



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 Tonmoy¹, S.M Mehediz Zaman¹, Vinija Jain^{3,4*}, Anku Rani², Vipula Rawte¹, Aman Chadha^{1,4*}, Amitava Das²

¹Islamic University of Technology, Bangladesh

²AI Institute, University of South Carolina, USA

³Stanford University, USA, ⁴Amazon 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:

- **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 rule-based



More statistical

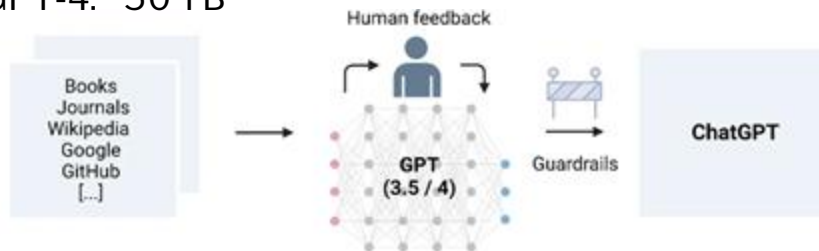
Each method may still be effective for some use cases!

What are LLMs, anyway?



$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$

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

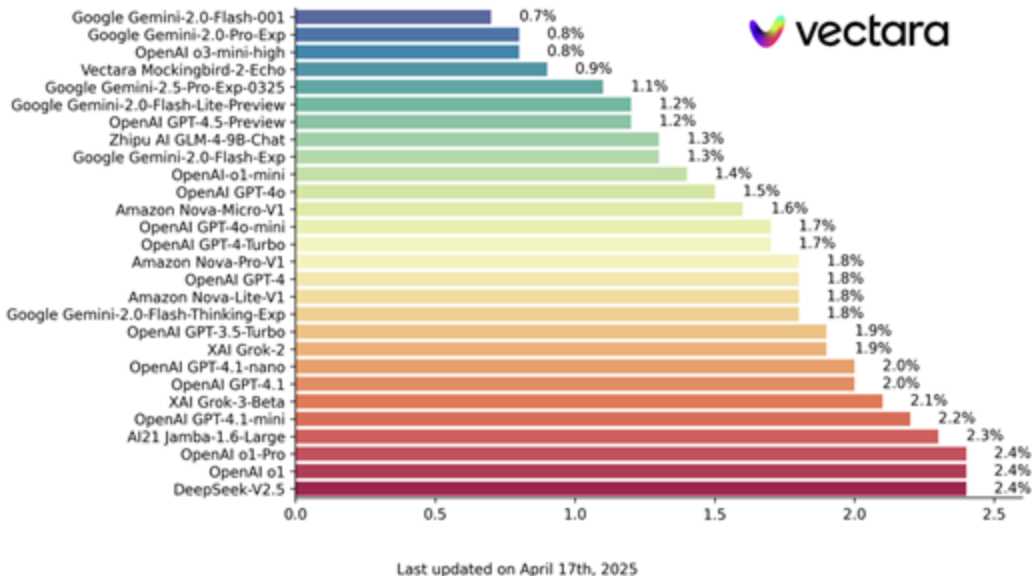


Figures adapted from Huyen (2019) article in *The Gradient*
(<https://thegradient.pub/understanding-evaluation-metrics-for-language-models/>)
and Clusmann et al. (2023) *Communications Medicine*



How often do LLMs fabricate knowledge?

Hallucination Rates for Top 25 LLMs



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**.

(Note: this evaluation is based on document summarization, and doesn't account for summary quality.)

