Grounding Large Language Models in Reality for Biomedical Information Extraction

J. Harry Caufield
Lawrence Berkeley National Laboratory
April 25 2025
EHLC Webinar

What will I be talking about?

- Are knowledge harmonization and discovery still jobs for humans?
  - Or is this a problem solved by AI?
- Either way, what tools can help?
- How can we focus on the strengths of AI approaches?
  - How may we complement human knowledge curation with AI rather than competing?

What will I be talking about?

- Are knowledge harmonization and discovery still jobs for humans?
    - Or is this a problem solved by AI?
- Either way, what tools can help?
- How can we focus on the strengths of AI approaches?
    - How may we complement human knowledge curation with AI rather than competing?

Guest

Why AI is a know-it-all know nothing

VentureBeat, Sep 28 2024

CORRESPONDENCE Open Access

Large language models and the perils of their hallucinations

Razvan Azamfirei[1*], Sapna R. Kudchadkar[1,2] and James Fackler[1,2]

Critical Care (2023)

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models

S.M Towhidul Islam Tonmoy[1], S M Mehedi Zaman[1], Vinija Jain[3,4*], Anku Rani[2], Vipula Rawte[2], Aman Chadha[3,4*], Amitava Das[2]

[1]Islamic University of Technology, Bangladesh
[2]AI Institute, University of South Carolina, USA
[3]Stanford University, USA, [4]Amazon AI, USA
towhidulislam@iut-dhaka.edu

arXiv (2024)

Where does knowledge come from?

It's the result of **repeated observations**.

Learning from and consistently **recording** these observations is a task in itself, but an **impossible** one at the scale we want.

How may we automate:

- Learning from literature?
- Comparing findings?
- Integrating observations?
  - Across different studies or replicates?
  - Across different knowledge bases?
  - Across different fields and disciplines?
  - Of similar concepts, even when described in different contexts?

Where does knowledge come from?

We need structured data.
This traditionally requires:

- Consistent data models
- Standards
- Ontologies and controlled vocabularies

They don’t do the work of structuring data for us.

For that we need:

- Human domain experts
- Access to data
- Tools (for data, standards, ontologies, …)

Where does knowledge come from?

Over the years, tools have included:

More rule-based

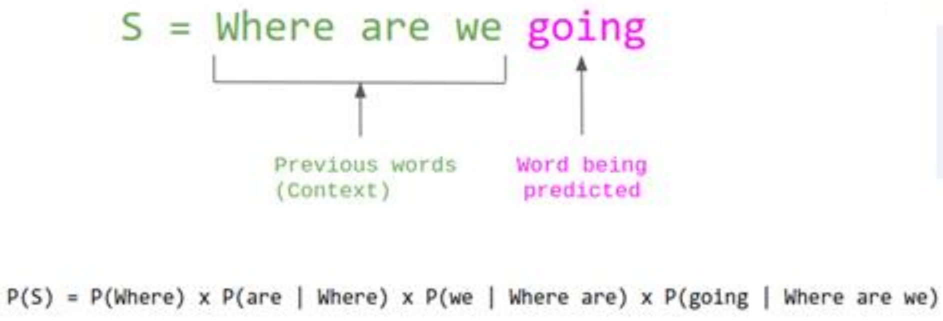
- **Rules, regular expressions, and parsers**
- **Rule-based extractors** like SemMedDB
- **Enrichment** of terms and/or annotations, like MELODI
- **Neural networks** for Natural Language Processing (e.g., LSTMs)
- Foundational **language models** (e.g., BERT)
- **Multi-task** learning (MTL) approaches

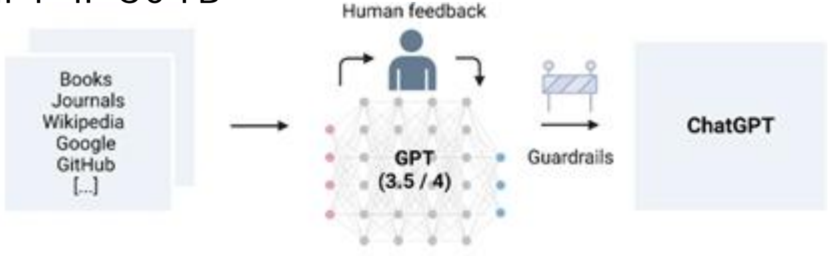
More statistical

Each method may still be effective for some use cases!

What are LLMs, anyway?

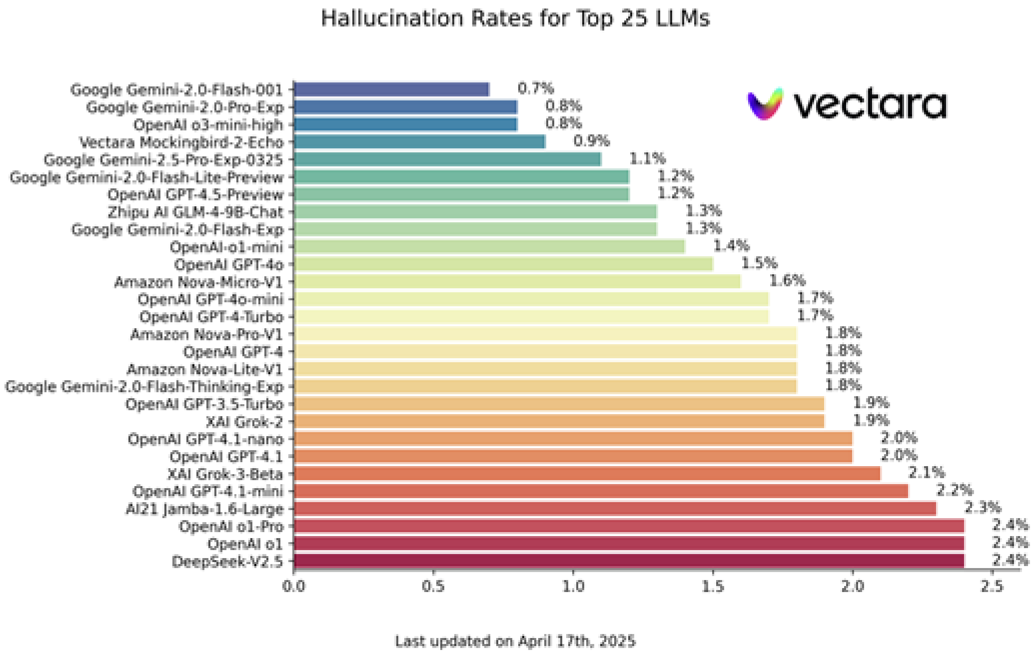
Input sizes:
GPT-3: ~570 GB
GPT-4: ~50 TB





Figures adapted from Huyen (2019) article in *The Gradient* (https://thegradient.pub/understanding-evaluation-metrics-for-language-models/)
and Clusmann et al. (2023) *Communications Medicine*

How often do LLMs fabricate knowledge?



