Effective Entity Augmentation By Querying External Data Sources

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

Oregon State University

Portland State University

IDEA Lab
Information and Data Management and Analytics

1. Drug Repositioning Can Save Lives

Patients with Castleman's disease
- ○ Rare
- ○ Potentially fatal: causes severe inflammation

Unfortunate reality:
✗ **Too rare**: no financial incentive for pharmaceutical companies to develop effective treatments

Alternative:
✓ Find an existing drug to treat Castleman's disease

Step 1: find a candidate drug

FDA-Approved Drugs

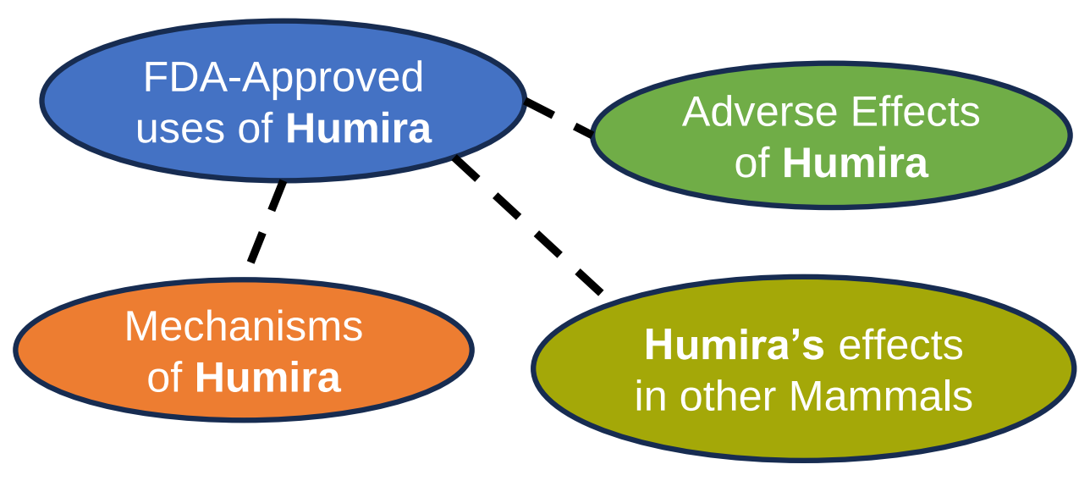
brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis, psoriatic arthritis …
Enbrel	TNF inhibitor	plaque psoriasis, ankylosing spondy…

Local Data Source

Biomedical Researcher

Humira is also used to treat conditions involving severe inflammation

Step 2: augment it with important information from external data sources


FDA-Approved uses of Humira
Adverse Effects of Humira
Mechanisms of Humira
Humira's effects in other Mammals

Manually Querying for External Information

Goal: Query for *data* about Humira from external sources

Challenge:
- ○ Data heterogeneity: each source = different representation
  - ○ Humira = Adalimumab = $C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$ = ???
- ○ Many sources to query!