<https://github.com/vectara/hallucination-leaderboard>

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
```

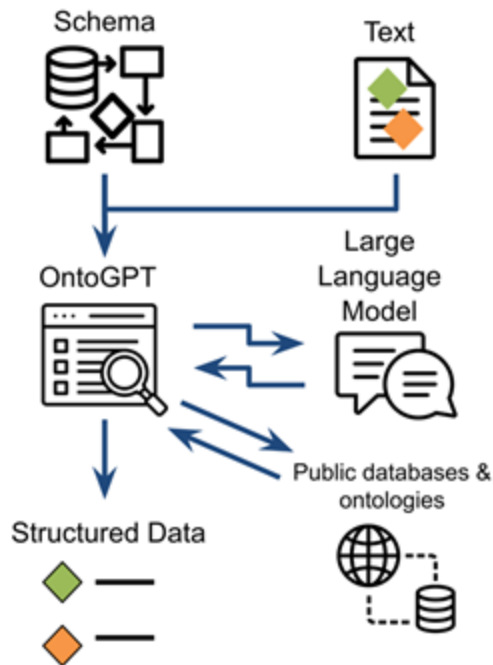
<https://chatgpt.com/share/67f35422-c290-8007-9a91-a4befdb4a6ef>

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^{1*}, Inas M. Masoud², Marwa M. Fekry³, Mohamed S. Abdel-Latif⁴, Hussein Abdel-Salam⁵, Mohamed Salem³ and Amany I. Shehata⁶

OntoGPT
 Template: `environmental_sample`
 (w/ Llama 4 Scout Instruct)



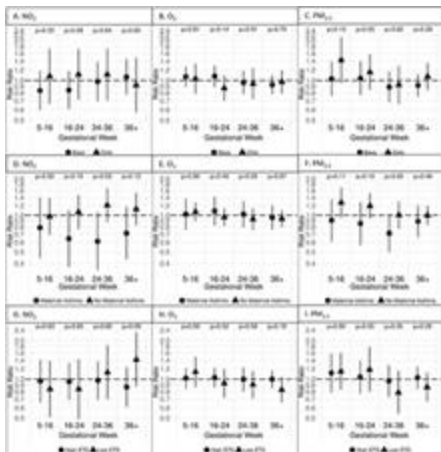
```

(results formatted and truncated for brevity)
extracted_object:
  location: Alexandria, Egypt
    - GAZ:00052491
  environmental_material: liquid water
    - ENVO:00002006
  environments: private swimming pool
    - ENVO:01000966
    - ENVTHES:20538 water
  causal_relationships: public swimming pool
    - cause: ENVO:01000967
      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 association with NO₂ in the first column (panels A, D, and G), O₃ in the second column (panels B, E, and H), and PM_{2.5} in the third column (panels C, F, and I). Estimate were reported per 5 ppb NO₂, 5 ppb O₃, and 2 μg/m³ PM_{2.5}. 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 p-interaction > 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 PM_{2.5}, those without maternal history of asthma tended to have higher risk ratios than those with a maternal history of asthma (e.g., p-value for interaction of NO₂ in the 2.4-3.6 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 (>14.3 ng/mL) and low ETS was defined as participants with a urinary cotinine value in the lowest three quartiles (≤14.3 ng/mL). No effect modification by ETS was observed (all p-interaction > 0.05).

(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 PM_{2.5}
- id: 2D
text: Effect estimates among those with maternal history of asthma for models including the full analytic sample (N = 1279) for NO₂

...

OntoGPT
Template: figure
(w/ Qwen Coder 2.5)





Can we use LLMs to get structured data?