(Note: this evaluation is based on document summarization, and doesn't account for summary quality.)
https://github.com/vectara/hallucination-leaderboard

Hallucinations (AKA confabulations)

Any output that looks believable but has no basis in reality (or only partial basis).

LLMs are grounded in **language**, not **fact**, so this is to be expected!

This has improved over time with newer models, but is still present.

Using **information from beyond the model** (retrieval augmented generation, or RAG) helps - especially as part of task-specific **agents**.

Can LLMs assign identifiers correctly?

If we ask GPT-4o: *Please provide the corresponding identifier from the Gene Ontology for each of the following terms.*
And then specify the desired format along with a list of terms…

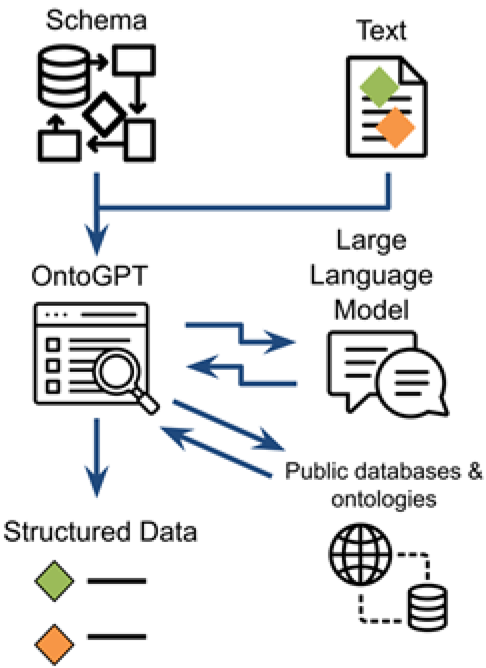
- >90% of the results are likely to be incorrect in some way.
- This happens with other ontologies as well.
- Also occurs with integrated web search (but search helps!)
- This is not a core strength of LLMs.

```
id: GO:0090729 label: ectopic germ cell programmed cell
death
id: GO:0070373 label: phosphothreonine residue binding
id: GO:2001316 label: positive regulation of tissue
kallikrein-kinin cascade
id: GO:0075047 label: haustorium neck formation
id: GO:1904382 label: negative regulation of
cytoplasmic transport
id: GO:0034128 label: regulation of toll-like receptor
21 signaling pathway
```

https://chatgpt.com/share/67f35422-c290-8007-9a91-a4befdb4a6ef

Previous evaluation results in Caufield et al. (2024) *Bioinformatics* and
https://github.com/monarch-initiative/ontogpt-experiments/blob/main/experiments/ground_compare/Comparing_Grounding.ipynb

Can we use LLMs to get structured data?



Can LLMs reliably **translate** unstructured scientific data directly into **knowledge**?

What if:

- We have specific **schemas** or data **models** we want to adhere to
- We need to link to external **unique identifiers**
  - And, ideally, not generate incorrect identifiers
- We want to ask **questions** about ontologies, sources, or both, in **natural language**

SPIRES: Structured Prompt Interrogation and Recursive Extraction of Semantics
(or, information extraction grounded in reality)

Available through OntoGPT:
https://github.com/monarch-initiative/ontogpt

See Caufield et al. (2024) Bioinformatics

Can we use LLMs to get structured data?

doi: 10.1186/s12889-023-15183-z


RESEARCH Open Access

Environmental health aspects and microbial infections of the recreational water

Microbial Infections and Swimming pools

Faika Hassanein, Inas M. Masoud, Marwa M. Fekry, Mohamed S. Abdel-Latif, Hussein Abdel-Salam, Mohamed Salem and Amany I Shehata

OntoGPT

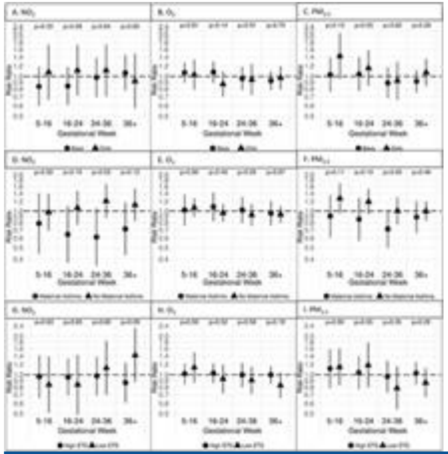
Template: `environmental_sample`

(w/ Llama 4 Scout Instruct)


```
(results formatted and truncated for brevity)
extracted_object:
  location:
    - GAZ:00052491         Alexandria, Egypt
  environmental_material:
    - ENVO:00002006        liquid water
  environments:
    - ENVO:01000966        private swimming pool
    - ENVTHES:20538        water
  causal_relationships:
    - cause: ENVO:01000967 public swimming pool
      effect: parasitic infection
    - cause: PATO:0001574  flow rate
      effect: parasitic infection
  variables:
    - ENVTHES:22023
    - fecal coliform
    - E. coli
    - parasitic infection (PI)
    - ENVO:01000967
    - PATO:0001574
    - PATO:0001736
    ...
```

Can we use LLMs to get structured data?

doi: 10.1016/j.ijheh.2024.114333



Effect modification of associations between prenatal air pollution and current asthma at age 8–9 by child sex, prenatal environmental tobacco smoke exposure, and maternal history of asthma. Risk ratios for current asthma and corresponding 95% confidence intervals are shown for associations with NO2 in the first column (panels A, D, and G), O3 in the second column (panels B, E, and H), and PM2.5 in the third column (panels C, F, and I). Estimates are reported per 5 ppb NO2, 5 ppb O3, and 2 μg/m3 PM2.5. All models are adjusted for child age, sex, study site, birth year, maternal education, household income*household count, maternal race, maternal smoking during pregnancy, maternal history of asthma, and Neighborhood Deprivation Index, as well as a product term between the air pollutant exposure and effect modifier of interest. P-values for the product interaction term are included at the top of each panel. In the first row (panels A–C), sex-specific effect estimates are shown for models including the full analytic sample (N = 1279). No evidence of effect modification by child sex was observed (all pinteraction >0.05). In the second row (panels D–F), effect estimates are shown among those with maternal history of asthma and those without maternal history of asthma for models including the full analytic sample (N = 1279). For NO2 and PM2.5, those without maternal history of asthma tended to have higher risk ratios than among those with a maternal history of asthma (e.g. p-value for interaction of NO2 in the 24–36 week window and maternal asthma = 0.03), though confidence intervals for strata-specific risk ratios all include the null. In the third row (panels G–I), effect estimates are shown for associations in a post-hoc analysis among those with high versus low environmental tobacco smoke (ETS) exposure, when the sample was restricted to non-smokers (N = 1155). High ETS was defined as participants with a urinary cotinine value in the highest quartile of the sample (>1.43 ng/mL) and low ETS was defined as participants with a urinary cotinine value in the lowest three quartiles (≤1.43 ng/mL). No effect modification by ETS was observed (all pinteraction >0.05).

