1st International Workshop on Databases and Large Language Models

Generating Data Augmentation Queries Using Large Language Models

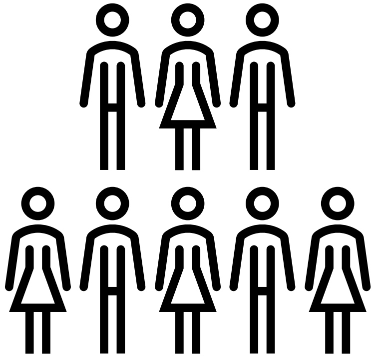
Christopher Buss, Jasmin Mousavi, Mikhail Tokarev, Arash Termehchy, David Maier, Stefan Lee

Oregon State University

Portland State University

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Drug Repositioning Can Save Lives



Patients with Castleman's disease

- Rare disease
- Potentially fatal: causes <u>severe inflammation</u>
- No effective treatments currently exist

?

Unfortunate reality:

✗ **Too rare**: no financial incentive for companies to develop treatments

Alternative:

✓ Find an existing drug to treat Castleman's disease

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

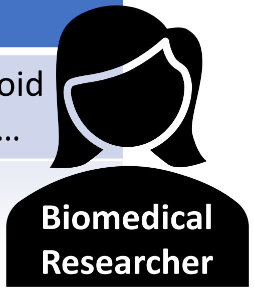
Identify a Candidate Drug

Find a candidate drug

FDA-Approved Drugs

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

Local Data Source



Castleman's causes severe inflammation…

Humira is used to treat conditions involving severe inflammation

Candidate drug: Humira

Next step: gather more information about Humuria:

- Will it help or hurt?

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Find External Sources

Local entity:

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …

FDA-Approved Drugs

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

Local Data Source

Biomedical Researcher

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with adalimumab …
Etanercept	Etanercept binds specifically to tumor …

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity to tumor …
$C_{2224}H_{3475}N_{621}O_{698}S_{36}$	There are two distinct receptors …

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

What we Want: Info Relevant to Humira

Local entity:

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …

FDA-Approved Drugs

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

Local Data Source

Biomedical Researcher

Relevant external entities

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with adalimumab …
Etanercept	Etanercept binds specifically to tumor …

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity to tumor …
$C_{2224}H_{3475}N_{621}O_{698}S_{36}$	There are two distinct receptors …

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Augment Humira With that Relevant Info

brand_name	class	uses	adverse_effects
Humira	TNF inhibitor	rheumatoid arthritis …	After treatment with adalimumab …

mechanisms
Binds with specificity to tumor …

FDA-Approved Drugs

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

Local Data Source

Biomedical Researcher

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with adalimumab …
Etanercept	Etanercept binds specifically to tumor …

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity to tumor …
$C_{2224}H_{3475}N_{621}O_{698}S_{36}$	There are two distinct receptors …

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Manually Querying for Relevant External Entities

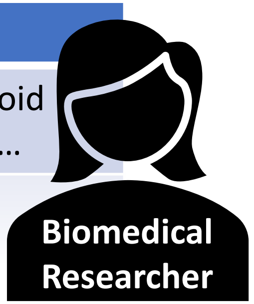
Challenges:

- Many external data sources
- Data heterogeneity: different representations
  - **Humira** = **Adalimumab** = $C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$ = ???

FDA-Approved Drugs

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

Local Data Source



Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with
Et	

Search

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity
C_{2}	

Search

External Data Source

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

1st Try: Query = Too Specific to Local Source

Query = no relevant entities!

FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheu arth
Enbrel	TNF inhibitor	pl