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

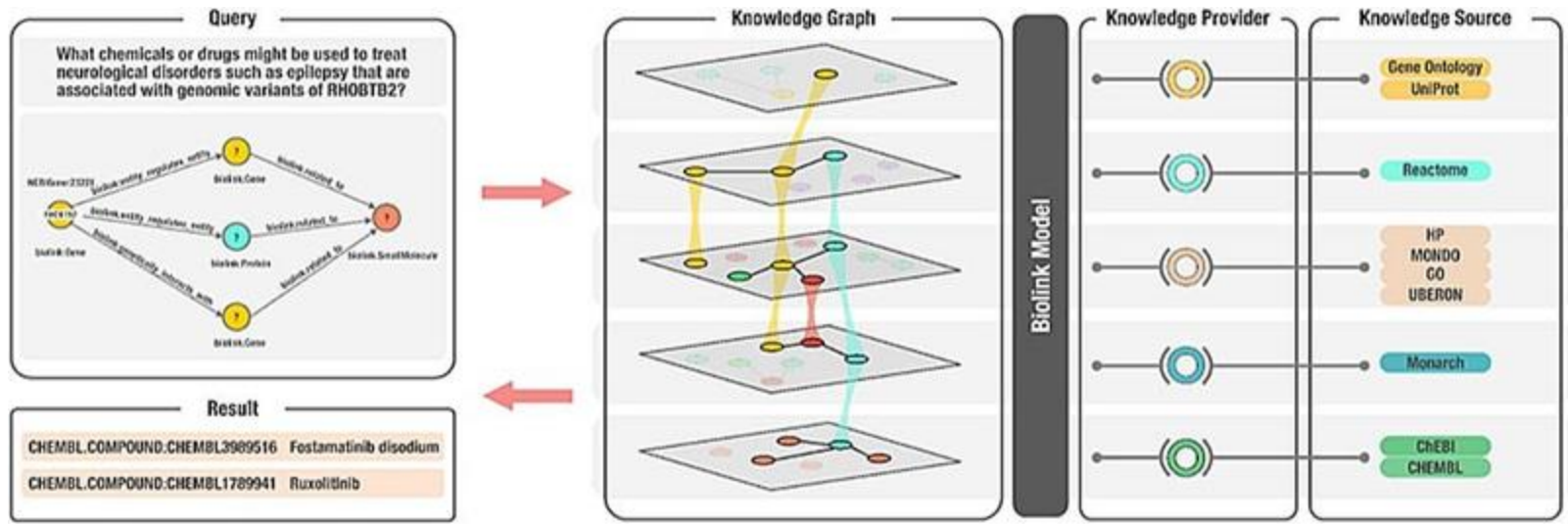


Figure from Unni and Moxon et al. (2022) *Clinical and Translational Science*



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 kg-hub.org