OntoGPT

Template: `figure`

(w/ Qwen Coder 2.5)


```
(results formatted and truncated for brevity)
extracted_object:
  title: Effect modification of associations between prenatal air pollution and current asthma at age 8–9 by child sex, prenatal environmental tobacco smoke exposure, and maternal history of asthma.
  subpanel:
    - id: 1A
      text: Sex-specific effect estimates for models including the full analytic sample (N = 1279) for NO2
    - id: 1B
      text: Sex-specific effect estimates for models including the full analytic sample (N = 1279) for O3
    - id: 1C
      text: Sex-specific effect estimates for models including the full analytic sample (N = 1279) for PM2.5
    - id: 2D
      text: Effect estimates among those with maternal history of asthma for models including the full analytic sample (N = 1279) for NO2
...
```

Can we use LLMs to get structured data?

Yes, back to structure again: we need a consistent data model, like **Biolink**

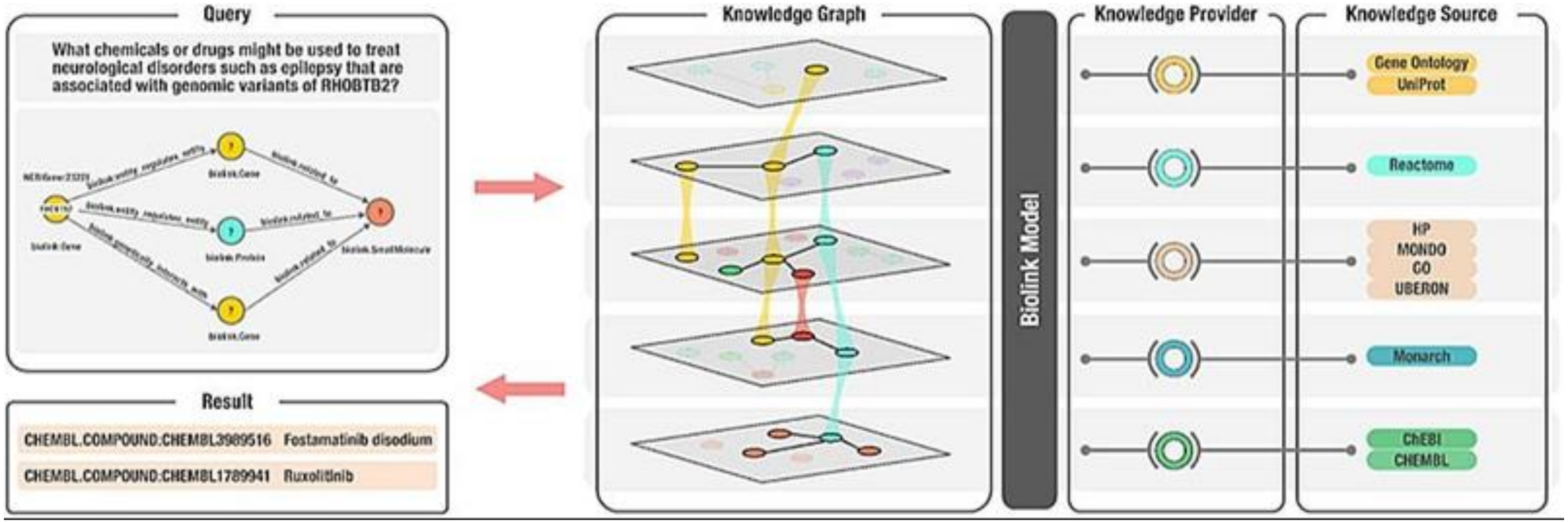


Figure from Unni and Moxon et al. (2022) *Clinical and Translational Science*

Can we use LLMs to get structured data?

The goal is often to create relationships to include in knowledge graphs.

Our framework for this is **KG-Hub**

See kghub.org

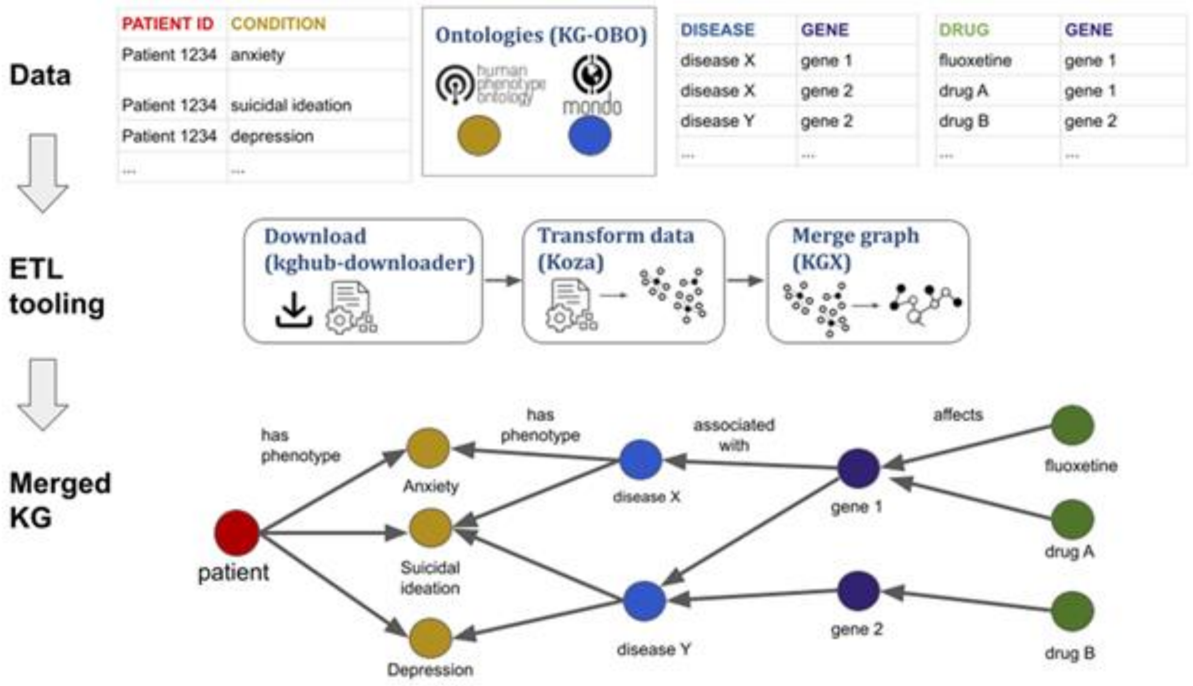
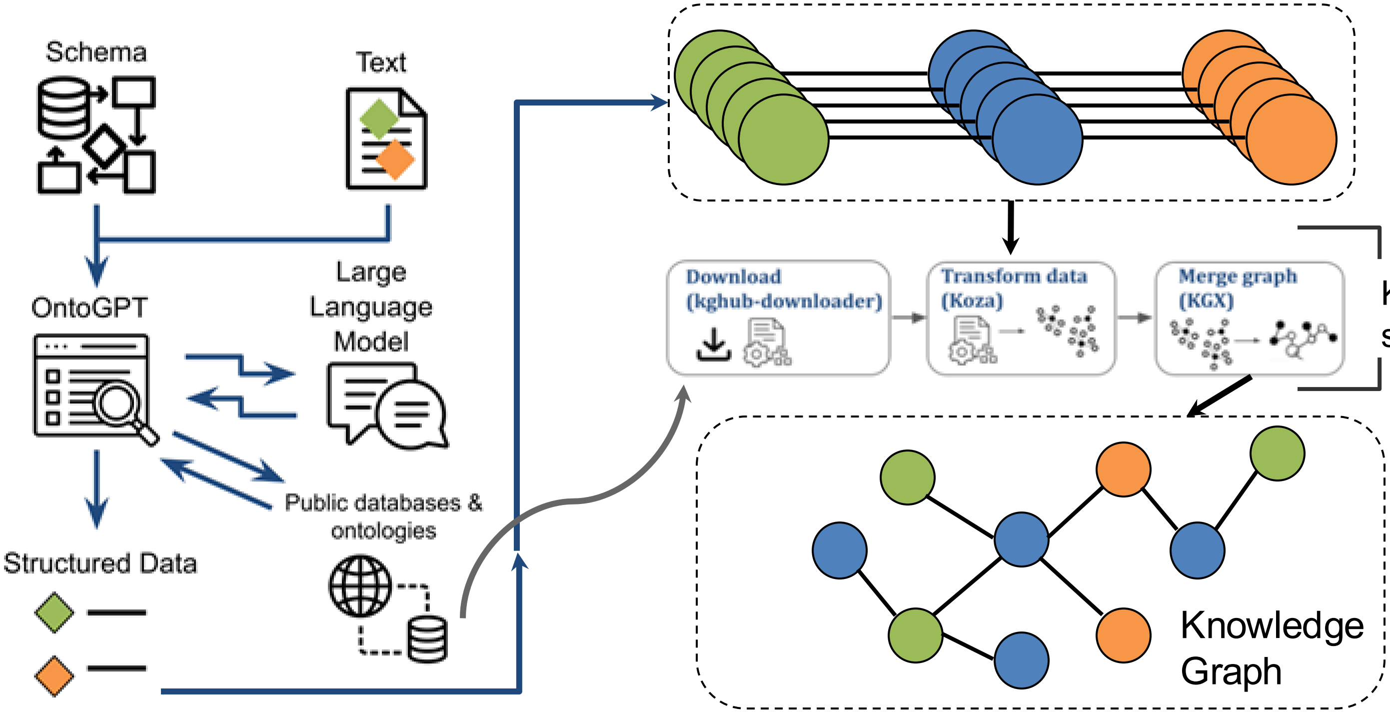