Local Data Source

Biomedical Researcher

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with
Et	

Humira | Search

0 results

External Data Source

Bio Compounds

formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity
C_{2}	

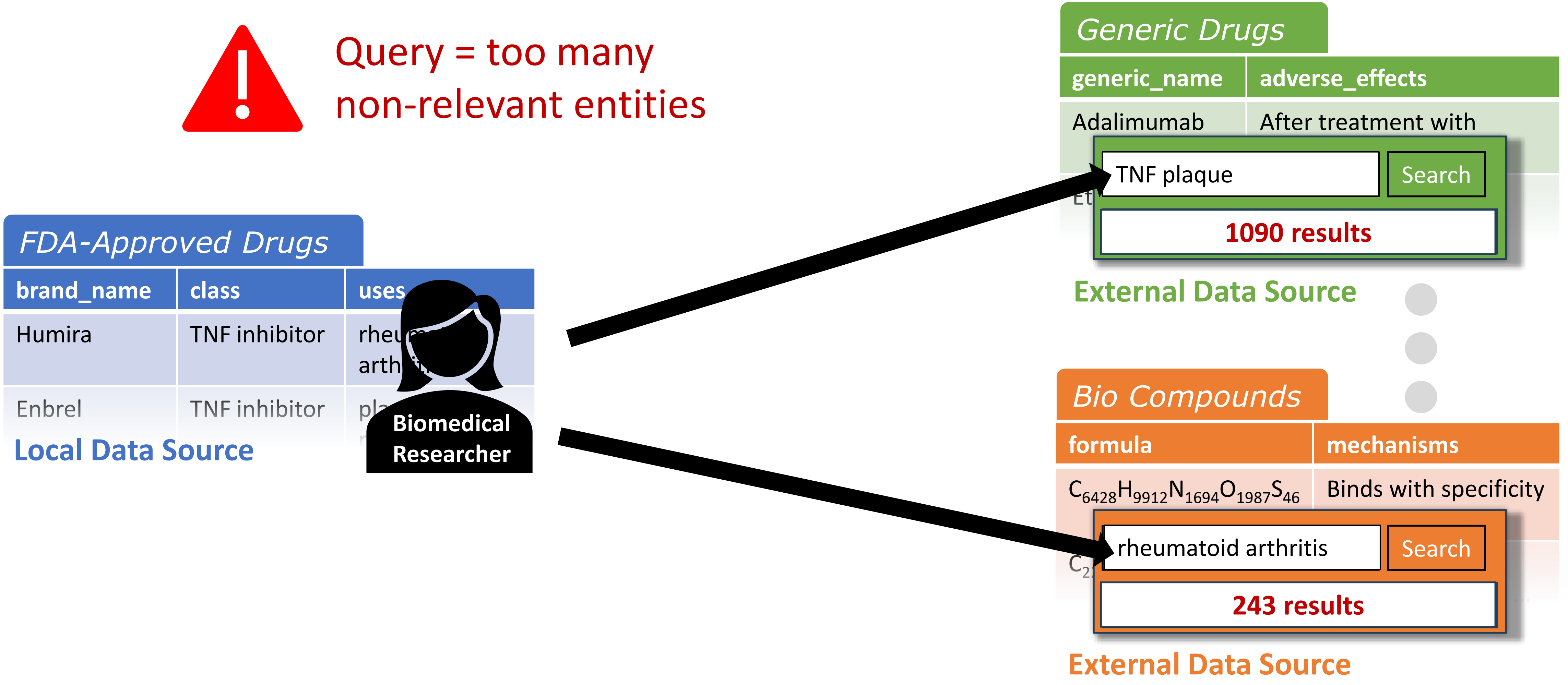
Humira | Search

0 results

External Data Source

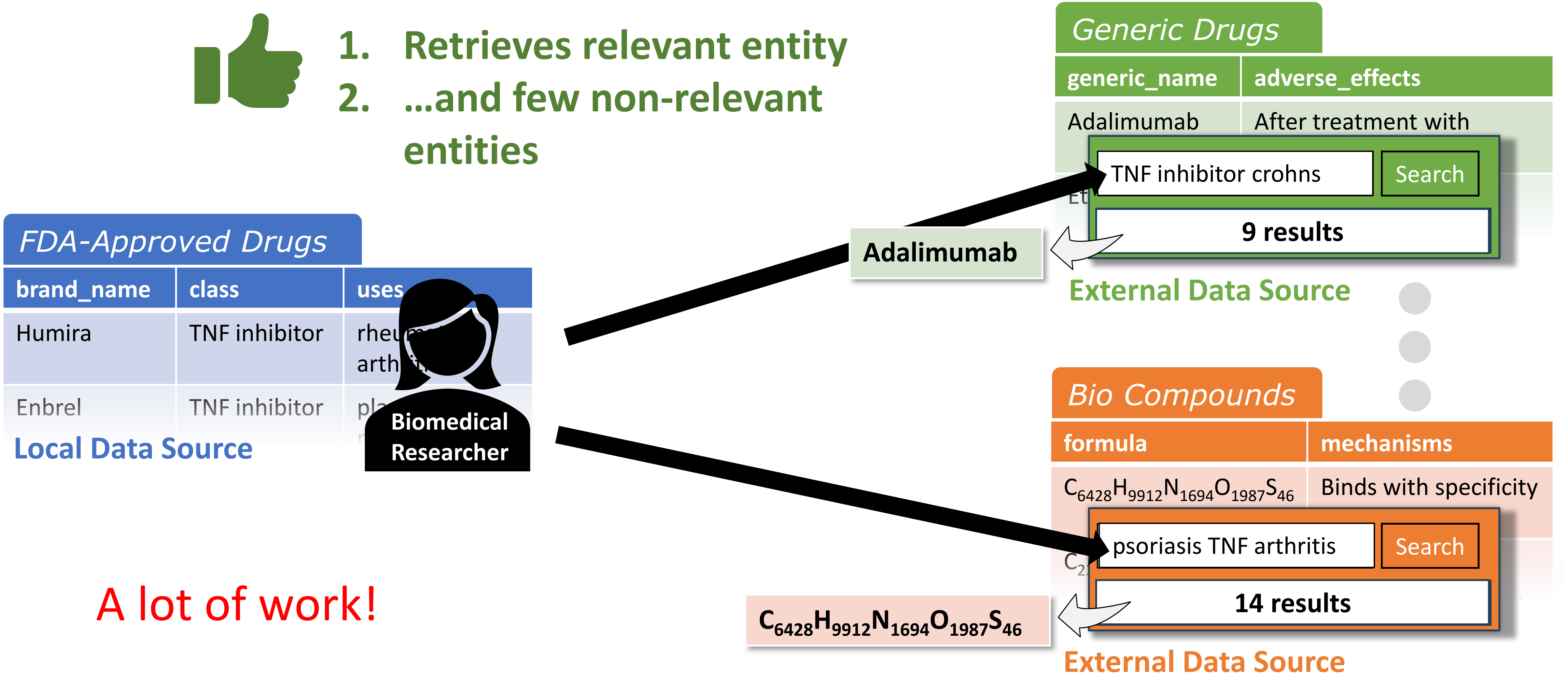
[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

2^{nd} Try: Query = Too General



Nth Try: Just Right!

1. Retrieves relevant entity
2. …and few non-relevant entities



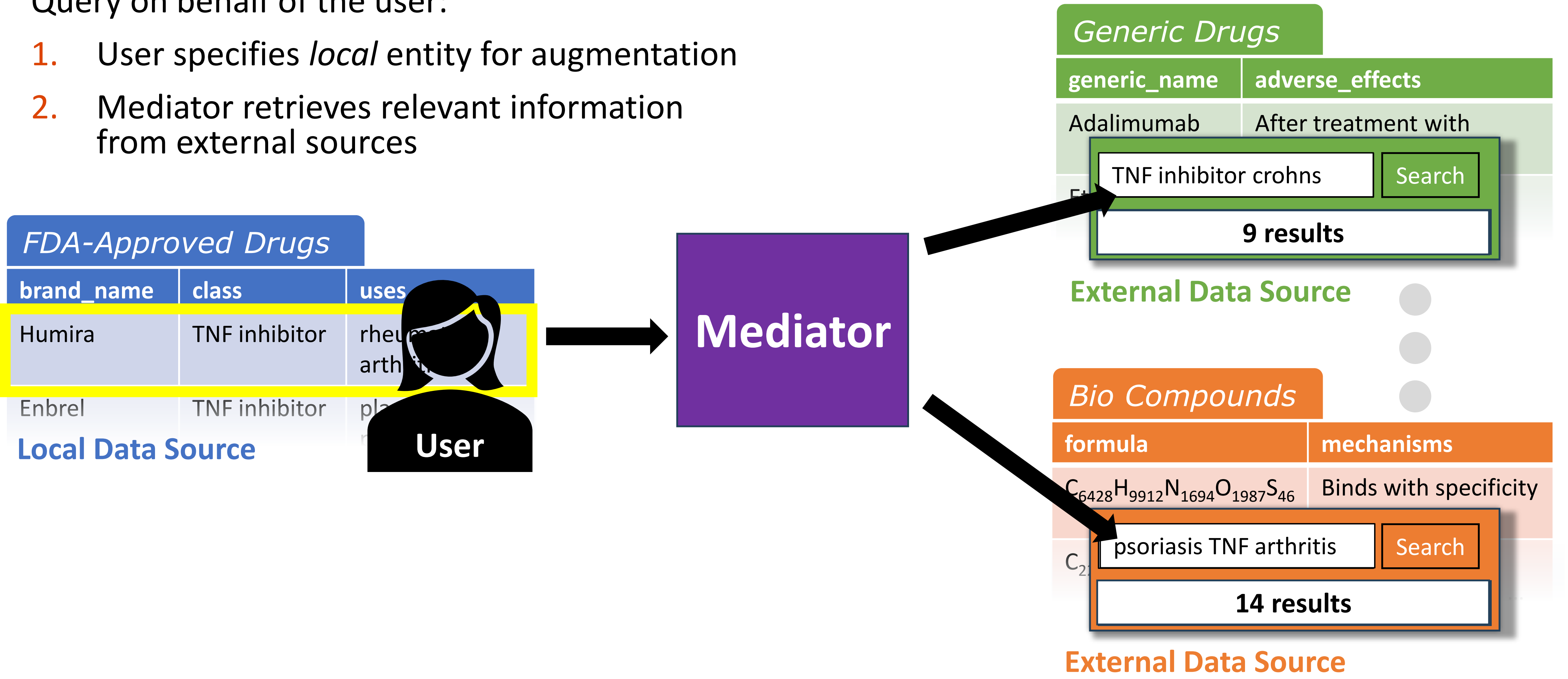
A lot of work!

[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Alternative: Use a Mediator

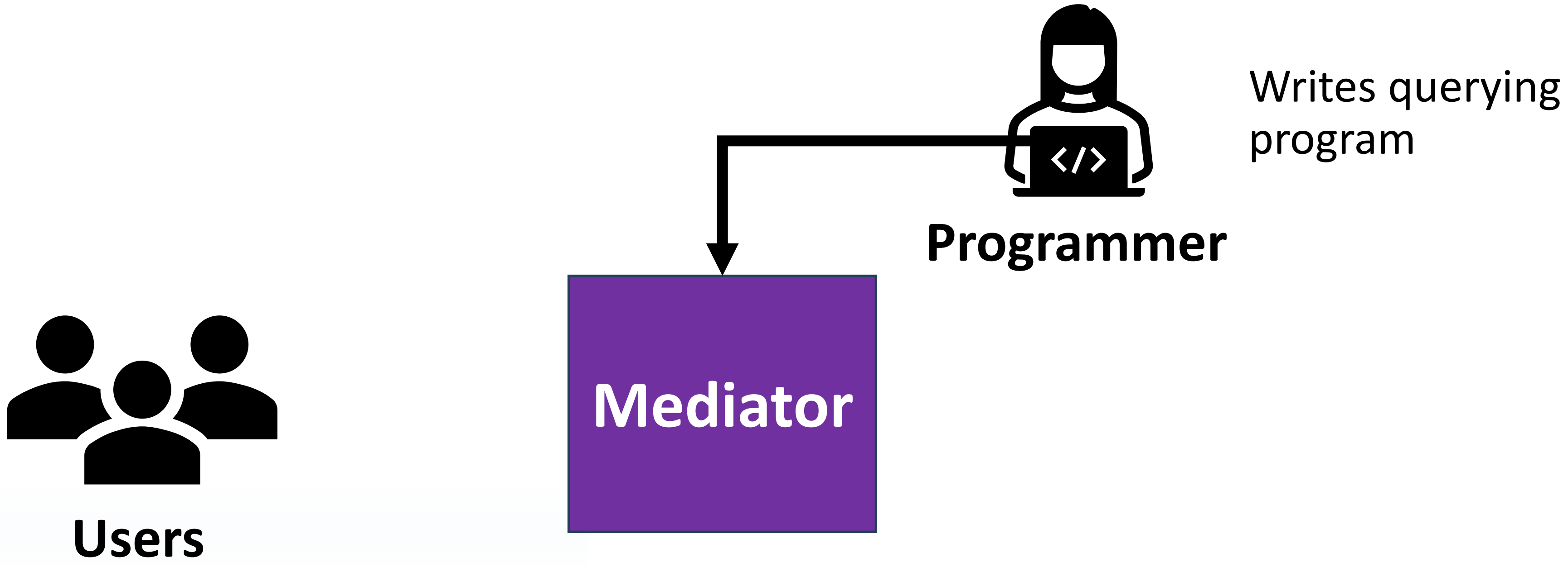
Query on behalf of the user:

1. User specifies *local* entity for augmentation
2. Mediator retrieves relevant information from external sources

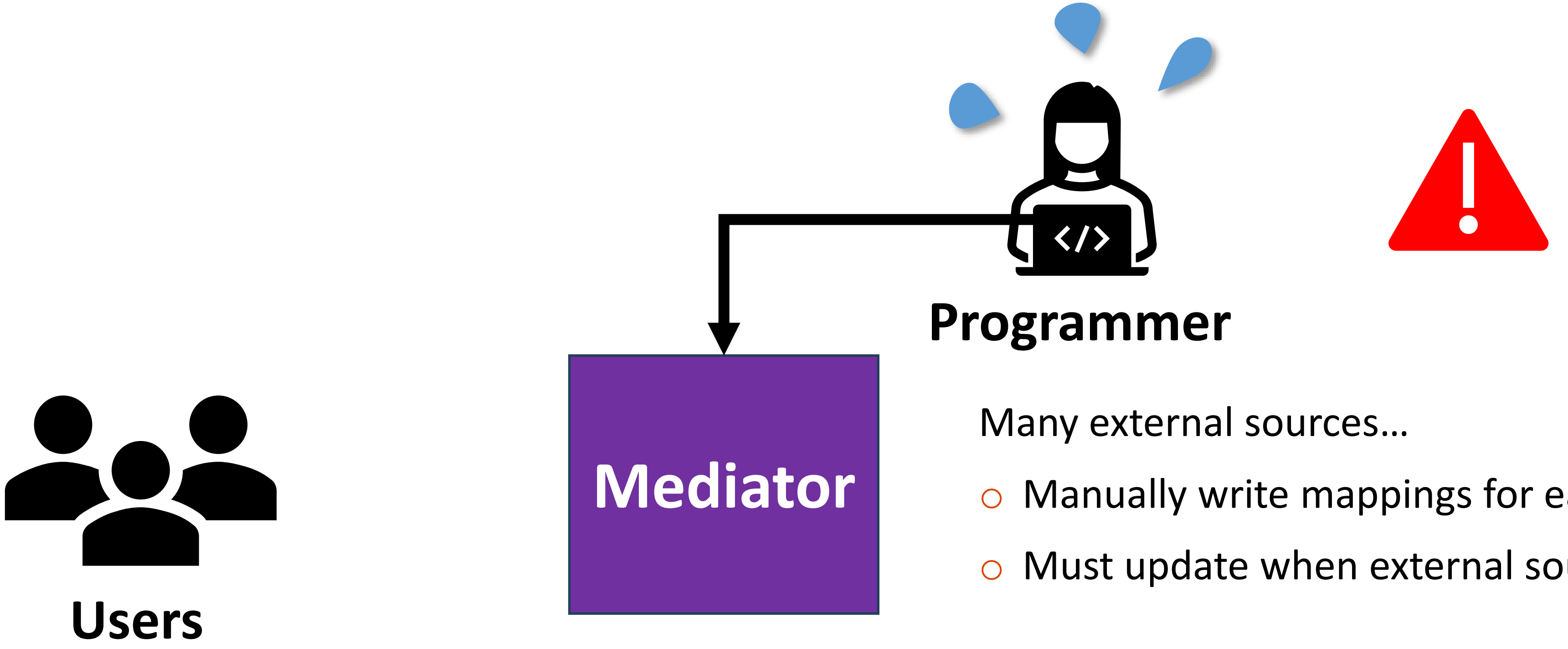


[Rare Daily Staff 2023] & [Ashburn & Thor 2004]

Existing Work: Mediator Written By Hand



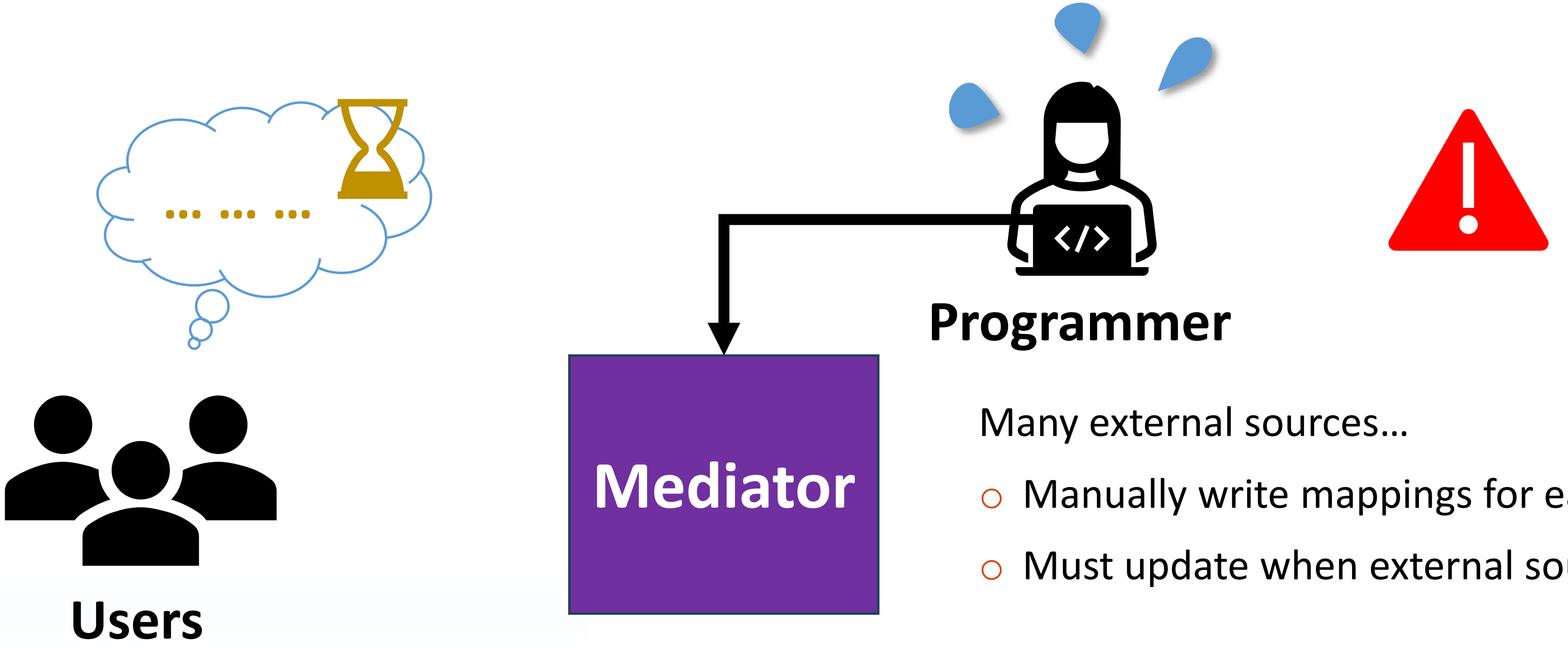
Existing Work: Lots of Work!