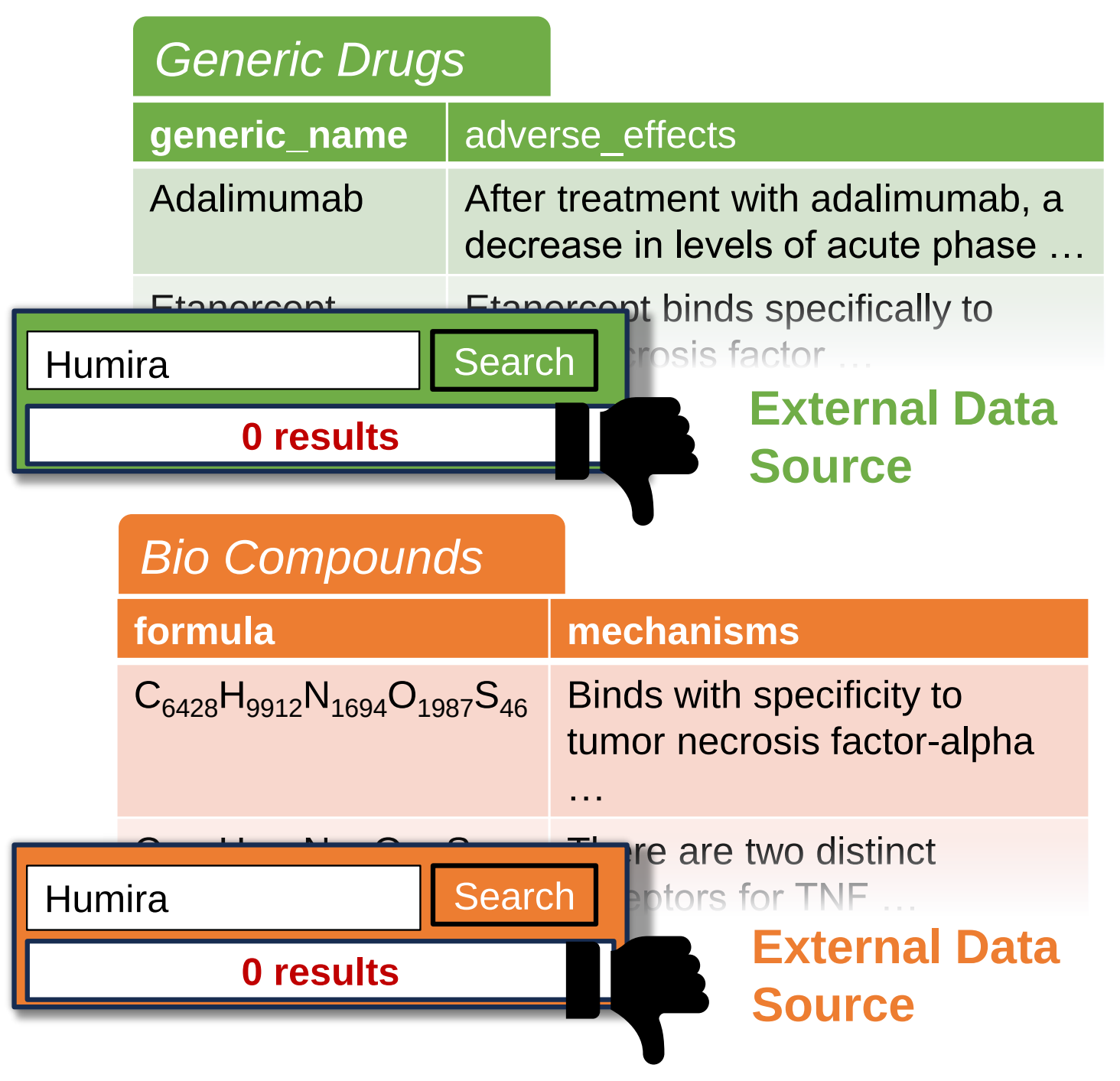
FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis, psoriatic arthritis …
Enbrel	TNF inhibitor	plaque psoriasis, ankylosing spon…

Local Data Source

Can Humira treat *Castleman Disease*?

Biomedical Researcher



Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with adalimumab, a decrease in levels of acute phase …
Etanercept	Etanercept binds specifically to …osis factor …