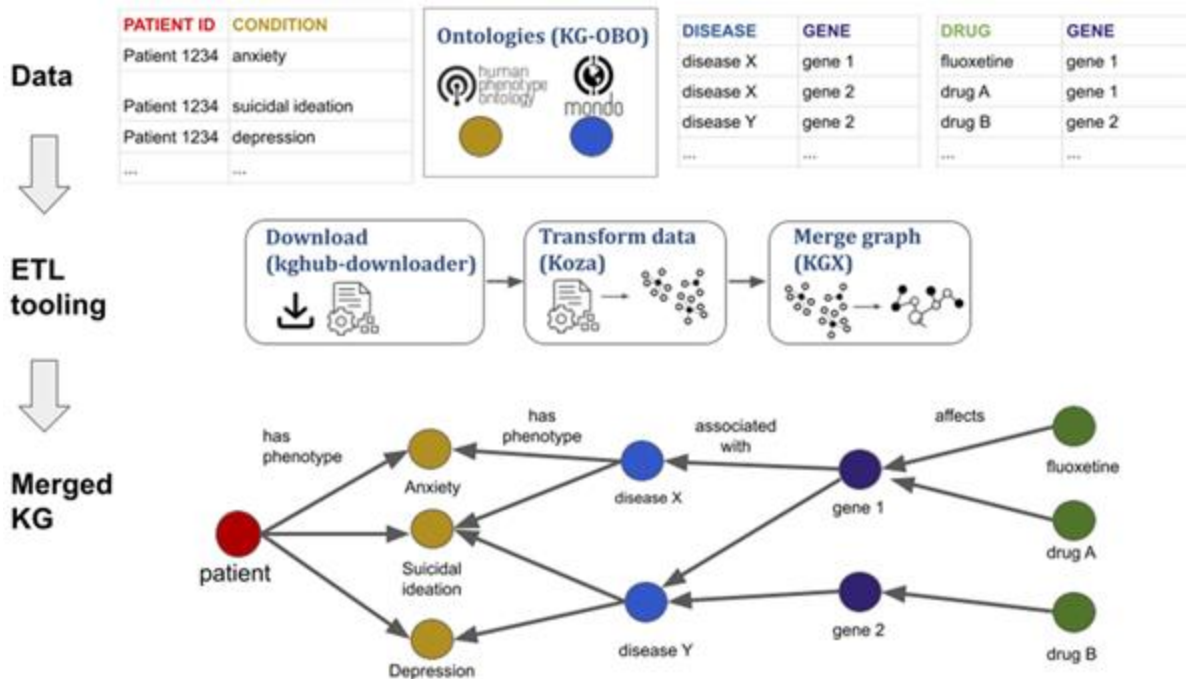
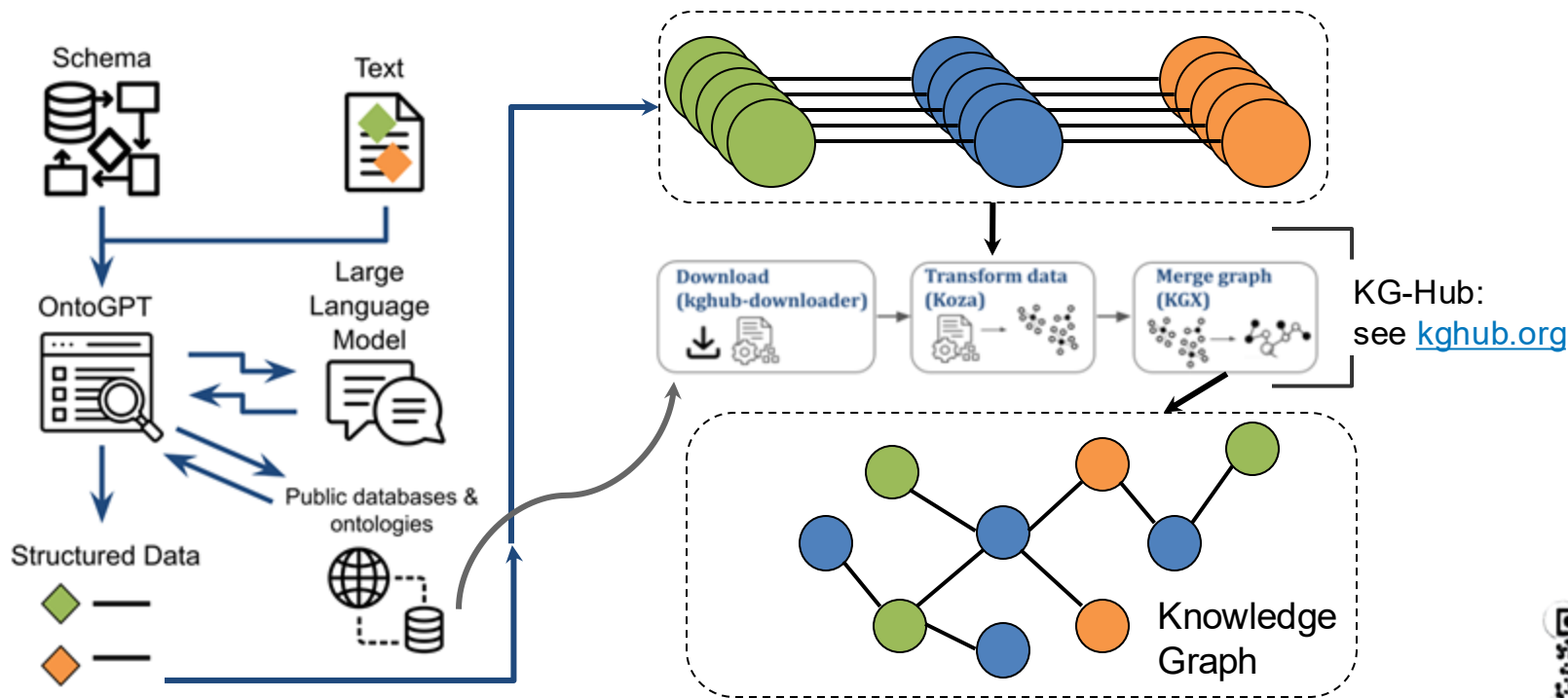


Figure from Caufield et al. (2023) *Bioinformatics*



Can we use LLMs to get structured data?



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).