Many external sources…

- Manually write mappings for each source
- Must update when external sources change

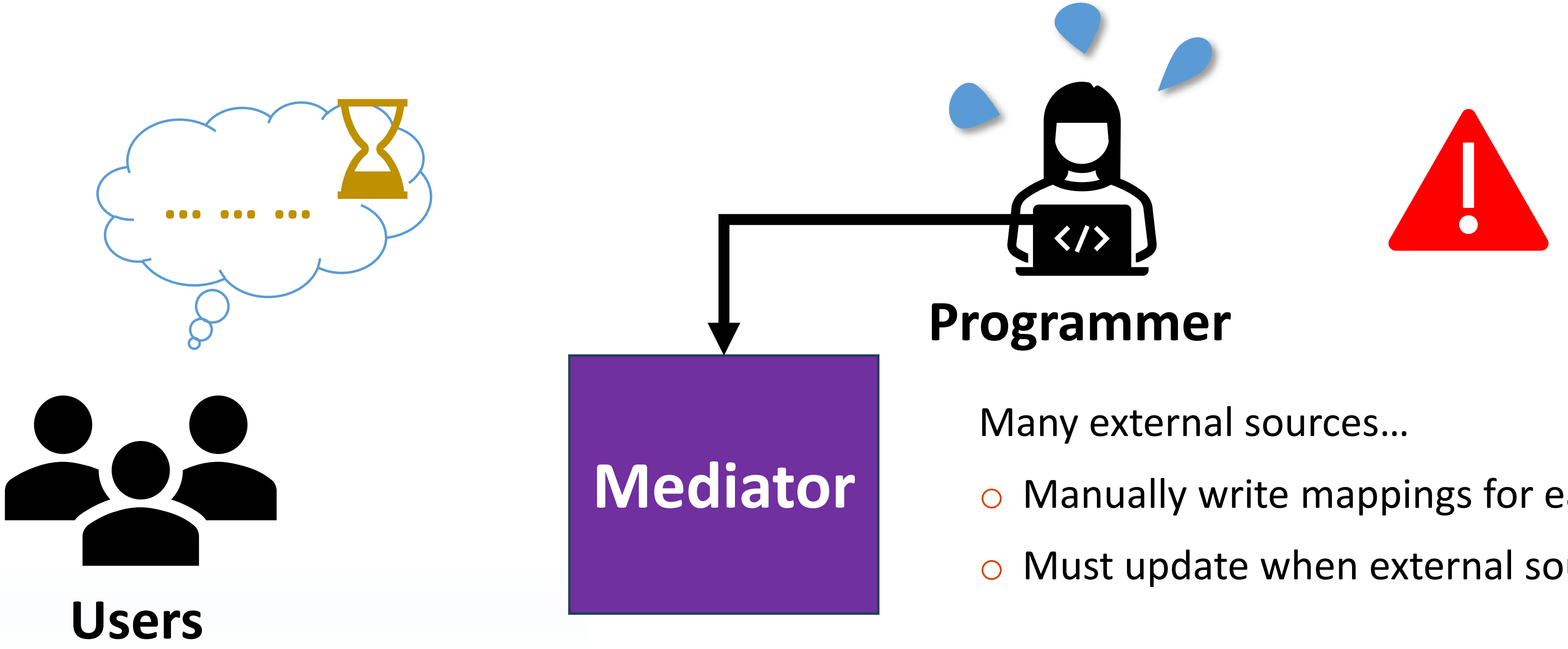
Existing Work: Information Delays



Many external sources…

- Manually write mappings for each source
- Must update when external sources change

Existing Work: Resource Intensive!



Many external sources…

- Manually write mappings for each source
- Must update when external sources change

For example: the NIH funds a consortium of such systems (~14 systems)

- Just one system has 73 external datasources and millions of entities
- Costs NIH **US$923 million per year!**

Our Approach: Learn a Mediator

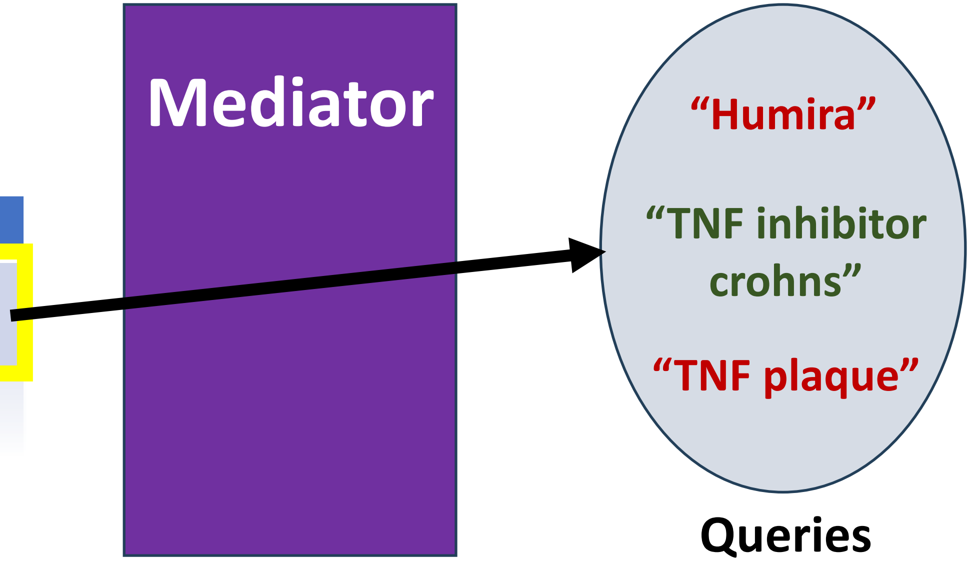
Reduce work and delays of writing the mediator *by hand*

Learn Mediator that maps **local entity** → **"Just right"** query

FDA-Approved Drugs

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

Local Data Source



keyword queries formed using words from **local entity's** content

How Do We Learn the Mediator?

Offline Learning:

1. Gather training data
2. Train mediator
3. Users query mediator

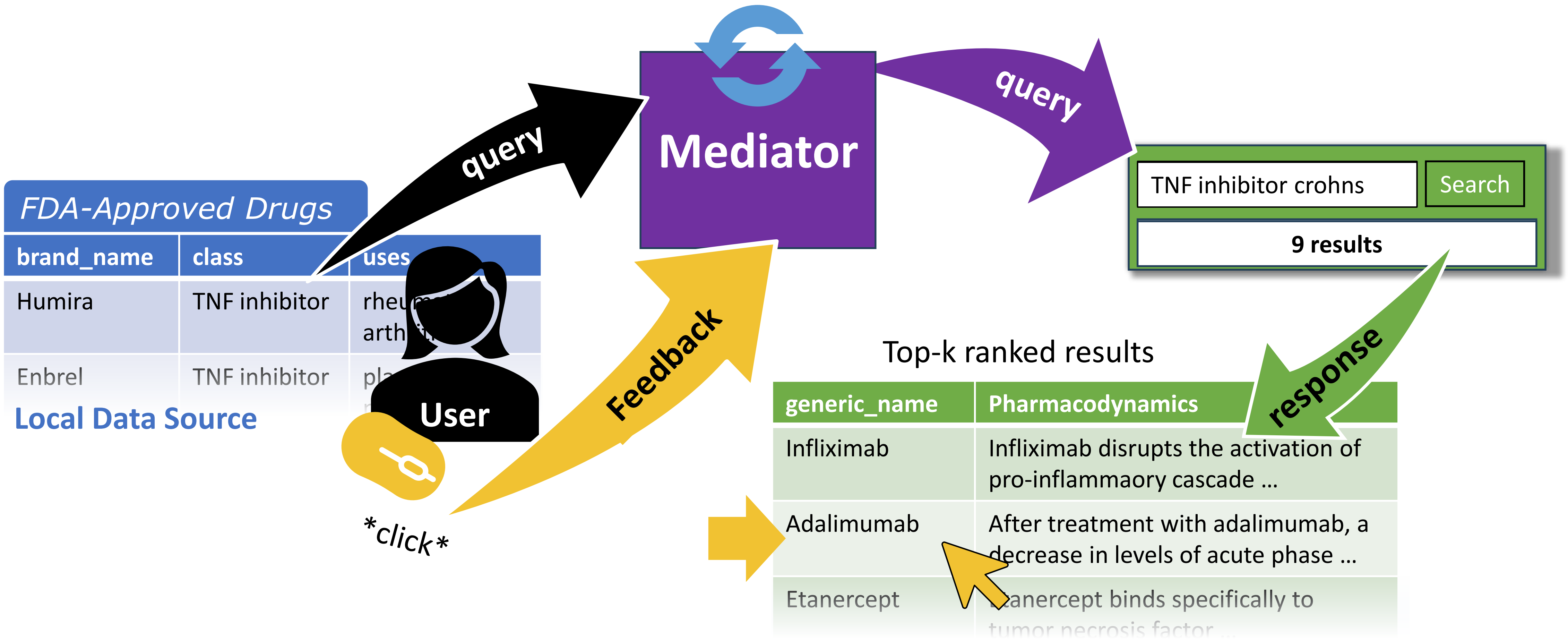
- Lots of expensive work
  - Hire domain experts to label data
  - External source updates → must repeat!
- Still delays…