External Data Source

Bio Compounds

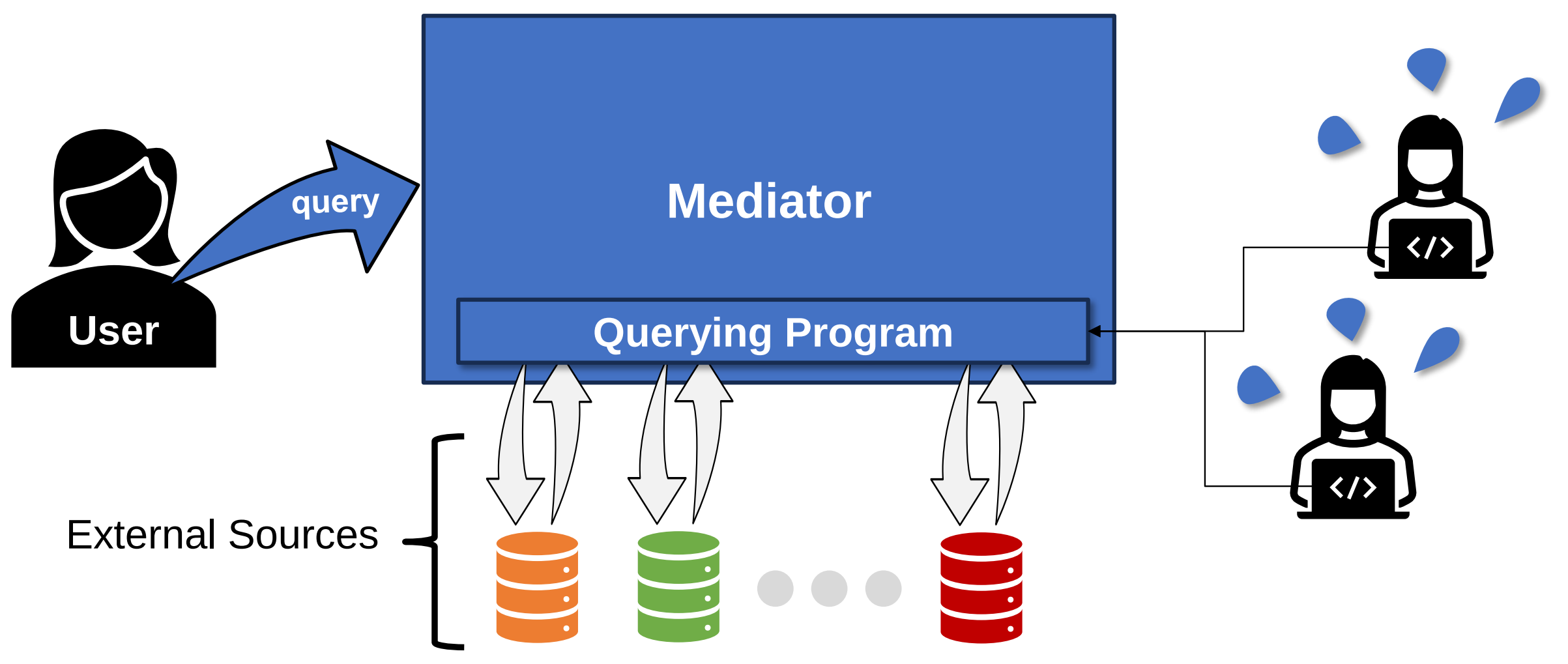
formula	mechanisms
$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$	Binds with specificity to tumor necrosis factor-alpha …
	There are two distinct …ptors for TNF …

External Data Source

2. Existing Work: Write Querying Programs By Hand

Use a **mediator** instead:
1. User specifies some *local* entity for augmentation (e.g., Humira)
2. Mediator returns entity augmented with *external* information

Mediator gathers external information using its own **Querying Program**



A lot of work for programmers
- ○ Write querying programs *for each* external source
- ○ Fix programs whenever sources change

Delayed entity augmentation
- ○ Takes time to build and maintain query programs
- ○ Users must wait for time-sensitive information

In an NIH-funded consortium of such systems (~14)
Just one system has…
- ○ 73 external sources
- ○ Millions of entities

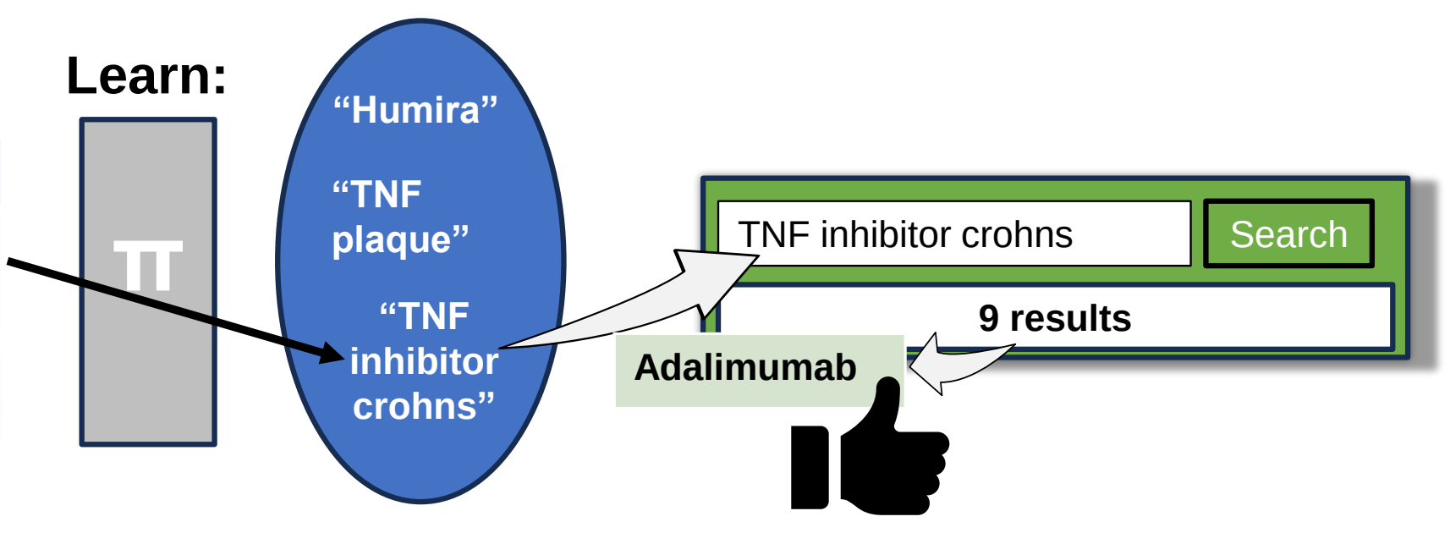
Total: US$923 million per year for development and maintenance