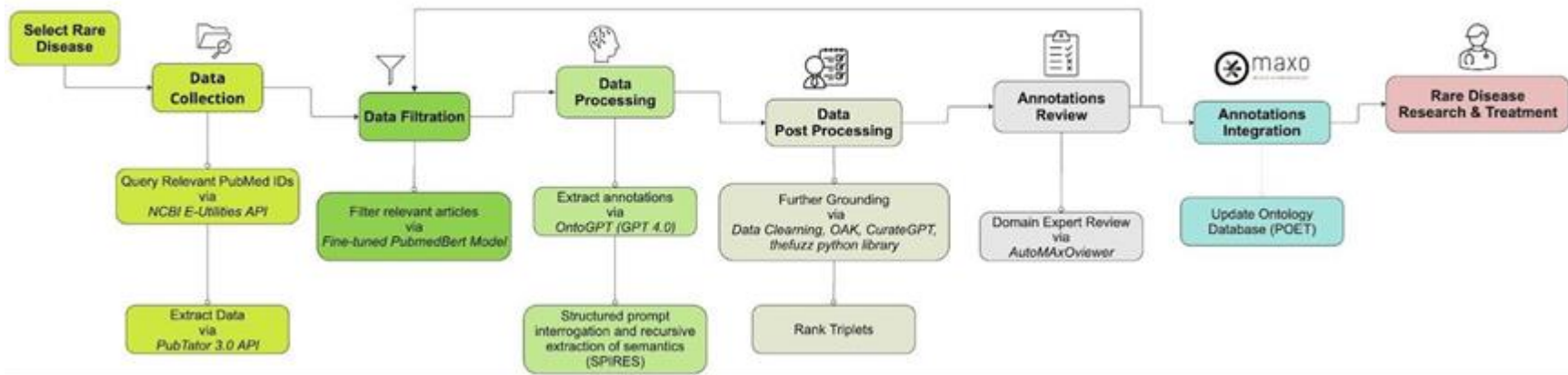
e.g.,

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

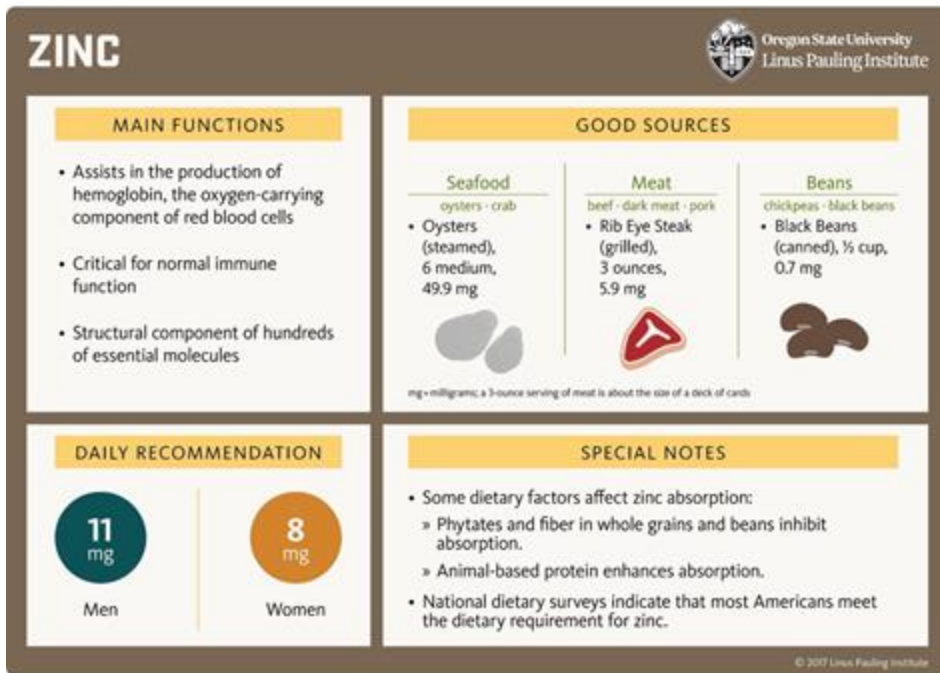
Nutrient to health status of
a body part or system