Online Learning:

- Train mediator *while* users query it

Online Learning Framework

Refine understanding of what makes a query good



Predicting Query Quality with f_w

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …

f_w

"Humira"

"TNF inhibitor crohns"

"TNF plaque"

Feedback is used to update w

Design Challenge:

Short-Run Success: find sufficiently good queries quickly
- Users must remain engaged with the system

Long-Run Success: should *continue* to improve over time
- Fit to domain/diversity of local entities

Leverage pretrained LLMs

Leveraging Pre-trained LLM Priors

High-level idea:

Entire tuple → **Embedding Layer (pre-trained LLM)** → **Prediction layer** → ***Prediction***

Online setting: only know the quality of queries tried
Exploration: try new queries that *may* be better
Exploitation: use queries known to be good

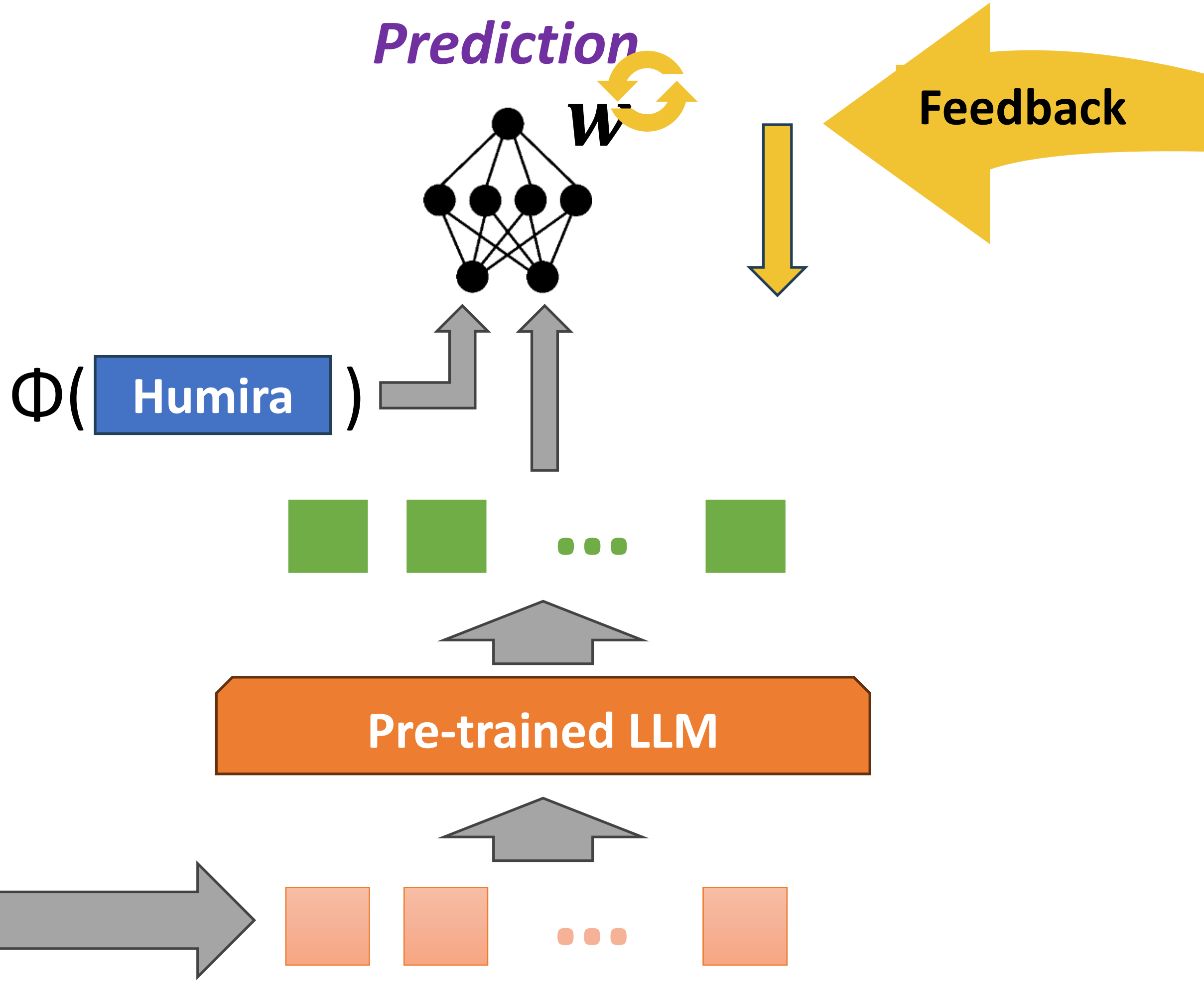
ε-greedy: with (1-ε) probability, select query with highest predicted quality
with ε probability, select random query

Model V. 1.0 (Prior Work)

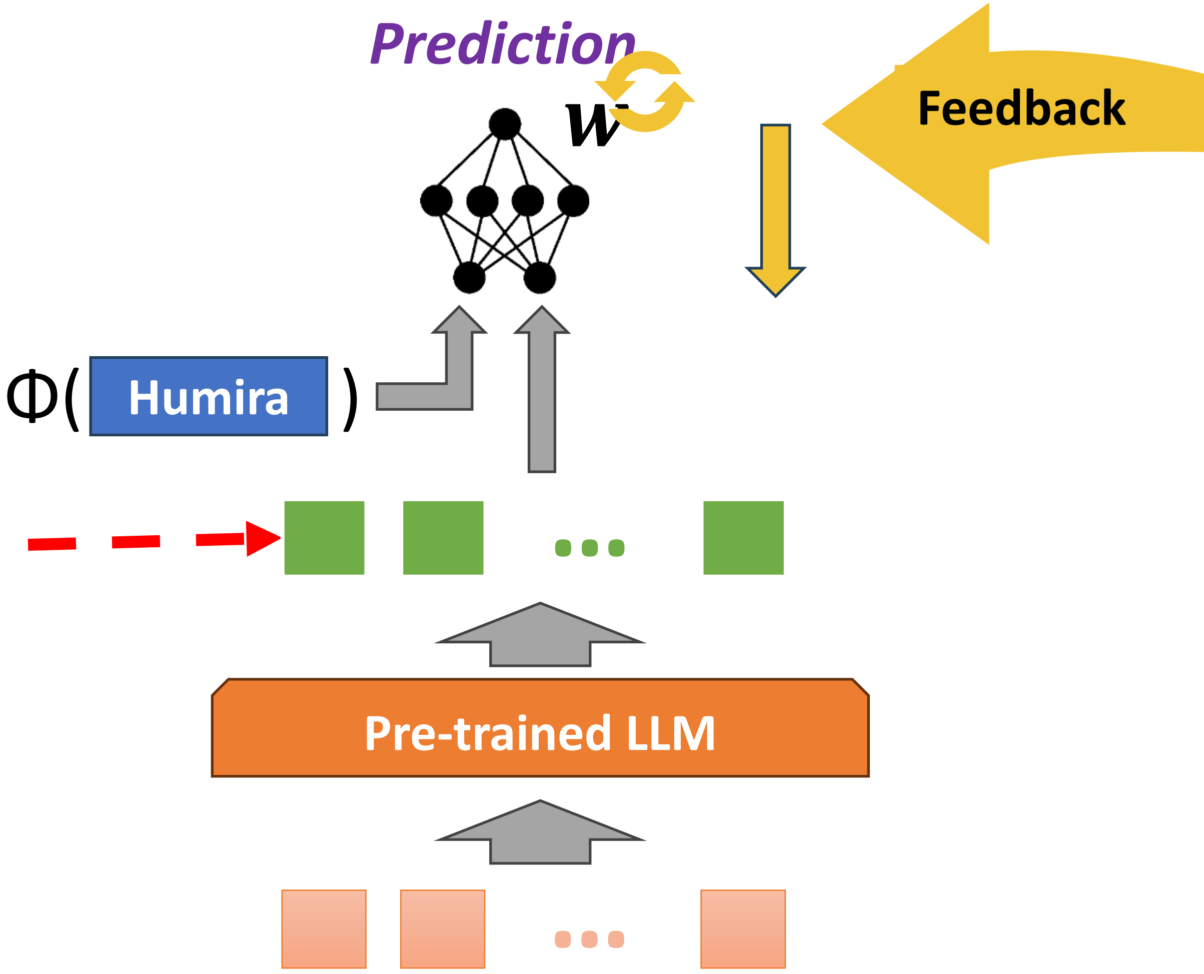
1. Concatenate terms into one string
2. Tokenize and embed
3. Get contextualized embedding
4. Inject features (lexical, distributional, and semantic)
   - Includes structural information
5. Predict quality using MLP head