Figure from Caufield et al. (2023) *Bioinformatics*

Can we use LLMs to get structured data?



KG-Hub:
see kghub.org

See https://arxiv.org/abs/2304.02711 and https://github.com/monarch-initiative/ontogpt

Can we use LLMs to get structured data?

Other recent OntoGPT applications:

- Medical Action Ontology (MaXO) extraction
- Micronutrient Information Center knowledge extraction (for Monarch Knowledge Graph)
- Pathology report summarization and categorization
- Phenopacket extraction
- Malnutrition prediction in pediatric oncology patients
- Harmonizing environmental science data sets (e.g. in ESS-DIVE)

See Niyonkuru et al. (2024) medRxiv

Oregon State University

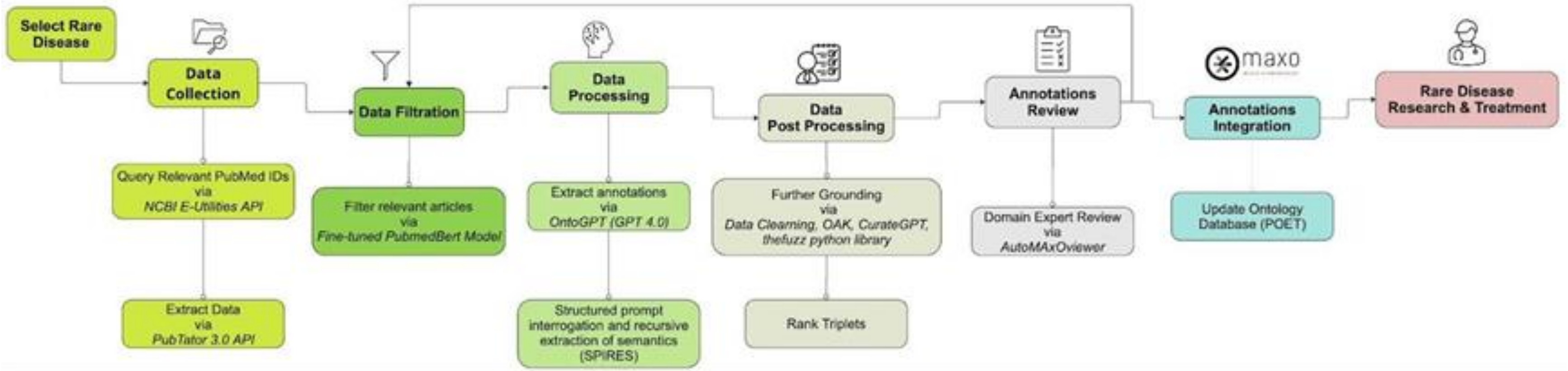
Linus Pauling Institute » Micronutrient Information Center

ESS-DIVE
Deep Insight for Earth Science Data

Can we use LLMs to get structured data?

Goal: identify candidate annotations for the Medical Action Ontology (MAxO).
e.g.,

```
Medical Action: copper chelator agent therapy [MAXO:0001224]
Relationship: PREVENTS
Phenotype: Cirrhosis [HP:0001394]
Disease: Wilson disease Anemia [MONDO:0010200]
```