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>

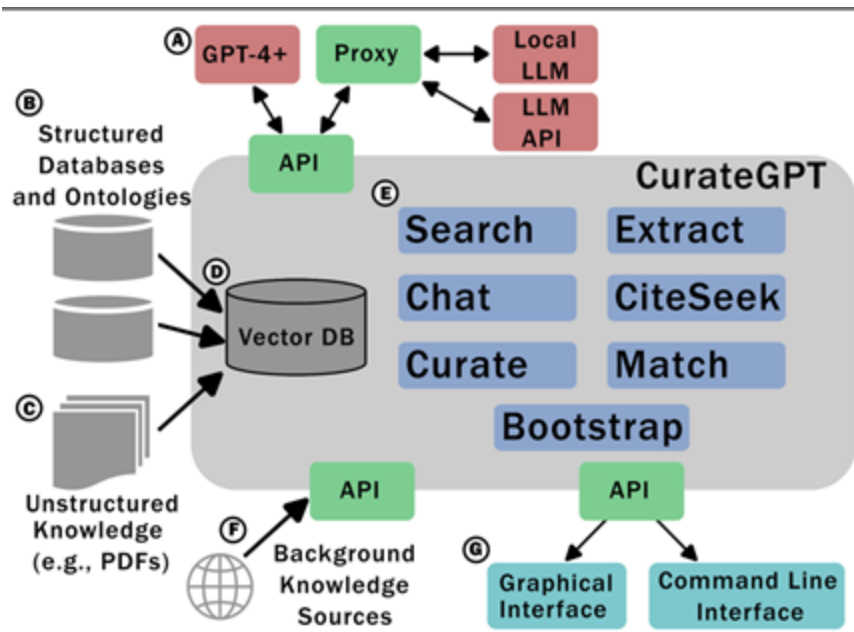


How may LLMs assist with curation?

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

Choose operation
Search

Choose collection
ont_cob

Choose model
gpt-4o

Extraction strategy
Basic

Background knowledge
No background collection

Cart: 0 items

Developed by the Monarch Initiative

CurateGPT! *alpha*

Search documents in *ont_cob*

Search by text

Relevance Factor

0.00 1.00

Search



How may LLMs assist with curation?

CurateGPT - try it out at curategpt.io

Choose operation
Chat

Choose collection
ont_cob

Choose model
gpt-4o

Extraction strategy
Basic

Background knowledge
No background collection

Cart: 0 items
Developed by the Monarch Initiative

CurateGPT! *alpha*

Chat with a knowledge base

Ask me anything (within the scope of ont_cob!)

What is a good term to describe a cell?

Detail
0 10 30

Expand query

Examples:
▶ Details
Chat

A good term to describe a cell would be a "cellular anatomical entity." According to Reference 1, a cellular anatomical entity is a part of a cellular organism with granularity above the level of a protein complex but below that of an anatomical system. This term can include both immaterial and material entities, capturing the complex structure and function of a cell as a component of a larger biological system 1.

References

Reference 1

```
id: CellularAnatomicalEntity
label: cellular anatomical entity
definition: A part of a cellular organism that is either an immaterial entity or a material entity with granularity above the level of a protein complex but below that of an anatomical system. Or, a substance produced by a cellular organism with granularity above the level of a protein complex.
relationships:
- predicate: HasPart
  target: Molecule
- predicate: subclassOf
  target: CellularComponent
- predicate: subclassOf
  target: MaterialAnatomicalEntity
original_id: GO:0110165
```

Add to cart 1



How may LLMs assist with curation?

CurateGPT - try it out at curategpt.io

The screenshot shows the CurateGPT! alpha web interface. On the left is a sidebar with configuration options: 'Choose operation' (CiteSeek), 'Choose collection' (ont_cob), 'Choose model' (gpt-4o), 'Extraction strategy' (Basic), and 'Background knowledge' (No background collection). The main area is titled 'CurateGPT! alpha' and 'Find citations for a claim'. It features a text input field containing 'A nucleic acid chain is a molecule.' Below this is a 'Detail' slider set to 10. A 'CiteSeek' button is visible. At the bottom, a text block provides evidence for the claim, citing Reference 1 and explaining that a nucleic acid chain is a subclass of a 'Molecule'.