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis …

"Humira TNF inhibitor rheumatoid arthritis … "



Embeddings Not Aligned With Domain/Task



Not adjusting embeddings

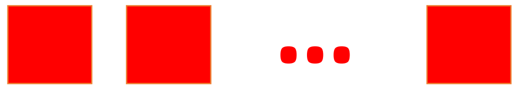
Fine-tuning = expensive;

- Update LLM weights
- Not fit for **short-run**

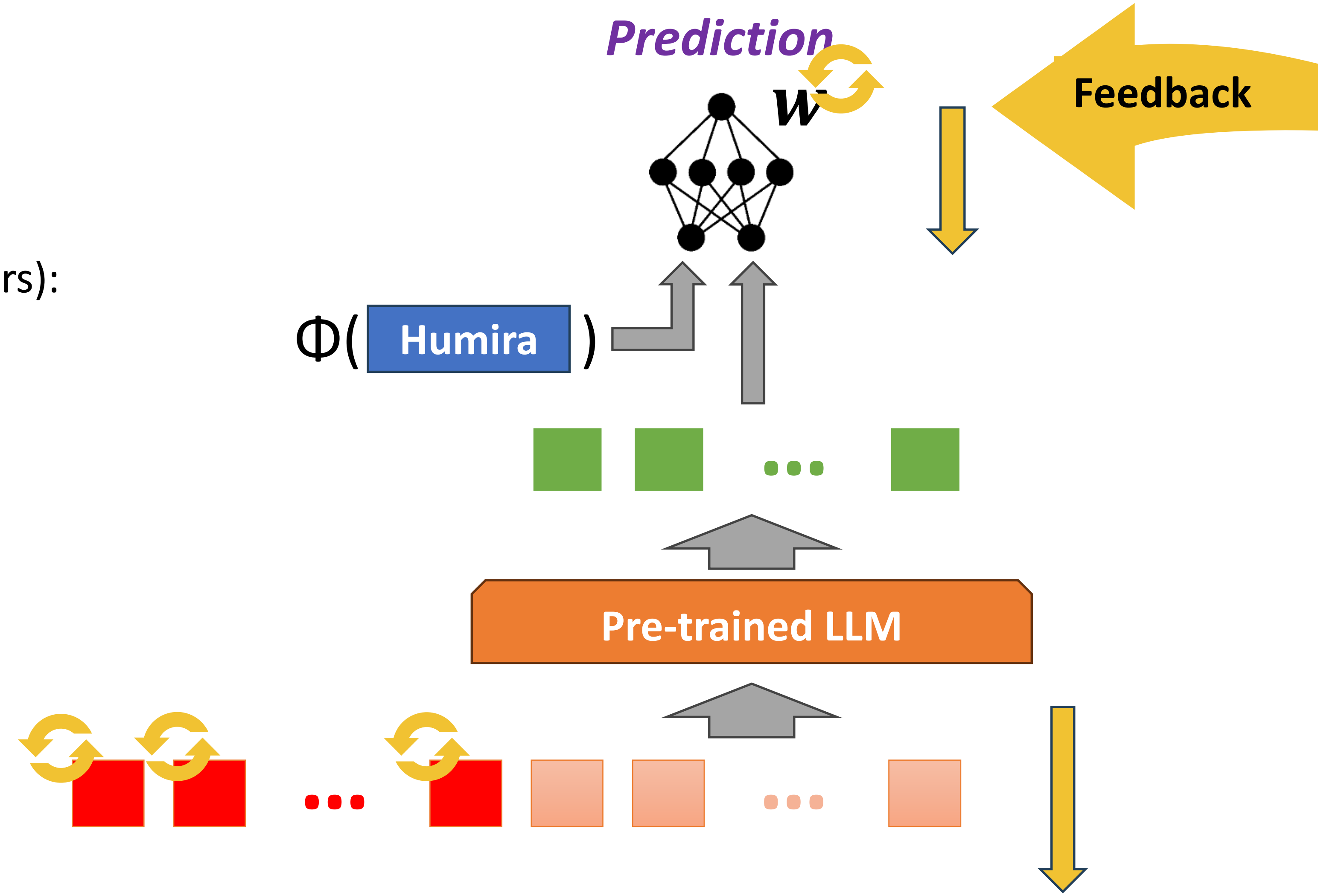
Prefix Tuning: Tuning for Domain and Task

Parameter-efficient approach:

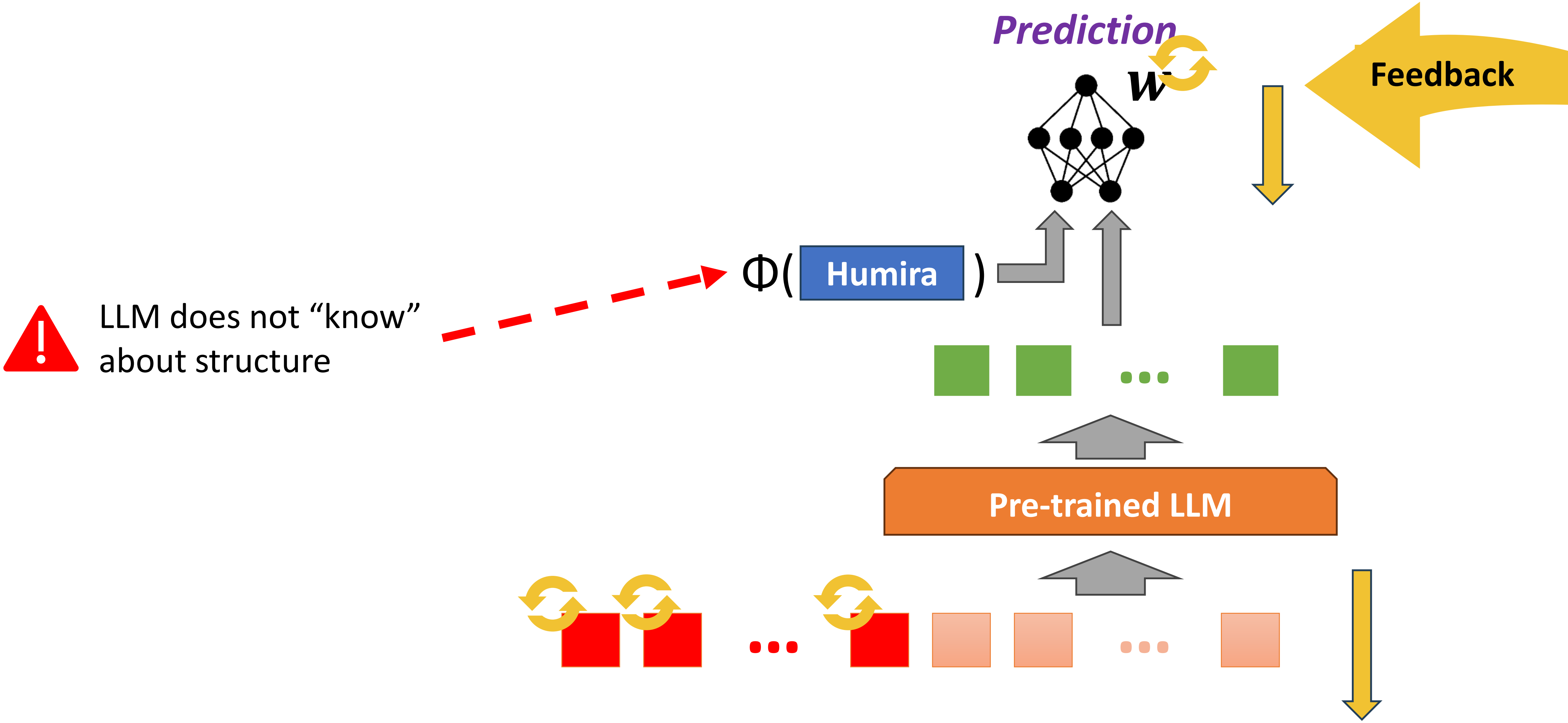
1. *p* “prompts” (continuous vectors):