Learn the Mediator Online

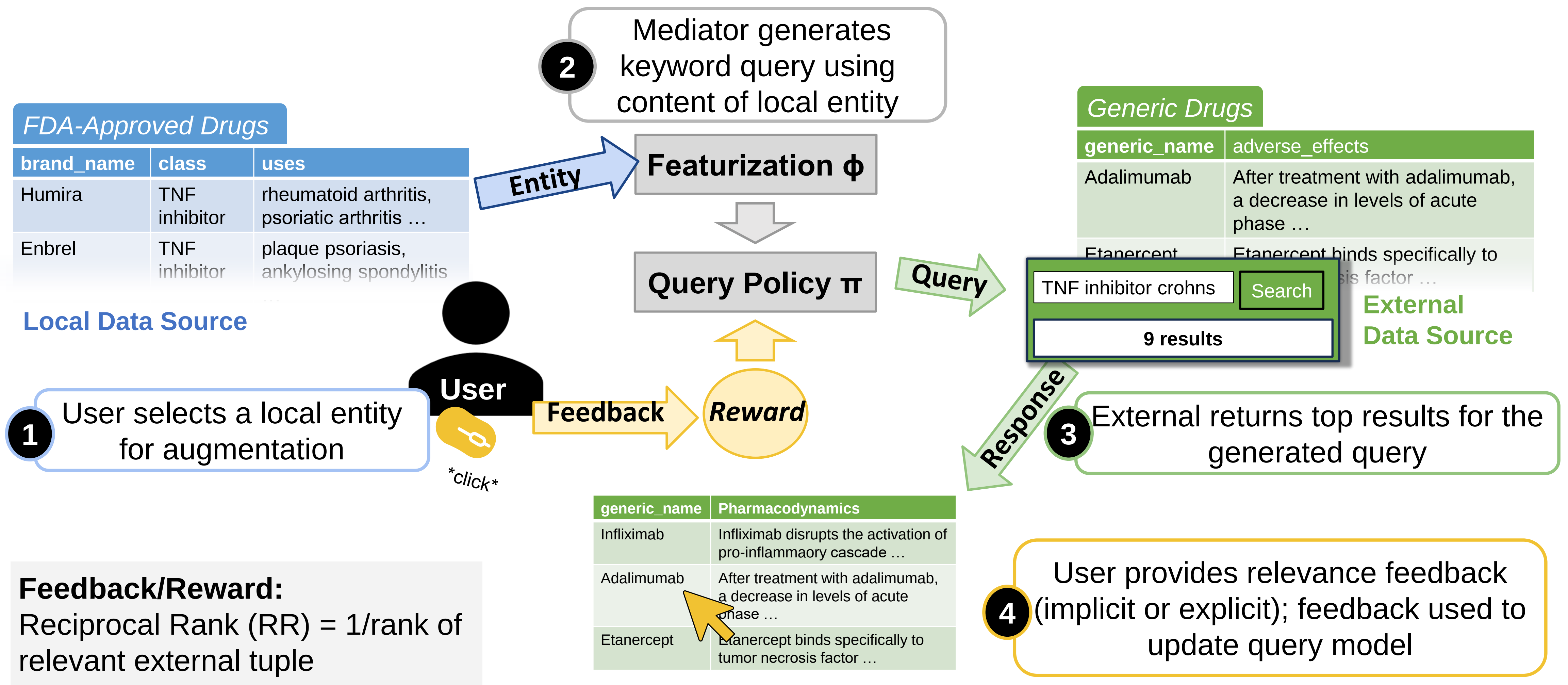
FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis, psoriatic arthritis …
Enbrel	TNF inhibitor	plaque psoriasis, ankylosing spondylitis …