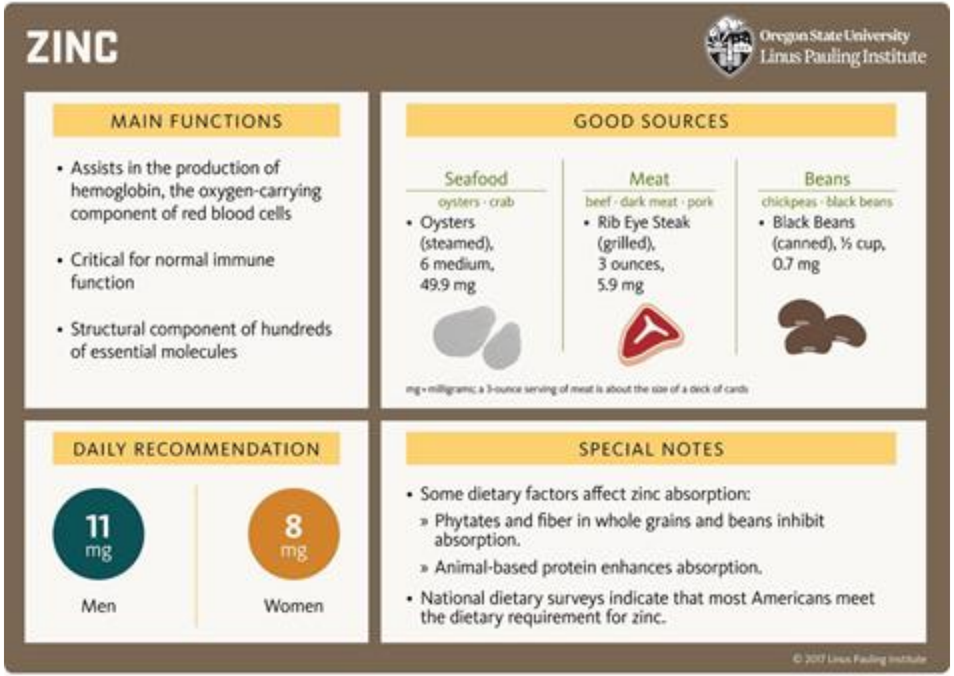
Niyonkuru et al. (2024) *medRxiv*

Can we use LLMs to get structured data?

Goal: extract structured, knowledge graph-ready relationships from the Micronutrient Information Center, including:

```
Nutrient to disease
Nutrient to phenotype
Nutrient to biological process
Nutrient to health status of
       a body part or system
Nutrient to food source
Nutrient to nutrient
```

See https://github.com/monarch-initiative/mic-ingest
for code, built with OntoGPT and the Koza data processing tool.

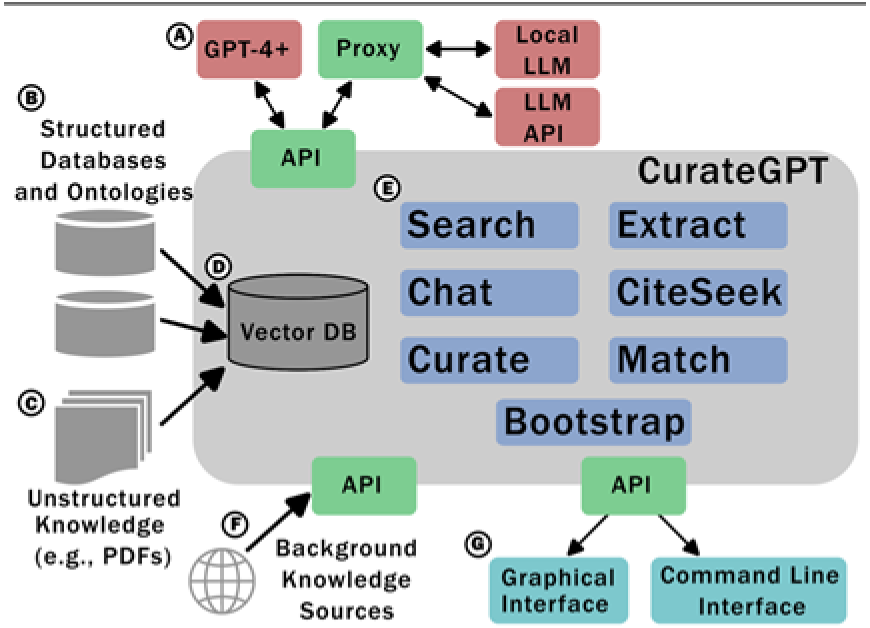


https://lpi.oregonstate.edu/mic/minerals/zinc

How may LLMs assist with curation?

CurateGPT - try it out at curategpt.io

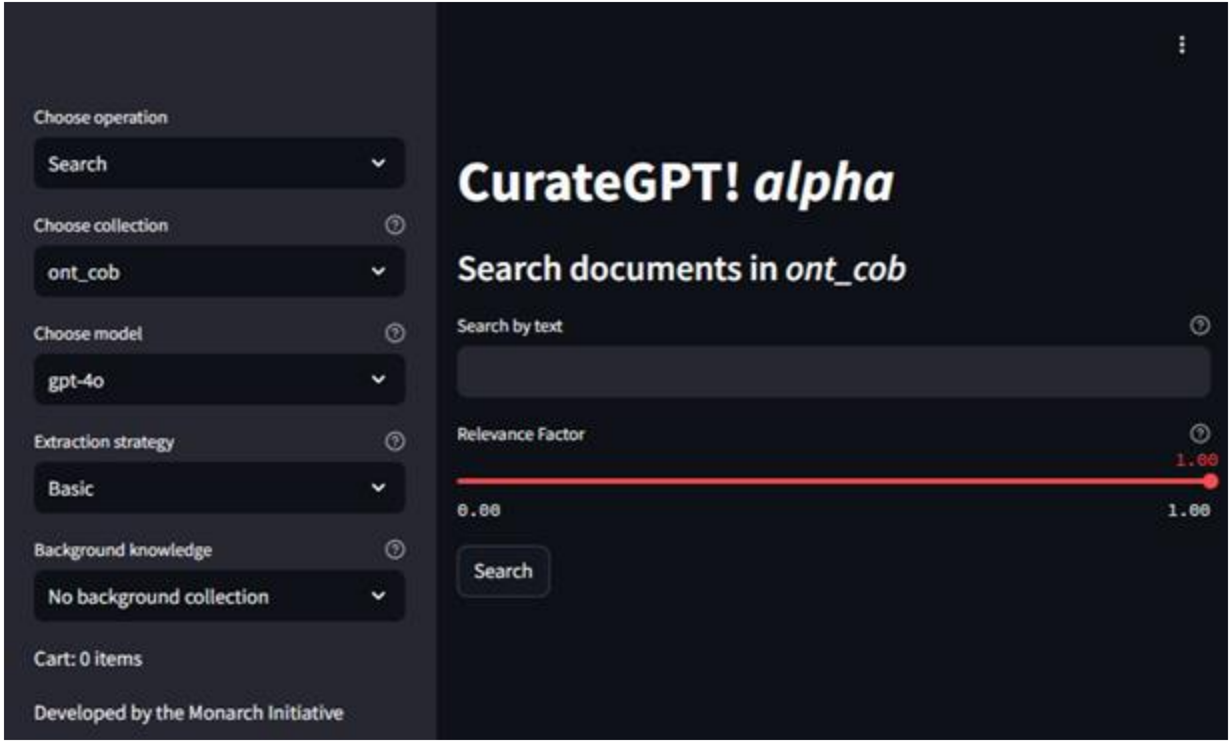
See https://github.com/monarch-initiative/curategpt
And Caufield et al. (2024) arXiv - https://arxiv.org/abs/2411.00046



Retrieval Augmented Generation (**RAG**) plus curation-focused **agents** plus a graphical interface.
Use your own resource (ontology, knowledge base, etc) to show the LLM what new entries should look like.