The screenshot shows the 'References' section of the interface. Under 'Reference 1', there is a code block containing the following information:

```
id: NucleicAcidChain
label: nucleic acid chain
definition: A molecule that consists of nucleotides (unmodified nucleotide residue
and/or modified nucleotide residues) linked by phosphodiester bonds or derivativ
of such bonds.
relationships:
- predicate: subclassOf
target: Molecule
original_id: COB:0000849
```




How may LLMs assist with curation?

CurateGPT - try it out at curategpt.io

Choose operation
CiteSeek

Choose collection
PubMed (via API)

Choose model
gpt-4o

Extraction strategy
Basic

Background knowledge
No background collection

Cart: 0 items

Developed by the Monarch Initiative

CurateGPT! *alpha*

Find citations for a claim

Enter YAML object to be verified by PubMed (via API)

A nucleic acid chain is a molecule.

Detail

0 1.0 30

CiteSeek

A nucleic acid chain is a molecule composed of nucleotide monomers linked by phosphodiester bonds, and it plays a crucial role in the storage and transmission of genetic information. Several references from the provided list offer evidence supporting this:

- Reference 4 discusses how Watson and Crick's work in 1953 led to the understanding of DNA's double-helix structure, which is made up of nucleic acid chains that contain genetic information [↗](#).
- Reference 9 explores the role of nucleic acids in early life, highlighting their function in genetic information storage and catalytic activities, suggesting that they are essential molecules in the RNA world hypothesis, where RNA acted as both genetic material and a catalyst [↗](#).

References

Reference 4

```
id: PMID:23622233
title: '1953: when genes became "information",'
abstract: In 1953, Watson and Crick not only described the double-helix structure of DNA, but also embraced the idea that genes contained a code that expresses in and thereby changed our view of life. This article traces how these ideas entered biological thinking and highlights the connections between different branches of science at the time, exploring the power of metaphor in science.
```

Reference 9

```
id: PMID:29233216
title: 'Nucleic acids: function and potential for abiogenesis,'
abstract: The emergence of functional cooperation between the three main classes of biomolecules - nucleic acids, peptides and lipids - defines life at the molecular level. However, how such mutually interdependent molecular systems emerged from prebiotic chemistry remains a mystery. A key hypothesis, formulated by Crick, Or and Woese over 40 years ago, posits that early life must have been simpler. Specifically, it proposed that an early primordial biology lacked proteins and DNA but instead relied on RNA as the key biopolymer responsible not just for genetic information
```

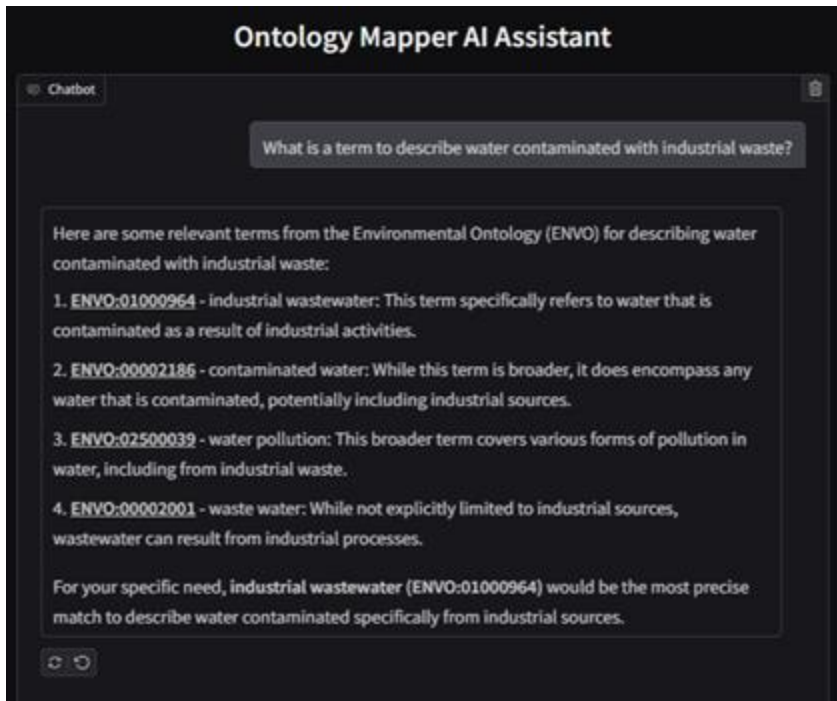


How may LLMs assist with curation?