Local Data Source



3. Online Autonomous Querying



FDA-Approved Drugs

brand_name	class	uses
Humira	TNF inhibitor	rheumatoid arthritis, psoriatic arthritis …
Enbrel	TNF inhibitor	plaque psoriasis, ankylosing spondylitis …

Local Data Source

1 User selects a local entity for augmentation

2 Mediator generates keyword query using content of local entity

3 External returns top results for the generated query

4 User provides relevance feedback (implicit or explicit); feedback used to update query model

Generic Drugs

generic_name	adverse_effects
Adalimumab	After treatment with adalimumab, a decrease in levels of acute phase …
Etanercept	Etanercept binds specifically to …sis factor …

External Data Source

generic_name	Pharmacodynamics
Infliximab	Infliximab disrupts the activation of pro-inflammatory cascade …
Adalimumab	After treatment with adalimumab, a decrease in levels of acute phase …
Etanercept	Etanercept binds specifically to tumor necrosis factor …

Feedback/Reward:
Reciprocal Rank (RR) = 1/rank of relevant external tuple

4. Challenges

A balancing act:

Short-Run Success: should learn sufficiently effective queries quickly
- ○ Users must remain engaged with the system

Long-Run Success: should continue to improve over time
- ○ Methods should not waste feedback
- ○ Avoid underfitting

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

5. Dataset-Level (Linear UCB)

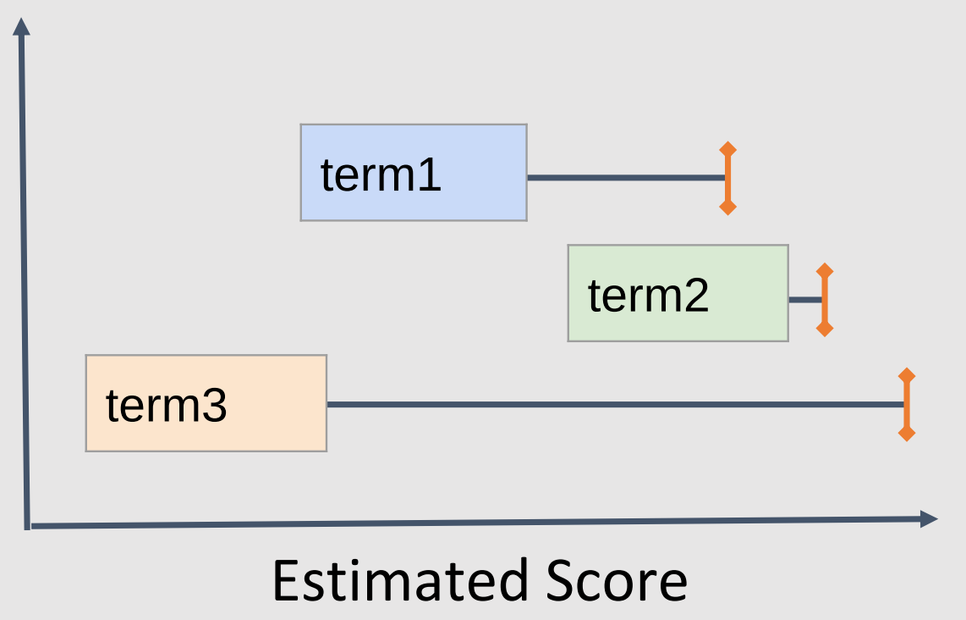
ϕ(inhibitor) **score** = $\theta_0 1\{\text{high IDF}\} + \ldots + \theta_n \text{term\_freq.}$

Learn Θ = $(\theta_0, \theta_1, \ldots, \theta_n)$ over *all* entities

Θ may underfit = poor long-run performance

linUCB
derive upper confidence bound on estimated scores

score = score + λ*UBC



6. Hybrid

Idea: When Θ overfits, diversify with more models!