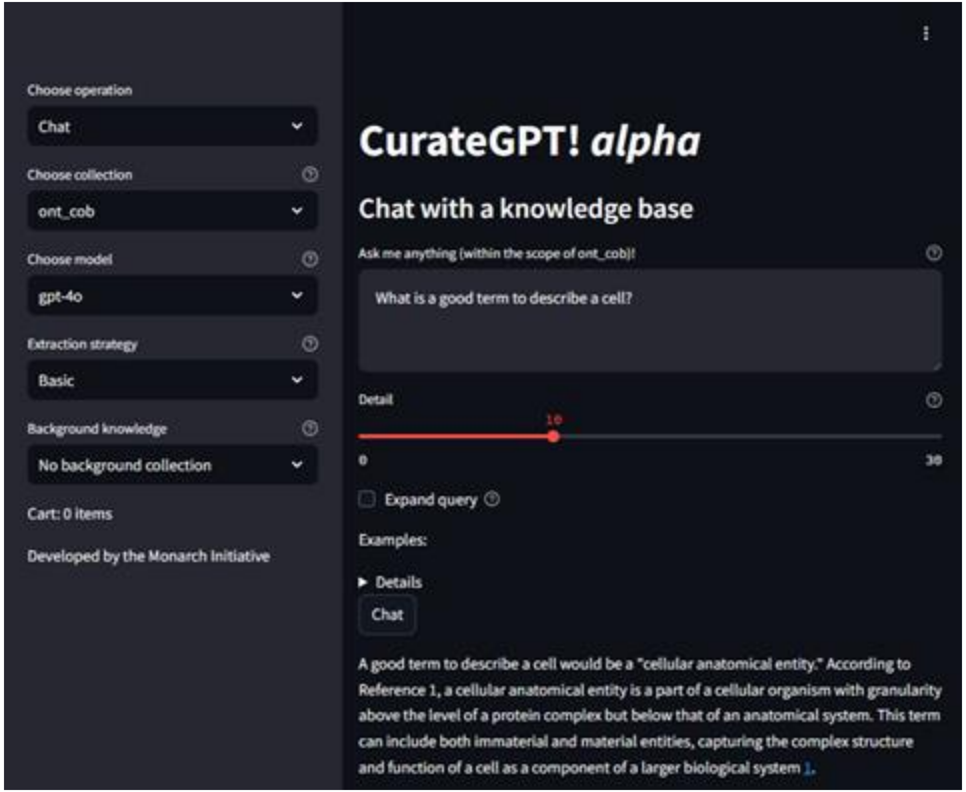
How may LLMs assist with curation?

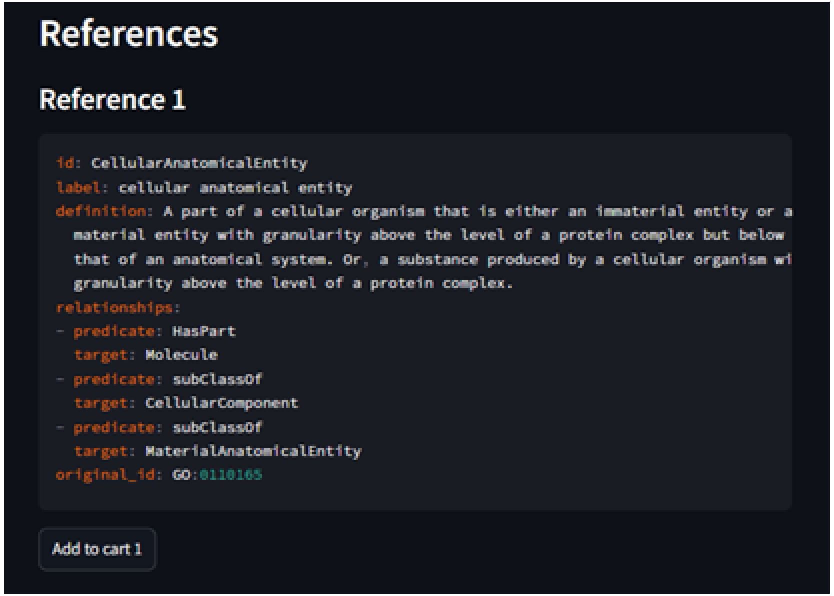
CurateGPT - try it out at curategpt.io



How may LLMs assist with curation?

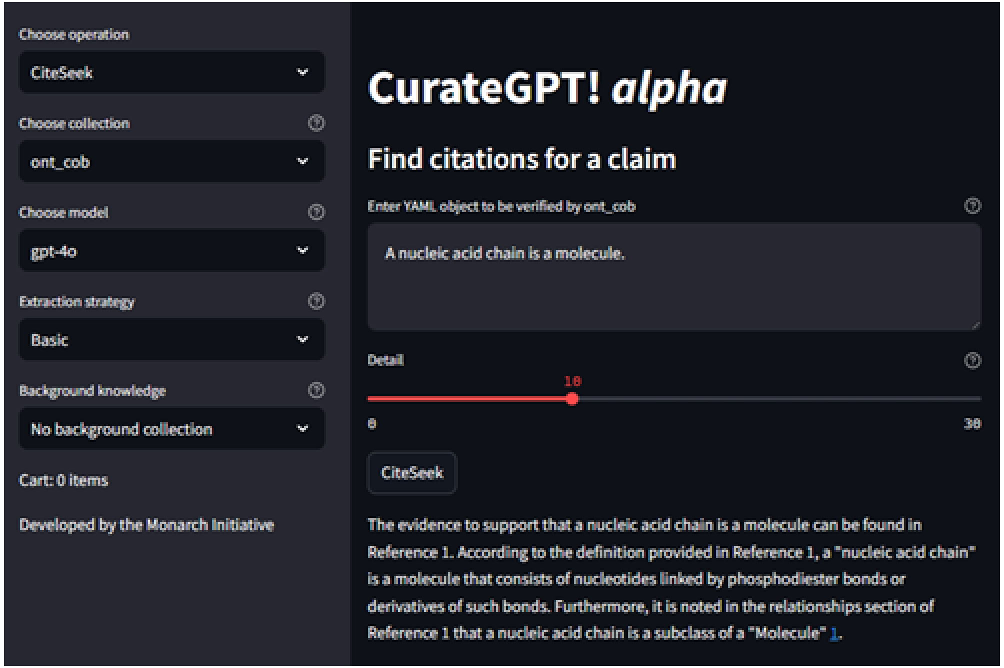
CurateGPT - try it out at curategpt.io

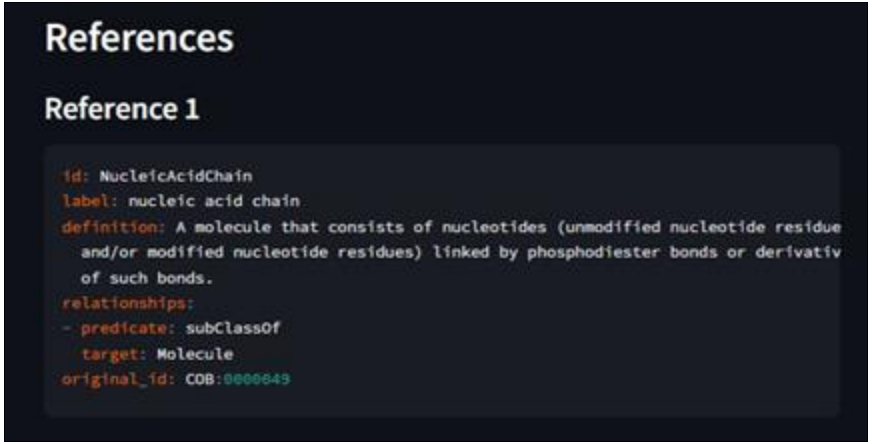




How may LLMs assist with curation?