2. Prepend onto pre-LLM input
   - Learned contextualization

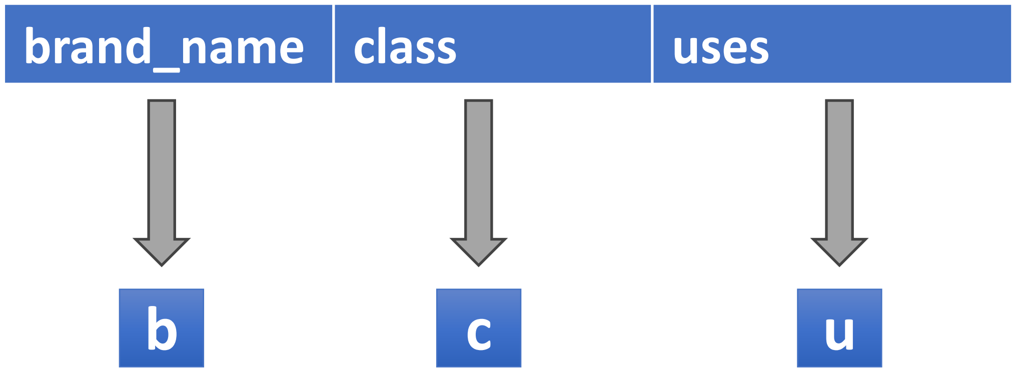


Structure Introduced Too Late!

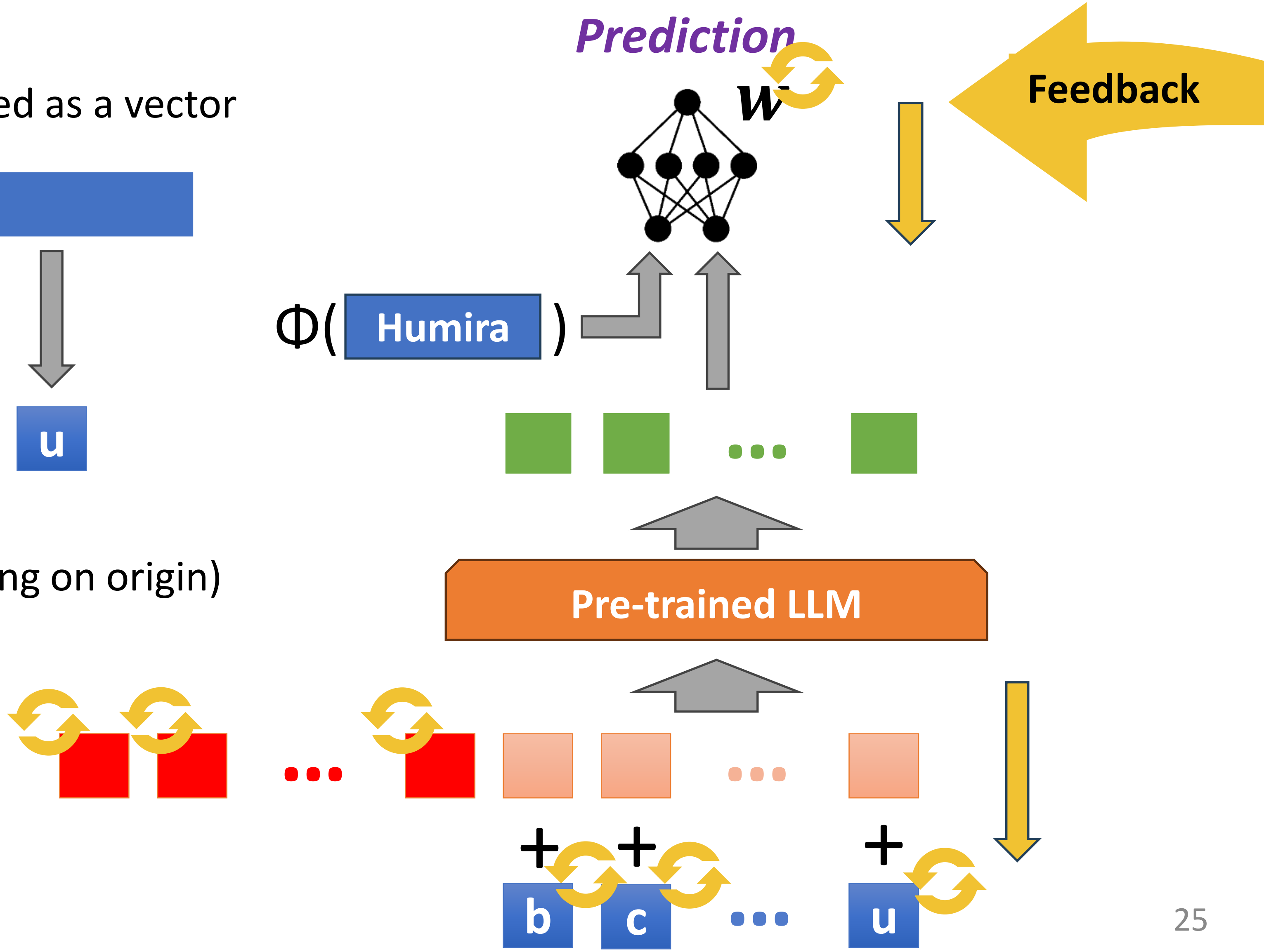


Attribute Encoding: Fusing Structure with Input

1. Each attribute (column) encoded as a vector



2. Add to pre-LLM input (depending on origin)



Empirical Study Setup

Dataset	Source	Desc.	#entities
Drugs	Local	Drug reviews	13,725
	External	Wikipedia summaries of drugs	46,976
WDC	Local	Products	57,109
	External	Products	55,247
ChEBI	Local	Molecular information specific to drugs	5,483
	External	Molecules and their effects on living organisms	189,467
CORD-19	Local	Abstract	250,575
	External	Title, authors, etc,.	340,826