Start: one Θ_{ALL} for all entities
Over Time: create new Θs for entities that Θ_{ALL} doesn't work for:

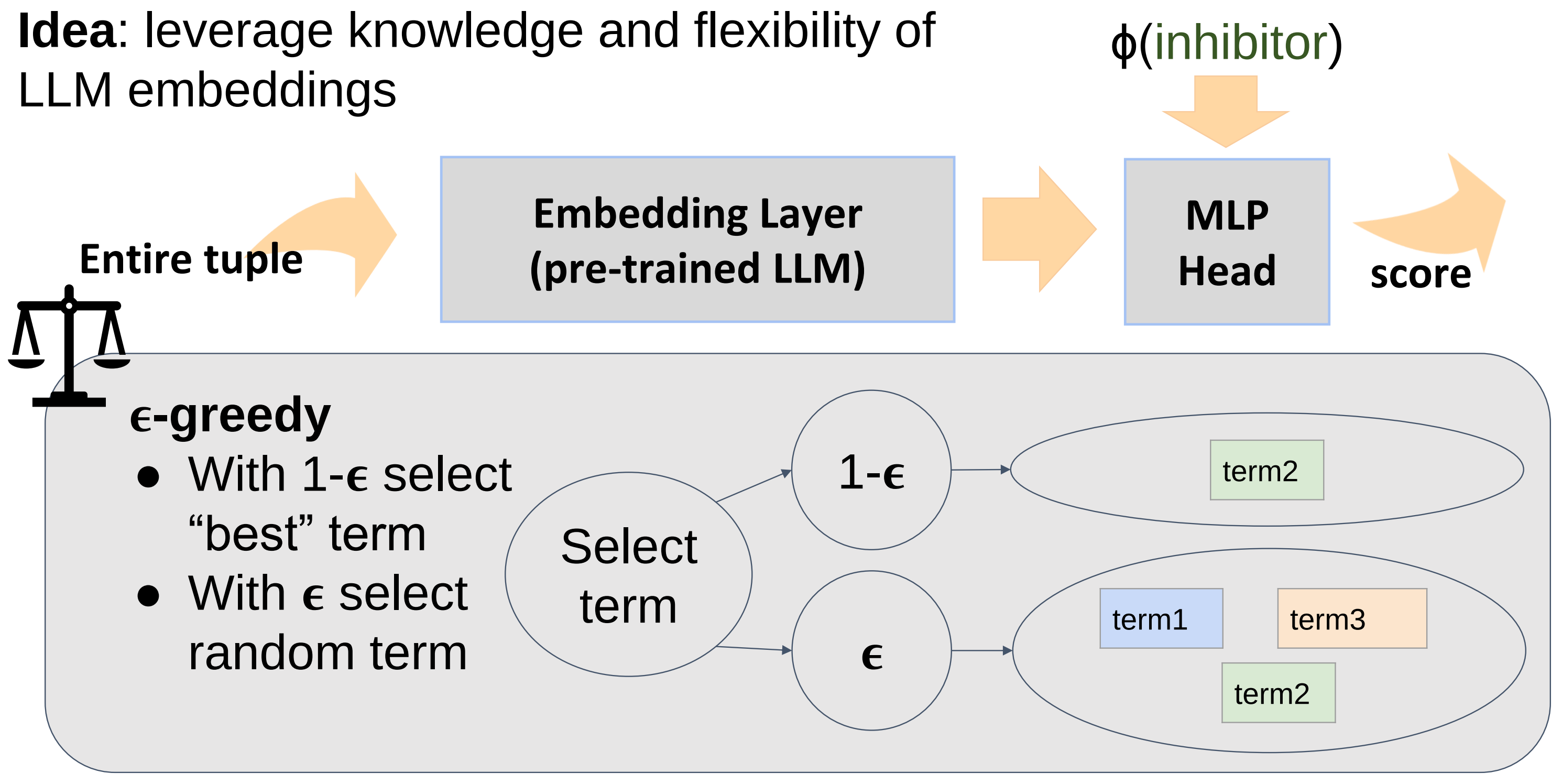
$$\Theta = (\Theta_{ALL}, \Theta_{Humira}, \ldots, \Theta_{Vyvanse})$$

Learn over all other terms

Learn over terms ∊ Humira

7. Leveraging LLM Priors (Longformer)

Idea: leverage knowledge and flexibility of LLM embeddings



ϵ-greedy
- With 1-ϵ select "best" term
- With ϵ select random term

8. Experimental Simulation (Dataset-Level vs. Longformer vs. Hybrid)

Simulate feedback with ground truth | **1 interaction** = 1 feedback cycle | Query length = 4 terms | 95% Confidence region for MRR (avg. over 5 runs)