Aurelian

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

The
mapper
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”

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



How may LLMs assist with curation?

Aurelian

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

The
chemistry
agent in
use

Chemistry AI Assistant

Chatbot

Interpret this SMILES: CC(=O)OC1=CC=CC=C1C(=O)O

The SMILES string "CC(=O)OC1=CC=CC=C1C(=O)O" corresponds to the compound 2-acetoxybenzoic acid. It contains the following functional groups:

1. Ester group (acetoxy) - represented by the -COOCH3 portion.
2. Carboxylic acid group - denoted by the -COOH component.

These functional groups are attached to a benzene ring in the structure.

What can you tell me about therapeutic applications of this chemical?

The SMILES string "CC(=O)OC1=CC=CC=C1C(=O)O" corresponds to the chemical structure of aspirin (acetylsalicylic acid). Here are its therapeutic applications:

1. Pain Relief: Aspirin is widely used as an analgesic to relieve minor aches and pains.
2. Anti-Inflammatory: It has anti-inflammatory properties, making it useful for conditions like arthritis.
3. Antipyretic: Aspirin can reduce fever.
4. Cardioprotective: Low doses are used to decrease the risk of heart attacks and strokes by preventing platelet aggregation.
5. Prevention of Blood Clots: Due to its effects on blood thinning, it is often used in patients at risk for blood clots.

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"

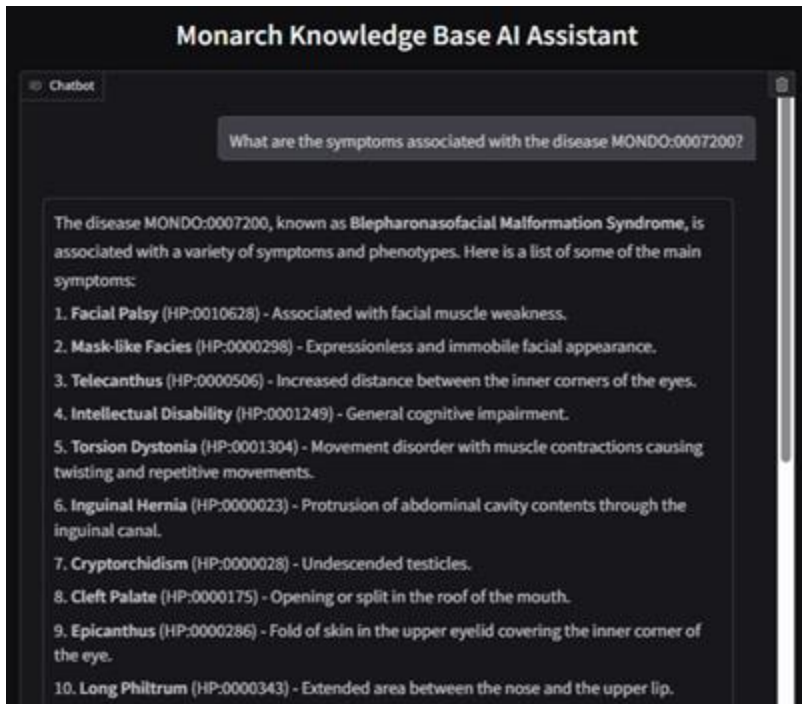


How may LLMs assist with curation?

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
Sujoy Patil
Justin Reese

Vincent Emonet (Maastricht Univ)
Nico Matentzoglou (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 Meier toberens and Nur Achmadi Yusuf (both % the Noun Project)
Graphics (unless otherwise specified) produced with Distillery (<https://followfox.ai/>)