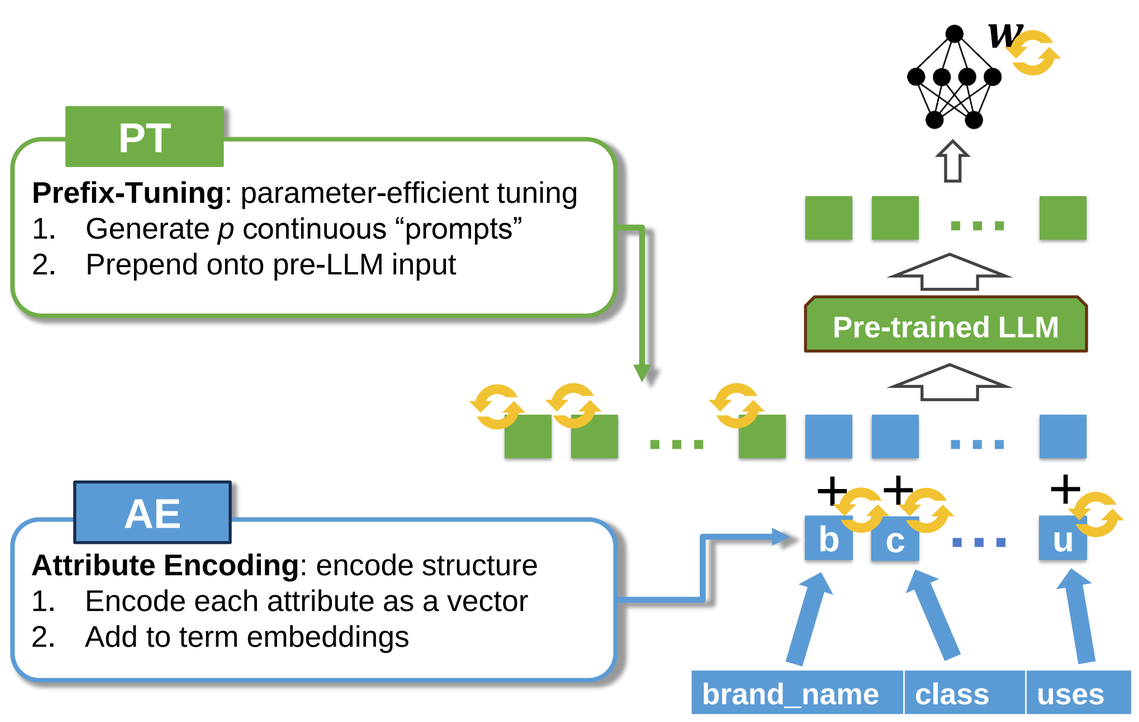
Online Setting: balances exploration & exploitation
- Model only knows quality of queries tried

Use e-greedy algorithm
- occasionally use a random query

7. Learning Query Quality using an LLM



8. Lightweight Tuning



8. Experimental Simulation

Query length: 4 terms

Average of 5 runs with 80% confidence interval

NoLLM: linear predictor using term features

PT: Prefix-Tuning with $p = 5$; **AE**: Attribute Encoding

CORD-19: research records related to COVID-19

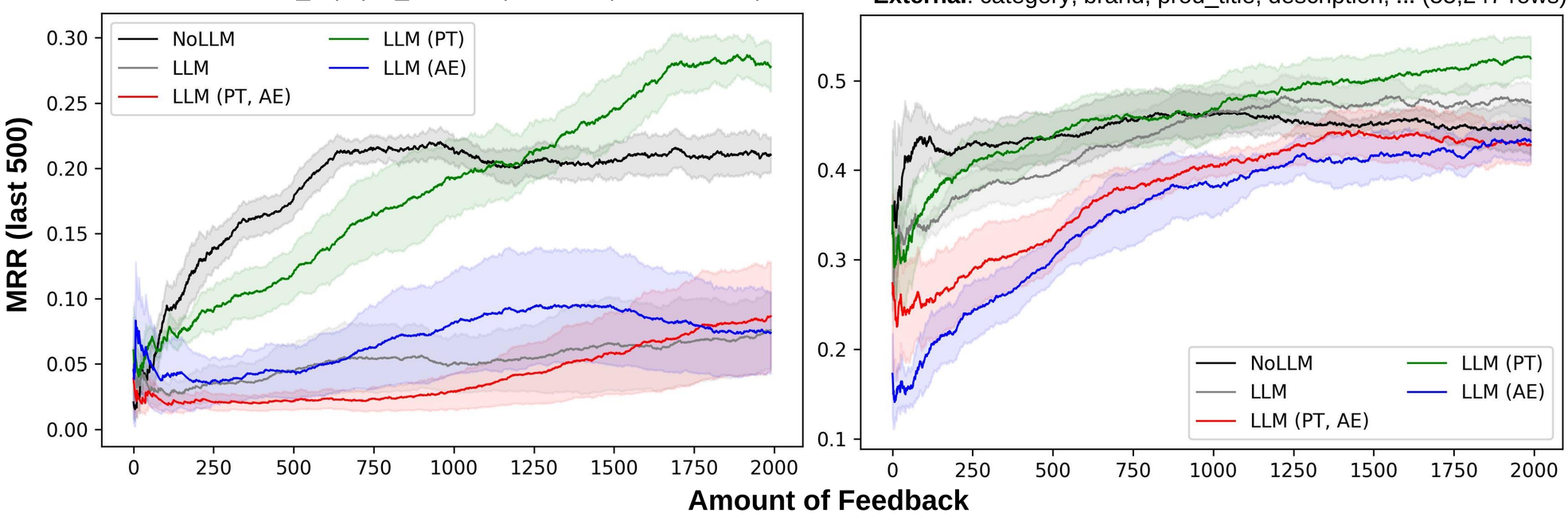
Local: abstract (250,575 rows)

External: sha, source_x, paper_title, doi, pmcid, ... (340,826 rows)

WDC: products scraped from various sites

Local: category, brand, prod_title, description, ... (57,109 rows)

External: category, brand, prod_title, description, ... (55,247 rows)



Results

Prefix Tuning:
- Small overhead
- Boosts long-run performance

Attribute encoding:
- Not as effective
- May harm performance

More

Models, techniques, results, source code and datasets