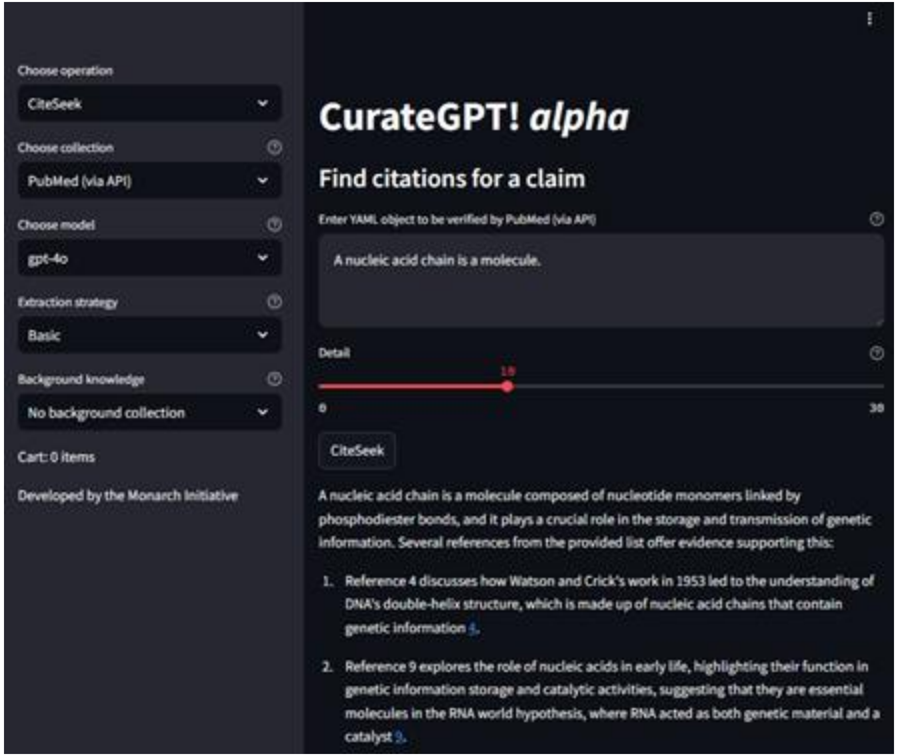
CurateGPT - try it out at curategpt.io

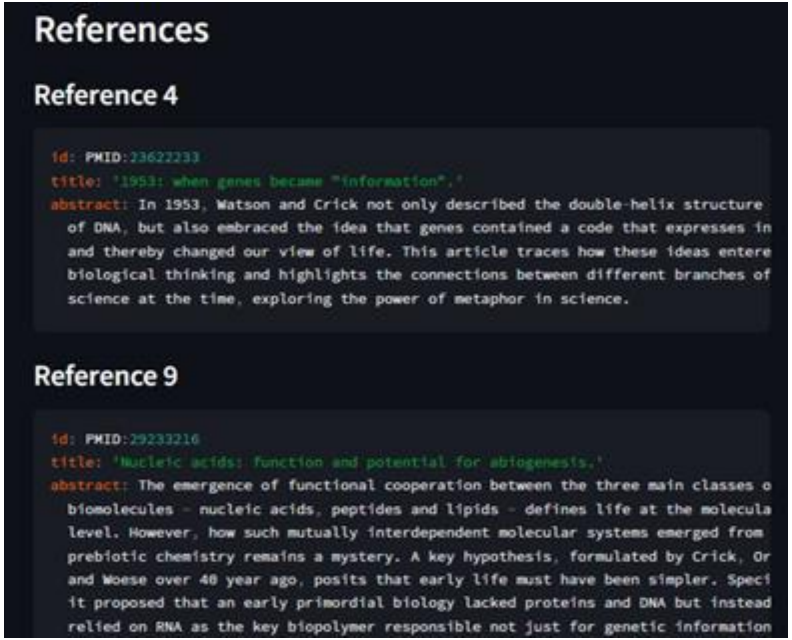




How may LLMs assist with curation?

CurateGPT - try it out at curategpt.io



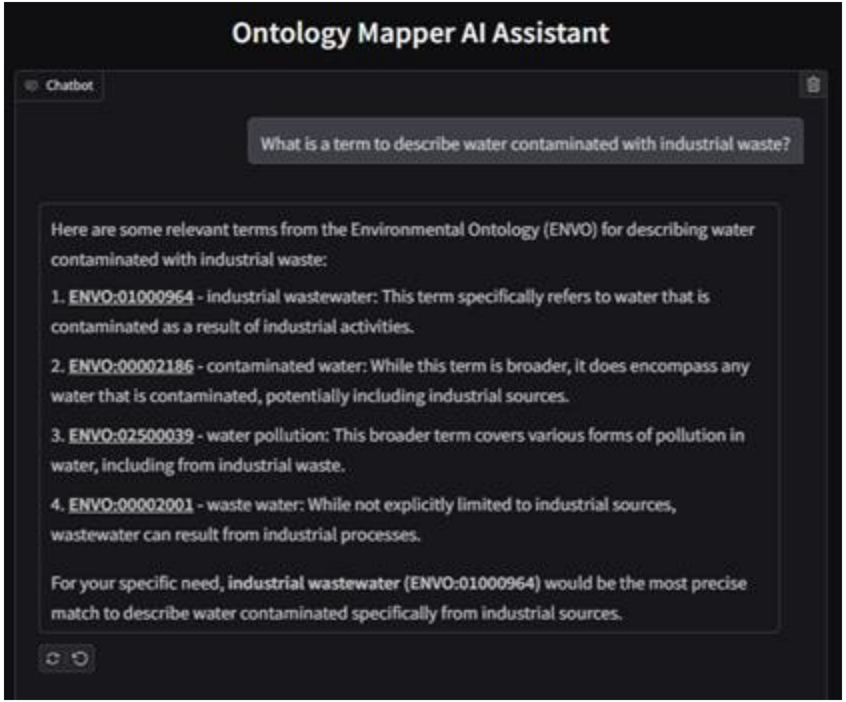


How may LLMs assist with curation?