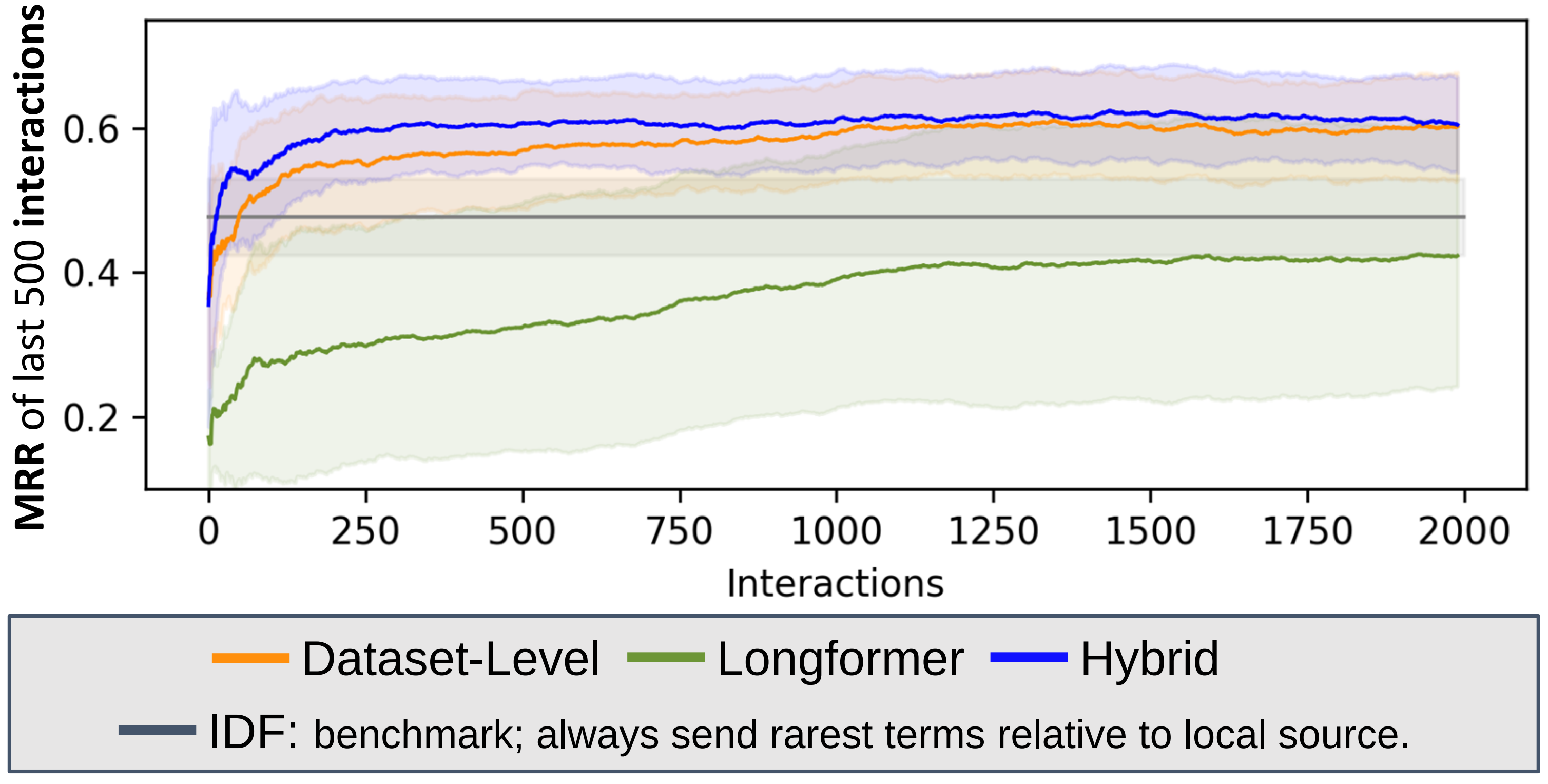
Model Comparison

Dataset-Level
- Good short-run performance
- Bad long-run performance

Longformer
- High variance in performance
- Slower to learn

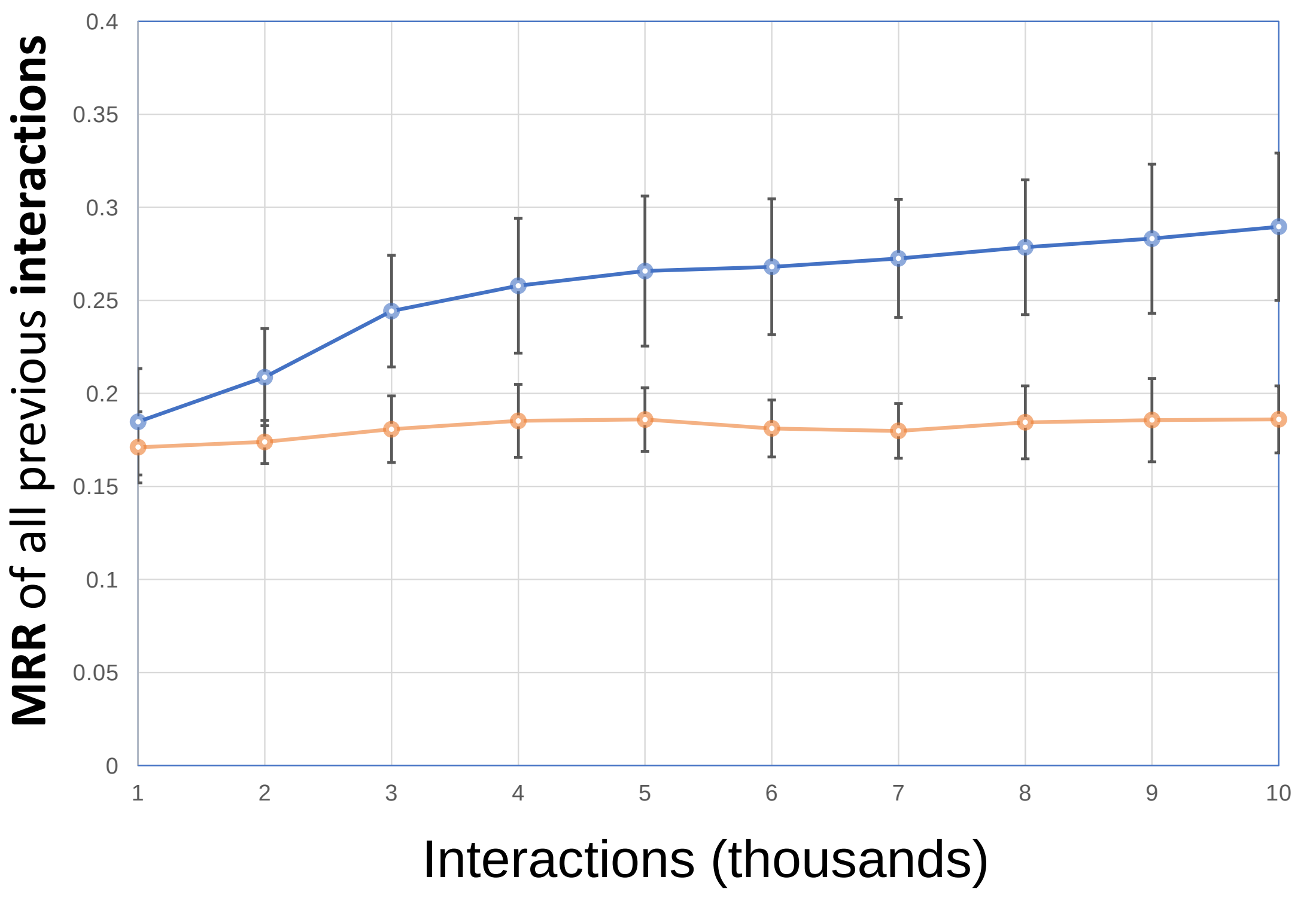
Hybrid
- Statistically insignificant increase over dataset-level

ChEBI



Dataset-level < Hybrid
- Hybrid significantly outperforms Dataset-Level over the same series of 10k tuples.
- Hybrid continues to learn over time = better long-run performance

CORD-19



See our paper for results over more datasets and query lengths. Also see additional techniques,

Term Borrowing: expand the kind of queries we can build *over time* as we interact with the external source
- Supervised: use terms from known external matches
- Unsupervised: find terms that allow us to discover new matches

Dynamic Query Length: adjust to the "sweet-spot" query length for external sources using a simple technique

LLaMA: use LLaMA as a pretrained model (as an alternative to Longformer)