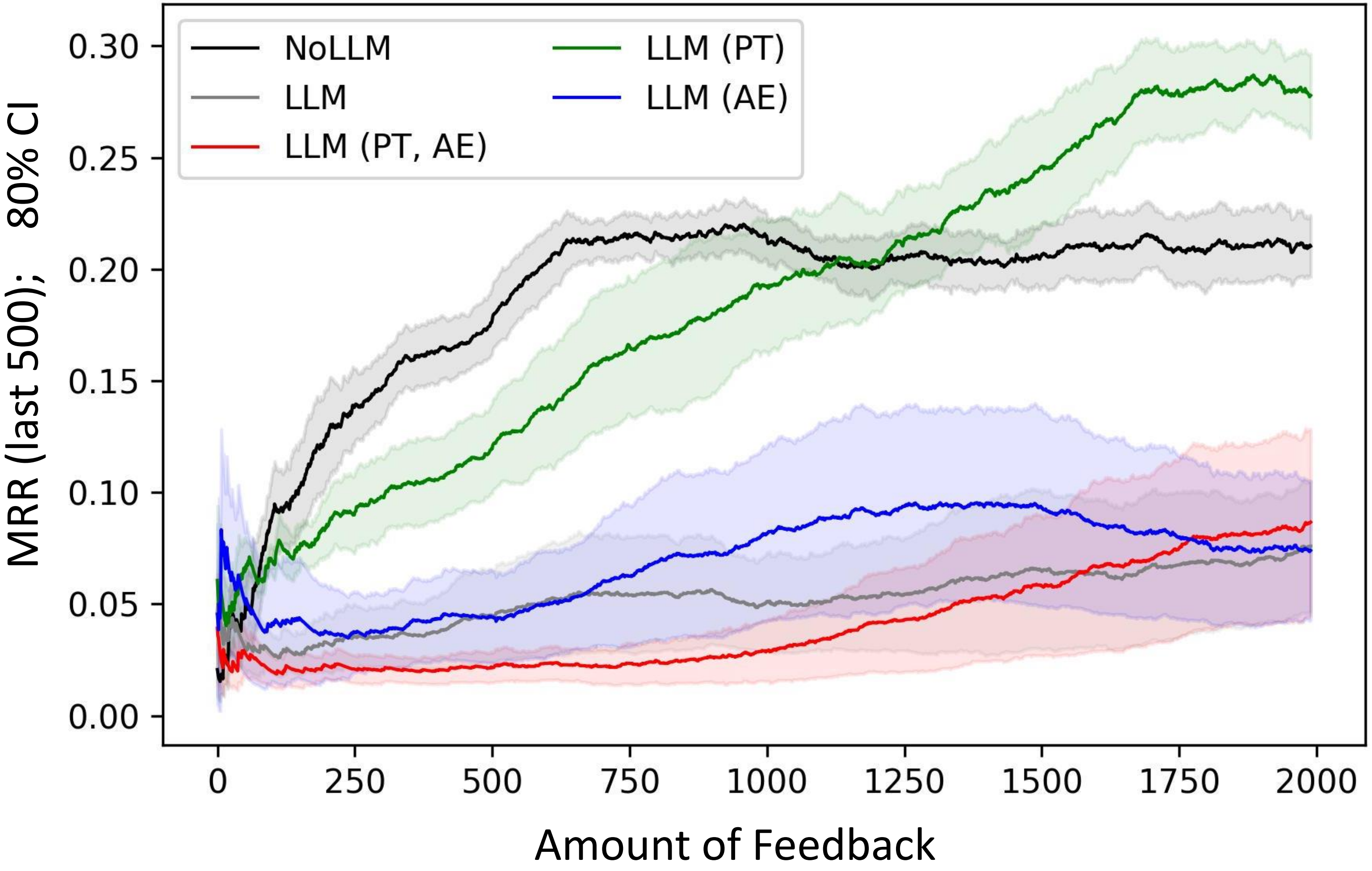
Run simulations over variety of domains

- Ground truth = feedback
- LLM = Longformer

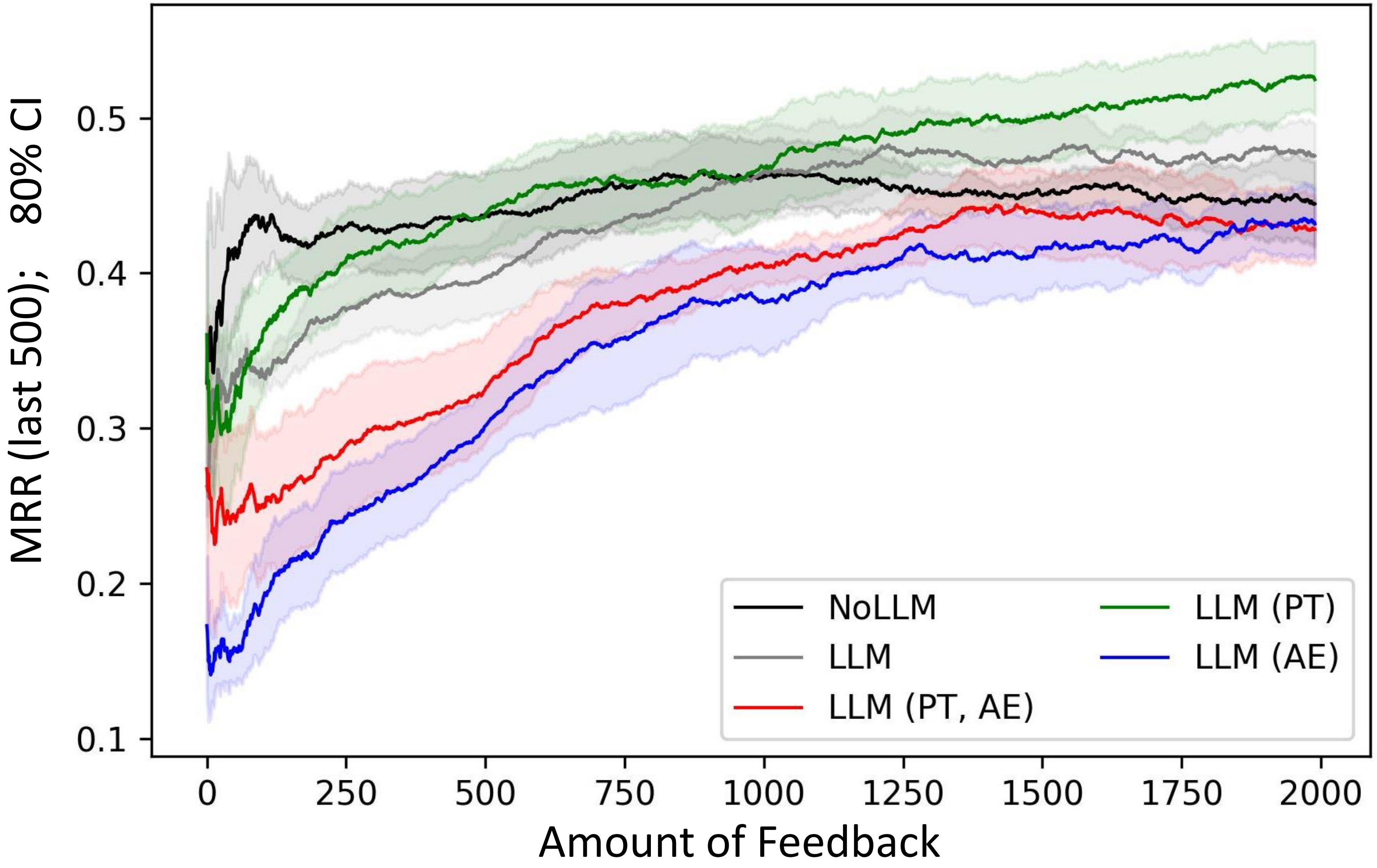
QUESTION:

- Does attribute encoding help?
- Does prefix-tuning help?

Comparing Enhancements: CORD-19



Comparing Enhancements: WDC



Future Work

- Better fuse domain-specific knowledge with pre-LLM input

- Generate structured queries (SQL, graph-based)
  - Weakly supervised semantic parsing + Short-run challenge = **very hard!!**
  - LLMs have strong performance for few-shot learning
    - Strong prior for complex queries

Takeaways

Motivation/Setup

- Mediators require a lot of resources to build/maintain by hand
- Learn the mediator online using user feedback!
- Pre-trained LLM (V. 1) prior

Enhancements

- Enhancements:
  - Prefix-tuning
  - Attribute encoding

Experiments

- Prefix-tuning beats or meets V. 1 performance
- Attribute encoding may degrade performance

Thank you!

Please share your questions!

Oregon State University

Portland State University