Aurelian

See https://github.com/monarch-initiative/aurelian

The `mapper` agent in use



Note the prompt here does not specify ENVO - the agent found that on its own

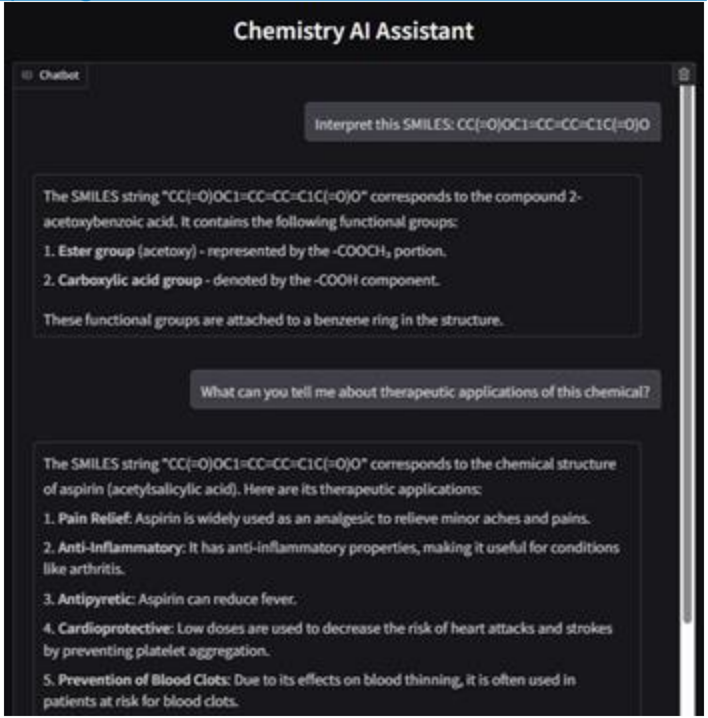
Specific **agents** for specific **tasks,** on the command line or in a graphical interface.
A task may involve:

- Domain knowledge
  - e.g., The LD50 of arsenic is ~1 mg/kg
- Technical knowledge
  - e.g., X tool may be used for Y
- Knowledge about a specific resource or study
  - e.g., there is a term in ENVO for “industrial wastewater”

How may LLMs assist with curation?

Aurelian

See https://github.com/monarch-initiative/aurelian



The chemistry agent in use

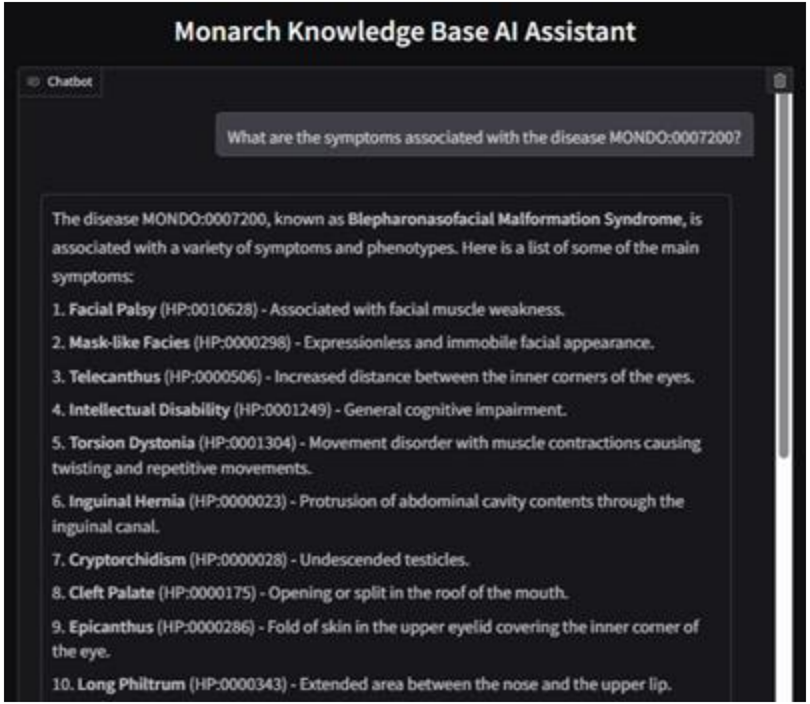
Specific **agents** for specific **tasks,** on the command line or in a graphical interface.
A task may involve:

- Domain knowledge
  - e.g., The LD50 of arsenic is ~1 mg/kg
- Technical knowledge
  - e.g., X tool may be used for Y
- Knowledge about a specific resource or study
  - e.g., there is a term in ENVO for "industrial wastewater"

How may LLMs assist with curation?

Aurelian

See https://github.com/monarch-initiative/aurelian



The `monarch` agent in use

Specific **agents** for specific **tasks,** on the command line or in a graphical interface.
A task may involve:

- Domain knowledge
  - e.g., The LD50 of arsenic is ~1 mg/kg
- Technical knowledge
  - e.g., X tool may be used for Y
- Knowledge about a specific resource or study
  - e.g., there is a term in ENVO for “industrial wastewater”

How may LLMs assist with curation?

Aurelian

See https://github.com/monarch-initiative/aurelian

The `talisman` agent in use

Specific **agents** for specific **tasks,** on the command line or in a graphical interface.
A task may involve:

- Domain knowledge
  - e.g., The LD50 of arsenic is ~1 mg/kg
- Technical knowledge
  - e.g., X tool may be used for Y
- Knowledge about a specific resource or study
  - e.g., there is a term in ENVO for “industrial wastewater”

Summary

- LLMs are improving, but still struggle with grounding in reality.
- Human-in-the-loop curation is still essential.

- Tools like OntoGPT and CurateGPT leverage LLMs to extract structured data.
- Specific curation tasks can be performed by different Aurelian agents.

Thank You

Email:
jhc@lbl.gov

Github:
@caufieldjh

Please contact me if you would like a demo of any resources mentioned in this presentation!

BBOP@LBL:
Chris Mungall (PI)
Seth Carbon
Nomi Harris
Harshad Hegde
Marcin Joachimiak
Patrick Kalita
Mark Miller
Sierra Moxon
Sujay Patil
Justin Reese

Vincent Emonet (Maastricht Univ)
Nico Matentzoglu (Semanticly)
HyeongSik Kim (Bosch Research)
Melissa Haendel (UNC)
Peter Robinson (JAX)
Carlo Kroll (UNC)

This work was supported by:

- National Institutes of Health National Human Genome Research Institute [RM1HG010860]
- National Institutes of Health Office of the Director [R24 OD011883]
- Office of Science, Office of Basic Energy Sciences, of the US Department of Energy [DE-AC0205CH11231]
- Bosch Research

Icons by Lars Meiertoberens and Nur Achmadi Yusuf (both ℅ the Noun Project)
Graphics (unless otherwise specified) produced with Distillery (https://followfox.ai/)